ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

MODELLING TIME SERIES OF COUNTS WITH AN APPLICATION ON DAILY CAR ACCIDENTS

By

Georgios Iordanis Sermaidis

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

2006

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Modelling time series of counts with an application on daily car accidents

By
GEORGIOS IORDANIS SERMAIDIS

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
March 2006

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Μοντελοποίηση χρονοσειρών για διακριτά δεδομένα με εφαρμογή σε ημερήσια ατυχήματα αυτοκινήτων

ΓΕΩΡΓΙΟΣ ΙΟΡΔΑΝΗΣ ΣΕΡΜΑΙΔΗΣ

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Μάρτιος 2006

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

A Thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Science

MODELLING TIME SERIES OF COUNTS WITH AN APPLICATION ON DAILY CAR ACCIDENTS

Georgios Iordanis Sermaidis

Approved by the Graduate Committee

D. Karlis	P. Dellaportas	E. Ioannidis
Assistant Professor	Associate Professor	Lecturer
Thesis Supervisor	Members of the Committee	

Athens, July 2006

Epameinondas Panas, Professor
Director of the Graduate Program

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor Dimitris Karlis for his guidance and support ever since my undergraduate studies. With his help and his patience I managed to attain all that was necessary to complete this thesis. I am also grateful to professor Evangelos Ioannidis for providing me a very detailed and insightful course on time series analysis. Finally, I would like to thank Tom Brijs for giving us the chance to work with such rich and well designed databases.

II

CURRICULUM VITAE

I was born in Athens in 1981. In 1999, I entered at the Athens University of Economics and Business, Department of Statistics, from where I graduated in 2004. Since then, I have been cooperating with an institute of biomedical sciences where I took part in the statistical analysis of projects in the field of biostatistics. The recent six months I have been working in the National School of Public Health as an external consultant in statistical problems. In my free time I like very much listening to music, going to the cinema and travelling.

ABSTRACT

Statistical theory provides various models for modelling continuous data which exhibit some form of dependence. Well known models such as the family of the autoregressive moving average models have proved of great use accounting for the dependence of the data. Such models, while suitable for continuous data, cannot be directly applied to discrete data due to their nature. This fact led to the need to define appropriate time series models to deal with discrete data. Models that capture the existing dependence in discrete data are categorized in two major classes, the observation and the parameter driven models.

In an observation driven model the conditional mean of the present observation is modelled through a function of the past observations and assumes a discrete distribution for the data. These models can account for the dependence as well as for covariates which may affect the marginal mean of the observation. By considering different distributions for the data, a variety of models arise; other assuming no overdispersion, such as the Poisson Integer Autoregressive model (INAR), and other accounting for overdispersion, such as the negative binomial model (NB).

In a parameter driven model the dependence arises from an unobserved process, mostly a time series. The distribution of the present observation is conditional on the latent process in a way that the marginal distribution of the observations accounts for the dependence among them. Similar to the observation driven models, parameter driven models are capable of modelling the mean of an observation as a function of covariates. These types of models are more flexible than the observation driven models because by assuming different structures of the latent process one can account for different structures of the dependence among the data as well as for possible overdispersion. On the other hand the estimation of the parameters is more complicated. A well known model of this family is Zeger's model (1988).

In our case, we used one model from each class to model daily car ac-

cidents counts in 27 areas of the Netherlands using weather variables from each area. The purpose was to measure and examine the effect of weather variables to the daily counts of accidents. We have selected a variety of different locations (in fact different road segments) to allow for variability in the weather conditions. We use two models; the Poisson INAR model from the observation driven class and Zeger's model from the parameter driven class. Diagnostic tests were considered in order to identify which model fitted the data best and subsequently the best model's estimated coefficients were used for a meta-analysis and a meta-regression. The latter analyses belong to the class of the meta-analytic methods which are used in order to obtain an overall measure of each covariate's effect and to examine the existence of factors that affect it.

Zeger's model was shown to provide a better fit to the data than the Poisson INAR model due to the fact that it accounts for overdispersion. The meta-analysis models provided a single overall effect for each covariate. The mean daily temperature was found to reduce the mean accidents by 0.8% (p-value $= 0.002$). Precipitation covariates, duration and intensity, were found to increase the mean accidents by 11.24% (p-value < 0.001) and 3.4% (p-value < 0.001) respectively. Temperature below zero showed to increase the mean accidents by 6.25%(p-value $= 0.08$). The other variables like wind velocity, windspeed and humidity, did not manage to explain any of the accidents variability. Note that detailed description of the weather effects is quite demanding as it is well known in the traffic literature that weather variables also affect the exposure, and exposure is highly correlated to the accident counts. Thus the reported effects do not indicate a causal effect and must be interpreted with care.

The meta-regression models identified variables which influence the effect of the covariates. It was shown that an increase of one unit of the maximum temperature decreases the effect of the mean temperature on the accidents by 0.004% (p-value $= 0.042$) and a decrease of a unit of the minimum tem-

perature increases the temperature below zero effect by 0.768% (p-value = 0.047). The effect of humidity covariate one the mean accidents becomes stronger when combined with lower minimum temperatures by 0.052% (p-value < 0.001). The rainfall intensity effect was related to the increase of the rainfall duration and it decreases by 25.67% (p-value = 0.09).

ΠΕΡΙΛΗΨΗ

Η θεωρία των χρονολογικών σειρών έχει καταλάβει ένα μεγάλο μέρος στη βιβλιογραφία της στατιστικής που αφορά τη μοντελοποίηση συνεχών δεδομένων με κάποια μορφή εξάρτησης. Γνωστά μοντέλα, όπως το μοντέλο αυτοπαλίνδρομου κινητού μέσου, έχουν βρει μεγάλη εφαρμογή διότι μπορούν να ανιχνεύσουν την εξάρτηση που υπάρχει στα δεδομένα. Τα μοντέλα αυτά, παρά το το γεγονός ότι είναι κατάλληλα για συνεχή δεδομένα, δε μπορούν να έχουν άμεση εφαρμογή σε διακριτά δεδομένα λόγω της φύσης τους. Αυτό οδήγησε στην ανάγκη να καθοριστούν κατάλληλα μοντέλα χρονολογικών σειρών για διακριτά δεδομένα. Μοντέλα που βρίσκουν την εξάρτηση σε διακριτά δεδομένα έχουν χωριστεί σε δύο κατηγορίες, τα observation και τα parameter driven μοντέλα.

Η κλάση των observation driven μοντέλων θεωρεί οτι η δεσμευμένη μέση τιμή της παρούσας τιμής δίνεται ως συνάρτηση των παρελθόντων τιμών και υποθέτουν διάφορες διακριτές κατανομές για τα δεδομένα. Τα μοντέλα αυτά λαμβάνουν υποψιν την υπο μελέτη εξάρτηση καθώς και πιθανές διαθέσιμες επεξηγηματικές μεταβλητές που μπορεί να επηρεάζουν τον περιθώριο μέσο των παρατηρήσεων. Υποθέτοντας διαφορετικές κατανομές για τα δεδομένα, εξάγονται πολλά μοντέλα, μερικά υποθέτουν οτι η διακύμανση είναι ίση με τη μέση τιμή, όπως το Poisson Integer Autoregressive model (INAR) και άλλα επιτρέπουν υπερδιακύμανση, όπως το Negative Binomial model (NB).

Στα parameter driven μοντέλα η εξάρτηση θεωρείται οτι προέρχεται απο μια μη παρατηρούμενη διαδικασία, συνήθως μια χρονοσειρά. Η κατανομή των παρούσων τιμών είναι δεσμευμένη στη διαδικασία αυτή με τέτοιο τρόπο ώστε η περιθώρια κατανομή των παρατηρήσεων να λαμβάνει υποψιν την εξάρτηση που υπάρχει μεταξύ τους. Τέτοιου τύπου μοντέλα είναι πιο ευέλικτα απο τα observation driven μοντέλα διότι υποθέτοντας διάφορες μορφές εξάρτησης στη λανθάνουσα διαδικασία εξάγονται και διαφορετικές μορφές εξάρτησης στην παρατηρηθείσα διαδικασία καθώς και διαφορετικού βαθμού υπερδιακύμανση. Ωστόσο, η εκτίμηση των παραμέτρων είναι πιο δύσκολη. Ένα από τα πιο

γνωστά μοντέλα της κλάσης αυτής είναι το μοντέλο του Zeger (1988).

Στην εφαρμογή μας, χρησιμοποιούμε ένα μοντέλο από κάθε κλάση για να περιγράψουμε ημερήσια ατυχήματα αυτοκινήτων σε 27 περιοχές της Ολλανδίας μέσω επεξηγηματικών μεταβλητών καιρικών συνθηκών από κάθε περιοχή. Ο σκόπος μας ήταν να εξετάσουμε την επίδραση καιρικών συνθηκών πάνω σε ημερήσια ατυχήματα. Διαλέξαμε μια πληθώρα διαφορετικών τοποθεσιών έτσι ώστε οι καιρικές συνθήκες να διαφοροποιούνται. Χρησιμοποιήσαμε δύο μοντέλα· το Poisson INAR μοντέλο από την observation driven κλάση και το μοντέλο του Zeger από την parameter driven κλάση. Διαγνωστικά τεστ εφαρμόστηκαν στα αποτελέσματα του κάθε μοντέλου για να βρεθεί ποιό μοντέλο προσαρμόζεται καλύτερα στα δεδομένα και στη συνέχεια οι εκτιμήσεις από το καλύτερο μοντέλο χρησιμοποιήθηκαν για μετα-ανάλυση και μετα-παλινδρόμηση. Αυτές οι μέθοδοι ανήκουν στην κατηγορία των μετα-αναλυτικών μεθόδων που χρησιμοποιούνται όταν ο ερευνητής ενδιαφέρεται να αποκτήσει μια μόνο εκτίμηση της επίδρασης της κάθε μεταβλητής και το πώς αυτή μπορεί να επηρεάζεται από διάφορους παράγοντες.

Το μοντέλο του Zeger έδειξε να προσαρμόζει καλύτερα τα δεδομένα μας από το Poisson INAR μοντέλο λόγω του οτι επιτρέπει υπερδιακύμανση. Οι μετα-αναλυτικές μέθοδοι μας έδωσαν μια και μόνο εκτίμηση για κάθε επεξηγηματική μεταβλητή. Η μέση ημερήσια θερμοκρασία έδειξε να μειώνει το μέσο αριθμό ατυχημάτων κατά 0.8% (p-value = 0.002). Οι μεταβλητές της βροχόπτωσης, η διάρκεια και η ένταση, φάνηκαν να αυξάνουν τους μέσους των ημερησίων ατυχημάτων κατά 11.24% (p-value < 0.001) and 3.4% (p-value < 0.001) αντίστοιχα. Θερμοκρασίες υπό του μηδενός φάνηκαν να αυξάνουν τα ατυχήματα κατά 6.25%(p-value = 0.08). Οι υπόλοιπες μεταβλητές όπως η ταχύτητα και η διεύθυνση του ανέμου καθώς και το ποσό της υγρασίας στην ατμόσφαιρα δεν κατάφεραν να εξηγήσουν σημαντικό ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής. Η περιγραφή των επιδράσεων των καιρικών συνθηκών είναι πολύπλοκη καθώς είναι γνωστό απο τη βιβλιογραφία ότι οι καιρικές συνθήκες επηρεάζουν την έκθεση στον κίνδυνο, η οποία είναι

υψηλά συσχετισμένη με τα ημερήσια ατυχήματα. Οπότε τα αναφερόμενα αποτελέσματα δε δείχνουν μια άμεση επίδραση και πρέπει να ερμηνευτούν με προσοχή.

Τα μοντέλα της μετα-παλινδρόμησης εντοπίσανε κάποιες μεταβλητές που επηρεάζουνε το αποτελέσμα των αρχικών επεξηγηματικών μεταβλητών. Συγκεκριμένα, αύξηση της μέγιστης θερμοκρασίας κατα μια μονάδα μειώνει την επίδραση της μέσης θερμοκρασίας στα ατύχηματα κατα 0.004% (p-value = 0.042) ενώ ελάττωση της ελάχιστης θερμοκρασίας κατα μια μονάδα αυξάνει την επίδραση των θερμοκρασιών υπό του μηδενός κατα 0.768% (p-value = 0.047). Επίσης, το ποσοστό υγρασίας στην ατμόσφαιρα έχει μεγαλύτερη επίδραση στο μέσο των ατυχημάτων κατα 0.052% (p-value $<$ 0.001) όταν η ελάχιστη θερμοκρασία μειωθεί κατα μια μόναδα. Η επίδραση της έντασης της βροχής σχετίζεται με τη διάρκεια της και μειώνεται κατα 25.67% (p-value = 0.09) όταν αυξάνεται η αντίστοιχη διάρκεια.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Car accidents modelling

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 accidents 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 accidents risk (Chang and Graham, 1993; Oppe, 1991). Other researchers have looked at month-to-month changes in accident levels (Van den Bossche *et al.*, 2004; Keeler, 1994; Fridstrøm and Ingebrigtsen, 1991). However, there are only few studies which have looked at changes in accident counts at a more disaggregate level. For instance, Levine *et al.* (1995b) and Jones *et al.* (1991) studied daily changes, whilst Ceder and Livneh (1982) examined hourly fluctuations in accidents. Both approaches, high-level or low-level data aggregation, have advantages and disadvantages. While changes in accident 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 accident counts. Several authors have therefore warned for biases being introduced by modelling accident counts at high levels of aggregation (Golob *et al.*, 1990; Jovanis and Chang, 1989). Therefore, in this thesis, we study the effects of weather conditions on daily accidents for 27 cities in the Netherlands in the year 2001. The use of weather conditions is motivated by earlier research where significant influences of weather conditions on accidents were found.

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

1.2 Observation driven models

Every potential model should take into account the large and significant autocorrelations in the data. Cox (1981) characterized two classes of time-dependent data: observation-driven and parameter-driven models. In an observation-driven model, the conditional distribution of Y_t is specified as a function of past observations $y_{t-1}, y_{t-2}, \ldots y_{t-k}$ for some value of k. The first-order autoregressive time-series model for Poisson distributed data (INAR) is an example of an observation driven model.

The Poisson INAR model was first developed by Al-Osh and Al-Zaid (1987) and McKenzie (1985). This model assumes that the present observation Y_t is the sum of two components. The first component is a random variable which is defined as the sum of the successes in an experiment with Y_{t-1} trials and probability of success α. This variable is notated as $\alpha \circ Y_{t-1}$ where " $\circ$ " is called the thinning operator. The second component is equal to a Poisson random variable R_t which is independent from the first component. Hence, the model assumes that $Y_t = \alpha \circ Y_{t-1} + R_t$. In fact the binomial thinning operator assumes that the random variable $\alpha \circ Y_{t-1}$ is a binomial random variable with success probability α and number of Bernoulli trials equal to Y_{t-1}. Joe (1996) generalized this approach by developing a method to define the thinning operator for cases where the marginal distribution is in the convolution-closed infinitely divisible class. This extension not only includes the Poisson case and many other models found in the literature, but also the Gaussian AR model defined below.

The model, we will work with, is based on the Poisson INAR model and is extended to a Poisson INAR regression model by letting the probability of success α and the Poisson random component R_t to depend on regressors. This model is estimated with the EM algorithm. Its procedure removes the need for other optimization algorithms that can be quite complicated for the problem with covariates, while at the same time offers interesting insights to the researcher. For example, the byproducts of the algorithm can be further used for predicting new values.

1.3 Parameter driven models

In the parameter-driven models autocorrelation is introduced through a latent process. Zeger (1988) proposed a model which belongs to this class where the dependence is arisen from an unobserved process, mostly a time series. The distribution of the present observation is conditional on the latent

process so that the marginal distribution of the observations accounts for the dependence among them. Similar to the observation driven models, parameter driven models are capable of modelling the mean of an observation as a function of explanatory variables. Specifically, Zeger models the conditional distribution of the present observation on a latent process ϵ_t as Poisson with mean equal to the product of the latent process value and the exponent of a linear function of the regressors, that is $Y_t \mid \epsilon_t \sim Poisson(\epsilon_t exp(\mathbf{x_t b}))$, where $\mathbf{x_t}$ is the vector of covariates at time t and $\mathbf{b}$ the vector of coefficients. This type of model is more flexible than the INAR model because by assuming different structures of the latent process one can account for different structures of the dependence among the data, not necessarily of autoregressive nature as the INAR model does, as well as for possible overdispersion. The estimation of the parameters provided is based on quasi-likelihood methods and leads to a Fisher scoring algorithm.

1.4 The structure of the thesis

In chapter 2 we present the theory and properties of the observation driven model Poisson INAR of order 1, which is a special case of the models defined by McKenzie (1985) and Al-Osh and Al-Zaid (1987). This model has very similar properties to a time series autoregressive model for continuous data (AR1). The model is later modified by adding covariates to either or both the components that comprise the observation. Subsequently, the estimation of the model's parameters are obtained by maximizing the likelihood function using the EM algorithm. The estimation procedure is thoroughly explained and it includes full implementation details.

Chapter 3 presents Zeger's parameter driven model (1988). We review the theory and the properties of this model and explicitly note the difficulty in the model's estimation by maximum likelihood methods. A more feasible method is applied, namely the quasi-likelihood, which leads to a Fisher scor-

ing algorithm. Further details are given for the estimation of the parameters of the underlying process which, as noted earlier, arises the autocorrelation in the observed data and a test for the existence of a latent process is provided. We also present some diagnostic methods for the goodness of fit of the models.

The data used in this thesis are daily accident count data for 27 sites in Netherlands for the year 2001. For all sites we have a full meteorological data taken from the nearby meteorological stations. This comprises 27 series of data with quite detailed weather conditions in a daily level. Note that the original data were in hourly basis but such data contain so many zeroes to be of practical use. The data themselves imply that we have to run our models in 27 different datasets and thus we need to synthesize the findings in order to be able to examine more thoroughly the weather effects. To do this we use standard techniques from biostatistics to synthesize results from different sources, namely we make use of meta-analysis procedures. Chapter 4 reviews the theory of meta-analytic methods and their purpose of use. The most common methods are described, namely the meta-analysis and meta-regression models. Estimation procedures are given for both fixed and random effects models as well as some criteria for choosing between them.

The data analysis, including data description, application and results of the fitted models, is fully explained in Chapter 5. The second section of this chapter is a preliminary analysis of the data that was conducted so that we can get acquainted with the data. The other sections consist of the estimated parameters of the two models and their interpretation, as well as a comparison between their results and goodness of fit. Subsequently, we apply the meta-analytic methods on the estimated coefficients of the models and obtain a common effect of each covariate on the mean accidents (meta-analysis) as well as the factors that this effect is related to (meta-regression). Finally, concluding remarks can be found in Chapter 6.

Chapter 2

Integer Autoregressive Models

2.1 Theory and Properties

Let's begin with the well-known autoregressive model AR(1) for continuous data. The model assumes that $Y_t = \phi Y_{t-1} + \epsilon_t$, where $|\phi| < 1$ and $\epsilon_t \sim N(0, \sigma^2)$ independently from Y_{t-1}. It can be shown that Y_t inherits its properties from ϵ_t, thus it is normal distributed. In the case of discrete data, Y_t is required to be an integer valued process making the normal distribution assumption inappropriate. Therefore this model cannot be used directly for discrete data. McKenzie (1985) and Al-Osh and Al-Zaid (1987) defined an analogous process for discrete data, called the Integer-valued autoregressive (INAR) process as follows:

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

$$Y_t = \alpha \circ Y_{t-1} + R_t, \quad t = 1, 2, \ldots \tag{2.1}$$

The operator " $\circ$ " denotes the binomial thinning operator defined by

$$\alpha \circ Y = \sum_{t=1}^{Y} Z_t,$$

where Z_t are independent Bernoulli random variables with $P(Z_t = 1) = \alpha = 1 - P(Z_t = 0)$, $\alpha \in [0, 1]$. Thus, conditional on Y_t, $\alpha \circ Y_t$ is a binomial random variable where Y_t denotes the number of trials and α the probability of success in each trial. Pavlopoulos and Karlis (2006) have summarized the elementary properties of the binomial thinning operator, which can be found in the Appendix A of this thesis. The term R_t is referred to as the innovation term and it is a sequence of uncorrelated non-negative integer valued random variables independent of Y_{t-1} with mean μ_R and variance σ_R^2.

One can easily see that the binomial operator mimics the multiplication used for the normal time series model so as to ensure that only integer values will occur. This implies that the INAR 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 follows a discrete distribution with mean μ_R and variance σ_R^2.

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 from it and depends only on the current time point. Thus, the first part represent the influence of previous time periods while the innovation term captures the effects of the present 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 thesis we will confine ourselves to the first-order case.

The mean of an INAR(1) process is given by the formulae

$$E(Y_t) = \alpha^t E(Y_0) + \mu_R \sum_{i=0}^{t-1} a^i$$

and the variance by

$$Var(Y_t) = a^{2t}Var(Y_0) + (1-a)\sum_{j=1}^{t} a^{2j-1}E(Y_{t-j}) + +\sigma_R^2 \sum_{j=1}^{t} a^{2(j-1)}$$

where μ_R and σ_R^2 are respectively the (assumed finite) mean and variance of the i.i.d. innovations. In order for second-order stationarity to hold, the initial value of the process, Y_0, must have:

$$E(Y_0) = \frac{\mu_R}{1-a} \qquad \text{and} \qquad Var(Y_0) = \frac{a\mu_R + \sigma_R^2}{1-a^2} \tag{2.2}$$

The auto-covariance function of a stationary INAR(1) process $\{Y_t\}_{t\in Z}$ is given by the formula

$$\gamma_Y(k) = Cov(Y_t, Y_{t-k}) = \alpha^{|k|}\ \gamma_Y(0) \quad , \quad k \in Z. \tag{2.3}$$

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

$$\rho(k) = \frac{\gamma_Y(k)}{\gamma_Y(0)} = \alpha^{|k|} \tag{2.4}$$

Thus, the autocorrelation decays exponentially with lag k and for $k = 1$ we obtain that the parameter α represents the correlation between successive time points. Note that this is the case for the time series model for continuous data AR(1). As is evident from (2.4), the model can account only for positive autocorrelation. By specifying the distributional form of the innovation term, a large number of different models can arise. The most common choice is to assume a Poisson distribution for the innovation term R_t. 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 an INAR Poisson regression model by adding covariates to both the innovation term and/or the autocorrelation parameter. The model then takes the form

$$\begin{aligned} Y_t &= \alpha_t \circ Y_{t-1} + R_t \\ R_t &\sim Poisson(\lambda_t) \end{aligned}$$

$$\begin{aligned} \log \lambda_t &= \mathbf{z}_t^{'}\beta \\ \log \frac{\alpha_t}{1-\alpha_t} &= \mathbf{w}_t^{'}\gamma \end{aligned} \tag{2.5}$$

where $\mathbf{z}_t$ and $\mathbf{w}_t$ are vectors of covariates at time t for the innovation term and the autocorrelation parameter respectively while $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the vector of the associated regression coefficients. Note that the covariate information for the two parts of the model are not necessarily the same.

The simple Poisson regression model corresponds to the case when $\alpha_t = 0$ for all t and thus it is in fact a special case of the model described in (2.5). The model can capture the autocorrelation present in time series data. It also assumes that the correlation between successive points may depend on some variables, i.e. it is not constant across time. We will derive some properties of this model by letting $\alpha_t = \alpha$ for all t since this is the case that we will use later. The mean and variance of the Poisson INAR(1) regression model described in (2.5) is

$$E(Y_t) = \alpha^t E(Y_0) + \sum_{i=0}^{t-1} \alpha^i \lambda_{t-i} \tag{2.6}$$

$$Var(Y_t) = \alpha^{2t} Var(Y_0) + \sum_{i=0}^{t-1} \alpha^{2i} \lambda_{t-i} + (1-\alpha) \sum_{i=1}^{t} \alpha^{2i-1} E(Y_{t-i}) \tag{2.7}$$

$$Cov(Y_t, Y_{t-h}) = \alpha Var(Y_{t-h}) \tag{2.8}$$

$$Cor(Y_t, Y_{t-h}) = \alpha\sqrt{\frac{Var(Y_{t-h})}{Var(Y_t)}} \tag{2.9}$$

Note that in order for the marginal mean to be equal with the marginal variance, implying that the marginal distribution of the variables is Poisson, the initial value Y_0 must be drawn from a Poisson process. If this is satisfied then the complicated formulae of the variance $Var(Y_t)$ reduces to the formulae of the mean $E(Y_t)$ as defined in (2.6). We can see that the model is no longer weak stationary, as is to be expected. Furthermore, the interpretation of the parameters have changed. For example, the correlation between two successive time points is not α anymore but weighted by the square root of their variance ratio as (2.9) shows. Considering the interpretation of a coefficient, if we fix t and two observations Y_t and $Y_t^{'}$ differ only in one covariate of $\mathbf{z}$, say $\mathbf{z_{k_t}}$ then from (2.6) we get that

$$\begin{aligned} E(Y_t^{'} - \alpha^t Y_0) &= \sum_{i=0}^{t-1} \alpha^i \mu_{R_{t-i}} \\ &= \sum_{i=0}^{t-1} \alpha^i exp(\mathbf{z}^{'}_{\mathbf{t}}\boldsymbol{\beta} + \beta_{\mathbf{k}}) \\ &= exp(\boldsymbol{\beta_k}) E(Y_t - \alpha^t Y_0) \end{aligned} \tag{2.10}$$

We see that the interpretation of a coefficient is not straightforward. Moreover, equation (2.10) shows that the effect of a covariate on the dependent variable is not constant across time. Nevertheless, if we let $t \to \infty$ then from (2.10) it is clear that the interpretation is exactly the same as the one in the simple Poisson regression.

Clearly, the above model offers great flexibility for modelling data with temporal serial correlation. For accident data, it is reasonable to assume correlation between successive time points as a result of sharing common elements like infrastructure and road conditions that may continue to have

effects on road safety during successive time periods (such as icy roads). However, it is also plausible to assume that this correlation changes across time due to changes in driving conditions, e.g. as a result of different weather conditions.

2.2 Estimation

The probability function of $Y_t \mid Y_{t-1}$ is the convolution of a Poisson with a binomial random variable (see, Shumway and Gurland, 1960). The conditional distribution for known values of the parameters α_t, λ_t takes the form

$$\begin{aligned} &P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha_t, \lambda_t) = \\ &= \sum_{k=s}^{y_t} \frac{exp(-\lambda_t)\lambda_t^k}{(k)!} \begin{pmatrix} y_{t-1} \\ y_t - k \end{pmatrix} \alpha_t^{y_t-k}(1-\alpha_t)^{y_{t-1}-y_t+k} \end{aligned} \tag{2.11}$$

where α_t and λ_t are defined previously and $s = \max(0, y_t - y_{t-1})$.

The likelihood for the model defined in (2.11), conditional on some initial value Y_0, takes the form

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{T} P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha_t, \lambda_t)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ denotes the vector of unknown parameters. The likelihood is complicated since it involves multiple summation making the maximization a difficult task. 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). Brijs *et al.* (2004) provided an EM algorithm for the Poisson INAR(1) model described in (2.5).

The EM algorithm is a general-purpose algorithm for maximum likeli-

hood estimation in a wide variety of problems that either containing missing values or they can be considered as containing missing values. For our formulation of the model, we can rewrite the observation at point t as $Y_t = A_t + R_t$ where $A_t = \alpha_t \circ Y_{t-1}$. In fact, we have observed data Y_t while we cannot observe the latent variables A_t and R_t. Note that if we could observe those values, then the estimation of the complete data (A_t, R_t) would be straightforward as it comprises simple maximum likelihood estimation in GLM models. Recall that $A_t \sim Binomial(Y_{t-1}, a_t)$ and $R_t \sim Poisson(\lambda_t)$.

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 A_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 these models. Hence the algorithm can be described as follows.

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

$$\begin{aligned} s_t &= E(R_t \mid y_t, y_{t-1}, \boldsymbol{\theta}^{old}) \\ &= \sum_{z=s}^{y_t} z P(R_t = z \mid y_t, y_{t-1}, \boldsymbol{\theta}^{old}) \\ &= \sum_{z=s}^{y_t} z \frac{P(R_t = z) P(Y_t = y_t - z)}{P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha_t^{old}, \lambda_t^{old})} \\ &= \sum_{z=s}^{y_t} z \frac{\frac{exp(-\lambda_t^{old})(\lambda_t^{old})^z}{z!}}{P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha_t^{old}, \lambda_t^{old})} P(Y_t = y_t - z) \\ &= \lambda_t^{old} \sum_{z=s}^{y_t} \frac{P(R_t = z-1) P(Y_t = y_t - z)}{P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha_t^{old}, \lambda_t^{old})} \\ &= \lambda_t^{old} \frac{P(Y_t = y_t - 1 \mid Y_{t-1} = y_{t-1}, \alpha_t^{old}, \lambda_t^{old})}{P(Y_t = y_t \mid Y_{t-1} = y_{t-1}, \alpha_t^{old}, \lambda_t^{old})} \end{aligned}$$

for $t = 1, \ldots, T$, where according to the model

$$\lambda_t^{old} = exp\left(\mathbf{z}_t \boldsymbol{\beta}^{old}\right) \text{ and } \alpha_t^{old} = \frac{exp(\mathbf{w}_t \boldsymbol{\gamma}^{old})}{1 + exp(\mathbf{w}_t \boldsymbol{\gamma}^{old})}$$

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

$$c_t = E(A_t \mid y_t, y_{t-1}, \boldsymbol{\theta}^{old}) = y_t - s_t.$$

- *M-Step*: Update the parameters in $\boldsymbol{\theta}$ by fitting two GLM models. Namely, update $\boldsymbol{\beta}$ by fitting a Poisson regression model with response variables c_t and design matrix $\mathbf{z}$, while $\boldsymbol{\gamma}$ 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.

Remark: The algorithm just described, ignoring the time series structure, below is in fact an algorithm that fits a Poisson-Binomial regression model. In addition one must be cautious at the M-step where the expectations are not necessarily integers. Thus fitting GLM models is not exactly true but just implies that we make use of the IRLS algorithm for ML estimation in GLM models.

The above algorithm has all the pros and cons of the standard EM. Initial values for $\boldsymbol{\beta}$ 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. Slow convergence can be improved if after few EM steps one uses other algorithms with better convergence properties like Newton-Raphson. Few iterations are sufficient to be close to the maximum and thus quite good initial values for other algorithms are available.

The model so far has been applied to many types of data. Franke and Seligmann (1993), for example, fitted an INAR process without covariates in epileptic seizure counts assuming that the innovation term was a finite mixture of Poisson random variables. A similar approach has been considered by Karlis and Xekalaki (2001) for modelling count data of fires in Greece for a specific time period.

2.3 Related material

A substantial volume of literature is available on the probabilistic properties of INAR(1) and generalizations to INAR(p) models of higher order (see e.g. Al-Zaid and Al-Osh, 1988, 1990; Du and Li, 1991; Al-Osh and Aly, 1992; DaSilva and Oliveira, 2004).

Statistical inference for parameters of INAR models is less developed than their probabilistic properties, motivated almost exclusively by (but also restricted to) specific cases of application on data of small counts, conducive to equidispersion or slight overdispersion. Al-Osh and Al-Zaid (1987) were

concerned with estimation of the two parameters of the Poisson INAR(1) model, where innovations follow a Poisson law. Based on Monte Carlo simulations of the model they assessed the behavior of bias and mean square error (MSE) of Yule-Walker type estimators (YW), obtained by the method of moments, and of estimators obtained by methods of conditional least squares (CLS) and conditional maximum likelihood (CML), conditioning on the initial observation in the series. Recently, Freeland and McCabe (2005) have rigorously addressed the asymptotic properties of YW and CLS estimators of the parameters of the Poisson INAR(1) model. They derived the asymptotic covariance matrix of CLS estimators explicitly, and showed asymptotic equivalence of the distributions of CLS and YW estimators, for large samples. Estimation by CML for the Poisson INAR(1) model is treated also by Freeland and McCabe (2004a), along with development of methodology for assessing the model's adequacy when fitted to time series of small counts. An overview on statistical inference for INAR(1) models, from the more general standpoint of conditional linear (CLAR) processes with discrete support, is provided by Jung *et al.* (2005). Recently Varin and Vidoni (2005) psoposed a composite likelihood approach for estimation of the INAR and other autoregressive models. The idea is close to the conditional maximum likelihood in the sense that they construct a composite likelihood by considering the joint distribution of pairs of observations.

Franke and Selingmann (1993) proved consistency and asymptotic normality of CML estimators of the vector of four parameters in the INAR(1) model with innovations following a m-mixture of Poisson components with $m = 2$. They referred to this model as *switching*-INAR(1), or SINAR(1) in short, and applied it to time series of slightly overdispersed data of daily counts of epileptic seizures. Clearly, the INAR(1) models treated in the present thesis are a special case of the SINAR(1) model for $m = 1$.

Thyregod *et al.* (1999) considered the INAR(1) and INAR(2) models with Poisson innovations, and also a *self-exciting threshold*-INAR(1) model

(SETINAR in short) consisting of two Poisson INAR(1) branches. For other applications of the INAR model one can see Cardinal *et al,* (1999) and Brijs *et al* (2004).

Prediction and forecasting by INAR models have been described in Freeland and McCabe (2004b), and on INAR(1) with Poisson, binomial, or negative binomial innovations, fitted by Bayesian methodology by McCabe and Martin(2005), while Jung and Tremayne (2006) have produced coherent forecasts by the Al-Zaid and Al-Osh (1990) version of the INAR(2) model with Poisson innovations, fitted by the method of moments. Integer forecasting is also considered in Pavlopoulos and Karlis (2006).

Extensions of the binomial thinning operator can be found in Al-Zaid and Al-Osh (1988, 1993), Al-Osh and Aly (1992). A large literature review can be found in Yiokari *et al.* (2001). Generalizations have been proposed by Brannas and Hellstrom (2001), Gourieroux and Jasiak (2003).


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 3

Zeger's model

3.1 Theory and Properties

A model which belongs to the class of the parameter-driven models is the one proposed by Zeger (1988). Let's suppose we have observed a time series of counts y_t, $t = 1, 2, ...T$, as well as a vector of covariates $\mathbf{x}_t$. Our goal is to describe $\mu_t = E(Y_t)$ as a function of the $p \times 1$ vector of covariates. With independent data, log-linear models can be used to achieve this. Furthermore, assuming that the distribution of y_t is Poisson, that is $y_t \sim Poisson(\mu_t)$, where $\mu_t = exp\ (\mathbf{x}_t'\mathbf{b})$, maximum likelihood method can be used to estimate the unknown vector of coefficients $\mathbf{b}$. In practice, quite often the sample variance exceeds the sample mean, providing evidence that an overdispersed relative to the Poisson distribution must be used. In this case quasi-likelihood methods which allow a variety of variance-mean relation is more appropriate. Two of the most common relations are (i) $Var(Y_t) = \phi\mu_t$ and (ii) $Var(Y_t) = \mu_t + \mu_t^2\sigma^2$, where ϕ and σ are unknown scale parameters.

With time series it is unlikely that the observations are independent. Extensions of log-linear models which account for dependence are necessary to obtain valid inference about the relationship of y_t and $\mathbf{x}_t$. Zeger suggested that if ϵ_t is an unobservable noise process then the conditional distribution

of y_t on ϵ_t is Poisson with mean equal to the product of the latent process value and the predictor as in a simple log-linear model. Therefore

$$Y_t \mid \epsilon_t \sim Poisson(\epsilon_t exp\ (\mathbf{x}_t^{'}\mathbf{b})) \tag{3.1}$$

Assume that ϵ_t is a non-negative time series with mean 1, autocovariance function $\gamma_\epsilon(h)$ and variance σ_ϵ^2. The assumption of non-negativity of the ϵ_t is clear in order to ensure that the conditional mean of Y_t is non-negative. The condition that $E(\epsilon_t) = 1$ is imposed for identifiability reasons; otherwise if $c = E(\epsilon_t) \neq 1$, then c can be absorbed into the intercept term in the exponent of μ_t.

In order to meet the non-negativity of the ϵ_t, it is often convenient to model the logarithms of the ϵ_t. Letting $\delta_t = \log \epsilon_t$, then the conditional mean of Y_t on ϵ_t can be written as

$$u_t = exp\ (\mathbf{x}_t^{'}\mathbf{b} + \delta_t) \tag{3.2}$$

Of course, in order for the corresponding ϵ_t to have mean 1, we must assume $E(\exp(\delta_t)) = 1$. Unless the δ_t is a stationary Gaussian process, there is not an explicit relationship between the ACVF's of ϵ_t and δ_t. In the case where ϵ_t is a stationary log-normal process, i.e., δ_t is a stationary Gaussian process with ACVF $\gamma_\delta(.)$ then there is a nice connection between the ACVF's of the two processes. First, in order to satisfy the identifiability requirement that $E(\epsilon_t) = E(\exp(\delta_t)) = 1$ it is required that $\delta_t \sim N(\frac{-\sigma_\delta^2}{2}, \sigma_\delta^2)$. Then, with this choice of mean and variance in the log-normal distribution, $\gamma_\epsilon(h) = E[exp(\delta_{t+h} - \delta_t) - 1] = exp(\gamma_\delta(h)) - 1$ for all h.

The latent process introduces both autocorrelation and overdispersion in Y_t. Specifically, the following can be derived from the model:

$$\mu_t = E(Y_t) = exp\ (\mathbf{x}_t^{'}\mathbf{b}) \ \text{ and } \ Var(Y_t) = \mu_t + \mu_t^2\sigma_\epsilon^2 \tag{3.3}$$

The autocovariance function of this process is given by the formulae :

$$\gamma_Y(h) = \mu_t \mu_{t+h} \gamma_\epsilon(h) \tag{3.4}$$

From the covariance function, it is easy to obtain the autocorrelation function as follows:

$$\begin{aligned} \rho_Y(h) &= \frac{\mu_{t+h}\mu_t\gamma_\epsilon(h)}{\sqrt{[\mu_{t+h} + \mu_{t+h}^2\sigma_\epsilon^2][\mu_t + \mu_t^2\sigma_\epsilon^2]}} \\ &= \frac{\rho_\epsilon(h)}{\sqrt{[1 + (\sigma_\epsilon^2\mu_{t+h})^{-1}][1 + (\sigma_\epsilon^2\mu_t)^{-1}]}} \end{aligned} \tag{3.5}$$

We can see from (3.3) that the marginal variance of Y_t is greater than its marginal mean providing this way a degree of overdispersion which depends on the variance of the latent process σ_ϵ^2. Another interesting property of this model is that the form of the autocorrelation of the observed counts inherits its structure from that of the latent process. It is also obvious from (3.5) that even if there is no significant autocorrelation in y_t, it does not necessarily mean that autocorrelation is not present in ϵ_t either, since $|\rho_y(h)| \leq |\rho_\epsilon(h)|$. This implies that the autocorrelation function of the observed count process will tend to underestimate that of the latent process, even in the simplest case where no regressors are present. Methods for estimating the underlying autocorrelation and to test if it is zero or not are provided in the following section. The interpretation of any element of the vector of coefficients $\mathbf{b}$ in the above model is the same as in a simple Poisson regression model.

3.2 Estimation

The estimation of the model's unknown parameters is a very difficult task if we consider maximum likelihood estimation. Having in mind the model

described in (3.1) the likelihood is the T-fold integral

$$P(y_1, y_2, ..., y_T) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{t=1}^{T} P(y_t \mid \epsilon_t) P(\epsilon_1, \epsilon_2 \ldots \epsilon_T)\, d\epsilon_1 \ldots d\epsilon_T$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} exp \left\{ \sum_{t=1}^{T} [\, \mathbf{x}_t' \mathbf{b} y_t - \epsilon_t exp\, (\mathbf{x}_t' \mathbf{b})\,] \right\}$$

$$\times \left(\prod_{t=1}^{T} \epsilon_t^{y_t} \right) P(\epsilon_1, \epsilon_2 \ldots \epsilon_T)\, d\epsilon_1 \ldots d\epsilon_T / \prod_{t=1}^{T} y_t!$$

As is evident, the likelihood of the model can not be written down in closed form making the maximization impossible by direct numerical methods. To overcome this difficulty, Chan and Ledolter (1995) proposed an algorithm, called Monte Carlo EM (MCEM), whose iterates converge to the maximum likelihood estimate. The algorithm shares the same principle with a simple EM with the difference that the vector of the means of the latent process variables conditional on the observed data is estimated via Monte Carlo simulation. The difficult step in the algorithm is the generation of replicates of the latent process given the observed data. Chan and Ledolter discuss the use of the Gibb's sampler for generating the desired replicates and give some guidelines on the implementation of the algorithm.

This section considers a simpler method of estimation of the regression parameters $\mathbf{b}$, given consistent estimators of the covariance parameters $\boldsymbol{\theta} = (\sigma_\epsilon^2, \boldsymbol{\theta}_\rho)'$, where $\boldsymbol{\theta}_\rho$ completely specifies the autocorrelation function $\rho_\epsilon(\mathrm{h})$. This estimation procedure is the one that was suggested by Zeger (1988). The estimation of $\mathbf{b}$ is based on a similar procedure as the one followed in quasi-likelihood for independent data.

The main idea of the quasi-likelihood method is based on the quantity

$$U_t = \frac{Y_t - \mu_t}{Var(Y_t)}$$

which has three similar properties to the log-likelihood functions, namely

$$E(U_t) = 0, \qquad Var(U_t) = \frac{1}{Var(Y_t)} \qquad \text{and} \qquad -E\left(\frac{dU_t}{d\mu_t}\right) = \frac{1}{Var(Y_t)}.$$

Since most first-order asymptotic theory connected with likelihood functions is founded on these three properties, it is reasonable that, to some extent, the integral

$$Q = \int_y^{\mu} \frac{y - z}{Var(Y)} dz$$

if it exists, should behave like a log-likelihood function for μ. The quasi-likelihood method leads to the maximum likelihood estimates for many of the generalized linear models. For more details in quasi-likelihood method one can look McCullagh and Nelder (1989).

Hence, in the case of independent data, $\hat{\mathbf{b}}$ is the root of the p equations

$$U(\hat{b}_r) = \sum_{t=1}^{T} \frac{d\mu_t}{db_r} v_t^{-1}(y_t - \mu_t) = 0, \qquad r = 1, 2, ..., p \tag{3.6}$$

The quasi-likelihood estimator is consistent and asymptotically Gaussian. This approach is also robust in that consistent inferences can be made given only that $E(Y_t) = \mu_t$ whether or not $v_t = Var(Y_t)$. Now, by letting

$$\mathbf{Y} = (y_1, y_2, ..., y_T)', \ \boldsymbol{\mu} = (\mu_1, \mu_2, ...\mu_T)', \ \mathbf{V} = Var(Y)$$

equation (3.6) can be rewritten

$$\mathbf{U}(\hat{\mathbf{b}}) = \mathbf{D}'\mathbf{V}^{-1}(\mathbf{Y} - \mu) = 0 \tag{3.7}$$

where $\mathbf{D}$ is a matrix whose components are $\mathbf{D}_{ir} = \frac{d\mu_i}{db_r}$. With independent

data $\mathbf{V}$ is diagonal. With time series data, $\mathbf{V}$ will include off-diagonal terms which depend on nuisance parameters. Specifically if $\mathbf{R}_\epsilon$ is a $n \times n$ matrix with j,k elements $\rho_\epsilon(|j-k|)$ then $\mathbf{V} = Var(\mathbf{Y}) = \mathbf{A} + \sigma_\epsilon^2 \mathbf{A}\mathbf{R}_\epsilon\mathbf{A}$, where $\mathbf{A} = diag(\mu_1, \mu_2, ...\mu_T)$. As we can see equation (3.7) depends on $\mathbf{b}$ but also on the nuisance parameters $\boldsymbol{\theta}$ through $\mathbf{V}$. To compute $\hat{\mathbf{b}}$ for a given value of $\hat{\boldsymbol{\theta}}$ we need to solve the system of equations

$$\mathbf{U}(\hat{\mathbf{b}}) = \mathbf{D}'\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}})\ (\mathbf{Y} - \boldsymbol{\mu}) = 0 \tag{3.8}$$

An iterative weighted procedure can be used as in the case of quasi-likelihood with independent data. The parameter estimates at the (j+1)st iteration, $\hat{\mathbf{b}}^{(j+1)}$, are given by

$$\hat{\mathbf{b}}^{(j+1)} = \hat{\mathbf{b}}^{(j)} + (\mathbf{D}'\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}})\ \mathbf{D})^{-1}\mathbf{U}(\hat{\mathbf{b}}^{(j)}) \tag{3.9}$$

where it can be shown that the inversed matrix is the asymptotic covariance matrix of $\hat{\mathbf{b}}$. So

$$Var(\hat{\mathbf{b}}) = (\mathbf{D}'\mathbf{V}^{-1}(\hat{\boldsymbol{\theta}})\ \mathbf{D})^{-1} \tag{3.10}$$

Given an estimation procedure for $\boldsymbol{\theta}$, $\hat{\mathbf{b}}$ is found by alternately solving (3.8) for $\hat{\mathbf{b}}^{(j+1)}$ given $\hat{\boldsymbol{\theta}}^{(j)}$, then using the updated $\hat{\mathbf{b}}^{(j+1)}$ to find $\hat{\boldsymbol{\theta}}^{(j+1)}$ until convergence.

A drawback of (3.9) is its solution requires inversion of the $n \times n$ covariance matrix, $\mathbf{V}$. We are unaware of an efficient algorithm for inverting matrices with the structure of $\mathbf{V}$. Hence, we consider an approximation to (3.8) that is computationally simpler and leads to nearly efficient estimators in many practical cases.

Inversion of $\mathbf{V}$ is difficult because the parameter-driven process does not have a stationary autocorrelation function. To simplify calculations, we approximate the actual autocorrelation matrix, $\mathbf{R}_\epsilon$, by a band diagonal matrix, corresponding to an autoregressive process. Let $\mathbf{B} = diag(\mu_t + \sigma_\epsilon^2\mu_t^2)$. We approximate $\mathbf{V}$ with $\mathbf{V_R} = \mathbf{B^{1/2}R}(\boldsymbol{\alpha})\mathbf{B^{1/2}}$, where $\mathbf{R}(\boldsymbol{\alpha})$ is the

autocorrelation matrix of a stationary autoregressive process with an $s \times 1$ vector of parameters, $\boldsymbol{\alpha}$. Let $\boldsymbol{\theta_R} = (\sigma_\epsilon^2, \boldsymbol{\alpha})$ and define $\hat{\mathbf{b}}_R$ to be the solution of the estimating equation

$$\mathbf{U}(\hat{\mathbf{b}}_\mathbf{R}) = \mathbf{D}'\mathbf{V}_\mathbf{R}^{-1}(\hat{\boldsymbol{\theta}}_\mathbf{R})\ (\mathbf{Y} - \boldsymbol{\mu}) = 0 \tag{3.11}$$

Note that the algorithm for finding $\hat{\mathbf{b}}_R$ given $\hat{\boldsymbol{\theta}}_\mathbf{R}$ is greatly simplified. The inverse of the matrix $\mathbf{V_R}$, ignoring edge effects, satisfies

$$\mathbf{V}_\mathbf{R}^{-1} = \mathbf{B}^{-1/2}\mathbf{L}\mathbf{L}'\mathbf{B}^{-1/2} \tag{3.12}$$

where $\mathbf{L}$ is the matrix which applies the autoregressive filter, i.e the elements of $\mathbf{Ly}$ are

$$y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} \quad (t > p)$$

The iterative procedure has now the form

$$\hat{\mathbf{b}}_\mathbf{R}^{(j+1)} = \hat{\mathbf{b}}_\mathbf{R}^{(j)} + (\mathbf{D}'\mathbf{B}^{-1/2}\mathbf{L}\mathbf{L}'\mathbf{B}^{-1/2}\mathbf{D})^{-1}\mathbf{U}(\hat{\mathbf{b}}_\mathbf{R}^{(j)}) \tag{3.13}$$

Under the assumption that ϵ_t is a stationary process and given $\sqrt{n}\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \mathbf{o}_p(1)$ for some fixed $\boldsymbol{\theta}$, then $\hat{\mathbf{b}}$ is asymptotically multivariate Gaussian with mean the true $\mathbf{b}$ and covariance matrix

$$Var(\hat{\mathbf{b}}_\mathbf{R}) = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1} \tag{3.14}$$

where $\mathbf{I_0} = \mathrm{D}'\mathrm{V}_\mathrm{R}^{-1}\mathrm{D}$ and $\mathbf{I_1} = \mathrm{D}'\mathrm{V}_\mathbf{R}^{-1}\mathrm{V}\mathrm{V}_\mathrm{R}^{-1}\mathrm{D}$. The more complicated form of the asymptotic covariance matrix is due to the fact that $\mathbf{V_R}$ is not the actual covariance matrix. '

3.3 Estimation of nuisance parameters

Zeger proposed estimation of the nuisance parameters by a method of moments. Note that $Var(Y_t) = \mu_t + \mu_t^2\sigma_\epsilon^2$. Hence σ_ϵ^2 can be estimated by

$$\hat{\sigma_\epsilon^2} = \frac{\sum_{t=1}^{T}[(y_t - \hat{\mu}_t)^2 - \hat{\mu}_t]}{\sum_{t=1}^{T}\hat{\mu}_t^2} \tag{3.15}$$

The autocorrelation function of ϵ_t can similarly be estimated by

$$\hat{\rho}_\epsilon(h) = \frac{\sum_{t=h+1}^{T}[(y_t - \hat{\mu}_t)(y_{t-h} - \hat{\mu}_{t-h})]}{\hat{\sigma}_\epsilon^2 \sum_{t=h+1}^{T}\hat{\mu}_t\hat{\mu}_{t-h}} \tag{3.16}$$

In many cases, $\rho_\epsilon(h)$ is fully specified in terms of fewer parameters $\boldsymbol{\theta}$. For example, if the latent process is assumed to be a integer autoregressive of order p with vector of coefficients θ then by the Yule-Walker equations we can estimate these parameters. One limitation of moment estimation is that $\hat{\sigma_\epsilon^2}$ can be negative and $\hat{\rho}_\epsilon(h)$ is not constrained to the interval (-1,1). When the sample size is small and $|\rho_\epsilon(h)|$ is large, a different approach may be needed. However, this is unlikely to happen if a test for the existence of a latent process is proved significant.

In order to test for the significance of autocorrelation in the latent process one needs the variance of the estimate of the autocovariance function $\hat{\gamma}_\epsilon(h)$ which is given by

$$Var(\hat{\gamma}_\epsilon(h)) = \frac{1}{(\sum_{t=1}^{T-h}\hat{\mu}_t\hat{\mu}_{t+h})^2}\sum_{t=1}^{T-h}\hat{\mu}_t^2\hat{\mu}_{t+h}^2(\hat{\sigma_\epsilon}^2 + \frac{1}{\hat{\mu}_t})(\hat{\sigma_\epsilon}^2 + \frac{1}{\hat{\mu}_{t+h}}) \tag{3.17}$$

Under the assumption that the latent process is white noise with positive

variance the $\hat{\gamma}_\epsilon(h)$ are asymptotically distributed as independent random variables with mean $\gamma_\epsilon(h)$ and variance estimated by (3.17). For more details of the inference on autocovariances of the latent process see Davis *et al* (1999).

The model was first applied by Zeger (1988) to modelling the trend in U.S. polio incidence by using linear and trigonometric functions of time as covariates. A similar analysis was carried out by Davis *et al.* (2000) with an application to daily asthma counts at a hospital in Sydney. As we will see later, this model will prove of great use in modelling daily accident counts as a function of weather variables.

3.4 Testing for the existence of a latent process

Prior to the estimation of the nuisance parameters it is reasonable to test for the existence of a latent process. Brannas and Johanson (1994) review the following statistic

$$S = \frac{\sum_{t=1}^{T}[(y_t - \hat{\mu}_t)^2 - y_t]}{\sqrt{2\sum_{t=1}^{T}\hat{\mu}_t^2}}$$

derived by several authors and based on a local alternative hypothesis or the Lagrange multiplier test of the Poisson distribution against a negative binomial distribution. A variant was introduced by Dean and Lawless (1989) in order to improve the small sample performance of the test

$$S_\alpha = \frac{\sum_{t=1}^{T}[(y_t - \hat{\mu}_t)^2 - y_t + \hat{h}_t\hat{\mu}_t]}{\sqrt{2\sum_{t=1}^{T}\hat{\mu}_t^2}}$$

where h_t is the tth diagonal element of the GLM "hat" matrix as defined in Fahrmeir and Tutz (1994), for example, as

$$\mathbf{H} = \mathbf{A}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}^{1/2}$$

where $\mathbf{A} = diag(\mu_1, \mu_2, ...\mu_T)$ and $\mathbf{X}$ the design matrix. This statistic is asymptotically distributed as a $N(0,1)$ variate under the null hypothesis of no latent process and it is used in a one sided test. Simulations have shown that S_α has better size properties in small samples.

Another test introduced in Davis *et al* (1999) was designed for overdispersion due to the existence of a latent process. Under the null hypothesis that there is no latent process the Pearson residuals

$$e_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\mu}_t}}$$

have approximately zero mean and unit variance. Hence the statistic

$$\mathcal{Q} = \frac{T^{-1}\sum_{t=1}^{T} e_t^2 - 1}{\hat{\sigma}_{\mathcal{Q}}}$$

where

$$\sigma_{\mathcal{Q}}^2 = (\frac{1}{T})(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\hat{\mu}_t} + 2)$$

can be used to test for a latent process. The expression for $\sigma_{\mathcal{Q}}^2$ can be derived using the fact that a Poisson random variable Y_t with mean μ_t has fourth central moment $E(Y_t - \mu_t)^4 = \mu_t + 3\mu_t^2$. Under the hypothesis that the variance of the hidden process is zero

$$\mathcal{Q} \sim N(0,1)$$

approximately. Studies based on simulated data in Davis *et al.* (1999) pro-

vided evidence of better size properties for S_α and we will use this for our analysis.

3.5 Diagnostics

This section considers measures of goodness of fit of the models based on the X^2 distribution. We will see that the types of diagnostics presented here, are similar to the residual diagnostics used in a standard GLM fit. We are referring to the Pearson residuals. These statistics are not the best way to assess the fit of the model described in this thesis. Nevertheless, we use them since we are mainly interested in a comparison between the two models.

According to the Poisson INAR(1) regression model, marginally each observation $Y_t \sim Poisson(\mu_t)$. Hence, we can obtain Pearson residuals by using

$$e_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{Var}(Y_t)}}$$

where $\hat{Var}(Y_t) = \hat{\mu}_t$. The well-known Pearson goodness of fit statistic is given as $\sum e_i^2$ which, under the assumption that the model fit is adequate, follows a X^2 distribution with degrees of freedom equal to the number of the observations minus the number of the estimated parameters (similarly to a standard GLM).

Another appealing definition of residual for the INAR model is described in Freeland and McCabe (2004a). By taking advantage of the decomposition of the observation in two unobservable processes, two sets of residuals are defined; one for each component. The natural way for this definition is as follows: for the continuation component let $\epsilon_{1t} = \alpha_t \circ Y_{t-1} - \hat{\alpha}_t Y_{t-1}$ and for the arrival component let $\epsilon_{2t} = R_t - \hat{\lambda}_t$. However, these definitions are not practical because $\alpha \circ Y_{t-1}$ and R_t are not observable. Nevertheless, they can be replaced with their conditional expectations given the observed values of Y_t and Y_{t-1}. These expectations have already been found through the imple-

mentation of the EM algorithm for the estimation of the Poisson INAR(1) regression model. Note that the sum of ϵ_{1t} and ϵ_{2t}, after the replacement with the conditional expectations, is equal to $y_t - \hat{\alpha}_t y_{t-1} - \hat{\lambda}_t$ which in fact is the difference between the observed and the fitted value.

The Pearson residual type for Zeger's model is the same as before with the difference that $\hat{Var}(Y_t) = \hat{\mu}_t + \hat{\mu}_t{}^2 \hat{\sigma}_\epsilon{}^2$. Other diagnostics for Zeger's model are extensively discussed in Davis it et al (1999). They propose use of standardized Pearson residuals which are defined as

$$\tilde{e}_t = \frac{e_t}{\sqrt{1 - \hat{h}_t}}$$

where e is the Pearson residual and $\hat{h}_t$ is the tth diagonal element of the GLM "hat" matrix.

The Pearson residuals can be further used to estimate the amount of the residual overdispersion; variability of the dependent variable that has not been explained by any of the available covariates. This can be achieved by dividing the Chi-square statistic with the respective degrees of freedom. The idea behind this is that under the hypothesis that there is no overdispersion the square of each residual is an estimate of unity. Hence their sum divided by the respective degrees of freedom should be a number lying close to unity. Large deviations from this are evidence of overdispersion.

While we will use this diagnostic in the sequel we must point out that asymptotic results are not valid, especially when the counts are very small. This is due to the fact that the denominator of Pearson residuals can be very small if the true counts are small which deviates from the asymptotic normality assumption. For this reason the diagnostic must be used with care.

Chapter 4

Meta-analytic methods

This chapter describes two of the most important models in the literature of meta-analytic methods; meta-analysis and meta-regression. These models are often used when the researcher is interested in combining specific results from related but independent studies and as we will see they are very similar to linear regressions. The following sections describe the theory and the properties of the meta-analysis and meta-regression models and present estimation procedures for each model's unknown parameters.

4.1 Meta-analysis

Meta-analysis can be defined as the quantitative review and synthesis of the results of related but independent studies. By combing information over different studies, an integrated analysis will have more statistical power to detect a specific effect than an analysis based on only one study. When several studies have conflicting conclusions, a meta-analysis can be used to estimate an average effect. Methods for providing such an overall estimate are well known, and have been extensively discussed from Fleiss (1993) from classical perspectives and from Smith *et al* (1995) from Bayesian perspectives. These methods are frequently used in the field of biostatistics where the researcher

combines several randomized control trials in order to detect the treatment effect produced by a specific therapy. In this case, a collection of studies related to this treatment is gathered and an overall effect is given by the methodology followed in the meta-analysis.

Once the primary studies have been collected, the meta-analyst needs to identify a summary measure common to all studies and subsequently combine the measure. Often there is little control over the choice of the summary measure because most of the decision is dictated by what was employed in the primary studies. For example if risk differences are reported in the primary studies instead of odds ratios, then the analyst has little choice but to utilize the average risk difference as the summary statistic in the meta-analysis. There are mainly two classes of measures. The first one consists of measures suitable for discrete outcome data such as difference in proportions and the second one for continuous data that may generally be thought of as means.

Another important issue to consider in a meta-analysis is the source of variation. There are at least three sources of variation to consider before combining summary statistics across studies. First, sampling error may vary among studies since different sample sizes may have been used in the primary studies resulting in estimated summaries with varying degrees of precision. Second, study level characteristics may differ creating reasons to believe that the effect is different among the studies. Third, there may exist inter-study variation. The fixed effects model presented in section 4.1.1 assumes that each study is measuring the same underlying parameter and that there is no inter-study variation. Conversely, the random-effects model introduced in section 4.1.3 assumes that each study is associated with a different but related parameter.

4.1.1 Fixed-effects model

A fixed-effects model assumes that each study summary statistic, Y_i is a realization from a population of study estimates with common mean θ. Let α be the central parameter of interest and assume there are $i = 1, 2, ..., k$ independent studies. Assume that Y_i is such that $E(Y_i) = \theta$ and let $Var(Y_i) = s_i^2$ be the variance of the summary statistic in the ith study. For moderately large study sizes, each Y_i should be normally distributed (by the central limit theorem) and approximately unbiased. Thus

$$Y_i \sim N(\alpha, s_i^2) \text{ for } i = 1, 2, ..., k \tag{4.1}$$

and s_i^2 assumed known. The central parameter of interest is α which quantifies the average effect.

4.1.1.1 Estimation

In order to estimate the common effect α, a maximum likelihood estimation will lead us to the obvious weighted average with weights $w_i = 1/s_i^2$. So

$$\hat{\alpha} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \tag{4.2}$$

Standard inferences about α are available using the fact $\hat{\alpha} \sim N(\alpha, (\sum_{i-1}^{k} w_i)^{-1})$.

4.1.2 Random-effects model

The random-effects model assumes that each study summary statistic Y_i is drawn from a distribution with a study-specific mean, α_i, and variance s_i^2.

$$Y_i \mid \alpha_i, s_i^2 \sim N(\alpha_i, s_i^2) \text{ for } i = 1, 2, ..., k \tag{4.3}$$

Furthermore, each study-specific mean α_i is assumed to have been drawn from some superpopulation of effects with mean α and variance τ^2 with

$$\alpha_i \mid \alpha, \tau^2 \sim N(\alpha, \tau^2) \tag{4.4}$$

The parameters α and τ^2 are to referred as hyperparameters and represent, respectively, the average effect and inter-study variation.

Note that given the hyperparameters, the distribution of each study summary measure, Y_i, after averaging over the study-specific effects, is normal with mean α and variance $s_i^2 + \tau^2$. As in the fixed-effects model, α is the parameter of central interest. However, the between-study variation, τ^2 plays an important role and must also be estimated. In addition, it is also possible to derive estimates of the study-specific effects, α_i, that are useful for inferences regarding identifying particularly effective studies. The distribution of α_i, conditional on the observed data and the hyperparameters is

$$\alpha_i \mid y_i, \alpha, \tau^2 \sim N(B_i\alpha + (1 - B_i)Y_i, s_i^2(1 - B_i)) \tag{4.5}$$

where $B_i = \frac{s_i^2}{s_i^2+\tau^2}$. This equation will prove very helpful in the estimation of the hyperparameters.

4.1.2.1 Estimation

The estimation of the parameters in the random-effects model is a little more complicated than the fixed-effects model due to the inter-study variation τ^2. Proposed methods for estimating τ^2 include methods of moments (DerSimonian and Laird (1986)) and restrictive maximum likelihood. We construct a very simple EM algorithm for the estimation of the hyperparameters and the estimates' standard errors are obtained via the expected information matrix. The algorithm can be described as follows.

- *E-step*: Using the current values of the estimates, say $\alpha^{old}, (\tau^{old})^2$, calculate

$$\begin{aligned} z_i &= E(\alpha_i \mid y_i, \alpha^{old}, (\tau^{old})^2) \\ &= B_i^{old}\alpha^{old} + (1 - B_i^{old})y_i \end{aligned}$$

where $B_i^{old} = \frac{s_i^2}{s_i^2 + (\tau^{old})^2}$. The above equation is obtained directly from equation (4.5) which shows the conditional distribution of α_i given the observed data and the hyperparameters.

- *M-step*: Update the central parameter α by

$$\alpha^{new} = \frac{\sum_{i=1}^{k} E(\alpha_i \mid y_i, \alpha^{old}, (\tau^{old})^2)}{k} = \frac{\sum_{i=1}^{k} z_i}{k}$$

and the between-study variation τ^2 by

$$\begin{aligned} (\tau^{new})^2 &= \frac{\sum_{i=1}^{k} E[(\alpha_i - \alpha^{old})^2 \mid y_i, \alpha^{old}, (\tau^{old})^2]}{k} \\ &= \frac{\sum_{i=1}^{k} [Var(\alpha_i \mid y_i, \alpha^{old}, (\tau^{old})^2) + (z_i - \alpha^{old})^2]}{k} \\ &= \frac{\sum_{i=1}^{k} [s_i^2(1 - B_i^{old}) + (z_i - \alpha^{old})^2]}{k} \end{aligned}$$

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

We saw earlier that the marginal distribution of y_i is normal with mean α and variance $s_i^2 + \tau^2$. This likelihood can be easily twice differentiated and we can derive the asymptotic standard errors of the estimated parameters by

the inverse of the expected information matrix. It can be easily shown that

$$Var(\hat{\alpha}) = [\ \sum_{i=1}^{k} 1/(s_i^2 + \tau^2)\ ]^{-1} \quad \text{and} \quad Var(\hat{\tau^2}) = 2\ [\ \sum_{i=1}^{k} 1/(s_i^2 + \tau^2)^2\ ]^{-1}$$

4.2 Meta-regression

The meta-analysis methods presented in the previous chapter attempt to combine the study-specific results in order to obtain a single summarized effect size. The observed effect in each study is an estimate, with some imprecision, of the true effect in that study. Statistical heterogeneity refers to the true effects in each study not being identical. Diversity among the studies included in a meta-analysis necessarily leads to statistical heterogeneity. In contrast to simple meta-analysis, meta-regression aims to relate the size of effect to one or more characteristics of the studies involved.

Various statistical methods for meta-regression have been published. For example, fixed effects meta-regression was described originally by Greenland (1987), a random effects model more recently by Berkley *et al.* (1995) and a fuller comparison of available methods made subsequently by Thompson and Sharp (1999). In the next sections we present the meta-regression models, both fixed and random effects and their estimation techniques through maximum likelihood estimation.

4.2.1 Fixed Effects Model

The fixed effects model is no different in nature from a simple linear regression. Similar to the meta-analysis, assume we have k studies plus some characteristics of each study. Let's denote the available information from i study, $i = 1, 2, ...k$ in the form of covariates, so that $\mathbf{x}_i$ is the vector of available information on the ith study. Also, let y_i and s_i be the summary statistic and its variance observed from the ith study. The fixed effects model

assumes that

$$Y_i \sim N(\alpha + \mathbf{x}_i'\boldsymbol{\beta}, s_i^2) \quad \text{for} \quad i = 1, 2, ..., k \tag{4.6}$$

where $(\alpha, \boldsymbol{\beta})$ are the parameters of interest. The main difference between this fixed effects meta-regression model and the fixed effects meta-analysis model is the additional vector of unknown coefficients $\boldsymbol{\beta}$ which describes the relation of the summary statistics with the explanatory variables. The model described in (4.6) is exactly the same with a linear regression model with heteroscedastic errors with known variances.

4.2.1.1 Estimation

The estimation of the model is feasible through maximum likelihood estimation and it leads, as expected, to the weighted least square estimator. By letting $Y = (y_1, y_2, ..., y_k)'$, $\mathbf{V} = diag(s_1^2, s_2^2, ..., s_k^2)$, $\mathbf{b} = (\alpha, \boldsymbol{\beta})$ and the $k \times p$ design matrix (including the constant) as $\mathbf{X}$ then

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$$

The inference on the coefficients can be easily extracted by the fact that $\hat{\mathbf{b}} \sim N_p(\mathbf{b}, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1})$. Due to the fact that the weights are considered known there is no need to multiply the standard errors of the coefficients with the mean square error, since it is equal to unity.

4.2.2 Random Effects Model

This model is an extension of the simple meta-analysis random effects model and it can account for residual heterogeneity even after the inclusion of significant covariates. The meta-regression random effects model assumes that the intercept is a random variable and follows a normal distribution with mean α and variance τ^2. These parameters, along with the unknown coefficients, are the parameters of main interest in this model. According to the

previous notation, the model assumes that

$$Y_i \mid \alpha_i \sim N(\alpha_i + \mathbf{x}_i'\boldsymbol{\beta}, s_i^2) \quad \text{for} \quad i = 1, 2, ..., k$$
$$\alpha_i \mid \alpha, \tau^2 \sim N(\alpha, \tau^2)$$

The marginal distribution of y_i is normal with mean $\alpha + \mathbf{x}_i'\boldsymbol{\beta}$ and variance $s_i^2 + \tau^2$. The between study variation τ^2 must be estimated and a EM algorithm will be provided in the next subsection for all the parameters' estimation. Similar to the meta-analysis random effects model, we can derive estimates of the study specific α_i given y_i. These estimates can be used in identifying studies with homogenous effects. The conditional distribution of α_i given y_i and the parameters is

$$\alpha_i \mid y_i, \alpha, \boldsymbol{\beta}, \tau^2 \sim N(B_i\alpha + (1 - B_i)(y_i - \mathbf{x}_i'\boldsymbol{\beta}), s_i^2(1 - B_i)) \tag{4.7}$$

where $B_i = \frac{s_i^2}{s_i^2 + \tau^2}$.

4.2.2.1 Estimation

Most estimation schemes for random effects model (not only for meta-regression) are based on restricted maximum likelihood (REML) due to the difficulty of the unobserved random variables. Nevertheless, maximum likelihood estimates can be derived with an EM algorithm. We will see that the two steps of the algorithm for the estimation of the meta-regression random effects model have many similarities with the steps of the estimation of the meta-analysis model. Note that if we could observe the random variables α_i for $i = 1, 2, ..., k$ then the estimation of the hyperparameters (α, τ^2) would be very easy; so would be the estimation of $\boldsymbol{\beta}$. Let's describe the two steps of the EM algorithm using the same notation as before.

- *E-step*: Using the current values of the estimates, say $\alpha^{old}, \boldsymbol{\beta}^{old}, (\tau^{old})^2$, calculate

$$\begin{aligned} z_i &= E(\alpha_i \mid y_i, \alpha^{old}, \boldsymbol{\beta}^{old}, (\tau^{old})^2) \\ &= B_i^{old}\alpha^{old} + (1 - B_i^{old})(y_i - \mathbf{x}_i\boldsymbol{\beta}^{old}) \end{aligned}$$

where $B_i^{old} = \frac{s_i^2}{s_i^2 + (\tau^{old})^2}$. The above equation is obtained directly from equation (4.7) which notes the conditional distribution of α_i given the observed data and the hyperparameters.

- *M-step*: Update the central parameter α by

$$\alpha^{new} = \frac{\sum_{i=1}^{k} E(\alpha_i \mid y_i, \alpha^{old}, \boldsymbol{\beta}^{old}, (\tau^{old})^2)}{k} = \frac{\sum_{i=1}^{k} z_i}{k}$$

the between-study variation τ^2 by

$$\begin{aligned} (\tau^{new})^2 &= \frac{\sum_{i=1}^{k} E[(\alpha_i - \alpha^{old})^2 \mid y_i, \alpha^{old}, \boldsymbol{\beta}^{old}, (\tau^{old})^2]}{k} \\ &= \frac{\sum_{i=1}^{k} [Var(\alpha_i \mid y_i, \alpha^{old}, \boldsymbol{\beta}^{old}, (\tau^{old})^2) + (z_i - \alpha^{old})^2]}{k} \\ &= \frac{\sum_{i=1}^{k} [s_i^2(1 - B_i^{old}) + (z_i - \alpha^{old})^2]}{k} \end{aligned}$$

By letting $\mathbf{X}$ denote the design matrix, **without** the inclusion of the intercept, the new parameter estimate of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}^{new} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{Z})$$

where $\mathbf{Z} = diag(z_1, z_2, ...z_k)$.

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

  The standard errors of the estimated parameters are found by the inverse of the information matrix.

4.3 Model selection

The analyst has to decide whether the inference of the combined effect will be based on the fixed or the random effects model. The fixed effects model assumes that the study-specific summary statistics share a common α. A statistical test for the homogeneity of the study means is equivalent to testing

$$\begin{aligned} H_0: &\quad \alpha_1 = \alpha_2 = \ldots = \alpha_k = \alpha \quad \text{against} \\ H_1: &\quad \text{At least one } \alpha_i \text{ different} \end{aligned}$$

Under H_0, for large sample sizes, the quantity $\mathrm{Q} = \sum_{i=1}^{k} w_i(y_i - \hat{\mu}_i)^2$ follows a X^2_{k-p}, where p and $\hat{\mu}_i$ are the number of unknown parameters and the fitted values of a fixed effects model and $w_i = 1/s_i^2$. If Q is greater than the $100(1-\alpha)$ percentile of the X^2_{k-p} distribution, then the meta-analyst may conclude that the study means arose from two or more distinct populations and proceed by either attempting to identify covariates that stratify studies into the homogenous populations or estimating a random effects model. If H_0 cannot be rejected the investigator would conclude that the studies share a common mean α. This test has low power against the alternative $var(\alpha_i) > 0$, this is why we will consider the heterogeneity significant at the 10% level instead of 5%.

Another criterion for choosing between the two models is the well-known Akaike Information Criterion (AIC) which is equal to minus twice the log-likelihood plus twice the number of the estimated parameters. This implies that the smaller the value of the AIC the better this model is. This criterion

can be easily evaluated for both fixed and random effects models since their respective likelihoods are the product of k independent normal distributions.

Chapter 5

Data Analysis

The aim of this chapter is to identify common points as well as possible differences between the two time-series models we described, the Poisson INAR(1) regression model and Zeger's model. We divide the current chapter into six sections. The first one is a description of the data and the explanatory variables that will be included in the time series models. The second one is a preliminary analysis of the data in order to discover some useful aspects of the data. The third section presents the results of the models we estimate with the two methods, INAR's and Zeger's respectively and a comparison between their estimates is made. The fourth section examines the goodness of fit of the two models for every station with the help of the Pearson residuals. The last two sections consist of the meta-analysis and meta-regression results respectively.

5.1 Data Description

This study is based on the daily accident counts that were obtained from the major roads covered by the surface of 27 big cities in the Netherlands in the year 2001. The cities were selected based on two criteria. Firstly, their proximity to some national weather stations 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. The location of each station in the Netherlands can be seen in the map (Figure 5.1).

With respect to weather conditions, the daily weather observations were obtained from the Dutch National Meteorological Institute. More specifically, the following variables were created from the data and considered for inclusion in the model, based on previous research where they have shown to be important/significant or at least hypothesized as being influential towards predicting the number of accidents. 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.* (1995b) and Baker and Reynolds (1992). The literature teaches 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 and the daily mean windspeed in 0.1 m/s. Note that we transformed the values of wind direction using the cosine of twice the wind degrees in order to equalize the effect of degrees that differed by 180 units.

- *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.

- *precipitation.* Rainfall has found to be a significant predictor for road accidents in many studies (see e.g. Fridstrøm *et al.*, 1995; Levine *et al.*, 1995b; Satterthwaite, 1976). We use precipitation duration in 0.1 hour

Figure 5.1: Location of the stations on the Netherlands

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.

- *humidity.* We refer to humidity as the percentage relative humidity which means how much moist is in the air. It depends on the season, temperature and the wind direction. If the humidity is high during winter periods, it can be dangerous because fog can start freezing on the road.

- *radiation.* Radiation refers to how much sun reaches the earth; on sunny days during summer the radiation is very high. So, this variable in some sense measures how intense the sun is burning on a day.

5.2 Preliminary analysis

Figures 5.2 and 5.3 show the accident series for the sites under study. It is apparent that the mean daily accident count is different between the sites. Table 5.1 shows some of the data characteristic for the sites. Clearly, there are differences between them. Firstly, for most of the sites the ratio of the variance to the mean is larger than 1 implying overdispersion relative to the simple Poisson distribution. An overdispersed model, like the negative binomial model could be used to account for the overdispersion. However, after fitting the Poisson INAR(1) regression model, it may turn out that the remaining overdispersion is no longer significant. The reason is that the covariates used for modelling the data will possibly explain the overdispersion to a large extent. Secondly, there is a large difference between the autocorrelation of the sites. Note also that in the majority of the sites the autocorrelation is not negligible and, thus, fitting a time series model is highly advocated. Also,

there seems to be a positive correlation between the mean of the accidents and the autocorrelations which can be seen in Figure 5.4. The autocorrelations reported are of the first order. Higher order autocorrelations were not large, apart from the autocorrelations for lags multiple of 7, which in some sense indicates the effect of the day. This is why we will include the days as dummy variables in the covariates using corner point parametrization and setting as reference level Sunday. Usage of these variables is also important as they provide information about the day specific traffic volume.

Figure 5.5 shows the boxplots for some of the weather variables for all the sites. The boxplots corresponds to the sites and each figure to a variable. The four variables presented in the plot are the mean temperature, the precipitation duration, the daily precipitation amount and the mean windspeed. It can be seen that there are some differences between the stations' covariates, especially for wind speed.

Before proceeding in the analysis we must point out some important aspects related to traffic accident analysis. First of all,many researchers have questioned whether accidents can be autocorrelated themselves based on the random nature of accidents. We emphasize that observed data show significant autocorrelation which can be interpreted due to sharing same environmental and infrastructure conditions. For example, à road with significant problems in its surface may be the reason for producing more accidents which are related at time because for example when it rains it is very slippery. Thus the observed autocorrelation can be attributed to the underlying conditions that produced the accidents. Other authors have reported negative relations between successive accident counts in road segments in the sense that if somewhere there were important accidents the next days the drivers are more alert and drive more carefully.

Moreover it is well established in the accident literature that exposure, i.e. the number of cars in the road is an important factor related to the accident counts. It is reasonable tat the more cars in the road the more

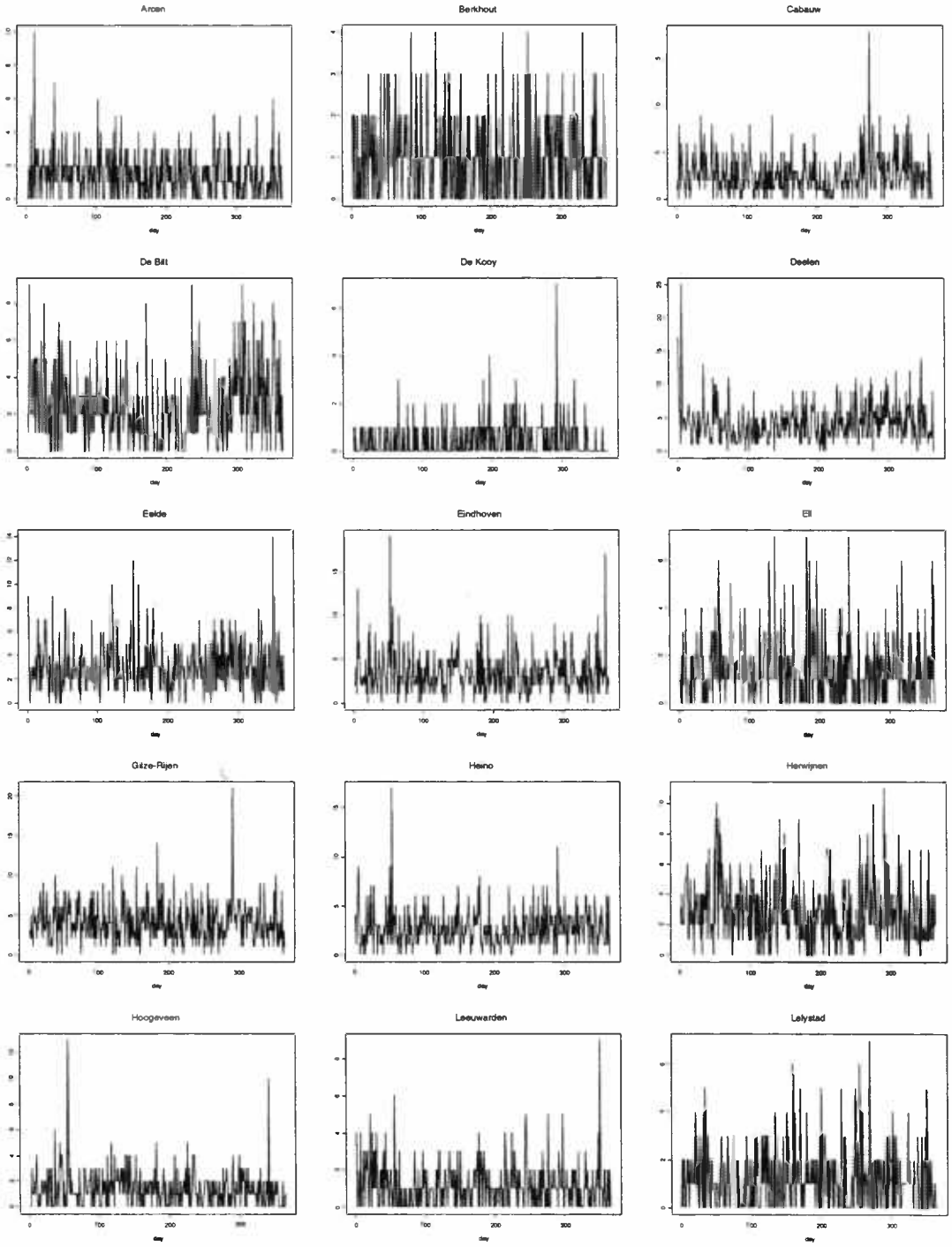

Figure 5.2: Accidents counts for the stations (a)

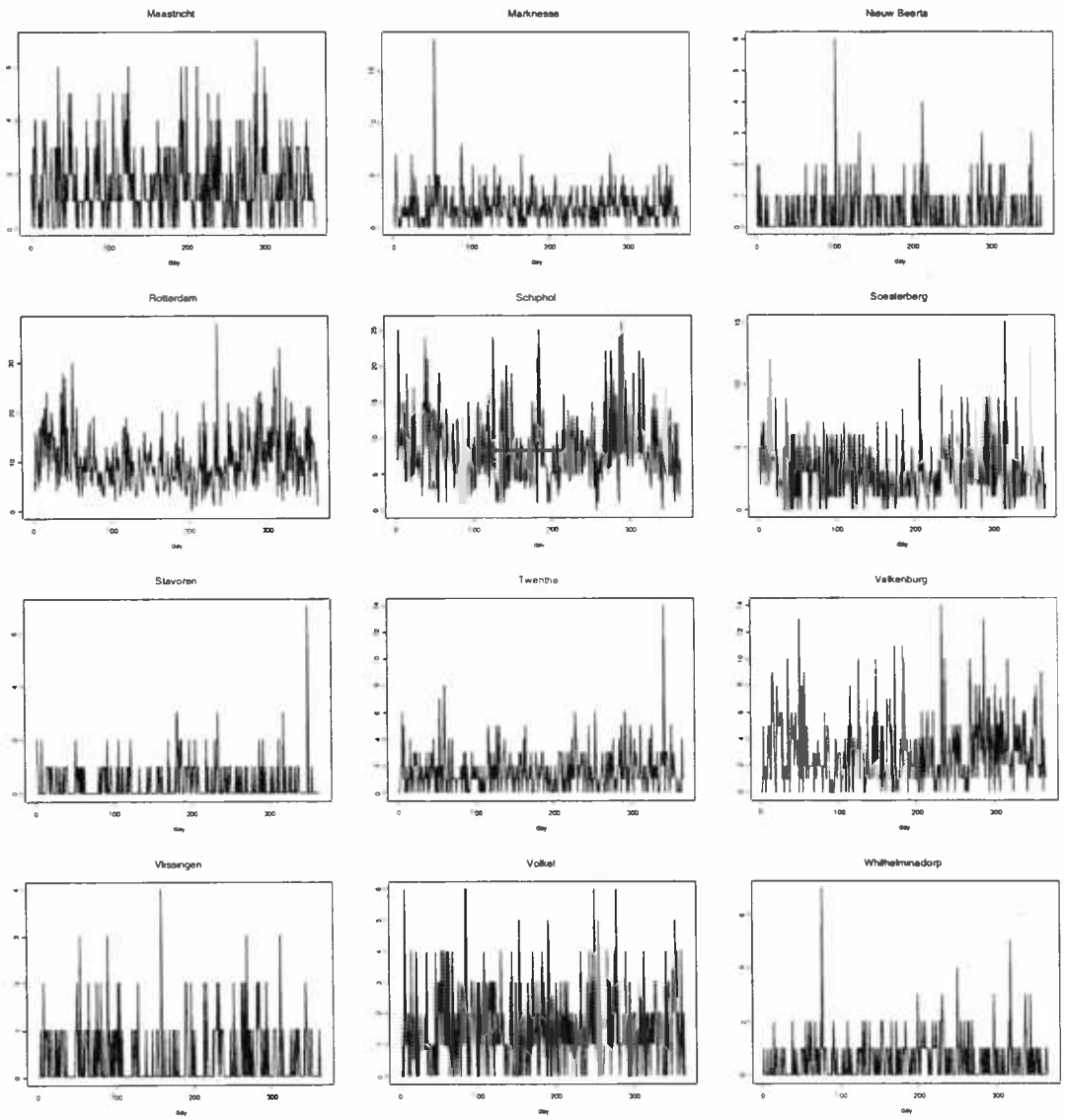

Figure 5.3: Accidents counts for the stations (b)

Station	mean	variance	autocorrelation	variance/mean
Arcen	1.463	1.969	-0.037	1.346
Berkhout	1.014	1.008	0.035	0.994
Cabauw	2.699	4.942	0.072	1.831
De Bilt	2.564	3.681	0.105	1.435
De Kooy	0.478	0.594	-0.039	1.242
Deelen	4.214	8.756	0.092	2.078
Eelde	3.170	4.455	-0.012	1.405
Eindhoven	3.438	6.379	0.012	1.855
Ell	1.581	2.107	0.040	1.333
Gilze-Rijen	4.088	6.630	0.035	1.622
Heino	2.710	4.031	0.067	1.488
Herwijnen	2.964	4.320	0.032	1.457
Hoogeveen	1.416	2.161	0.116	1.526
Leeuwarden	1.134	1.485	0.007	1.309
Lelystad	1.296	1.594	0.014	1.230
Maastricht	1.638	2.094	0.037	1.278
Marknesse	1.940	3.293	0.021	1.698
Nieuw Beerta	0.460	0.551	-0.018	1.198
Rotterdam	10.153	35.141	0.152	3.461
Schiphol	8.781	24.60	0.193	2.802
Soesterberg	3.370	5.673	0.070	1.684
Stavoren	0.381	0.506	-0.018	1.328
Twenthe	1.578	2.530	0.015	1.603
Valkenburg	3.044	6.537	0.012	2.147
Vlissingen	0.416	0.474	-0.026	1.139
Volkel	1.521	1.849	0.009	1.216
Wilhelminadorp	0.586	0.710	0.024	1.211

Table 5.1: Descriptive measures for the stations

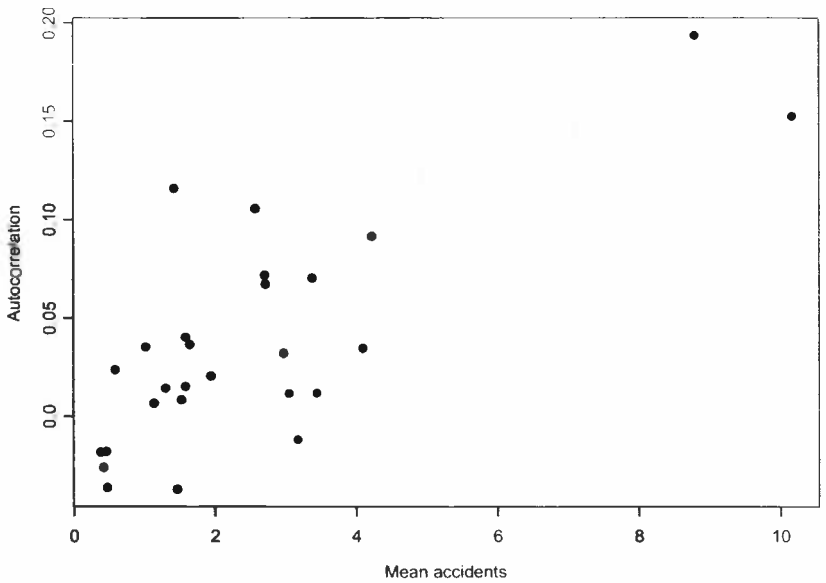

Figure 5.4: Plot of the mean accidents versus the autocorrelation

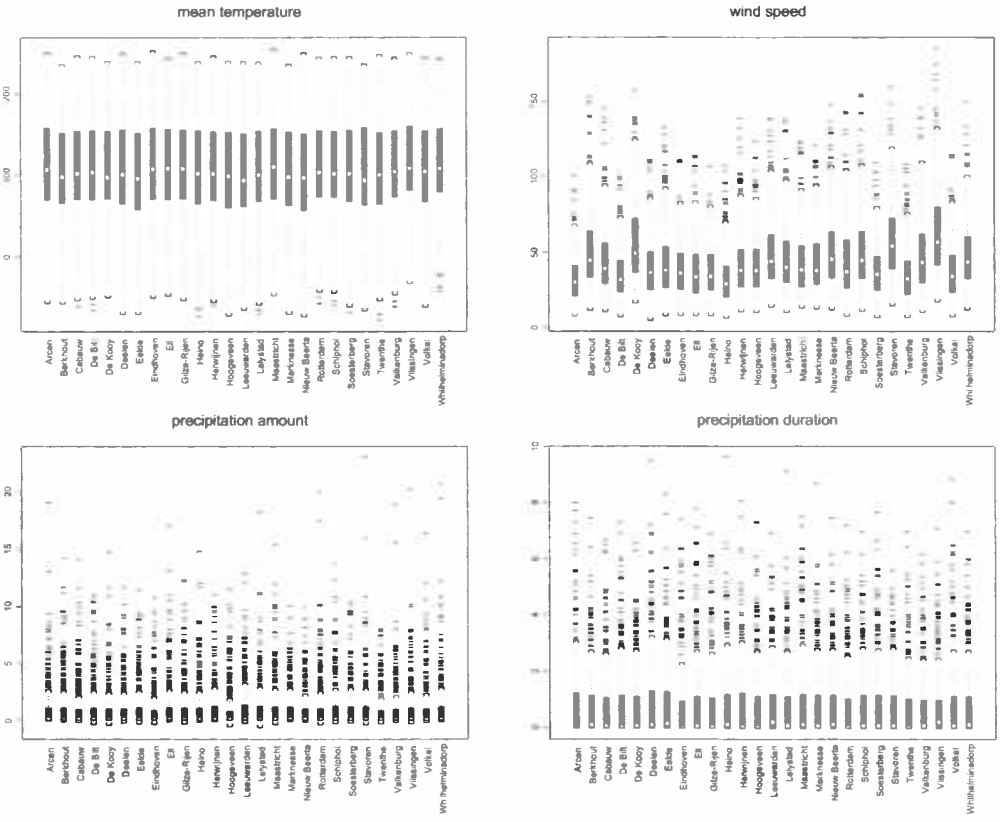

Figure 5.5: Boxplots of some of the weather variables for the stations

accidents we observe, but again a sort of congestion may occur where due to small speed the accidents can be reduced. So, exposure is important but unfortunately we do not have data related to traffic volumes for all areas in a daily basis. To account for this we use as proxy variables the days since it is observed that traffic volumes relate to the day significantly. However one must be cautious when interpreting observed effects. For example high temperatures are usually related to more cars in the road and thus we expect more accidents. So, from one hand the exposure is higher and thus we expect more accidents. So the effect of the high temperature is that it may increase the accidents but not directly but indirectly via the increase of the exposure.

Concluding interpretation of the observed effects must be very cautious since car accidents are very complicated phenomena.

5.3 INAR and Zeger results

In this chapter we present the results of the fitted models applying both INAR and Zeger regressions and we note significant differences between them. To be more specific, we will use all the methodology described in chapter 2 and 3. For Zeger's model, we assume that the underlying process δ_t that passes through the autocorrelation to the observed process is a stationary log-normal autoregressive model of first order with autocorrelation parameter ϕ and that $\epsilon_t = exp(\delta_t)$ with autocorrelation parameter of first order α. The estimated parameters of the models for each of the sites are placed side by side so that the reader can make a direct comparison between the two models' results. The bottom part of all the tables consists of the estimated autocorrelation parameters α for the INAR's and Zeger's model and the variance σ_ϵ^2 of the latent process. The inference of the autocorrelation parameter in Zeger's model is based on the asymptotical normality of the estimated autocovariance of order 1 with mean the true autocovariance and variance defined in (3.17). So, the standard errors and t-values reported for this parameter are referred to the autocovariance of first order and not to the autocorrelation parameter α directly. We suggestively present the results only for four of the stations. The results for all the stations can be found in Appendix B.

We begin the analysis with the test for the existence of a latent process, as described in section 3.4. Specifically, we fitted a Poisson regression model to the data ignoring the possible presence of a latent process and then evaluated the Q statistic. The third column of Table 5.2 shows the significance of the test for each station. The test is highly significant for all stations except for Berkhout, implying both the presence of a latent process and overdispersion. Hence, Zeger's model is highly advocated.

Let's begin the comparison between the two models for site De Bilt (Table 5.3). The last row of the table shows the inference on the estimated autocorrelation parameter α as well as the estimated value of the variance of the latent process σ_ϵ^2. The INAR model suggests that the dependence among the data is not significant ($\hat{\alpha} = 0.0435$, p-value $= 0.326$) while Zeger's model shows a large correlation on the hidden process ($\hat{\alpha} = 0.8199$, p-value $= 0.034$). This diversity is explained by the fact that the autocorrelation of the latent process will always be greater than the autocorrelation of the observed process, as noted in Chapter 3. Zeger's model also assumes the presence of overdispersion due to the existence of the latent process. Its estimated variance is equal to 0.0618 which, although being so small, has a significant effect on the standard errors of the estimated coefficients. Note that the standard errors of Zeger's estimated parameters are systematically greater than the INAR's respective. However, both models have indicated significant the same explanatory variables. Specifically, INAR model suggests that if the temperature reaches below zero, it increases the mean accidents by 27.978% (p-value $= 0.087$) while Zeger implies a larger increase by 32.152% (p-value $= 0.072$). Also, an increase of the rainfall duration by one unit increases significantly the mean accidents by 8.804% (p-value $= 0.003$) and by 11.427% (p-value < 0.001), according to INAR and Zeger's model respectively. Moreover, the inclusion of the effects of the weekdays are highly significant and they indicate a bigger number of accidents with respect to Sunday.

Table 5.4 depicts the estimated parameters for the two time-series models for Deelen's station. Both models suggest that significant autocorrelation is present and should be accounted for. Zeger's model has provided a larger estimate of the autocorrelation parameter ($\hat{a} = 0.2547$, p-value $= 0.041$) than the INAR's ($\hat{a} = 0.0764$, p-value $= 0.022$) for the same reason as for De Bilt's station. The variance of the latent process is quite larger, $\hat{\sigma}_\epsilon^2 = 0.1808$ and we can see how it decreases the precision of the estimated coefficients relatively to the INAR model. Highly significant, besides the weekdays, is the

precipitation duration effect which, according to the INAR model, increases the number of the accidents by 8.797% (p-value < 0.001) and by 8.256% (p-value $= 0.005$) according to Zeger's model.

Carrying on to Rotterdam's station, we can see from Table 5.5 that both the INAR and Zeger's regression models have identified significant presence of autocorrelation. The standard errors of the estimates of Zeger's model are still larger than the INAR's due to the extra variability that this parameter driven model assumes. Considering the effect of the covariates, INAR model shows that an increase of the mean temperature by one unit will lead to a decrease of the mean accidents by 0.111% (p-value $= 0.036$) and Zeger's model by 0.131% (p-value $= 0.062$). Radiation is significantly reducing the mean accidents by 0.492% (p-value < 0.001), as shown from INAR model, and by 0.393% (p-value $= 0.005$) from Zeger's model. Precipitation's duration and intensity are found to be highly significant for this station increasing the mean accidents by 14.509% and 14.201% for INAR and by 15.021% and 13.69% for Zeger's model.

Finally, Table 5.6 presents the results for Schiphol's station. The estimated values of the autocorrelation parameters for both models are highly significant. Similar to the previous analyses, the standard errors of Zeger's estimated parameters are greater than the INAR's respective. The INAR estimated coefficient of the humidity covariate is significant and results in a decrease of the mean accidents by 0.652%, which may seem contradicting. This estimate can be misinterpreted since it implies that higher percentages of moist in the air leads to less mean accidents. However, we can not be certain of this result due to the absence of exposure data during that time. This observation will be further discussed in Chapter 6 where we note the limitations that rise from the absence of exposure variables. Radiation is also significant, as the INAR model shows. Specifically, an increase of the daily mean radiation by one unit decreases the mean accidents 0.232% (p-value $= 0.059$). Moreover, rainfall duration and intensity have proven significant

from both models, increasing the mean accidents by 19.346% and 4.557% (INAR model) and by 18.242% and 4.708% (Zeger's model).

5.4 Diagnostics

This section assesses the goodness of fit of the two models by means of Pearson residuals and plots of the fitted and observed values for the sites that were presented in the previous chapter. Figures 5.6 and 5.7 show the plots of the observed and fitted values for the two models for each site analysed in the previous section. The thick lines represents the fitted values while the thin lines the observed data. We can see that both models' fitted values follow the same pattern with the actual data. Also, the fitted values of the two models do not seem to differ significantly. Recall that the estimated coefficients of both INAR and Zeger model were different mostly between their standard errors.

Figure 5.8 shows the plots of the Pearson residuals versus time. Both INAR and Zeger residuals are plotted in the same graph so that the reader can compare easily the two residual series; the thick one is for Zeger's model and the thin one for the INAR model. We can see from the graphs that the INAR residuals are larger, in absolute value, from Zeger's residuals. This is explained by the fact that Zeger's model assumes that the data are overdispersed. Hence, we expect larger estimated values of variances of the data by Zeger's model, thus smaller absolute residual values. The residual series from Zeger's model does not exhibit any outliers apart from one observation in De Bilt's station and in Deelen's station.

Table 5.7 consists of the Chi-square statistic constructed by the Pearson residuals and the goodness of fit test which is obtained by comparing this value with the 95% percentile of the X^2 distribution with each model degrees of freedom. The third column is produced by dividing this Chi-square statistic with the respective degrees of freedom, which are 348 for both models.

Station	S_α	p-value
Arcen	3.653	<0.001
Berkhout	-0.003	0.501
Cabauw	6.201	<0.001
De Bilt	2.635	0.004
De Kooy	2.454	0.007
Deelen	11.263	<0.001
Eelde	3.483	<0.001
Eindhoven	5.362	<0.001
Ell	2.502	0.006
Gilze-Rijen	5.652	<0.001
Heino	5.412	<0.001
Herwijnen	3.392	<0.001
Hoogeveen	6.809	<0.001
Leeuwarden	3.037	0.001
Lelystad	2.344	0.010
Maastricht	2.767	0.003
Marknesse	5.143	<0.001
Nieuw Bierta	2.410	0.008
Rotterdam	11.047	<0.001
Schiphol	12.205	<0.001
Soesterberg	7.099	<0.001
Stavoren	3.310	<0.001
Twenthe	6.161	<0.001
Valkenburg	7.590	<0.001
Vlissingen	1.777	0.038
Volkel	2.135	0.016
Whilhelminadorp	2.170	0.015

Table 5.2: Test for the existence of a latent process for each station

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.2297	0.6375	0.360	0.719	0.3406	0.6575	0.518	0.605
Wind								
Direction	0.0166	0.0511	0.324	0.746	0.0122	0.0530	0.231	0.817
Speed	0.0009	0.0026	0.367	0.714	0.0010	0.0028	0.367	0.713
Mean Temp	-0.0014	0.0011	-1.361	0.173	-0.0014	0.0013	-1.142	0.253
Temp below zero	0.2467	0.1443	1.709	0.087	0.2788	0.1550	1.799	0.072
Humidity	0.0031	0.0061	0.507	0.612	0.0011	0.0065	0.168	0.867
Radiation	-0.0008	0.0024	-0.348	0.728	0.0005	0.0026	0.187	0.851
Precipitation								
Duration	0.0844	0.0287	2.944	0.003	0.1082	0.0306	3.533	<0.001
Intensity	-0.0040	0.0239	-0.168	0.867	-0.0002	0.0238	-0.010	0.992
Weekday								
Monday	0.5094	0.1430	3.564	<0.001	0.5109	0.1389	3.679	<0.001
Tuesday	0.4305	0.1450	2.969	0.003	0.4459	0.1412	3.159	0.002
Wednesday	0.6997	0.1373	5.097	<0.001	0.7131	0.1353	5.270	<0.001
Thursday	0.4714	0.1450	3.251	0.001	0.4971	0.1416	3.512	<0.001
Friday	0.7558	0.1374	5.499	<0.001	0.7640	0.1351	5.656	<0.001
Saturday	0.0774	0.1623	0.477	0.634	0.1270	0.1535	0.827	0.408
Other parameters								
α	0.0435	0.0443	0.981	0.326	0.8199	0.0239	2.122	0.034
σ_ε^2					0.0618	-	-	-

Table 5.3: Results based on the fitted INAR and Zeger's regression model for De Bilt

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.9799	0.4691	2.089	0.037	0.4578	0.5929	0.772	0.440
Wind								
Direction	-0.0159	0.0416	-0.382	0.703	-0.0279	0.0518	-0.539	0.590
Speed	-0.0009	0.0018	-0.481	0.630	0.0003	0.0023	0.138	0.890
Mean Temp	0.0005	0.0008	0.599	0.549	0.0001	0.0010	0.094	0.925
Temp below zero	0.0898	0.1105	0.813	0.416	0.1601	0.1380	1.160	0.246
Humidity	-0.0009	0.0045	-0.201	0.841	0.0052	0.0058	0.910	0.363
Radiation	-0.0021	0.0021	-1.007	0.314	0.0002	0.0026	0.065	0.948
Precipitation								
Duration	0.0843	0.0210	4.022	<0.001	0.0793	0.0282	2.816	0.005
Intensity	-0.0068	0.0241	-0.281	0.779	-0.0122	0.0299	-0.409	0.683
Weekday								
Monday	0.4979	0.1141	4.364	<0.001	0.4803	0.1285	3.737	<0.001
Tuesday	0.3628	0.1180	3.075	0.002	0.4235	0.1348	3.143	0.002
Wednesday	0.5415	0.1131	4.788	<0.001	0.5184	0.1344	3.858	<0.001
Thursday	0.5225	0.1156	4.518	<0.001	0.5163	0.1364	3.785	<0.001
Friday	0.7073	0.1103	6.413	<0.001	0.6938	0.1307	5.310	<0.001
Saturday	0.0831	0.1313	0.633	0.527	0.1521	0.1370	1.110	0.267
Other parameters								
α	0.0764	0.0333	2.297	0.022	0.2547	0.0225	2.046	0.041
σ_ε^2					0.1808	-	-	-

Table 5.4: Results based on the fitted INAR and Zeger's regression model for Deelen

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	2.2787	0.3045	7.485	<0.001	2.2825	0.3772	6.051	<0.001
Wind								
Direction	0.0410	0.0262	1.563	0.118	0.0392	0.0339	1.156	0.248
Speed	-0.0014	0.0008	-1.774	0.076	-0.0012	0.0011	-1.092	0.275
Mean Temp	-0.0011	0.0005	-2.092	0.036	-0.0013	0.0007	-1.864	0.062
Temp below zero	-0.0866	0.0857	-1.010	0.312	-0.0932	0.1089	-0.856	0.392
Humidity	-0.0034	0.0031	-1.119	0.263	-0.0030	0.0039	-0.768	0.443
Radiation	-0.0049	0.0011	-4.429	<0.001	-0.0039	0.0014	-2.802	0.005
Precipitation								
Duration	0.1355	0.0150	9.043	<0.001	0.1399	0.0210	6.652	<0.001
Intensity	0.1328	0.0177	7.485	<0.001	0.1283	0.0234	5.486	<0.001
Weekday								
Monday	0.5681	0.0729	7.796	<0.001	0.5372	0.0835	6.430	<0.001
Tuesday	0.3985	0.0734	5.429	<0.001	0.3992	0.0872	4.576	<0.001
Wednesday	0.5231	0.0715	7.318	<0.001	0.5116	0.0864	5.922	<0.001
Thursday	0.6407	0.0718	8.929	<0.001	0.6226	0.0868	7.170	<0.001
Friday	0.5432	0.0714	7.606	<0.001	0.5444	0.0859	6.335	<0.001
Saturday	0.1962	0.0780	2.515	0.012	0.1826	0.0888	2.057	0.040
Other parameters								
α	0.0542	0.0268	2.026	0.043	0.2185	0.0096	1.649	0.099
σ_ϵ^2					0.0722	-	-	-

Table 5.5: Results based on the fitted INAR and Zeger's regression model for Rotterdam

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	2.1993	0.3154	6.974	<0.001	2.1841	0.4051	5.392	<0.001
Wind								
Direction	-0.0690	0.0307	-2.250	0.025	-0.0555	0.0368	-1.508	0.132
Speed	-0.0011	0.0009	-1.266	0.206	-0.0006	0.0012	-0.534	0.594
Mean Temp	-0.0001	0.0006	-0.108	0.914	-0.0005	0.0008	-0.575	0.565
Temp below zero	0.1143	0.1012	1.129	0.259	0.0361	0.1296	0.278	0.781
Humidity	-0.0065	0.0032	-2.052	0.040	-0.0050	0.0042	-1.183	0.237
Radiation	-0.0023	0.0012	-1.886	0.059	-0.0013	0.0016	-0.801	0.423
Precipitation								
Duration	0.1769	0.0171	10.359	<0.001	0.1676	0.0226	7.407	<0.001
Intensity	0.0446	0.0184	2.420	0.016	0.0460	0.0236	1.954	0.051
Weekday								
Monday	0.4469	0.0774	5.775	<0.001	0.4186	0.0842	4.972	<0.001
Tuesday	0.3660	0.0797	4.589	<0.001	0.3949	0.0896	4.405	<0.001
Wednesday	0.4069	0.0789	5.155	<0.001	0.4083	0.0916	4.455	<0.001
Thursday	0.5232	0.0776	6.740	<0.001	0.5139	0.0907	5.665	<0.001
Friday	0.4543	0.0780	5.827	<0.001	0.4785	0.0889	5.383	<0.001
Saturday	-0.1654	0.0995	-1.663	0.096	-0.0386	0.0926	-0.417	0.677
Other parameters								
α	0.1241	0.0283	4.382	0.000	0.4395	0.0115	3.598	0.000
σ_ϵ^2					0.0937	-	-	-

Table 5.6: Results based on the fitted INAR and Zeger's regression model for Schiphol

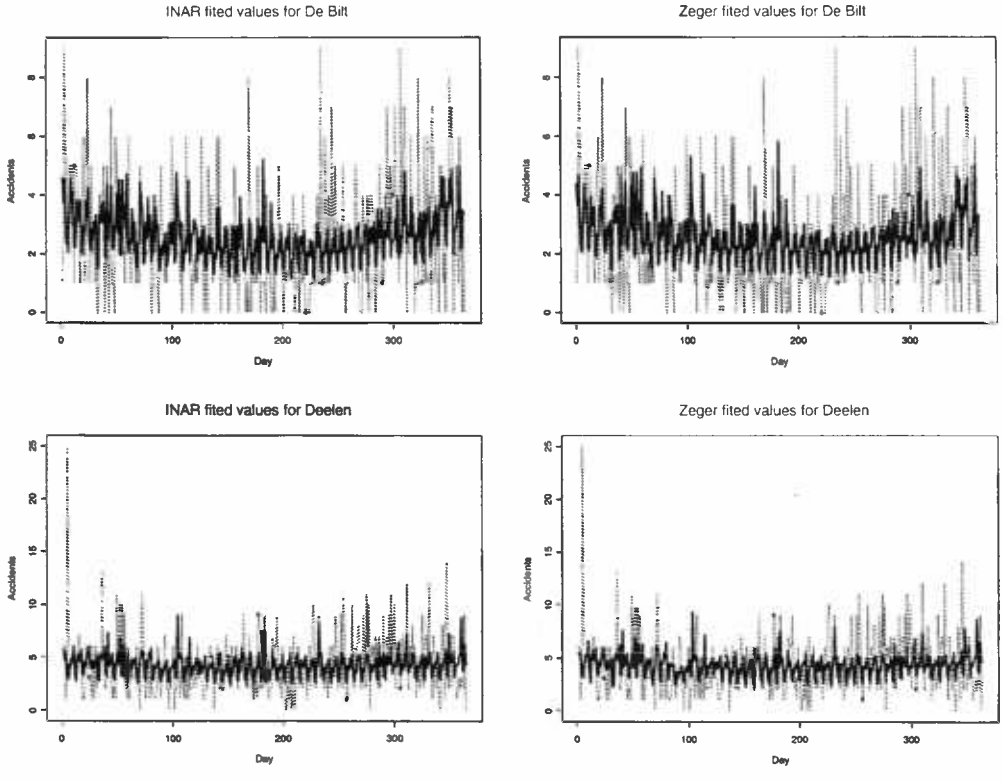

Figure 5.6: Plot of observed and fitted values versus time for each model for De Bilt and Deelen

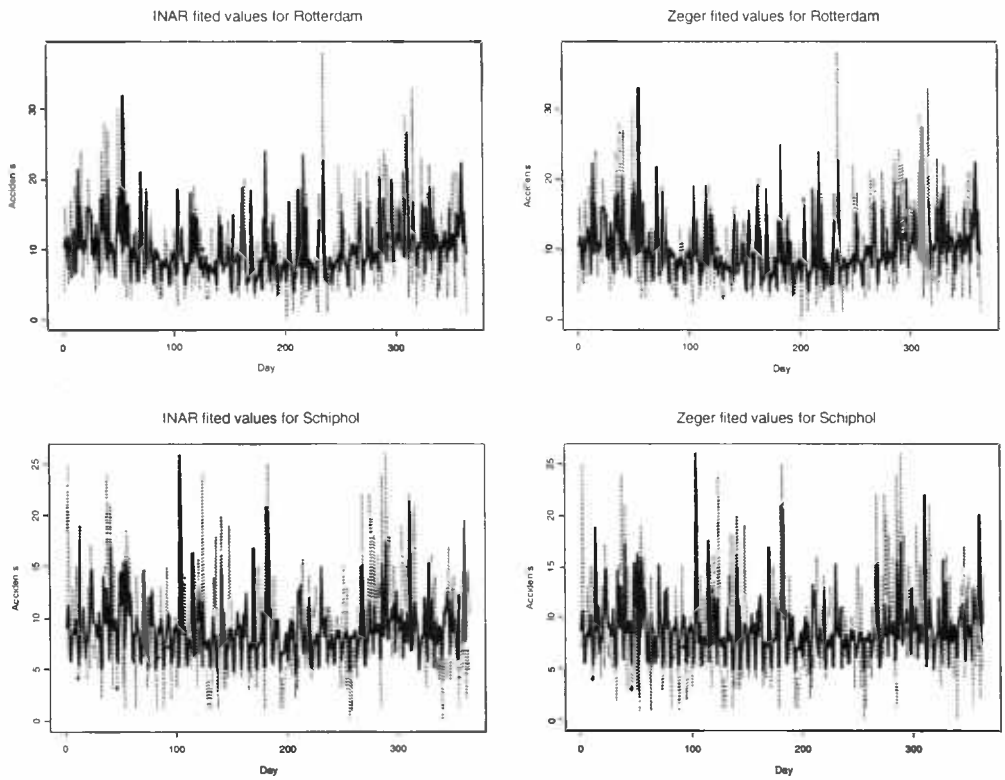

Figure 5.7: Plot of observed and fitted values versus time for each model for Rotterdam and Schiphol

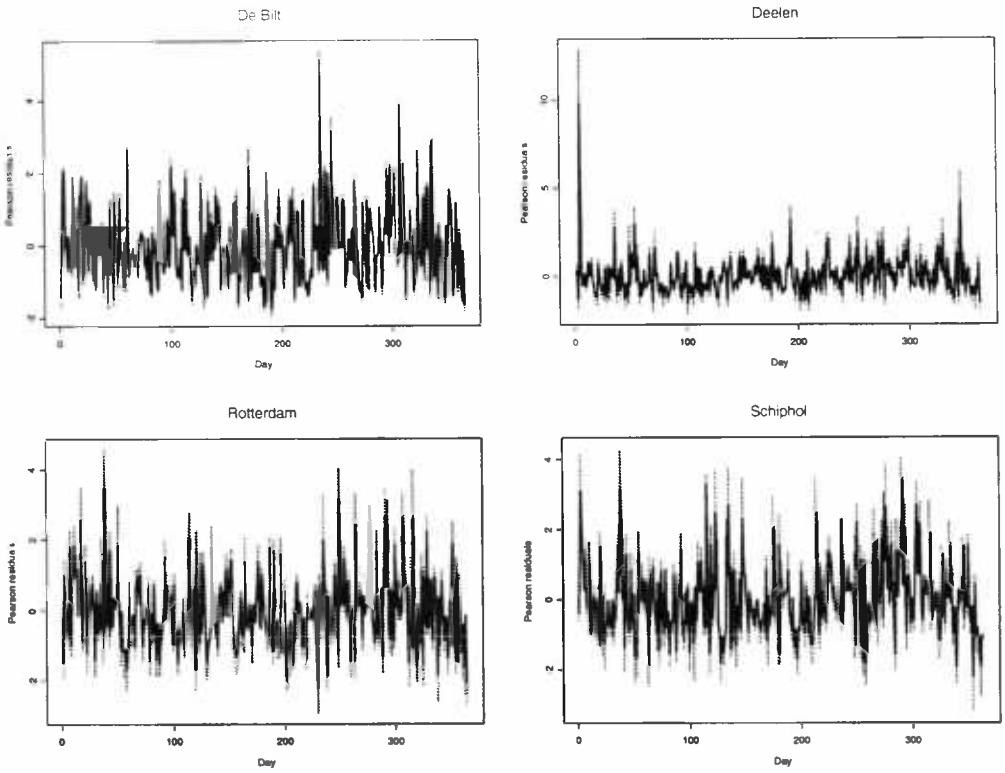

Figure 5.8: Plot of the residual series of both models versus time

	INAR			Zeger		
	Chi-square	p-value	overdispersion	Chi-square	p-value	overdispersion
Arcen	458.350	<0.001	1.317	381.261	0.106	1.096
Berkhout	354.677	0.391	1.019	370.304	0.197	1.064
Cabauw	495.741	<0.001	1.425	354.526	0.393	1.019
De Bilt	432.400	0.001	1.243	381.178	0.107	1.095
De Kooy	418.813	0.004	1.214	370.648	0.164	1.074
Deelen	673.577	<0.001	1.936	399.133	0.030	1.147
Eelde	417.540	0.006	1.200	347.306	0.500	0.998
Eindhoven	489.325	<0.001	1.406	365.984	0.243	1.052
Ell	405.336	0.018	1.165	355.437	0.380	1.021
Gilze-Rijen	496.423	<0.001	1.427	363.177	0.277	1.044
Heino	471.970	<0.001	1.356	346.133	0.518	0.995
Herwijnen	444.455	<0.001	1.277	370.831	0.192	1.066
Hoogeveen	504.860	<0.001	1.451	351.576	0.436	1.010
Leeuwarden	427.105	0.002	1.227	365.663	0.247	1.051
Lelystad	422.840	0.004	1.215	378.239	0.127	1.087
Maastricht	413.000	0.009	1.187	355.302	0.382	1.021
Marknesse	446.437	<0.001	1.283	334.311	0.692	0.961
Nieuw Beerta	400.533	0.027	1.151	354.767	0.390	1.019
Rotterdam	621.615	<0.001	1.786	358.697	0.335	1.031
Schiphol	673.639	<0.001	1.936	375.797	0.146	1.080
Soesterberg	548.421	<0.001	1.576	375.232	0.151	1.078
Stavoren	409.443	0.013	1.177	348.004	0.490	1.000
Twenthe	502.347	<0.001	1.444	363.693	0.270	1.045
Valkenburg	558.119	<0.001	1.604	381.079	0.107	1.095
Vlissingen	407.378	0.015	1.171	376.467	0.141	1.082
Volkel	405.197	0.019	1.164	366.781	0.234	1.054
Wilhelminadorp	385.459	0.081	1.108	347.515	0.497	0.999

Table 5.7: Goodness of fit tests and diagnostics for INAR and Zeger's model

This way we obtain a measure of the residual overdispersion; variability of the dependent variable that has not been explained by any of the available covariates. The left part of the table contains these diagnostics for the INAR fitted models and the right part for Zeger's fitted models.

Under the assumption of Poisson distributed data and the inclusion of every significant covariate, each residual should have unit variance. Therefore the sum of the squared residuals divided by the respective degrees of freedom (residual overdispersion) must be a number close to unity. However, we can see from the third column of the table that the INAR regression model is not a good fit for the data, or at least as good as Zeger's model, since there exists a large amount of variability which remains unexplained from the regression. The residual overdispersions are highly significant in the sense that they are significantly larger than the unity, implying that the Poisson distribution assumption is not plausible for the data. In contrast, Zeger's model, which accounts for overdispersion assuming that $Var(Y_t) > E(Y_t)$, seems to provide a good fit for almost all of the sites since the residual overdispersions are very close to unity and they are no longer significant. Zeger's model succeeded in fitting the data better than the INAR model due to fact that even in the cases where there was not statistical presence of autocorrelation, it still accounts for overdispersion (remember that $Var(Y_t) = \mu_t + \sigma_\epsilon^2 \mu_t^2$, which is independent of α); something that the Poisson INAR did not (if the autocorrelation parameter is set to zero then the INAR regression model reduces to a simple Poisson regression). Hence, we conclude that Zeger's method provides more reliable estimated effects than INAR's (regarding to the data under study) and any further analysis and inference of the factors should be based on Zeger's model results.

5.5 Meta-analysis results

We are interested in combining all the available information we have on every factor using the meta-analysis. We saw earlier that the researcher needs to identify the independent studies which quantify the effect of the covariates. In our case, each station represents a study since every regression model applied to the sites has estimated the effects of these factors. Another important task of the meta-analysis is the identification of the summary measure common to all studies. It is obvious that the suitable measure which quantifies the effect of each and every covariate is its estimated coefficients. The last thing left to do is to combine the estimated effects from each study by choosing between fixed or random effects model. We present the pooled estimates for both type of models so that the reader can see possible differences between them.

The comparison between the goodness of fit of the INAR and Zeger's model showed that the estimates obtained from the parameter-driven model are more reliable, as mentioned in section 5.4. Therefore, we will use these estimated effects as the common measure of the studies. The regression model through which the pooled estimate and its inference is obtained assumes that the observations are drawn from normal distributions. The estimated coefficients satisfy this assumption as we noted in section 3.2. Hence, we can carry on with the estimation of the fixed and random effects models for the covariates.

A very simple and nice way to begin a meta-analysis is by creating a weighted forest plot. This figure plots the common measures of every study along with their respective confidence intervals giving us the ability to identify at once which study produced a statistical significant effect and what is the direction of this effect. Hence, the researcher gets a very good idea of the studies' variability and of the amount of heterogeneity. The bottom point of this graph is the pooled estimate with its confidence interval. As we saw before, the length of this interval depends on whether we use fixed or

random effects model. This plot was named as weighted because the size of the plotted points, representing each study estimate, differ with respect to the standard error of the estimate; the smaller the standard error the larger the point gets. The name forest was given because the intersections of the confidence intervals with the y-axis produce a graph which looks like a tree.

This section presents the meta-analyses of the covariates that proved statistical significant. The rest of the analyses can be found in the Appendix C. The first covariate which will be analyzed is the mean temperature. Its weighted forest plot (Figure 5.9) shows that most of the estimated coefficients are negative which means that higher mean temperatures have an effect of decreasing the mean accidents. The overall mean temperature effect has been found highly significant for both fixed and random effects models with an estimate equal to 0.7968% (Table 5.8). The test for homogeneity of the effects across the stations did not exhibit any statistical significance implying that there is no need for the random effects model. We derive the same conclusion if we choose the best model according to the AIC for these models.

Figure 5.10 presents the weighted forest plot for the temperature below zero indicator. We can see that only three stations exhibited statistical significance at the 5% level for this covariate and all showing an effect of increasing the mean accidents. The other stations' estimated effects were either negative (negative coefficient) or positive (positive coefficient) but without any significance. The above means that probably the heterogeneity of the studies will not be significant. Indeed, we can see from Table 5.9 that the Q quantity is not statistical significant at the 10% level. We arrive at the same conclusion if we select the best model with the AIC. The fixed effects model estimated coefficient is equal to 0.0606 which means that when the temperature reaches below zero, it increases the mean accidents by 6.247% (p-value = 0.077).

The next variable considered in the meta-analysis is the precipitation duration. This covariate was found to be statistical significant for almost

every station; some with greater effects than the others (Figure 5.11). The fact that many of the studies provided us with significant coefficients of this covariate implies that the estimated pooled effect will be significant too. Considering though the variation of these estimated effects, there is a big chance that the between-study variation will be significant. The test of zero between-study variation is significant at the 10% and the estimated τ^2 is equal to 0.0011. Moreover, the AIC of the random effects model is smaller than that of the fixed effects model. Therefore, the inference of the common estimate will be based on the random effects model. Table 5.10 shows that the estimated common coefficient is highly significant and equal to 0.1065. Thus, if the precipitation duration is increased by one unit we expect an increase of the mean accidents by 11.238%.

Figure 5.12 presents the weighted forest plot for the precipitation intensity, which was defined as the ratio of the rainfall amount to rainfall duration. Most of the estimated coefficients are positive but some of them exhibit a greater effect than the others. This can be seen in the figure where there area few large positive values. Hence, the common effect will probably be estimated by the random effects model. The test for homogeneity across the study-specific estimates is significant at the 10% level, implying that the random effects model should be preferred to the fixed effects model; so does the AIC test. Table 5.11 shows that the estimate of the between study variation is equal to 0.0006 and the estimated pooled coefficient is equal to 0.0334, which is highly significant. This means that the mean accidents are increased by 3.396% for an extra unit of rainfall intensity.

5.6 Meta-regression results

We saw in the meta-analysis results that some of the coefficients presented significant heterogeneity. This means that there could be two or more populations from whom the coefficients arise. A simple way to incorporate this

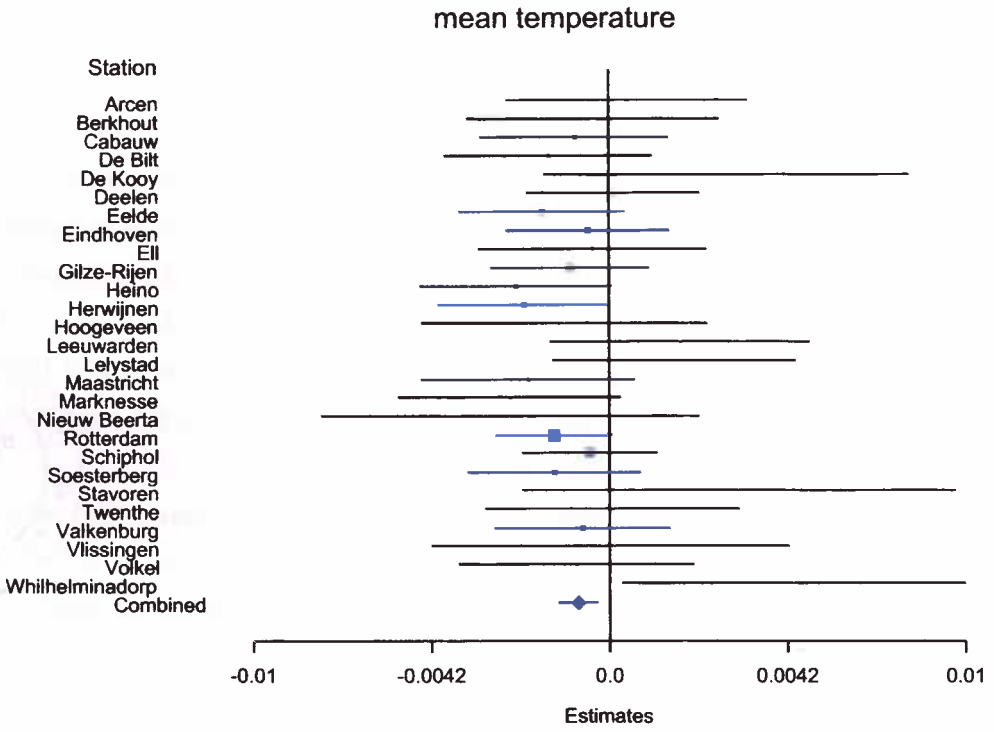

Figure 5.9: Weighted forest plot for the mean temperature

	Fixed effects		Random effects		Test of homogeneity
	α	τ^2	α	τ^2	Q
estimate	-0.0008	0.0000	-0.0008	0.0000	25.3876
standard error	0.0002	-	0.0002	0.0000	-
t-value	-3.1936	-	-3.1936	0.0013	-
p-value	0.0014	-	0.0014	0.9990	0.4971
AIC	-280.0027		-278.0027		

Table 5.8: Estimated common parameter and inter-variation for the mean temperature

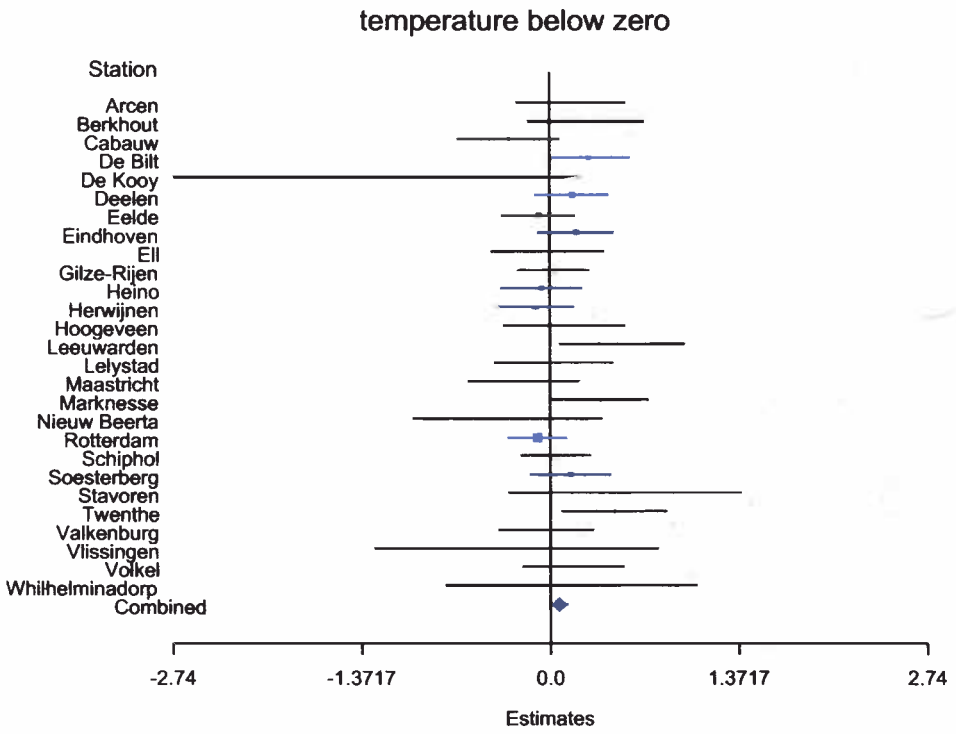

Figure 5.10: Weighted forest plot for the temperature below zero indicator

	Fixed effects		Random effects		Test of homogeneity
	α	τ^2	α	τ^2	Q
estimate	0.0606	0.0000	0.0635	0.0036	32.5678
standard error	0.0342	-	0.0368	0.0086	-
t-value	1.7703	-	1.7288	0.4152	-
p-value	0.0767	-	0.0839	0.6780	0.175
AIC	-0.7606		1.0536		

Table 5.9: Estimated common parameter and inter-variation for the temperature below zero indicator

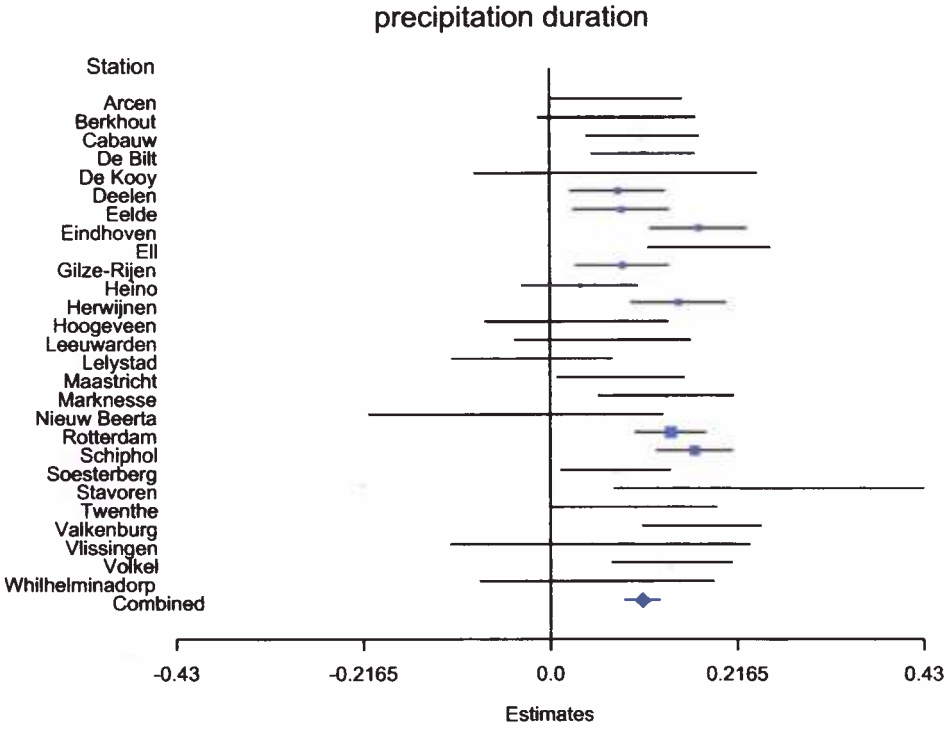

Figure 5.11: Weighted forest plot for the precipitation duration

	Fixed effects		Random effects		Test of homogeneity
	α	τ^2	α	τ^2	Q
estimate	0.1129	0.0000	0.1065	0.0011	51.4029
standard error	0.0069	-	0.0100	0.0007	-
t-value	16.4699	-	10.6482	1.6740	-
p-value	<0.0001	-	<0.0001	0.0941	0.0021
AIC	-69.5379		-74.9582		

Table 5.10: Estimated common parameter and inter-variation for the precipitation duration

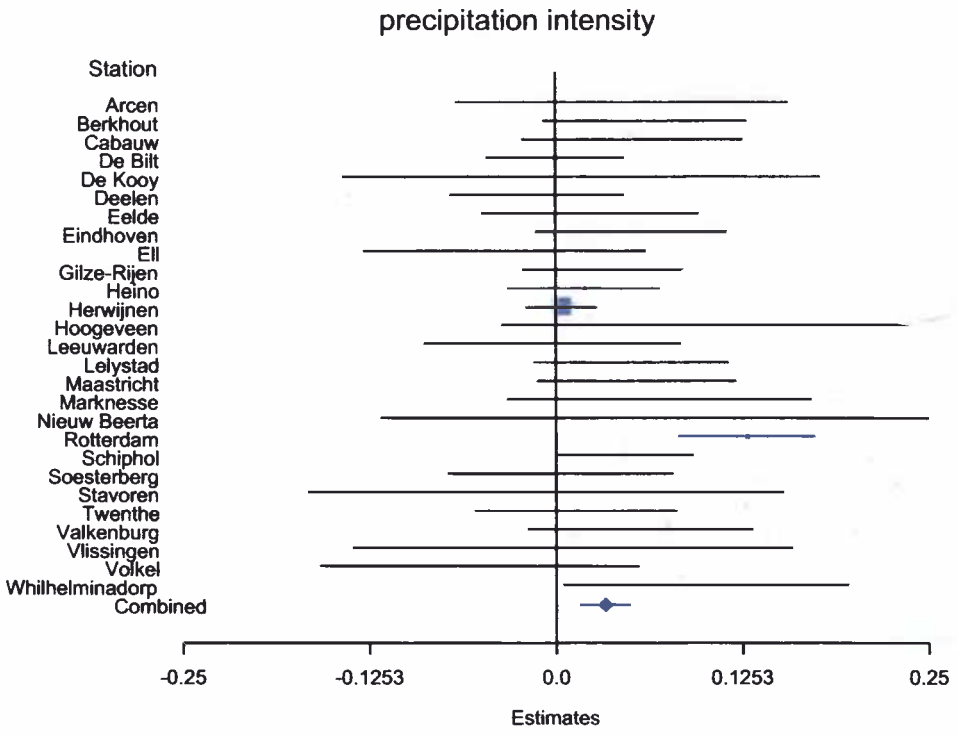

Figure 5.12: Weighted forest plot for the precipitation intensity

	Fixed effects		Random effects		Test of homogeneity
	α	τ^2	α	τ^2	Q
estimate	0.0295	0.0000	0.0334	0.0006	38.6616
standard error	0.0062	-	0.0086	0.0005	-
t-value	4.7751	-	3.8765	1.2750	-
p-value	<0.0001	-	0.0001	0.2023	0.0525
AIC	-83.4549		-88.3415		

Table 5.11: Estimated common parameter and inter-variation for the precipitation intensity

into a model is by adding covariates which reflect some of the characteristics of each study. Therefore, significant regressors will result in lowering the between-study variation τ^2 rendering him insignificant. In other words, the addition of covariates in the model will probably explain an amount of the heterogeneity (Thompson and Sharp (1999)).

The purpose of this chapter is the identification of such regressors using meta-regression analysis. We will consider only the random effects model since we want to make a direct comparison of the heterogeneity before and after the meta-regression. The covariates used for the models were constructed from each site's annual data. Hence, we included regressors such as the minimum, maximum and average of wind speed, temperature, radiation, humidity and precipitation's duration and intensity. A dummy variable was also added to the analysis indicating the proximity of each station to the sea.

Given the large amount of explanatory variables in each study and the problem of multicollinearity associated with it, a stepwise model selection procedure was carried out. Specifically, the selection of the variables for the models was based on a backward 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 the highest correlation with the response variable. Regressors that belonged to the same set but had small correlation between them were included in the analysis. After estimating these models and evaluating the significance of each of the included variables, we started removing the most insignificant covariates. This procedure continued until no further covariate could be removed or added to the models. The entry/remove tests were based on likelihood ratio tests.

Table 5.12 depicts the estimated parameters of a random effects model for the wind speed covariate on the annual minimum temperature and the annual mean humidity. The coefficient of the temperature's covariate is negative which means that the effect that the wind speed has on the mean accidents

is weaken by the increase of the temperature. On the other hand, higher percentage of humidity in the air increase the effect of the wind speed on the mean accidents.

The mean temperature effect on the mean accidents shows a significant relation to the annual highest temperature (Table 5.13). Specifically, the coefficient of the maximum temperature's covariate is equal to -0.00004, meaning that an increase of a unit of the maximum temperature decrease the effect of the mean temperature on the accidents by 0.0044%. This relation was expected to be found, since we had seen that most of the mean temperature coefficients were negative, implying that higher temperature leads to less mean accidents.

A plausible relation between the effect of the temperature below zero indicator and the minimum temperature covariate is shown in Table 5.14. The model estimated a negative coefficient of the minimum temperature's covariate equal to -0.0077. This means that an increase of one unit of the minimum temperature decrease the effect of the temperature below zero indicator on the mean accidents by 0.7677%. This was of course expected since as the minimum temperature increases, the risk of ice presence on the roads is reduced.

The humidity effect on the mean accidents becomes stronger as the minimum temperature decreases as is shown from Table 5.15. If the minimum temperature is decreased by one unit, then the effect of the humidity variable on the mean accidents is increased by 0.052%. This is explained by the fact that the danger of an amount of humidity in the atmosphere is higher when combined with low temperatures due to the fact that fog can possibly start freezing on the roads and the roads are more slippery.

The effect of radiation seem to depend on the maximum wind speed and the mean temperature. Table 5.16 shows the estimated coefficients of these regressors from the random effects model. Recall from the simple meta-analysis for the site-specific radiation effects that there was significant

presence of heterogeneity. The meta-regression model succeeded in lowering this amount by including these two covariates. This can be seen from the reduction of the Q statistic equal to 44.2473 (p-value = 0.0142) for the meta-analysis to 20.0780 (p-value = 0.6923) for the meta-regression model.

The precipitation duration results (Table 5.17) need to be interpreted with caution. We can see that the effect of an intense rainfall on the mean accidents is dependent of three covariates, the maximum temperature, the mean precipitation duration and the maximum precipitation intensity. The estimated negative coefficient of the max temperature weakens the effect of the intensity.

Despite the inclusion of these covariates, there is still a significant amount of heterogeneity not explained by the model. The Q statistic before the meta-regression was equal to 51.4029 (p-value = 0.0021) and now is equal to 32.1980 (p-value = 0.0961). This is evidence that there still may be an important explanatory variable (unavailable, though) that should be included in the model in order to account for the observed residual heterogeneity.

The negative coefficient of the mean rainfall duration produces a decrease on the intensity effect by 25.6704% as the duration increases (Table 5.18). From a first view, we should expect the opposite; higher precipitation duration should probably lead to stronger intensity effects on the mean accidents. However, we must keep in mind that rainfall duration and intensity are negatively correlated, e.g. a long time rainfall will be accompanied by weak intensity. High intensity rainfall will probably appear with short rainfall duration. This is why the mean precipitation duration covariate has a negative coefficient. Recall also that rainfall duration influences also the traffic volume, the more rain the less cars in the road.

Moreover, the significant variation between the sites that had appeared in the meta-analysis has now been explained by the additional regressors that were included in the meta-regression model. The residual overdispersion is no longer significant ($Q = 20.5942$, p-value = 0.6086).

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	-0.0522	0.0246	-2.1186	0.0341
Min Temp	-0.0001	0.0001	-2.3137	0.0207
Mean Humidity	0.0005	0.0003	1.6485	0.0993
τ^2	0.0000	0.0000	0.0000	1.0000
Test of homogeneity				
Q	20.7764	-	-	0.6519

Table 5.12: Meta-regression random effects model for the wind speed

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	0.0136	0.0071	1.9231	0.0545
Max Temp	-0.00004	0.0000	-2.0303	0.0423
τ^2	0.0000	0.0000	0.0000	1.0000
Test of homogeneity				
Q	21.2654	-	-	0.6777

Table 5.13: Meta-regression random effects model for the mean temperature

The meta-regression random effects model for the wind direction effect did not identify any significant factors that relate to it. This implies that the model includes only the intercept term and thus it is reduced to the simple meta-analysis model.

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	-0.5812	0.3253	-1.7864	0.0740
Min Temp	-0.0077	0.0039	-1.9836	0.0473
τ^2	0.0000	0.0073	0.0001	0.9999
Test of homogeneity				
Q	28.6330	-	-	0.2795

Table 5.14: Meta-regression random effects model for the temperature below zero indicator

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	-0.0407	0.0123	-3.3095	0.0009
Min Temp	-0.0005	0.0001	-3.4733	0.0005
τ^2	0.0000	0.0000	0.0000	1.0000
Test of homogeneity				
Q	15.8411	-	-	0.9195

Table 5.15: Meta-regression random effects model for the humidity

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	0.0738	0.0169	4.3600	0.0000
Max wind speed	-0.00004	0.0000	-1.9257	0.0541
Mean Temp	-0.0006	0.0002	-3.7491	0.0002
τ^2	0.0000	0.0000	0.0001	0.9999
Test of homogeneity				
Q	20.0780	-	-	0.6923

Table 5.16: Meta-regression random effects model for the radiation

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	-0.7148	0.2468	-2.8959	0.0038
Mean Temp	0.0062	0.0024	2.6140	0.0089
Min Humidity	0.0046	0.0022	2.0949	0.0362
Max Radiation	0.0001	0.0000	1.9314	0.0534
τ^2	0.0000	0.0003	0.1387	0.8897
Test of homogeneity				
Q	32.1980	-	-	0.0961

Table 5.17: Meta-regression random effects model for the precipitation duration

	Random Effects Model			
	estimate	st.error	t-value	p-value
α	0.7546	0.2313	3.2618	0.0011
Max Temp	-0.0014	0.0007	-1.9915	0.0464
Mean Precipitation duration	-0.2967	0.1766	-1.6803	0.0929
Max Precipitation intensity	-0.0010	0.0005	-2.1079	0.0350
τ^2	0.0000	0.0002	0.0000	1.0000
Test of homogeneity				
Q	20.5492	-	-	0.6086

Table 5.18: Meta-regression random effects model for the precipitation intensity


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 6

Conclusions

The models described in this thesis serve the same purpose; to model time series of counts accounting for the presence of autocorrelation between them. It is important to note that the literature contains very few publications on fitting time series models in accident count data and usually the models fitted were not correct. So in this thesis we provide correct models for the problem at hand. The Poisson INAR(1) regression model failed in fitting the accident counts of the stations under analysis due to the fact that it does not allow for overdispersion.

Therefore, in order to fit an observation driven model in the data, one needs to assume a different distribution for the process $\{R_t\}$ than the Poisson, such as to account for possible overdispersion. Another solution described thoroughly in Pavlopoulos and Karlis (2006) is to assume that the innovation process is a finite mixture of m independent Poisson processes. Note that the Poisson INAR model presented in this thesis is a special case of the finite mixture model for $m = 1$. By letting $m \geq 2$ the model becomes able of accounting for overdispersion. Another issue that possibly leads to an inadequate fit of the INAR model is the existence of significant covariates (unavailable though), such as exposure, that have not been included in the regression. This of course suggests the following limitations.

Due to the absence of data on exposures for the different days of the week and for each of the 27 cities studied, we cannot measure the possibly confounding effect of weather conditions on accidents, directly or indirectly through exposure. Indeed, weather may have both a direct effect on accidents due to a change of the risk per unit of exposure and an indirect effect through exposure, since the amount and means of traffic may change due to differing weather conditions. In this thesis, we only model the direct effect of weather conditions on accidents. However, we believe that this limitation does not have serious effects on the results, given the type of roads we consider in the data. Indeed, this analysis focusses on the number of accidents on the major roads network and we believe that the effect of weather on exposure is much more dominant on the underlying road network than on the major roads network.

Also, the use of climatological weather data instead of using accident 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 accidents 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.

Finally, our model does not distinguish between different types of injuries, such as fatalities, severe and light injuries. In fact, earlier research has shown 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. A recent work on this type of model has been carried out by Heinen and Rengifo (2004), who present a multivariate autoregressive model of time series of count data using copulas.

Appendix A

Properties of Binomial Thinning

Let X and Y denote two independent non-negative integer-valued random variables, and the symbol $\stackrel{D}{=}$ denotes equality of probability distributions. We provide some properties of the binomial thinning based on the results of Pavlopoulos and Karlis (2006).

A1. The characteristic function of $\alpha \circ X$, is $\Phi_{\alpha\circ X}(u) = E\left(e^{iu\cdot(\alpha\circ X)}\right) = E\left\{(1-\alpha+\alpha e^{iu})^X\right\}$, for $u \in \mathbb{R}$. This is derived from taking double expectation of $\alpha \circ X$ given X, which is a binomial random variable with X trials and probability of success α. Direct consequences of **A1** are the following properties.

A2. $0 \circ X \stackrel{D}{=} 0 \quad \& \quad 1 \circ X \stackrel{D}{=} X$.

A3. $\alpha_1 \circ (\alpha_2 \circ X) \stackrel{D}{=} (\alpha_1\alpha_2) \circ X$, for every $\alpha_1, \alpha_2 \in [0, 1]$.

A4. $\alpha \circ (X+Y) \stackrel{D}{=} (\alpha\circ X) + (\alpha\circ Y)$, for every $\alpha \in [0, 1]$.

Non-central moments $\mu'_r(\alpha \circ X) = E[(\alpha \circ X)^r]$ of $\alpha \circ X$ are obtained directly from **A1**, provided that $\mu'_r(X) = E(X^r) < \infty$, for $r = 1, 2, 3, 4$ respectively:

A5. $\mu'_1(\alpha \circ X) = \alpha\mu'_1(X)$,

A6. $\mu'_2(\alpha \circ X) = \alpha^2\mu'_2(X) + \alpha(1-\alpha)\mu'_1(X)$,

A7. $\mu'_3(\alpha \circ X) = \alpha^3\mu'_3(X) + 3\alpha^2(1-\alpha)\mu'_2(X) + \alpha(1-3\alpha+2\alpha^2)\mu'_1(X)$,

A8. $\mu'_4(\alpha\circ X) = \alpha^4\mu'_4(X)+6\alpha^3(1-\alpha)\mu'_3(X)+\alpha^2(1-\alpha)(7-11\alpha)\mu'_2(X)+\alpha(1-\alpha)(6\alpha^2-6\alpha+1)\mu'_1(X)$.

Central moments $\mu_r(\alpha \circ X) = E[(\alpha \circ X - E(\alpha \circ X))^r]$, for $r = 1, 2, 3, 4$, are

obtained by the non-central moments using the standard formulas. Tedious algebra leads to the following formulas for the central moments:

A9. $\mu_2(\alpha \circ X) = Var\,(\alpha \circ X) = \alpha^2 Var\,(X) + \alpha(1-\alpha)E(X)$,

A10. $\mu_3(\alpha \circ X) = \alpha^3\mu_3(X) + 3\alpha^2(1-\alpha)Var(X) + \alpha(1-3\alpha+2\alpha^2)E(X)$,

A11. $\mu_4(\alpha \circ X) = \alpha^4\mu_4(X) + 6\alpha^3(1-\alpha)E(X^3) + \alpha^2(1-\alpha)(7-11\alpha)E(X^2)$ $+\alpha(1-\alpha)(6\alpha^2 - 6\alpha + 1)E(X) - 12\alpha^3(1-\alpha)E(X)E(X^2) - 4\alpha^2(1-3\alpha+2\alpha^2)[E(X)]^2 + 6\alpha^3(1-\alpha)[E(X)]^3$.

Appendix B

INAR and Zeger results

In this appendix we provide detailed results for all sites when fitting both the INAR and Zeger's models.

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-1.2621	0.8637	-1.461	0.144	-1.0714	0.8691	-1.233	0.218
Wind								
Direction	0.0074	0.0681	0.108	0.914	0.0283	0.0746	0.379	0.705
Speed	0.0033	0.0038	0.873	0.383	0.0027	0.0041	0.649	0.516
Mean Temp	0.0004	0.0013	0.325	0.745	0.0004	0.0015	0.290	0.772
Temp below zero	0.1822	0.1920	0.949	0.343	0.1538	0.2039	0.754	0.451
Humidity	0.0104	0.0083	1.253	0.210	0.0083	0.0085	0.979	0.328
Radiation	0.0016	0.0032	0.484	0.628	0.0014	0.0034	0.406	0.685
Precipitation								
Duration	0.0701	0.0350	2.005	0.045	0.0782	0.0389	2.009	0.045
Intensity	0.0396	0.0521	0.760	0.447	0.0446	0.0571	0.782	0.434
Weekday								
Monday	0.4182	0.1867	2.240	0.025	0.4096	0.2049	1.999	0.046
Tuesday	0.4353	0.1864	2.336	0.020	0.4119	0.1990	2.070	0.038
Wednesday	0.6084	0.1787	3.405	<0.001	0.6042	0.1941	3.112	0.002
Thursday	0.5294	0.1839	2.878	0.004	0.5402	0.1999	2.702	0.007
Friday	0.7366	0.1780	4.137	<0.001	0.7338	0.1901	3.860	<0.001
Saturday	0.3719	0.1892	1.966	0.049	0.3765	0.2094	1.798	0.072
Other parameters								
α	0.0000	0.0445	0.000	1.000	-0.3807	0.0440	-1.298	0.194
σ_ε^2					0.1501	-	-	-

Table 6.1: Results based on the fitted INAR and Zeger's regression model for Arcen

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-0.8392	1.0465	-0.802	0.423	-0.7971	0.9553	-0.834	0.404
Wind								
Direction	-0.0303	0.0855	-0.354	0.723	-0.0327	0.0793	-0.412	0.680
Speed	0.0024	0.0026	0.924	0.356	0.0022	0.0024	0.933	0.351
Mean Temp	-0.0004	0.0016	-0.244	0.807	-0.0004	0.0015	-0.247	0.805
Temp below zero	0.2505	0.2349	1.066	0.286	0.2610	0.2175	1.200	0.230
Humidity	0.0039	0.0102	0.381	0.703	0.0041	0.0095	0.431	0.666
Radiation	0.0045	0.0033	1.370	0.171	0.0043	0.0031	1.399	0.162
Precipitation								
Duration	0.0825	0.0498	1.655	0.098	0.0779	0.0468	1.665	0.096
Intensity	0.0625	0.0376	1.663	0.096	0.0600	0.0350	1.712	0.087
Weekday								
Monday	0.0910	0.2030	0.448	0.654	0.1023	0.1881	0.544	0.587
Tuesday	0.1814	0.2008	0.904	0.366	0.1953	0.1884	1.037	0.300
Wednesday	0.1244	0.2035	0.611	0.541	0.1207	0.1911	0.632	0.528
Thursday	-0.0484	0.2131	-0.227	0.820	-0.0339	0.1987	-0.171	0.865
Friday	0.0918	0.2045	0.449	0.654	0.0851	0.1940	0.439	0.661
Saturday	-0.1376	0.2194	-0.627	0.531	-0.1211	0.2029	-0.597	0.551
Other parameters								
α	0.0454	0.0554	0.82	0.412	-0.7127	0.0492	0.612	0.535
σ_ϵ^2					0.0000	-	-	-

Table 6.2: Results based on the fitted INAR and Zeger's regression model for Berkhout

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.5007	0.6109	0.820	0.412	0.5547	0.7009	0.791	0.429
Wind								
Direction	-0.0565	0.0482	-1.172	0.241	-0.0535	0.0575	-0.930	0.352
Speed	-0.0004	0.0017	-0.226	0.821	-0.0003	0.0020	-0.136	0.892
Mean Temp	-0.0007	0.0009	-0.788	0.431	-0.0008	0.0011	-0.721	0.471
Temp below zero	-0.2706	0.1615	-1.676	0.094	-0.3009	0.1874	-1.606	0.108
Humidity	0.0000	0.0060	0.003	0.998	-0.0006	0.0070	-0.080	0.937
Radiation	-0.0037	0.0021	-1.710	0.087	-0.0036	0.0025	-1.426	0.154
Precipitation								
Duration	0.1086	0.0262	4.142	<0.001	0.1081	0.0333	3.248	0.001
Intensity	0.0499	0.0317	1.572	0.116	0.0516	0.0380	1.358	0.175
Weekday								
Monday	0.6940	0.1417	4.897	<0.001	0.6803	0.1600	4.252	<0.001
Tuesday	0.6891	0.1403	4.913	<0.001	0.6787	0.1598	4.247	<0.001
Wednesday	0.7269	0.1395	5.210	<0.001	0.7258	0.1595	4.552	<0.001
Thursday	0.8510	0.1394	6.107	<0.001	0.8396	0.1592	5.274	<0.001
Friday	0.8317	0.1383	6.016	<0.001	0.8381	0.1582	5.297	<0.001
Saturday	0.2388	0.1570	1.521	0.128	0.2414	0.1730	1.395	0.163
Other parameters								
α	0.0000	0.0360	0.000	1.000	-0.0007	0.0279	-0.003	0.997
σ_ϵ^2					0.1457	-	-	-

Table 6.3: Results based on the fitted INAR and Zeger's regression model for Cabauw

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-0.7917	1.1265	-0.703	0.482	-0.8516	1.1811	-0.721	0.471
Wind								
Direction	0.1319	0.1153	1.144	0.253	0.1067	0.1223	0.872	0.383
Speed	-0.0049	0.0034	-1.443	0.149	-0.0047	0.0035	-1.319	0.187
Mean Temp	0.0026	0.0021	1.220	0.223	0.0028	0.0022	1.261	0.207
Temp below zero	-1.2576	0.7516	-1.673	0.094	-1.2628	0.7555	-1.672	0.095
Humidity	0.0001	0.0115	0.009	0.993	0.0006	0.0121	0.047	0.963
Radiation	0.0005	0.0038	0.138	0.890	0.0001	0.0040	0.026	0.979
Precipitation								
Duration	0.0789	0.0782	1.009	0.313	0.0764	0.0837	0.913	0.361
Intensity	0.0280	0.0747	0.375	0.708	0.0175	0.0819	0.214	0.831
Weekday								
Monday	-0.1539	0.2883	-0.534	0.594	-0.1436	0.3144	-0.457	0.648
Tuesday	0.0182	0.2763	0.066	0.948	0.0381	0.2922	0.130	0.896
Wednesday	-0.2030	0.2958	-0.686	0.493	-0.1980	0.3161	-0.626	0.531
Thursday	-0.2154	0.3041	-0.708	0.479	-0.2081	0.3184	-0.654	0.513
Friday	0.3801	0.2553	1.489	0.137	0.3929	0.2749	1.429	0.153
Saturday	-0.3744	0.3204	-1.169	0.243	-0.3604	0.3378	-1.067	0.286
Other parameters								
α	0.0000	0.0517	0.000	1.000	-0.3920	0.1253	-0.870	0.384
σ_ϵ^2					0.2780	-	-	-

Table 6.4: Results based on the fitted INAR and Zeger's regression model for De Kooy

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-0.0826	0.6181	-0.134	0.894	-0.1782	0.6199	-0.288	0.774
Wind								
Direction	0.0001	0.0449	0.003	0.997	-0.0038	0.0494	-0.076	0.939
Speed	0.0008	0.0017	0.455	0.649	0.0010	0.0019	0.514	0.607
Mean Temp	-0.0016	0.0009	-1.785	0.074	-0.0016	0.0010	-1.581	0.114
Temp below zero	-0.1269	0.1273	-0.997	0.319	-0.0873	0.1347	-0.649	0.517
Humidity	0.0090	0.0059	1.520	0.129	0.0098	0.0062	1.582	0.114
Radiation	0.0035	0.0021	1.685	0.092	0.0036	0.0023	1.582	0.114
Precipitation								
Duration	0.0814	0.0253	3.222	0.001	0.0834	0.0286	2.916	0.004
Intensity	0.0228	0.0340	0.670	0.503	0.0233	0.0373	0.624	0.532
Weekday								
Monday	0.5893	0.1219	4.833	<0.001	0.5935	0.1289	4.605	<0.001
Tuesday	0.4311	0.1221	3.531	<0.001	0.4636	0.1311	3.537	<0.001
Wednesday	0.4209	0.1236	3.407	<0.001	0.4201	0.1334	3.148	0.002
Thursday	0.3985	0.1243	3.207	0.001	0.3983	0.1338	2.977	0.003
Friday	0.5386	0.1216	4.429	<0.001	0.5390	0.1297	4.156	<0.001
Saturday	0.2139	0.1286	1.664	0.096	0.2107	0.1392	1.514	0.130
Other parameters								
α	0.0000	0.0472	0.000	1.000	-0.0733	0.0202	-0.241	0.810
σ_ϵ^2					0.0663	-	-	-

Table 6.5: Results based on the fitted INAR and Zeger's regression model for Eelde

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.2742	0.5088	0.539	0.590	0.3922	0.5606	0.700	0.484
Wind								
Direction	-0.0509	0.0436	-1.166	0.244	-0.0410	0.0506	-0.809	0.418
Speed	0.0001	0.0019	0.044	0.965	-0.0006	0.0023	-0.277	0.782
Mean Temp	-0.0003	0.0009	-0.369	0.713	-0.0005	0.0010	-0.509	0.611
Temp below zero	0.2177	0.1232	1.766	0.077	0.1899	0.1413	1.344	0.179
Humidity	0.0034	0.0050	0.695	0.487	0.0028	0.0056	0.497	0.619
Radiation	0.0016	0.0020	0.781	0.435	0.0016	0.0024	0.693	0.488
Precipitation								
Duration	0.1711	0.0235	7.293	<0.001	0.1728	0.0287	6.022	<0.001
Intensity	0.0492	0.0276	1.779	0.075	0.0510	0.0326	1.566	0.117
Weekday								
Monday	0.4147	0.1182	3.508	<0.001	0.4032	0.1342	3.005	0.003
Tuesday	0.5057	0.1175	4.302	<0.001	0.4798	0.1331	3.605	<0.001
Wednesday	0.4660	0.1176	3.964	<0.001	0.4524	0.1338	3.380	<0.001
Thursday	0.5550	0.1189	4.668	<0.001	0.5433	0.1346	4.037	<0.001
Friday	0.7649	0.1139	6.714	<0.001	0.7340	0.1291	5.685	<0.001
Saturday	0.1847	0.1276	1.448	0.148	0.1660	0.1428	1.163	0.245
Other parameters								
α	0.0000	0.0376	0.000	1.000	-0.0607	0.0211	-0.287	0.774
σ_ϵ^2					0.0995	-	-	-

Table 6.6: Results based on the fitted INAR and Zeger's regression model for Eindhoven

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.6128	0.8270	0.741	0.459	0.5636	0.8791	0.641	0.522
Wind								
Direction	0.1274	0.0630	2.024	0.043	0.1227	0.0670	1.832	0.067
Speed	-0.0007	0.0030	-0.221	0.826	-0.0005	0.0032	-0.144	0.886
Mean Temp	-0.0003	0.0013	-0.217	0.829	-0.0004	0.0014	-0.286	0.775
Temp below zero	-0.0170	0.1988	-0.085	0.932	-0.0180	0.2100	-0.086	0.932
Humidity	-0.0087	0.0083	-1.055	0.291	-0.0082	0.0089	-0.922	0.356
Radiation	0.0032	0.0030	1.076	0.282	0.0034	0.0032	1.065	0.287
Precipitation								
Duration	0.1867	0.0335	5.574	<0.001	0.1848	0.0365	5.068	<0.001
Intensity	-0.0361	0.0455	-0.793	0.428	-0.0340	0.0485	-0.701	0.483
Weekday								
Monday	0.3672	0.1653	2.221	0.026	0.3632	0.1753	2.072	0.038
Tuesday	0.0936	0.1809	0.518	0.605	0.1040	0.1879	0.553	0.580
Wednesday	0.3911	0.1650	2.370	0.018	0.3993	0.1752	2.279	0.023
Thursday	0.5514	0.1666	3.309	<0.001	0.5453	0.1772	3.077	0.002
Friday	0.5762	0.1630	3.536	<0.001	0.5787	0.1735	3.336	<0.001
Saturday	0.2535	0.1771	1.431	0.152	0.2527	0.1865	1.355	0.175
Other parameters								
α	0.0018	0.0409	0.043	0.966	0.0420	0.0379	0.096	0.924
σ_ϵ^2					0.0863	-	-	-

Table 6.7: Results based on the fitted INAR and Zeger's regression model for Ell

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.7042	0.5085	1.385	0.166	0.7301	0.5678	1.286	0.199
Wind								
Direction	-0.0614	0.0405	-1.516	0.129	-0.0597	0.0465	-1.283	0.200
Speed	0.0007	0.0021	0.352	0.725	0.0007	0.0024	0.281	0.779
Mean Temp	-0.0009	0.0008	-1.156	0.248	-0.0009	0.0010	-0.971	0.331
Temp below zero	0.0244	0.1180	0.207	0.836	0.0236	0.1349	0.175	0.861
Humidity	0.0027	0.0049	0.554	0.580	0.0027	0.0057	0.474	0.636
Radiation	0.0029	0.0019	1.502	0.133	0.0029	0.0023	1.298	0.194
Precipitation								
Duration	0.0836	0.0233	3.588	<0.001	0.0843	0.0279	3.024	0.003
Intensity	0.0304	0.0239	1.275	0.202	0.0322	0.0277	1.161	0.246
Weekday								
Monday	0.4689	0.1047	4.478	<0.001	0.4602	0.1163	3.956	<0.001
Tuesday	0.3225	0.1081	2.982	0.003	0.3155	0.1203	2.621	0.009
Wednesday	0.3831	0.1056	3.627	<0.001	0.3708	0.1198	3.096	0.002
Thursday	0.4887	0.1047	4.668	<0.001	0.4839	0.1189	4.069	<0.001
Friday	0.4896	0.1048	4.673	<0.001	0.4837	0.1186	4.077	<0.001
Saturday	-0.0543	0.1227	-0.442	0.659	-0.0517	0.1287	-0.402	0.688
Other parameters								
α	0.0179	0.0369	0.486	0.627	0.1583	0.0177	0.799	0.424
σ_ϵ^2					0.0896	-	-	-

Table 6.8: Results based on the fitted INAR and Zeger's regression model for Gilze-Rijen

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.5323	0.6882	0.773	0.439	0.5411	0.7190	0.753	0.452
Wind								
Direction	-0.0489	0.0519	-0.942	0.347	-0.0463	0.0570	-0.813	0.416
Speed	0.0014	0.0027	0.505	0.613	0.0014	0.0030	0.489	0.625
Mean Temp	-0.0021	0.0010	-2.068	0.039	-0.0022	0.0012	-1.910	0.056
Temp below zero	-0.0822	0.1360	-0.604	0.546	-0.0580	0.1494	-0.388	0.698
Humidity	-0.0006	0.0066	-0.095	0.925	0.0007	0.0072	0.090	0.928
Radiation	0.0013	0.0024	0.565	0.572	0.0017	0.0026	0.656	0.512
Precipitation								
Duration	0.0415	0.0301	1.381	0.167	0.0353	0.0343	1.028	0.304
Intensity	0.0223	0.0230	0.972	0.331	0.0193	0.0260	0.741	0.459
Weekday								
Monday	0.6410	0.1522	4.211	<0.001	0.5640	0.1481	3.809	<0.001
Tuesday	0.4541	0.1524	2.980	0.003	0.4299	0.1535	2.801	0.005
Wednesday	0.5421	0.1514	3.580	<0.001	0.4858	0.1537	3.160	0.002
Thursday	0.6277	0.1497	4.192	<0.001	0.5746	0.1515	3.792	<0.001
Friday	0.7220	0.1477	4.888	<0.001	0.6715	0.1487	4.515	<0.001
Saturday	0.4673	0.1522	3.070	0.002	0.4278	0.1516	2.822	0.005
Other parameters								
α	0.0690	0.0436	1.583	0.113	0.1953	0.0264	0.957	0.339
σ_ϵ^2					0.1293	-	-	-

Table 6.9: Results based on the fitted INAR and Zeger's regression Heino

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	1.0132	0.5850	1.732	0.083	0.9584	0.6269	1.529	0.126
Wind								
Direction	0.0338	0.0468	0.723	0.470	0.0377	0.0518	0.728	0.467
Speed	-0.0006	0.0017	-0.344	0.731	-0.0007	0.0019	-0.344	0.731
Mean Temp	-0.0020	0.0009	-2.159	0.031	-0.0020	0.0010	-1.944	0.052
Temp below zero	-0.1130	0.1286	-0.879	0.380	-0.1029	0.1416	-0.727	0.468
Humidity	-0.0046	0.0058	-0.797	0.426	-0.0040	0.0064	-0.626	0.531
Radiation	0.0034	0.0020	1.704	0.088	0.0037	0.0022	1.658	0.097
Precipitation								
Duration	0.1467	0.0246	5.967	<0.001	0.1494	0.0282	5.305	<0.001
Intensity	0.0037	0.0111	0.339	0.735	0.0041	0.0122	0.340	0.734
Weekday								
Monday	0.4988	0.1285	3.881	<0.001	0.4962	0.1333	3.723	<0.001
Tuesday	0.3703	0.1299	2.850	0.004	0.3738	0.1385	2.700	0.007
Wednesday	0.5267	0.1263	4.172	<0.001	0.5226	0.1350	3.872	<0.001
Thursday	0.5612	0.1266	4.434	<0.001	0.5582	0.1361	4.100	<0.001
Friday	0.6290	0.1257	5.004	<0.001	0.6198	0.1342	4.619	<0.001
Saturday	0.2982	0.1333	2.237	0.025	0.2928	0.1411	2.075	0.038
Other parameters								
α	0.0012	0.0408	0.029	0.977	0.1995	0.0214	0.630	0.529
σ_ϵ^2					0.0676	-	-	-

Table 6.10: Results based on the fitted INAR and Zeger's regression model for Herwijnen

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-0.6984	0.9273	-0.753	0.451	-0.4417	1.0970	-0.403	0.687
Wind								
Direction	-0.0863	0.0736	-1.174	0.241	-0.0981	0.0839	-1.169	0.242
Speed	0.0014	0.0030	0.455	0.649	-0.0005	0.0036	-0.141	0.888
Mean Temp	-0.0010	0.0014	-0.724	0.469	-0.0011	0.0017	-0.608	0.543
Temp below zero	0.1414	0.1910	0.740	0.459	0.1032	0.2265	0.456	0.649
Humidity	0.0064	0.0090	0.709	0.478	0.0062	0.0109	0.566	0.572
Radiation	0.0037	0.0034	1.085	0.278	0.0026	0.0040	0.658	0.510
Precipitation								
Duration	0.0479	0.0470	1.020	0.308	0.0310	0.0543	0.571	0.568
Intensity	0.0936	0.0597	1.566	0.117	0.1009	0.0702	1.439	0.150
Weekday								
Monday	0.0019	0.1978	0.010	0.992	-0.0093	0.2043	-0.045	0.964
Tuesday	0.3203	0.1833	1.747	0.081	0.2514	0.1994	1.261	0.208
Wednesday	0.1694	0.1907	0.888	0.374	0.1566	0.2070	0.756	0.449
Thursday	0.3374	0.1829	1.845	0.065	0.2983	0.2020	1.477	0.140
Friday	0.2860	0.1845	1.550	0.121	0.2594	0.2004	1.295	0.196
Saturday	0.3277	0.1837	1.784	0.075	0.2666	0.1939	1.375	0.169
Other parameters								
α	0.0693	0.0395	1.753	0.080	0.3491	0.0547	2.083	0.037
σ_ϵ^2					0.3265	-	-	-

Table 6.11: Results based on the fitted INAR and Zeger's regression model for Hoogeveen

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-2.4049	0.9673	-2.486	0.013	-2.2310	0.9685	-2.303	0.021
Wind								
Direction	-0.0117	0.0808	-0.145	0.885	-0.0152	0.0846	-0.179	0.858
Speed	0.0060	0.0027	2.242	0.025	0.0061	0.0029	2.124	0.034
Mean Temp	0.0014	0.0015	0.956	0.339	0.0017	0.0016	1.069	0.285
Temp below zero	0.4795	0.2237	2.144	0.032	0.5247	0.2335	2.247	0.025
Humidity	0.0167	0.0093	1.803	0.071	0.0154	0.0096	1.611	0.107
Radiation	0.0026	0.0033	0.797	0.426	0.0021	0.0035	0.618	0.537
Precipitation								
Duration	0.0642	0.0490	1.312	0.190	0.0610	0.0522	1.169	0.243
Intensity	0.0008	0.0410	0.020	0.984	-0.0022	0.0441	-0.050	0.960
Weekday								
Monday	0.8108	0.2066	3.924	<0.001	0.7633	0.2030	3.760	<0.001
Tuesday	0.2831	0.2218	1.276	0.202	0.3114	0.2195	1.419	0.156
Wednesday	0.3532	0.2207	1.600	0.110	0.3281	0.2213	1.482	0.138
Thursday	0.3479	0.2238	1.554	0.120	0.3072	0.2235	1.375	0.169
Friday	0.4626	0.2196	2.107	0.035	0.4195	0.2174	1.929	0.054
Saturday	0.3713	0.2198	1.690	0.091	0.3480	0.2188	1.591	0.112
Other parameters								
α	0.0380	0.0443	0.859	0.390	0.1778	0.0553	0.495	0.620
σ_ϵ^2					0.1540	-	-	-

Table 6.12: Results based on the fitted INAR and Zeger's regression model for Leeuwarden

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-1.2776	0.8973	-1.424	0.155	-1.1738	0.9245	-1.270	0.204
Wind								
Direction	0.0383	0.0705	0.543	0.587	0.0376	0.0746	0.504	0.614
Speed	0.0062	0.0025	2.473	0.013	0.0059	0.0027	2.198	0.028
Mean Temp	0.0016	0.0014	1.134	0.257	0.0015	0.0015	1.043	0.297
Temp below zero	0.0559	0.2105	0.266	0.791	0.0260	0.2214	0.117	0.907
Humidity	0.0088	0.0088	0.995	0.320	0.0079	0.0093	0.850	0.395
Radiation	0.0007	0.0029	0.251	0.802	0.0006	0.0031	0.181	0.856
Precipitation								
Duration	-0.0243	0.0450	-0.540	0.589	-0.0208	0.0476	-0.438	0.662
Intensity	0.0481	0.0310	1.552	0.121	0.0508	0.0335	1.515	0.130
Weekday								
Monday	0.2724	0.1931	1.411	0.158	0.2737	0.1972	1.388	0.165
Tuesday	0.1352	0.1956	0.691	0.490	0.1163	0.2038	0.571	0.568
Wednesday	0.3425	0.1889	1.813	0.070	0.3472	0.1957	1.775	0.076
Thursday	0.2226	0.1939	1.148	0.251	0.2257	0.2017	1.119	0.263
Friday	0.6608	0.1804	3.663	<0.001	0.6594	0.1863	3.539	<0.001
Saturday	0.2452	0.1917	1.279	0.201	0.2496	0.2000	1.248	0.212
Other parameters								
α	0.0062	0.0429	0.144	0.886	0.0406	0.0458	0.087	0.931
σ_ϵ^2					0.0981	-	-	-

Table 6.13: Results based on the fitted INAR and Zeger's regression model for Lelystad

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.3761	0.7317	0.514	0.607	0.3721	0.7614	0.489	0.625
Wind								
Direction	-0.0680	0.0688	-0.989	0.323	-0.0663	0.0728	-0.910	0.363
Speed	-0.0028	0.0025	-1.106	0.269	-0.0027	0.0027	-0.986	0.324
Mean Temp	-0.0018	0.0012	-1.509	0.131	-0.0019	0.0013	-1.497	0.135
Temp below zero	-0.1784	0.1991	-0.896	0.370	-0.1900	0.2094	-0.907	0.364
Humidity	0.0017	0.0075	0.230	0.819	0.0021	0.0079	0.270	0.787
Radiation	0.0012	0.0030	0.399	0.690	0.0013	0.0032	0.405	0.686
Precipitation								
Duration	0.0847	0.0354	2.391	0.017	0.0819	0.0380	2.155	0.031
Intensity	0.0546	0.0319	1.711	0.087	0.0543	0.0342	1.587	0.113
Weekday								
Monday	0.1827	0.1589	1.150	0.250	0.1821	0.1654	1.101	0.271
Tuesday	0.1569	0.1612	0.973	0.330	0.1635	0.1678	0.975	0.330
Wednesday	-0.0504	0.1679	-0.300	0.764	-0.0433	0.1736	-0.249	0.803
Thursday	0.1127	0.1653	0.681	0.496	0.1090	0.1738	0.627	0.531
Friday	0.5060	0.1507	3.357	<0.001	0.5017	0.1600	3.136	0.002
Saturday	-0.2062	0.1868	-1.104	0.270	-0.1837	0.1804	-1.019	0.308
Other parameters								
α	0.0230	0.0473	0.485	0.628	0.2114	0.0375	0.543	0.588
σ_ϵ^2					0.0961	-	-	-

Table 6.14: Results based on the fitted INAR and Zeger's regression model for Maastricht

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-1.2899	0.7860	-1.641	0.101	-1.3260	0.8629	-1.537	0.124
Wind								
Direction	0.0290	0.0585	0.495	0.621	0.0362	0.0672	0.538	0.590
Speed	0.0068	0.0024	2.838	0.005	0.0068	0.0028	2.433	0.015
Mean Temp	-0.0026	0.0012	-2.259	0.024	-0.0024	0.0013	-1.751	0.080
Temp below zero	0.3411	0.1569	2.174	0.030	0.3582	0.1816	1.972	0.049
Humidity	0.0116	0.0076	1.533	0.125	0.0121	0.0086	1.413	0.158
Radiation	0.0105	0.0026	4.053	<0.001	0.0103	0.0030	3.478	<0.001
Precipitation								
Duration	0.1419	0.0333	4.261	<0.001	0.1343	0.0401	3.346	<0.001
Intensity	0.0704	0.0437	1.611	0.107	0.0696	0.0522	1.332	0.183
Weekday								
Monday	0.1644	0.1541	1.067	0.286	0.1548	0.1734	0.893	0.372
Tuesday	0.2930	0.1533	1.912	0.056	0.2740	0.1714	1.598	0.110
Wednesday	-0.1516	0.1687	-0.898	0.369	-0.1684	0.1873	-0.899	0.369
Thursday	0.3494	0.1537	2.274	0.023	0.3342	0.1704	1.961	0.050
Friday	0.6185	0.1430	4.326	<0.001	0.6146	0.1618	3.797	<0.001
Saturday	0.4420	0.1480	2.986	0.003	0.4004	0.1699	2.357	0.018
Other parameters								
α	0.0016	0.0385	0.041	0.967	-0.0259	0.0369	-0.117	0.907
σ_ϵ^2					0.1674	-	-	-

Table 6.15: Results based on the fitted INAR and Zeger's regression model for Marknesse

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-2.8480	1.5424	-1.847	0.065	-2.6485	1.5906	-1.665	0.096
Wind								
Direction	0.1038	0.1133	0.916	0.360	0.1049	0.1210	0.867	0.386
Speed	0.0016	0.0043	0.373	0.709	0.0011	0.0046	0.237	0.813
Mean Temp	-0.0023	0.0021	-1.101	0.271	-0.0023	0.0023	-1.027	0.304
Temp below zero	-0.2952	0.3356	-0.880	0.379	-0.3114	0.3515	-0.886	0.376
Humidity	0.0167	0.0149	1.124	0.261	0.0148	0.0155	0.953	0.341
Radiation	0.0081	0.0048	1.691	0.091	0.0078	0.0051	1.535	0.125
Precipitation								
Duration	-0.0432	0.0820	-0.527	0.598	-0.0398	0.0871	-0.457	0.648
Intensity	0.0686	0.0871	0.787	0.431	0.0664	0.0939	0.707	0.479
Weekday								
Monday	0.7655	0.3283	2.332	0.020	0.7830	0.3432	2.282	0.023
Tuesday	0.2938	0.3556	0.826	0.409	0.2733	0.3714	0.736	0.462
Wednesday	0.4149	0.3483	1.191	0.234	0.4193	0.3638	1.153	0.249
Thursday	0.7475	0.3264	2.290	0.022	0.7587	0.3436	2.208	0.027
Friday	0.7893	0.3221	2.451	0.014	0.7859	0.3398	2.313	0.021
Saturday	0.4367	0.3454	1.264	0.206	0.4455	0.3619	1.231	0.218
Other parameters								
α	0.0000	0.0573	0.000	1.000	-0.0892	0.1302	-0.195	0.845
σ_ϵ^2					0.2848	-	-	-

Table 6.16: Results based on the fitted INAR and Zeger's regression model for Nieuw Beerta

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.6713	0.5032	1.334	0.182	0.6389	0.5845	1.093	0.274
Wind								
Direction	0.0603	0.0459	1.314	0.189	0.0568	0.0541	1.050	0.294
Speed	-0.0044	0.0021	-2.097	0.036	-0.0046	0.0025	-1.812	0.070
Mean Temp	-0.0013	0.0009	-1.558	0.119	-0.0013	0.0010	-1.256	0.209
Temp below zero	0.1382	0.1239	1.115	0.265	0.1472	0.1490	0.988	0.323
Humidity	0.0039	0.0048	0.794	0.427	0.0046	0.0058	0.793	0.428
Radiation	0.0003	0.0020	0.164	0.870	0.0006	0.0024	0.233	0.816
Precipitation								
Duration	0.0767	0.0263	2.913	0.004	0.0756	0.0326	2.317	0.021
Intensity	0.0011	0.0328	0.035	0.972	0.0030	0.0386	0.078	0.938
Weekday								
Monday	0.4911	0.1229	3.997	<0.001	0.4725	0.1365	3.461	<0.001
Tuesday	0.4548	0.1222	3.721	<0.001	0.4473	0.1391	3.216	0.001
Wednesday	0.4841	0.1210	3.999	<0.001	0.4782	0.1388	3.445	<0.001
Thursday	0.5683	0.1210	4.697	<0.001	0.5580	0.1385	4.029	<0.001
Friday	0.4819	0.1213	3.973	<0.001	0.4781	0.1389	3.443	<0.001
Saturday	0.2773	0.1283	2.162	0.031	0.2723	0.1435	1.898	0.058
Other parameters								
α	0.0262	0.0373	0.701	0.483	0.1117	0.0232	0.671	0.502
σ_ϵ^2					0.1396	-	-	-

Table 6.17: Results based on the fitted INAR and Zeger's regression model for Soesterberg

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-4.6585	1.7977	-2.591	0.010	-4.4495	1.9216	-2.316	0.021
Wind								
Direction	-0.1138	0.1329	-0.856	0.392	-0.1338	0.1457	-0.918	0.358
Speed	0.0018	0.0041	0.435	0.663	0.0018	0.0044	0.411	0.681
Mean Temp	0.0039	0.0024	1.601	0.109	0.0031	0.0026	1.168	0.243
Temp below zero	0.7410	0.3908	1.896	0.058	0.5415	0.4337	1.249	0.212
Humidity	0.0286	0.0175	1.638	0.101	0.0283	0.0188	1.501	0.133
Radiation	0.0052	0.0049	1.079	0.281	0.0048	0.0052	0.930	0.352
Precipitation								
Duration	0.2777	0.0854	3.253	0.001	0.2532	0.0917	2.761	0.006
Intensity	-0.0203	0.0752	-0.270	0.787	-0.0068	0.0817	-0.084	0.933
Weekday								
Monday	0.2824	0.3196	0.884	0.377	0.2095	0.3495	0.599	0.549
Tuesday	-0.0279	0.3525	-0.079	0.937	0.0187	0.3623	0.052	0.959
Wednesday	0.0943	0.3315	0.285	0.776	0.0517	0.3621	0.143	0.886
Thursday	0.0881	0.3452	0.255	0.799	0.0508	0.3638	0.140	0.889
Friday	0.1757	0.3327	0.528	0.598	0.1300	0.3572	0.364	0.716
Saturday	0.2104	0.3422	0.615	0.539	0.1790	0.3618	0.495	0.621
Other parameters								
α	0.0000	0.0452	0.000	1.000	-0.1317	0.1680	-0.398	0.691
σ_ϵ^2					0.5071	-	-	-

Table 6.18: Results based on the fitted INAR and Zeger's regression model for Stavoren

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-1.8085	0.7245	-2.496	0.013	-1.7417	0.8413	-2.070	0.038
Wind								
Direction	-0.0537	0.0669	-0.804	0.422	-0.0536	0.0794	-0.675	0.500
Speed	0.0044	0.0031	1.437	0.151	0.0039	0.0037	1.053	0.293
Mean Temp	-0.0001	0.0013	-0.077	0.939	0.0001	0.0015	0.042	0.966
Temp below zero	0.4672	0.1616	2.891	0.004	0.4627	0.1963	2.358	0.018
Humidity	0.0140	0.0071	1.973	0.049	0.0134	0.0084	1.602	0.109
Radiation	0.0041	0.0029	1.396	0.163	0.0039	0.0035	1.113	0.266
Precipitation								
Duration	0.0951	0.0398	2.389	0.017	0.0980	0.0486	2.017	0.044
Intensity	0.0137	0.0288	0.474	0.635	0.0135	0.0347	0.389	0.697
Weekday								
Monday	0.7359	0.1829	4.023	<0.001	0.7238	0.2062	3.510	<0.001
Tuesday	0.5704	0.1889	3.020	0.003	0.5321	0.2140	2.487	0.013
Wednesday	0.5286	0.1900	2.782	0.005	0.5392	0.2146	2.513	0.012
Thursday	0.8983	0.1802	4.986	<0.001	0.8995	0.2052	4.383	<0.001
Friday	0.8724	0.1788	4.879	<0.001	0.8609	0.2054	4.193	<0.001
Saturday	0.5677	0.1889	3.005	0.003	0.5500	0.2117	2.598	0.009
Other parameters								
α	0.0000	0.0389	0.000	1.000	0.0356	0.0476	0.187	0.852
σ_ϵ^2					0.2504	-	-	-

Table 6.19: Results based on the fitted INAR and Zeger's regression model for Twenthe

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.0369	0.4993	0.074	0.941	0.1070	0.5751	0.186	0.852
Wind								
Direction	0.0458	0.0471	0.973	0.331	0.0571	0.0584	0.978	0.328
Speed	-0.0002	0.0014	-0.153	0.879	-0.0003	0.0018	-0.160	0.873
Mean Temp	-0.0003	0.0009	-0.289	0.773	-0.0006	0.0011	-0.607	0.544
Temp below zero	0.0148	0.1501	0.099	0.921	-0.0304	0.1779	-0.171	0.865
Humidity	0.0067	0.0049	1.357	0.175	0.0064	0.0058	1.105	0.269
Radiation	-0.0020	0.0016	-1.250	0.211	-0.0014	0.0019	-0.741	0.459
Precipitation								
Duration	0.1710	0.0267	6.411	<0.001	0.1752	0.0353	4.958	<0.001
Intensity	0.0591	0.0309	1.912	0.056	0.0566	0.0386	1.466	0.143
Weekday								
Monday	0.5896	0.1230	4.795	<0.001	0.5480	0.1492	3.673	<0.001
Tuesday	0.4641	0.1239	3.746	<0.001	0.4218	0.1476	2.857	0.004
Wednesday	0.4272	0.1254	3.407	<0.001	0.4105	0.1491	2.753	0.006
Thursday	0.7427	0.1205	6.166	<0.001	0.7146	0.1445	4.945	<0.001
Friday	0.5583	0.1227	4.550	<0.001	0.5523	0.1464	3.773	<0.001
Saturday	0.1130	0.1372	0.823	0.410	0.0786	0.1621	0.485	0.628
Other parameters								
α	0.0016	0.0327	0.048	0.962	-0.1252	0.0271	-0.746	0.456
σ_ϵ^2					0.1613	-	-	-

Table 6.20: Results based on the fitted INAR and Zeger's regression model for Valkenburg

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	2.6276	1.1885	2.211	0.027	2.6049	1.2236	2.129	0.033
Wind								
Direction	-0.0345	0.1270	-0.272	0.786	-0.0427	0.1318	-0.324	0.746
Speed	-0.0058	0.0031	-1.879	0.060	-0.0058	0.0032	-1.807	0.071
Mean Temp	0.0000	0.0021	0.009	0.993	0.0000	0.0022	0.009	0.993
Temp below zero	-0.1790	0.5159	-0.347	0.729	-0.2494	0.5268	-0.473	0.636
Humidity	-0.0348	0.0123	-2.821	0.005	-0.0344	0.0127	-2.708	0.007
Radiation	-0.0101	0.0042	-2.390	0.017	-0.0102	0.0043	-2.336	0.020
Precipitation								
Duration	0.0706	0.0844	0.836	0.403	0.0577	0.0884	0.652	0.514
Intensity	0.0042	0.0745	0.057	0.955	0.0112	0.0755	0.149	0.882
Weekday								
Monday	0.1052	0.3103	0.339	0.735	0.1017	0.3288	0.309	0.757
Tuesday	-0.1661	0.3317	-0.501	0.617	-0.1883	0.3408	-0.553	0.581
Wednesday	0.2011	0.3050	0.659	0.510	0.2072	0.3171	0.654	0.514
Thursday	0.1793	0.3034	0.591	0.555	0.1636	0.3181	0.514	0.607
Friday	-0.0530	0.3179	-0.167	0.868	-0.0477	0.3263	-0.146	0.884
Saturday	0.1266	0.3070	0.412	0.680	0.1088	0.3261	0.334	0.739
Other parameters								
α	0.0000	0.0491	0.000	1.000	-0.5592	0.1358	-0.856	0.392
σ_ϵ^2					0.2079	-	-	-

Table 6.21: Results based on the fitted INAR and Zeger's regression model for Vlissingen

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	0.7703	0.8326	0.925	0.355	0.9605	0.8533	1.126	0.260
Wind								
Direction	-0.0008	0.0637	-0.013	0.990	0.0026	0.0662	0.039	0.969
Speed	-0.0051	0.0029	-1.740	0.082	-0.0055	0.0031	-1.803	0.071
Mean Temp	-0.0009	0.0013	-0.672	0.501	-0.0008	0.0014	-0.559	0.576
Temp below zero	0.1816	0.1802	1.007	0.314	0.1659	0.1895	0.876	0.381
Humidity	-0.0054	0.0081	-0.665	0.506	-0.0072	0.0084	-0.848	0.397
Radiation	0.0019	0.0032	0.580	0.562	0.0014	0.0033	0.429	0.668
Precipitation								
Duration	0.1373	0.0336	4.083	<0.001	0.1405	0.0356	3.943	<0.001
Intensity	-0.0531	0.0534	-0.995	0.320	-0.0514	0.0546	-0.942	0.346
Weekday								
Monday	0.2518	0.1661	1.516	0.130	0.2400	0.1696	1.415	0.157
Tuesday	0.1038	0.1736	0.598	0.550	0.0849	0.1774	0.479	0.632
Wednesday	0.1720	0.1679	1.025	0.306	0.1683	0.1733	0.971	0.332
Thursday	0.1774	0.1711	1.037	0.300	0.1808	0.1758	1.029	0.304
Friday	0.4577	0.1606	2.850	0.004	0.4512	0.1653	2.729	0.006
Saturday	-0.0318	0.1827	-0.174	0.862	-0.0286	0.1840	-0.156	0.876
Other parameters								
α	0.0192	0.0476	0.403	0.687	0.2399	0.0386	0.458	0.647
σ_ϵ^2					0.0736	-	-	-

Table 6.22: Results based on the fitted INAR and Zeger's regression model for Volkel

	INAR				Zeger			
	estimate	st.error	t-value	p-value	estimate	st.error	t-value	p-value
Constant	-1.2200	1.2375	-0.986	0.324	-1.1864	1.3160	-0.902	0.367
Wind								
Direction	0.1138	0.1024	1.112	0.266	0.1195	0.1089	1.098	0.272
Speed	0.0009	0.0036	0.251	0.802	0.0009	0.0039	0.220	0.826
Mean Temp	0.0045	0.0019	2.359	0.018	0.0044	0.0021	2.102	0.036
Temp below zero	0.2081	0.4488	0.464	0.643	0.1497	0.4685	0.320	0.749
Humidity	-0.0008	0.0126	-0.065	0.949	-0.0012	0.0135	-0.089	0.929
Radiation	-0.0012	0.0038	-0.313	0.754	-0.0010	0.0041	-0.254	0.800
Precipitation								
Duration	0.0511	0.0652	0.783	0.433	0.0542	0.0692	0.783	0.434
Intensity	0.0983	0.0441	2.228	0.026	0.1007	0.0489	2.057	0.040
Weekday								
Monday	0.2880	0.2504	1.150	0.250	0.2877	0.2621	1.098	0.272
Tuesday	-0.0345	0.2730	-0.126	0.900	-0.0277	0.2838	-0.097	0.922
Wednesday	-0.3197	0.2937	-1.089	0.276	-0.3114	0.3065	-1.016	0.310
Thursday	0.2770	0.2576	1.075	0.282	0.2823	0.2718	1.039	0.299
Friday	0.1752	0.2620	0.669	0.504	0.1879	0.2737	0.687	0.492
Saturday	0.1868	0.2596	0.720	0.472	0.1928	0.2711	0.711	0.477
Other parameters								
α	0.0000	0.0423	0.000	1.000	0.2261	0.0989	0.426	0.670
σ_ϵ^2					0.1860	-	-	-

Table 6.23: Results based on the fitted INAR and Zeger's regression model for Whilhelminadorp

Appendix C

Meta-analysis results

This appendix provides the weighted forest plots and fitted meta-analysis fixed and random regression models for the covariates that were not statistical significant.

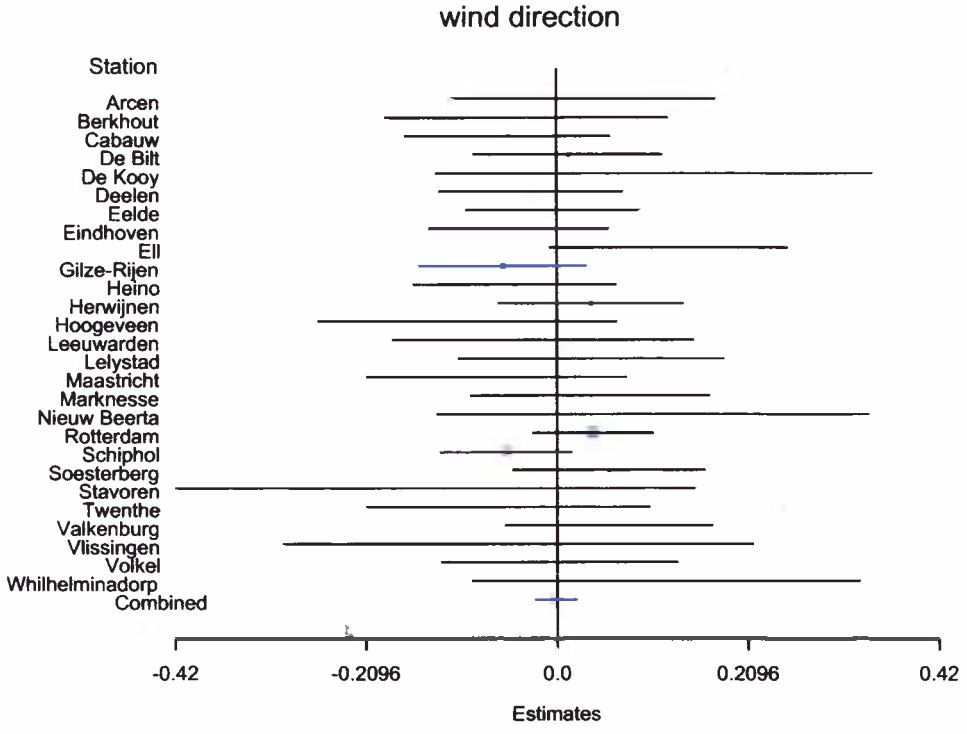

Figure 6.1: Weighted forest plot for the wind direction

	Fixed effects		Random effects		Test of homogeneity
	θ	τ^2	θ	τ^2	Q
estimate	-0.0023	0.0000	-0.0023	0.0000	20.8967
standard error	0.0117	-	0.0117	0.0008	-
t-value	-0.1995	-	-0.1995	0.0000	-
p-value	0.8419	-	0.8419	1.000	0.7473
AIC	-72.295		-70.2926		

Table 6.24: Estimated common parameter and inter-variation for the wind direction

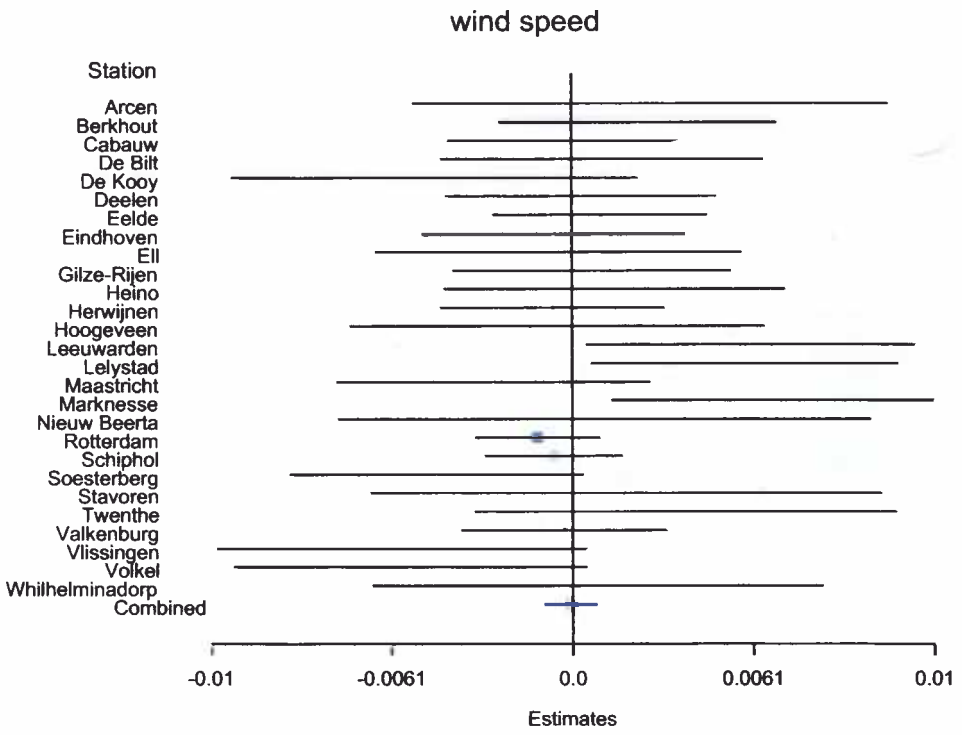

Figure 6.2: Weighted forest plot for the wind speed

	Fixed effects		Random effects		Test of homogeneity
	θ	τ^2	θ	τ^2	Q
estimate	-0.0001	0.0000	-0.0001	0.0000	32.9156
standard error	0.0004	-	0.0004	0.0000	-
t-value	-0.1666	-	-0.1664	0.0004	-
p-value	0.8677	-	0.8678	0.9997	0.1646
AIC	-235.5824		-233.5804		

Table 6.25: Estimated common parameter and inter-variation for the wind speed

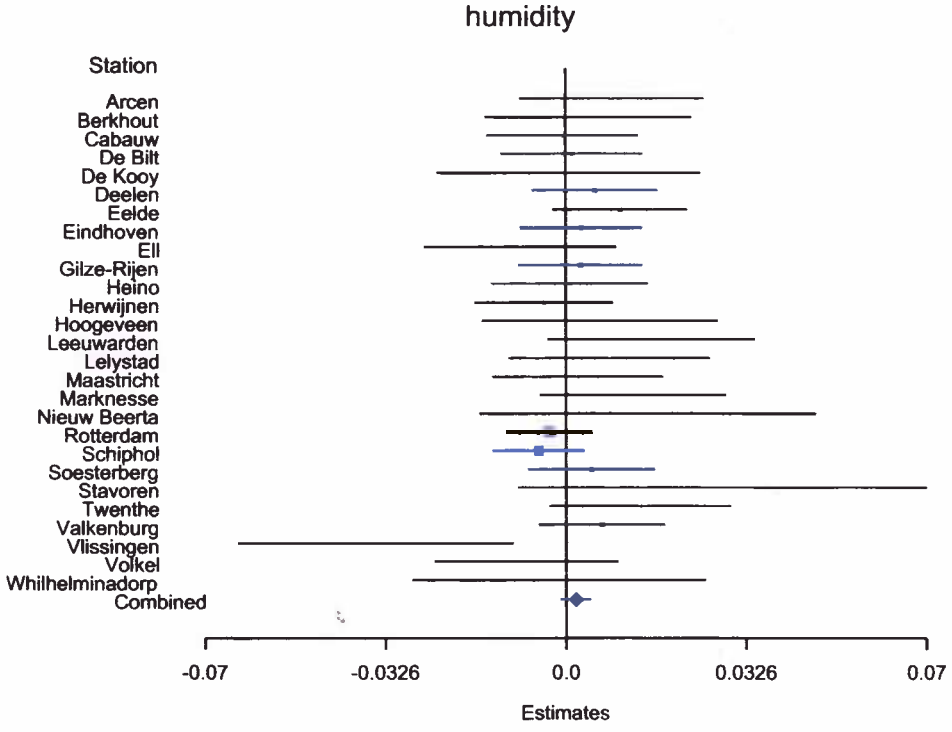

Figure 6.3: Weighted forest plot for the humidity

	Fixed effects		Random effects		Test of homogeneity
	θ	τ^2	θ	τ^2	Q
estimate	0.0018	0.0000	0.0018	0.0000	27.9048
standard error	0.0014	-	0.0014	0.0000	-
t-value	1.2892	-	1.2907	0.0090	-
p-value	0.1973	-	0.1968	0.9928	0.3632
AIC	-181.0422		-179.0401		

Table 6.26: Estimated common parameter and inter-variation for the humidity

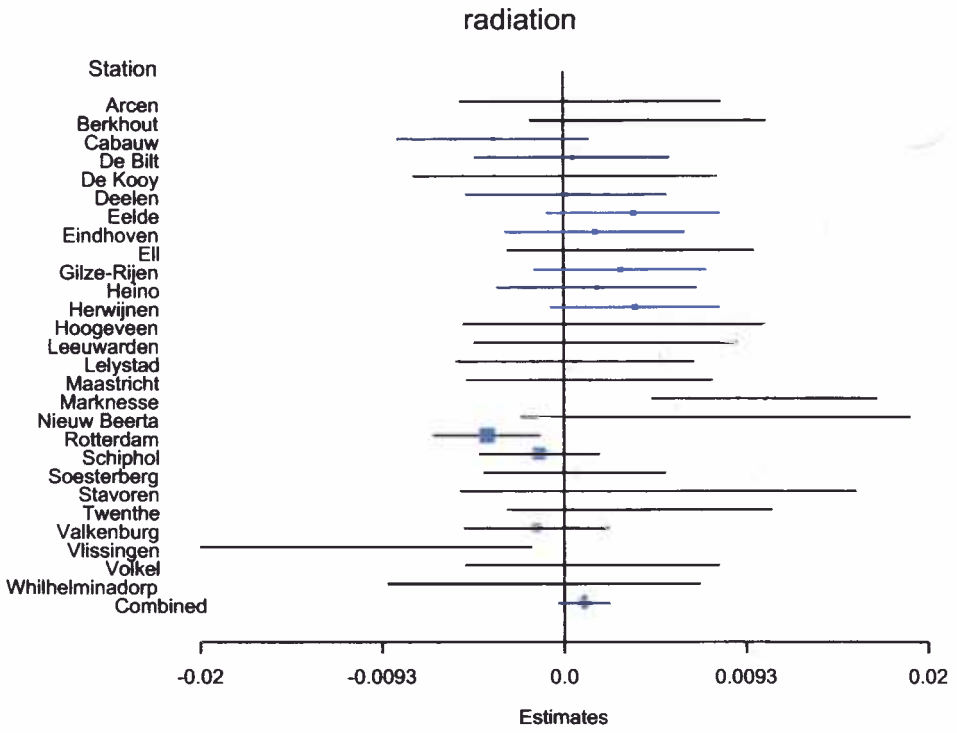

Figure 6.4: Weighted forest plot for the radiation

	Fixed effects		Random effects		Test of homogeneity
	θ	τ^2	θ	τ^2	Q
estimate	0.0006	0.0000	0.0010	0.0000	44.2473
standard error	0.0005	-	0.0007	0.0000	-
t-value	1.1385	-	1.5354	1.3699	-
p-value	0.2549	-	0.1247	0.1707	0.0142
AIC	-219.340		-223.7851		

Table 6.27: Estimated common parameter and inter-variation for the radiation

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.

Al-Osh M.A. and Aly E.A.A. (1992). First-Order Autoregressive Time Series with Negative Binomial and Geometric Marginals, *Communications in Statistics - Theory and Methods*, **21**, 2483-2492.

Al-Zaid A.A. and Al-Osh M.A. (1988). First Order Integer Valued Autoregressive (INAR(1)) Process: Distributional and Regression Properties, *Statistica Neerlandica*, **42**, 53-62.

Al-Zaid A.A. and Al-Osh M.A. (1990). An Integer-Valued pth-order Autoregressive Structure (INAR(p)) Process, *Journal of Applied Probability*, **27**, 314-324.

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

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

Berkley C.S., Hoaglin D.C., Mosteller F. and Colditz G.A. (1995). A Random Effects Regression Model for Meta-Analysis, *Statistics in Medicine*, **14**, 395-411.

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

Brannas K. and Hellstrom J. (2001). Generalized Integer-Valued Autoregression, *Econometric Reviews*, **20**, 425-443.

Brannas K. and Johansson (1994). Time Series Count Data Regression, *Communications in Statistics - Theory and Methods*, **23(10)**, 2907-2925.

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.

Brijs T., Karlis D. and Wets G. (2004). An Integer Autoregressive Model to Study the Effect of Weather Conditions on Daily Car Accident Counts, submitted.

Böckenholt U. (1999). Mixed INAR(1) Poisson Regression Models: Analyzing Heterogeneity and Serial Dependence in Longitudinal Count Data, *Journal of Econometrics*, **89**, 317-338.

Cardinal M., Roy R. and Lambert J. (1999). On the Application of Integer-Valued Time Series Models for the Analysis of Disease Incidence, *Statistics in Medicine*, **18(15)**, 2025-2039.

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

Chan K.S. and Ledolter J. (1995). Monte Carlo EM estimation for time series involving counts, *Journal of the American Statistical Association*, **90**, 242-251.

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.

Cox, D.R. **(1981).** Statistical analysis of time series, some recent developments, *Scandinavian Journal of Statistics*, **8**, 93-115.

DaSilva M.E. and Oliveira V.L. (2004). Difference Equations for the Higher-Order Moments and Cumulants of the INAR(1) Model, *Journal of Time Series Analysis*, **25(3)**, 317-333.

Davis R.A., Dunsmuir W.T. and Wang Y. (1999). Modelling Time Series of Count Data, *Asymptotics, Nonparametrics and Time Series*, 63-114.

Davis R.A., Dunsmuir W.T. and Wang Y. (2000). On Autocorrelation in a Poisson Regression Model, *Biometrika*, **87**, 491-505.

Dean C. and Lawless J.F. (1989). Tests for detecting overdispersion in Poisson regression modes, *Journal of the American Statistical Association*, **84**, 467-472.

DerSimonian R. and Laird N. (1986). Meta-Analysis in Clinical Trials, *Controlled Clinical Trials*, **7**, 177-188.

Du J.G. and Li Y. (1991). The Integer-Valued (INAR(p)) model, *Journal of Time Series Analysis*, **12(2)**, 129-142.

Fahrmeir L. and Tutz G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer-Verlag, New York, Chapters 6-8.

Fleiss J.L. (1993). The statistical basis of meta-analysis, *Statistical Methods in Medical Research*, **2**, 121-145.

Franke J. and Seligmann (1993). Conditional Maximum Likelihood Estimates for INAR(1) processes and their application to modelling epileptic seizure counts , *In: Developments in Time Series Analysis*, Chapman and Hall, London.

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

Freeland R.K. and McCabe B.P.M. (2004a). Analysis of Low Count Time Series Data by Poisson Autoregression, *Journal of Time Series Analysis*, **25(5)**, 701-722.

Freeland R.K. and McCabe B.P.M. (2004b). Forecasting Discrete Valued Low Count Time Series, *International Journal of Forecasting*, **20**, 427-434.

Freeland R.K. and McCabe B.P.M. (2005). Asymptotic Properties of CLS Estimators in the Poisson AR(1) Model, *Statistics and Probability Letters*, **73(2)**, 147-153.

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

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

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

Gourieroux C. and Jasiak J. (2003). Heterogeneous INAR(1) model with application to car insurance, *Insurance Mathematics and Economics*, **33(2)**, 419-419 .

Greenland S. (1987). Quantitative Methods in the Review of Epidemiologic Literature, *Epidemiologic Reviews*, **9**, 1-30.

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.

Heinen A. and Rengifo E. (2004). Multivariate Autoregressive Modelling of Time Series Count Data Using Copulas, *Econometric Society 2004 Far Eastern Meetings*, **755**, Econometric Society.

Jin-Guan D. and Yuan L. (1991). The Integer-Valued Autoregressive (INAR(p)) Model, *Journal of Time Series Analysis*, **12**, 129-42.

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.

Jung R.C., Ronning G. and Tremayne A.R. (2005). Estimation in Conditional First Order Autoregression with Discrete Support, *Statistical Papers*, **46**, 195-224.

Jung R.C. and Tremayne A.R. (2006). Coherent Forecasting in Integer Time Series Models, *International Journal of Forecasting*, in press.

Karlis D. and Xekalaki E. (2001). ML Estimation For Integer Valued Time Series Models, *Proceedings of the 5th Hellenic-European Conference on Computer Mathematics and its Applications*, Athens, Greece.

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

McCabe B.P.M. and Martin G.M. (2005). Bayesian Predictions of Low Count Time Series, *International Journal of Forecasting*, **21**, 315-330.

McCullagh P. and Nelder J.A. (1989). *Generalized Linear Models.* 2nd Edition, Chapman and Hall, London.

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). Modelling 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, *Traffic Engineering*, **43**, 14-20.

Pavlopoulos H. and Karlis D. (2006). On Time Series of Overdispersed Counts: Modelling, Inference, Simulation and Prediction, *Technical Report, Department of Statistics, Athens University of Economics*, No221.

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, *Accidents 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.

Smith T.C., Spiegelhalter D.J. and Thomas A. (1995). Bayesian Approaches to meta-analysis: a comparative study, *Statistics in Medicine*, **14**, 2685-2699.

Thompson S.G. and Sharp S.J. (1999). Explaining Heterogeneity in Meta-Analysis: A Comparison of Methods, *Statistics in Medicine*, **18**, 2693-2708.

Thyregod P., Carstensen J., Madsen H. and Arnbjerg-Nielsen (1999). Integer-Valued Autoregressive Models for Tipping Bucket Rainfall Measurements, *Environmetrics*, **10**, 395-411.

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.

Varin C. and Vidoni P. (2005). Pairwise likelihood inference for ordinal categorical time series, submitted.

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 Meeting*, Washington D.C., USA.

Yiokari N., Xekalaki E. and Karlis D. (2001). On Some Discrete-Valued Time Series Models Based on Mixtures and Thinning, *Proceedings of the 5th Hellenic-European Conference on Computer Mathematics and its Applications*, Athens, Greece.

Zeger S.L. (1988). A Regression Model for Time Series of Counts, *Biometrika*, **75(4)**, 621-9.

Δωρεά.