Time series analysis of count data with an application to the incidence of cholera

Author
Jennifer Patricia Holloway

Supervisor
Professor Linda Haines

Co-supervisors
Dr. Kerry Leask and Dr. Chris Elphinstone

Submitted to the Department of Statistical Sciences
in fulfillment of the requirements for the degree of

Master of Science in Mathematical Statistics
at the
University of Cape Town

May 30, 2011

Abstract

This dissertation comprises a study into the application of count data time series models to weekly counts of cholera cases that have been recorded in Beira, Mozambique. The study specifically looks at two classes of time series models for count data, namely observation-driven and parameter-driven, and two models from each of these classes are investigated. The autoregressive conditional Poisson (ACP) and double autoregressive conditional Poisson (DACP) are considered under the observation-driven class, while the parameter-driven models used are the Poisson-gamma and stochastic autoregressive mean (SAM) model. An in-depth case study of the cholera counts is presented in which the four selected count data time series models are compared. In addition the time series models are compared to static Poisson and negative binomial regression, thereby indicating the benefits gained in using count data time series models when the counts exhibit serial correlation. In the process of comparing the models, the effect of environmental drivers on the outbreaks of cholera are observed and discussed.

Plagiarism Declaration

1. This dissertation is my own work. It has not been submitted before for any degree or examination to any other University.
2. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
3. Each significant contribution to, and quotation in, this dissertation from the work of other people has been cited and referenced.

Signature: _____

Date: _____

Acknowledgements

I would like to thank my supervisor, Professor Linda Haines, for firstly supervising my Masters studies from a distance and secondly for her patience and support in helping me tackle the selected project. Her continuous and dedicated efforts to help me grasp the theoretical aspects of the study and her enthusiasm have been invaluable. I would also like to thank my co-supervisor from UCT, Dr. Kerry Leask for her assistance, particularly with regard to certain computational hurdles that were encountered, and my co-supervisor from CSIR, Dr. Chris Elphinstone, for his endless encouragement and practical advice.

I would like to give a special thanks to the Department of Health of the city of Beira city and Sofala province in Mozambique and the CHAEM laboratory in Beira who supplied the CSIR with the cholera data for the initial study.

I would also like to thank Professor Andrew Harvey of Cambridge University for his response to questions regarding the Poisson-gamma model and to Professor Robert Jung from the University of Tübingen for supplying an extended version of his article, together with his GAUSS program for estimating the SAM model and the asthma data. His program was invaluable in helping me to understand the many computational steps involved in estimating the SAM model.

Lastly, I would like to thank the CSIR and in particular my manager, Theo Stylianides, for financial support and encouragement throughout the duration of my studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General overview

A research project was recently conducted by the Council for Scientific and Industrial Research (CSIR) to gain an understanding of the ecology of the cholera bacterium, i.e. *Vibrio cholerae*, found in the rivers and coastal waters of Mozambique. One component of this project involved the mathematical and statistical analyses of cholera count data that had been recorded at the hospitals and clinics in Beira. In terms of the statistical analysis, two approaches were used to model the cholera case data, of which the first involved standard techniques applicable to count data and the second involved techniques suitable for continuous time series data. The application of these techniques highlighted a gap in the analysis, namely the need to investigate time series methods applicable to count data. This dissertation provided an opportunity to research such techniques.

Initial studies of the literature revealed a large number of potential count data time series models. Although the topic is more recent than that relating to the well known continuous time series and standard packages for fitting the appropriate models are not typically available, numerous articles have been written which document both the theory and application of such models. What was specifically of interest was the successful application of some of the documented techniques to patient count data for other medical conditions or diseases, such as asthma and polio. In this dissertation an overview of some of these methods is provided, but with emphasis on the models which fall in either the observation-driven or parameter-driven classes as initially defined by Cox (1981). Two observation-driven models, namely the autoregressive conditional Poisson (ACP) and the double autoregressive conditional Poisson (DACP) both developed by Heinen (2003), and two parameter-driven models, referred to as the Poisson-gamma (Harvey and Fernandes, 1989a) and stochastic autoregressive mean (SAM) (Jung

et al., 2006), were selected for an in-depth study. This dissertation documents the formulation and theory of these four models and the application of these models to the cholera data. Other models that are documented in the literature as having specifically been used to model cholera case data are also noted. The four selected models are all fitted to the cholera data using the available climatic variables as drivers. The fits of these models are compared to each other as well as to the static Poisson and negative binomial regression models and a discussion of the overall comparison is provided. Although the significance and choice of the climatic variables in predicting cholera cases is still of interest in this study, it is the comparison of the selected discrete time series models and their relative performance when fitted to cholera case data that is of main importance in this dissertation.

1.2 Overview of cholera problem

As part of the cholera research project, initiated by the CSIR, data on cholera outbreaks in Beira were made available. The city of Beira in Mozambique is plagued with cholera almost every year and various environmental conditions have been suggested as contributing to these outbreaks. As a result, the first objective of the initial study was to establish whether relationships between the environmental factors and the outbreaks of cholera exist and whether these relationships support the hypothesis that climatic conditions drive the proliferation of cholera cases. The second objective was to develop an early warning system to predict future outbreaks of cholera. The time series data that were made available for this study consisted of weekly cholera counts together with weekly air temperature, rainfall and humidity, spanning the six year period from January 1999 to December 2004.

As mentioned in Section 1.1, two approaches were initially considered in addressing these objectives from a statistical modelling point of view, firstly using available static count data techniques and secondly using dynamic continuous time series techniques. Some of this statistical modelling has been reported in detail in Van der Berg *et al.* (2008). The first approach acknowledged the fact that cholera cases are counts rather than measurements on a continuous scale which can become negative and the techniques applied in this approach were Poisson regression and negative binomial regression. These had the disadvantage of assuming independence of the observations and thus ignoring any correlation over time amongst the cholera counts. In contrast to these techniques, the second approach assumed that the counts could be approximated on a continuous scale and hence that the data could be treated as a time series where data points may be dependent over time, i.e. may exhibit serial correlation. ARIMA and dynamic regression models, as defined by Pankratz (1991), were used in this approach. In most cases,

techniques for continuous data can be used on count data where the values of the events are far from zero but in this case study there were many time periods with no reported cholera cases. As a result, although the dynamic regression models fitted the data well, several forecasted points included small negative values.

Taking into consideration the disadvantages of both the above-mentioned techniques, the use of time series methods for count data was deemed relevant for further study. The aim in this study is not to address the same objectives of the initial cholera project, but rather to explore models that can better represent the relationship between the number of cholera cases and climatic variables, taking into account the inherent properties that exist in such time series data of counts.

1.3 Structure of the dissertation

In order to document the research study, this dissertation is divided into five chapters. A literature survey which firstly gives an overview of count data time series models and secondly looks at the numerical techniques that have previously been used in studies of cholera is provided in Chapter 2. The detailed theory of the four selected count data time series models is presented in Chapter 3, with the section on the simpler observation-driven ACP and DACP models described first, followed by the section on the observation-driven Poisson-gamma and SAM models. This chapter describes the formulation and implementation of these models, together with diagnostic and forecasting aspects. Chapter 4 looks at the actual cholera case data, starting with a description and exploratory analysis of the data and moving to the results obtained from the fitted models. Here the results of the static Poisson and negative binomial regression models are given for comparison purposes, followed by the results from the observation-driven ACP and DACP models and then the results of the two selected parameter-driven models, namely the Poisson-gamma and the SAM. This chapter ends with a comparison of the fits of these models to the cholera data. In the concluding chapter, Chapter 5, an initial summary is given followed by a general comparison of the models, including comments on the computational aspects of the models. Overall conclusions for the cholera case study are also provided and the chapter ends with some final remarks and recommendations.

Chapter 2

Literature overview

2.1 Introduction to literature review

This chapter covers an overview of the literature on time series models for count data and a literature survey on models that have been applied to cholera data. Section 2.2 briefly describes some of the time series models for count data that were reviewed, together with the models that were finally selected for further study. Section 2.3 looks at both mathematical and statistical models that have been used in the analysis of cholera data and some of the findings with respect to the relationships between environmental drivers and cholera counts. An overall summary is provided at the end of this chapter.

2.2 Review of time series models for count data

When analysing count data together with explanatory variables the starting point typically involves the use of Poisson regression but for count data that are recorded in the form of a time series, the assumption regarding independence of observations becomes a problem. There are numerous possible ways of introducing dependency into time series models for count data and many such models have been developed. There have been several reviews of time series count data models, including those of Fahrmeir and Tutz (1994), Brockwell and Davis (1996), MacDonald and Zucchini (1997) and McKenzie (2003). Cameron and Trivedi (1998) also devoted a chapter of their book to reviewing several of the methods that have appeared in various articles. For the purpose of this overview, more focus is placed on the review by Cameron and Trivedi (1998) but with the inclusion of a few models from some more recent papers. The system initially proposed by Cox (1981) will also be used in classifying certain types of models. Specifically, Cox (1981) referred to the models as either observation-driven or parameter-driven but it should be noted that there are also some models that do not fall into either of these

categories.

In order to find models that could be suitable for the modelling of cholera counts, various observation-driven and parameter-driven models were first considered. For the purpose of this study it was decided that only models falling in either of these two categories would be selected and that the two categories of models would later be compared. An overview of these models is provided in the next two subsections, together with the models chosen for the cholera case study. The third subsection of this chapter describes some other models that do not fall into either of these classifications, but no models from this group were selected as they were deemed to be beyond the scope of the study.

2.2.1 Observation-driven models

In the observation-driven models, the observations are typically assumed to follow a Poisson distribution but in addition, lagged values of the observed variable are incorporated directly into the calculation of the mean function. Considering observations from a time series of counts y_t, for $t = 1, ..., T$, the generic format of observation-driven models can be written as

$$y_t|Y_{t-1} \sim Poisson(\mu_t),$$

where Y_{t-1} denotes the information on the observations up to $t-1$, and with the logarithm of the conditional mean at time t comprising a linear combination of explanatory variables and functions of the lagged observations. The generic format of the conditional mean is expressed as

$$\ln(\mu_t) = x_t'\delta + \sum_{i=1}^{p} \phi_i f(y_{t-i}), \tag{2.1}$$

where x_t is a vector of explanatory variables at time t, δ is the vector of unknown parameters associated with these explanatory variables, $f(y_{t-i})$ is a function of the lagged observations, p indicates the number of lags and ϕ_i is the parameter associated with the function $f(y_{t-i})$.

In the observation-driven class of models, the most basic model is the autoregressive model where the function $f(y_{t-i})$ in equation (2.1) is simply set equal to y_{t-i} (Cameron and Trivedi, 1998). The conditional mean, μ_t, is therefore specified as

$$\mu_t = \exp\left(x_t'\delta + \sum_{i=1}^{p} \phi_i y_{t-i}\right)$$

and describes the inclusion of lagged values of the observed counts as explanatory variables in a Poisson regression. However, for $p = 1$, with

$\mu_t = \exp(x_t'\delta + \phi y_{t-1})$, Zeger and Qaqish (1988) and Cameron and Trivedi (1998) indicated that the process cannot be stationary for $\phi > 0$ and is therefore not practically useful since $\phi \leq 0$ would imply that there can be no positive dependence on past observations.

The autoregressive model, initially developed by Zeger and Qaqish (1988) for generalised linear models and which is referred to as the multiplicative AR(1) model by Cameron and Trivedi (1998), makes more practical sense. This model specifies the function of lagged observations in equation (2.1) as $f(y_{t-1}) = \ln(y^*_{t-1})$, with the number of lags, p, set equal to 1 and where y^*_{t-1} is a transformation of y_{t-1} in order to ensure that y^*_{t-1} is always greater than zero. This allows the conditional mean of the model to be expressed as

$$\mu_t = \exp(x_t'\delta)(y^*_{t-1})^{\phi}. \tag{2.2}$$

An alternative formulation given by Zeger and Qaqish (1988) is that in which $f(y_{t-1}) = \ln(y^*_{t-1}) - x'_{t-1}\delta$ and the lag p is also of order 1. Substituting this function of the lagged observations into equation (2.1) results in the following expression for the conditional mean:

$$\mu_t = \exp(x_t'\delta)\left(\frac{y^*_{t-1}}{\exp(x'_{t-1}\delta)}\right)^{\phi}. \tag{2.3}$$

This is referred to as a multiplicative AR(1) error model by Cameron and Trivedi (1998) while Zeger and Qaqish (1988) referred to the models in (2.2) and (2.3) as Markov models of order 1.

Details of the transformations of y_{t-1} to y^*_{t-1} used in the models (2.2) and (2.3) are given in Zeger and Qaqish (1988) and summarised in Cameron and Trivedi (1998). An example of one of these transformations is given as

$$y^*_{t-1} = y_{t-1} + c,$$

where the constant c is defined such that $0 < c < 1$. However, Cameron and Trivedi (1998) also pointed out that the down side of such transformations for data comprising zero values of y_t is that they are fairly ad-hoc. An added caution is that, when introducing explanatory variables, the effect of these variables on the change in conditional mean is harder to evaluate when transformations of the lagged observations are also included. These points are also noted by Davis *et al.* (1999). Although the implementation of these models is said to be relatively simple, they were not considered further due to the high number of zeros in the cholera data being analysed.

A further group of autoregressive models, sometimes referred to as generalized linear autoregressive moving average (GLARMA) models, also falls into

the observation-driven class of models. These models were initially proposed by Shephard (1995) and introduce the past count values into the function $f(y_{t-i})$ of the conditional mean in (2.1) via a linear combination of weighted residuals, $e_{t-i} = \dfrac{(Y_{t-i} - \mu_{t-i})}{\mu_{t-i}}$. Davis *et al.* (1999, 2003) suggested an adaption to this GLARMA model in which e_{t-i} is expressed as a scaled residual through the addition of a parameter λ such that $e_{t-i} = \dfrac{(Y_{t-i} - \mu_{t-i})}{(\mu_{t-i})^{\lambda}}$. This gives the expression for the conditional mean as

$$\mu_t = \exp\Big(x_t'\delta + \sum_{i=1}^{p} \phi_i \frac{(Y_{t-i} - \mu_{t-i})}{(\mu_{t-i})^{\lambda}}\Big).$$

Davis *et al.* (2003) indicated that this model is easy to fit via conditional maximum likelihood if the parameter λ is taken as fixed, and is additionally easy to use in forecasting. The down side, however, is that, similarly to the AR(1) models, when explanatory variables are added the interpretation of the effects of these variables on the mean may be difficult (Davis *et al.*, 1999, 2003). This is due to the manner in which past observations are included in the mean function. Davis *et al.* (1999) applied their adapted GLARMA model to polio counts and asthma data while Davis *et al.* (2003) used it to model the same asthma counts with the additional inclusion of pollution and meteorological effects. This model was, however, not selected for the analysis of the cholera data due to the issues regarding interpretation of the explanatory variables.

The autoregressive conditional Poisson (ACP) model, which is a model proposed fairly recently by Heinen (2003), is similar to the previous models but in the conditional mean of equation (2.1), the $\sum_{i=1}^{p} \phi_i f(y_{t-i})$ component is given as $\ln(\omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \mu_{t-j})$ where $\omega, \alpha_j, \beta_j$ are all unknown parameters with the condition that $\omega > 0$ and $\alpha_j, \beta_j \geq 0$. Hence the conditional mean for the ACP(p, q) model can be expressed as

$$\mu_t = \exp(x_t'\delta)(\omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \mu_{t-j}).$$

Heinen (2003) also proposed an extension to this model, called the double autoregressive conditional Poisson (DACP) model which accommodates both under-dispersion and over-dispersion in its conditional distribution. In essence it should actually be called an autoregressive conditional double Poisson (ACDP) since it does not have a double autoregressive component

but rather replaces the Poisson distribution with the double Poisson distribution introduced by Efron (1986). Jung *et al.* (2006) indicated that the advantages of both the GLARMA and ACP models is that the introduction of explanatory variables is fairly straightforward and that the implementation of the models by maximum likelihood (ML) techniques is easy to do. The GLARMA model, however, has the problem of interpretation of covariate effects and therefore the ACP model is favoured in the present study. Both Heinen (2003) and Jung *et al.* (2006) successfully applied the ACP model to data relating to patient counts of diseases or illnesses, these being polio and asthma counts respectively. Consequently, this model, together with the DACP variation, were selected as the observation-driven models to be applied to the cholera count data. The detailed theory of these models is given in Section 3.2.

2.2.2 Parameter-driven models

As with the observation-driven models, the parameter-driven models have observations that are typically assumed to follow a Poisson distribution. The mean function, however, is determined by a latent dynamic process which evolves independently of the past observations (Davis *et al.*, 2003). The implementation of parameter-driven models was initially considered to be too computationally intensive to be feasible but with the advancement of fast computers this is no longer a problem (Davis *et al.*, 2003; Cameron and Trivedi, 1998).

Taking a time series of observed counts y_t, for $t = 1, ..., T$, the generic format of parameter-driven models can be written as

$$y_t|\mu_t \sim Poisson(\mu_t),$$

where the logarithm of the conditional mean at time t, μ_t, consists of a linear combination of explanatory variables and the logarithm of a random variable ε_t. The conditional mean is therefore given by

$$\ln(\mu_t) = x_t'\delta + \ln(\varepsilon_t), \tag{2.4}$$

where x_t defines the vector of explanatory variables at time t and δ is the corresponding vector of parameters associated with these variables. The random variable ε_t of this parameter-driven model can be defined and modelled using two types of methods. The first of these methods defines ε_t as a multiplicative error term while the second method follows a state space formulation with ε_t described by a transition equation. These two methods are now described.

Multiplicative error terms

We consider here the parameter-driven models with a multiplicative error term where $\mu_t = \exp(x_t'\delta)\varepsilon_t$. The most well known parameter-driven model is the one developed by Zeger (1988) in which the serial correlation in y_t was introduced by taking the latent variable ε_t in equation (2.4) to be a stationary process with mean 1 and variance σ^2. This model is referred to as a serially correlated error model by Cameron and Trivedi (1998) and has also been called a marginal model due to the fact that μ_t is not conditional on lagged values of y_t. Many authors have studied Zeger's model further, including Brännäs and Johansson (1994), and it has frequently been used as a benchmark for comparing other parameter-driven models. Since no distributional assumptions are made, Zeger (1988) adopted a quasi-likelihood approach to parameter estimation. Zeger (1988) applied this model to monthly US polio counts and Campbell (1994) used Zeger's model to determine the effect of temperature on counts of sudden infant death syndrome (SIDS).

Davis *et al.* (1999, 2000) considered the same parameter-driven model used by Zeger (1988) but introduced a stationary Gaussian process through the parameter $\lambda_t = \ln(\varepsilon_t)$. Specifically, λ_t is taken to follow a $N(-\frac{\sigma_\lambda^2}{2}, \sigma_\lambda^2)$ distribution which implies that the mean is half the variance. This model satisfies the condition in the model of Zeger (1988) whereby $E(\varepsilon_t) = E(e^{\lambda_t}) = 1$ (Davis *et al.*, 2000) and also implies that ε_t is lognormal. The mean function for this model can be written as

$$\mu_t = \exp(x_t'\delta + \lambda_t).$$

Davis *et al.* (1999, 2000) used a standard generalised linear model (GLM) approach in order to estimate the model. Davis and Wu (2009) extended the framework of this model, from the Poisson log-linear regression model to include a negative binomial logit regression model.

State space

Harvey and Fernandes (1989a) proposed a parameter-driven model that can be expressed in terms of a state space, or more specifically a structural model formulation, but which uses the convenience of natural conjugates from the Bayesian paradigm. Various distributions of the observations were considered by Harvey and Fernandes (1989a) but, according to Cameron and Trivedi (1998), the most meaningful and attractive of these models is the Poisson-gamma in which the observations are considered to be taken from a Poisson distribution and the mean is assumed to follow a gamma distribution, which is the natural conjugate of the Poisson. For this model, Harvey and Fernandes (1989a), by invoking the state space formulation, based the transition equation for ε_t of (2.4) on an approach used by Smith and Miller

(1986) and used the Kalman filter and ML techniques to estimate the model. Although this Poisson-gamma model does not appear to have been fitted to applications involving disease counts, the estimation of the model and forecasting of future values is reasonably straightforward and consequently it is studied in more depth in Chapter 3 and used in fitting the cholera data in Chapter 4.

The stochastic autoregressive mean (SAM) model implemented by Jung *et al.* (2006) is similar to the Poisson-gamma model in the use of state space formulation but in this model the conjugacy is relaxed. Taking the same mean equation as in (2.4), and taking $\lambda_t = \ln(\varepsilon_t)$, they describe λ_t by means of a transition equation involving a Gaussian first-order autoregressive process, such that

$$\lambda_t = \gamma\lambda_{t-1} + \nu\epsilon_t,$$

where $\epsilon_t \sim NID(0,1)$ and γ and ν are unknown parameters. This model is based on the original model of Zeger (1988) but differs in the formulation of λ_t. The estimation of this model is not straightforward due to the dynamic latent process and the resulting high dimensional integrals over the latent variables which are required to evaluate the likelihood function. As a result, Jung *et al.* (2006) used efficient importance sampling (EIS) techniques to estimate the SAM model. Jung *et al.* (2006) also compared this parameter-driven model to the observation-driven ACP model of Heinen (2003) using asthma data and obtained similar results for both models when comparing the estimated effect of the explanatory variables on the observed series. Although the estimation of the SAM model is not easy to implement, the interesting manner in which EIS is applied to obtain maximum likelihood estimates, together with its successful application to asthma count data by Jung *et al.* (2006) and its simple and natural formulation, were the reasons why it was selected as one of the models for analysing the cholera data in the present study. This model is discussed in detail in Chapter 3 and its application to the cholera data is presented in Chapter 4.

Durbin and Koopman (1997) applied a model to count data that is similar in structure to the SAM model since it can also be expressed in a state space form and has as a starting point the Poisson regression model. In this model the mean of the Poisson is defined as $\mu_t = \exp(x_t'\delta_t + \lambda_t + \gamma_t)$, where γ_t is introduced as a seasonal term and the trend parameter λ_t changes over time according to a random walk given as

$$\lambda_t = \lambda_{t-1} + \epsilon_t,$$

with $\epsilon_t \sim N(0, \Sigma_t)$. This model is referred to by Cameron and Trivedi (1998) as a state space model with normally distributed parameters but there is no closed form solution for the unknown parameters so estimation can only be

achieved numerically and is computer-intensive. Shephard and Pitt (1997) used a Markov Chain Monte Carlo (MCMC) method to evaluate the likelihood of this model while Durbin and Koopman (1997) calculated the likelihood using Kalman filter techniques for an approximating linear Gaussian model and then adjusted it to obtain the true likelihood. The SSPIR package in the R software (R Development Core Team, 2009) accommodates the state space model of Durbin and Koopman (1997).

2.2.3 Other models

There are various other models for time series of count data that have been developed but which do not specifically fit into the same formulation as that of the observation-driven and parameter-driven models. Some of the more well known models include the integer valued autoregressive moving average (INARMA) model, the normally distributed parameter model, hidden Markov models and the autoregressive conditional ordered probit (ACOP) model. These models are mentioned below and some are described briefly. However, none of these models were considered further in terms of analysing the cholera counts since they would constitute an intense study in themselves and do not fall into the observation-driven and parameter-driven classes which have been selected for the scope of this study.

The integer valued AR and ARMA models (INAR and INARMA) were proposed by both McKenzie (2003) and Al-Osh and Alzaid (1987). The integer valued autoregressive process of lag 1 (INAR(1)) for the observations y_t can be written as

$$y_t = \rho_t o y_{t-1} + \varepsilon_t, \tag{2.5}$$

where $0 < \rho_t < 1$, ε_t is a discrete random variable independent of the observations and the symbol o represents the binomial thinning operator of Steutel and van Harn (1979). This implies that each observation y_t is modelled as a combination of a discrete random variable and a "thinned" value of y_{t-1}. Brännäs (1995) extended the INAR(1) model to include explanatory variables and this extended model is referred to as a Poisson INAR(1) regression model. The process for the observations remains the same as that defined in (2.5). However, the explanatory variables are introduced through the random term ε_t and the binomial thinning parameter ρ_t. Specifically, this is done by considering the random term

$$\varepsilon_t \sim Poisson(z_t'\delta)$$

and taking the thinning parameter to be a logistic function given by

$$\rho_t = \frac{1}{1 + \exp^{-x_t'\gamma}}$$

to ensure that $0 < \rho_t < 1$. The vectors of parameters δ and γ are associated with the explanatory variables z_t and x_t respectively and are unknown. This means that, for the Poisson INAR(1) regression model, the observation y_t is essentially being modelled as $Poisson(z_t'\delta)$ with an added component of earlier observations y_{t-1} that are "thinned", or reduced in magnitude, through the use of a thinning parameter. Various methods of estimation for this model have been proposed including conditional nonlinear least squares (NLS), conditional weighted least squares (WLS) (Cameron and Trivedi, 1998), and the generalised method of moments (GMM) (Brännäs, 1995). However, this model is not particularly well established or well tested and the dependence on the explanatory variables can be rather complicated to interpret.

Jacobs and Lewis (1978a,b, 1983) defined a class of models called the discrete ARMA (DARMA) models. However, these models were not used together with explanatory variables.

Another group of count data time series models are the models presented in the book by Zucchini and MacDonald (2009) and referred to as hidden Markov time series models. With hidden Markov models, various parametric models are specified in different regimes, where the unobserved regimes evolve over time according to a Markov chain. Details of these models can be found in Zucchini and MacDonald (2009) and a summary is given in Cameron and Trivedi (1998). These models are powerful and broad in scope.

Jung *et al.* (2006) also extended the ordered probit model discussed in Cameron and Trivedi (1998) to include explanatory variables and to allow for positive and negative serial correlation. This model is referred to as the autoregressive conditional ordered probit (ACOP) model. Jung *et al.* (2006) also fitted the ACOP model to the asthma counts and compared the results to both the ACP and SAM models.

2.3 Review of quantitative analyses done on cholera data

There have been various applications of count data time series models to data involving counts of specific medical conditions but, it would seem from a review of the literature, that no such models have been fitted to cholera counts. Zeger (1988), Davis *et al.* (1999, 2000), Heinen (2003), Davis and Wu (2009) and other authors have applied count data time series models to polio counts while numerous authors have looked at the modelling of disease

counts, other than cholera, and their relationship to environmental variables such as pollution and weather. These include Campbell (1994), who considered the effect of temperature on sudden infant death syndrome (SIDS), and Davis *et al.* (1999, 2000) and Jung *et al.* (2006) who looked at the effect of pollution on asthma counts in Sydney.

Despite there not being applications of discrete time series models to cholera data, the literature on cholera and more specifically the analysis of cholera counts, is extensive. Many authors working on cholera have been modelling time series comprising of counts but using mathematical and statistical models which do not explicitly accommodate both the time series aspects and the discreteness in the data. For the purpose of this study, where the effect of environmental drivers on cholera counts is of interest, some of the types of analyses that have been performed on cholera counts are reviewed and important findings regarding such relationships are noted.

2.3.1 Mathematical models

Numerous articles have been written linking seasonal cycles and climatic conditions, particularly rainfall and temperature, to the occurrence of cholera in endemic areas. Pascual *et al.* (2002) reviewed many of these papers but the main cholera model that they described is mathematical and deterministic, using differential equations to model infection rates relative to several parameters including flow and drainage rates of rivers. Pascual *et al.* (2002) pointed out, however, that further work is required to determine whether the associations between cholera outbreaks and environmental variables are sufficiently strong to allow for predictions of outbreaks.

Other deterministic techniques that have frequently been used in modelling cholera counts include singular spectrum analysis (Pascual *et al.*, 2000; Rodó *et al.*, 2002; Van der Berg *et al.*, 2008), spectral analysis with a fast Fourier transform (Fernández *et al.*, 2009; Van der Berg *et al.*, 2008) and cross-wavelet analysis (Constantin de Magny *et al.*, 2006; Cazelles *et al.*, 2007; Van der Berg *et al.*, 2008). The former two techniques have specifically been utilised for detecting trend and periodicity in cholera counts, while the latter approach was used to explore the links between cholera counts and environmental variables through pairwise comparisons, that is comparing only one explanatory variable at a time to the cholera time series data. These techniques have typically been used for exploratory analysis as a prelude to further statistical modelling (Van der Berg *et al.*, 2008; Fernández *et al.*, 2009) and are not used as models for predicting cholera outbreaks.

2.3.2 Statistical models applied to cholera data

Poisson regression is one of the statistical techniques that has been most commonly used in modelling the relationship between cholera counts and environmental variables. It has been applied by Huq *et al.* (2005) and Masahiro *et al.* (2008) to cholera count data collected in Bangladesh, by Constantin de Magny *et al.* (2008) to cholera from sites in both Bangladesh and India and by Fernández *et al.* (2009) to cholera epidemics in Zambia. Cholera count data typically exhibits a large degree of over-dispersion and the latter two papers have accommodated this by adjusting the standard errors from the Poisson regression according to the degree of over-dispersion in the model. The study by Van der Berg *et al.* (2008), which is discussed in Chapters 1 and 4 as part of the background to this dissertation, also acknowledged the high degree of over-dispersion and included a fit of the negative binomial instead of the Poisson regression model. Negative binomial regression was also used in Emch *et al.* (2008) to study whether season, latitude or the interaction of the two have an effect on the incidence of cholera cases across the world, using data from 140 countries.

These applications of Poisson regression and negative binomial regression have all taken into account the fact that cholera counts are discrete but have generally ignored the time dependency in the data. The exception to this was the study by Fernández *et al.* (2009) which, to some degree, incorporated an autoregressive component into the model by including a one week lag of the number of cholera cases as one of the regressors. A drawback of this study, as noted by the authors, was the unavailability of data in between the epidemics which prevented the use of time series models. The authors, however, indicated that models such as ARIMA would have been preferred had the full time series been collected. The study by Van der Berg *et al.* (2008) fitted both univariate ARIMA and dynamic regression models, which incorporate explanatory variables into an ARIMA, to the Beira cholera dataset, thereby assuming that the response was continuous and Gaussian. In the process of fitting these models, however, the authors acknowledged that using such techniques on count data, particularly with a large number of zeros, was clearly not satisfactory and indeed resulted in some negative values being forecasted.

Non-linear, non-parametric time series models have also been used by Pascual and Ellner (2000) and Pascual *et al.* (2000) with the former authors including a feedback neural network (FNN). The study by Pascual *et al.* (2000) showed a relationship between cholera incidence in Bangladesh and El-Niño-Southern Oscillation (ENSO), seasonality and previous levels of cholera, while Koelle and Pascual (2004) used a semi-parametric time series model which is appropriate for modelling disease dynamics with temporary

immunity. The latter model, being an epidemic model, also required that the size of the susceptible population be known and therefore would not have been applicable in most of the other studies mentioned.

2.3.3 General findings on relationships between cholera counts and explanatory variables

Lipp *et al.* (2002) pointed out that the relationship between climate and health has been a topic of study for a very long time but that recent technologies have led to more interesting observations with regard to how the environment, including weather, plays a role in infectious diseases. They also stated that progress is being made towards developing predictive models for cholera using climatic factors. Most of the literature, covered in the previous sections on mathematical and statistical models, is focused on establishing relationships between cholera counts and environmental factors but only a few of the studies attempted to develop predictive models for cholera epidemics.

In terms of predictive models, the studies by Huq *et al.* (2005), Van der Berg *et al.* (2008), Constantin de Magny *et al.* (2008) and Fernández *et al.* (2009) used either Poisson or negative binomial regression to develop models that could be used to predict cholera epidemics from environmental drivers. Although the results of the predictive models developed by Huq *et al.* (2005) were not consistent across all locations in Bangladesh in terms of lag periods; water temperature, air temperature and rainfall were among the environmental factors that affected the counts of cholera cases most frequently. This finding is similar to those obtained by Van der Berg *et al.* (2008) and Fernández *et al.* (2009) in Mozambique and Zambia respectively, where lagged values of rainfall and temperature were found to be drivers of cholera counts, that is increases in rainfall and temperature in a given week resulted in increases in the number of cholera infections a few weeks later. Masahiro *et al.* (2008), however, suggested that river level is actually the causal link between the rainfall and cholera relationship in Bangladesh. Lobitz *et al.* (2004) and Gil *et al.* (2004) both found a correlation between sea surface temperature and cholera counts in the Bay of Bengal and coastal areas of Peru respectively, while Constantin de Magny *et al.* (2008) built a predictive model using sea surface temperature, together with chlorophyll concentration and rainfall. Models developed by Pascual *et al.* (2000) also indicated the significance of ENSO in cholera dynamics in Bangladesh, together with seasonality and previous disease levels. In contrast to these last four papers, Koelle and Pascual (2004) found no correlations of cholera counts with sea surface temperature, the Southern Oscillation Index or ENSO years and instead suggested that cycles of cholera counts can be associated with temporary immunity, seasonality and noise.

Of the aforementioned prediction model studies by Huq *et al.* (2005), Van der Berg *et al.* (2008), Constantin de Magny *et al.* (2008) and Fernández *et al.* (2009) which rely solely on environmental factors and do not involve measures of susceptible population or immunity levels, none have successfully reported an actual implementation of their early warning models. Specifically, the model developed by Constantin de Magny *et al.* (2008), despite being developed for prediction, would not actually serve the purpose of an early warning system since the equation for predicting the number of cholera cases included the explanatory variables at lag zero. Although such attempts have been made to predict outbreaks from climatic factors, Fernández *et al.* (2009, p142), citing Pascual *et al.* (2002), actually states: "climate factors are not enough to understand the size and timing of cholera outbreaks. To improve our insight into cholera epidemics, immunity levels of the population in the region should be taken into account."

2.4 Summary of literature review

The literature review in Section 2.2 has provided a brief overview of some of the time series models that have been developed for count data. The models that were chosen from either the observation-driven or parameter-driven classes, for the purpose of analysing the cholera count data, have also been highlighted and certain reasons have been given regarding the choice. The following chapter, Chapter 3, describes the four selected models in detail, namely the ACP, DACP, Poisson-gamma and SAM models.

From the literature review of models previously applied to cholera counts, it is clear that time series models for count data have not been utilised in such studies and more specifically, the time series properties of the counts are largely ignored. This highlights the need to further study some of the models described in Section 2.2 and the importance of using such models in the present application of cholera count data. The literature review has also highlighted some of the typical relationships between environmental drivers and incidences of cholera that have been found through the application of various mathematical and statistical techniques. These findings are of interest for the case study in Chapter 4 where the relationship between cholera counts in Beira and selected climatic variables is explored through the application of the four selected models.

Chapter 3

Count data time series models

3.1 Introduction to selected models

The previous chapter provided an overview of various time series models for count data, including a brief description of the four models that were chosen for the purpose of this study. This chapter focuses on the detailed theory of these four selected models. In the same manner as the model overview, this chapter is divided into observation-driven and parameter-driven models, with the former, more simple, models being described first. The section on observation-driven models contains details on the ACP and DACP models while the parameter-driven model section describes the Poisson-gamma and SAM models. Since both of the parameter-driven models that have been chosen are described in terms of a state space or structural model formulation, a brief description of these general model structures is provided at the start of the relevant section. For all of the models in this chapter, the following features are addressed: the basic formulation or description of the model, the estimation of the parameters by maximum likelihood, model diagnostics and the forecasting of future values.

3.2 Observation driven models

3.2.1 Autoregressive Conditional Poisson (ACP) model

Description of the ACP model

One of the characteristics of a Poisson distribution is equi-dispersion, where the mean is equal to the variance. However, most time series involving count data are over-dispersed with the variance greater than the mean (Jung *et al.*, 2006). These data often also exhibit serial correlation. By taking the counts

to be from a Poisson distribution and modelling the mean as an autoregressive process, where the mean is conditional on previous observations and previous means, over-dispersion and serial correlation can be accommodated by the model. This is the basis for the Autoregressive Conditional Poisson model (ACP) introduced by Heinen (2003). This ACP model falls in the category of observation-driven models since the conditional mean depends on past observations.

Taking a time series of counts, $y_1, \ldots, y_T$, let Y_{t-1} denote the information on the time series of counts up to time $t-1$. Then for the ACP model with no explanatory variables, the counts, conditional on past observations, are modelled using a Poisson distribution as follows:

$$y_t|Y_{t-1} \sim Poisson(\mu_t), \tag{3.1}$$

with an autoregressive conditional mean given as

$$E[y_t|Y_{t-1}] = \mu_t = \omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \mu_{t-j}, \tag{3.2}$$

and $\omega > 0$ and $\alpha_j, \beta_j \geq 0$ are unknown parameters. This formulation in (3.2) is referred to as the ACP(p, q) model by Heinen (2003) where p describes the number of lags on the observed variable that are incorporated into the model and q indicates the lags of previous means.

The inclusion of the lagged terms y_{t-j} in equation (3.2) is the reason why the ACP model is referred to as an observation-driven model. The non-negative values of all α, β and ω ensure that the Poisson mean μ_t remains positive.

In the present study, only the commonly used ACP(1,1) model will be considered and not the more general ACP(p,q) model. Setting $p = q = 1$ implies that the past observations and past means are only taken up to lag 1 and therefore the mean equation is simplified to

$$\mu_t = \omega + \alpha y_{t-1} + \beta \mu_{t-1}. \tag{3.3}$$

Heinen (2003) shows that, provided $\alpha + \beta < 1$, the ACP(1,1) is stationary and its unconditional mean and variance are given by

$$E[y_t] = \mu = \frac{\omega}{1 - (\alpha + \beta)} \tag{3.4}$$

and

$$Var[y_t] = \sigma^2 = \frac{\mu(1 - (\alpha + \beta)^2 + \alpha^2)}{1 - (\alpha + \beta)^2} \tag{3.5}$$

respectively. Since $\alpha + \beta$ is taken to be less than 1, it can be deduced from equation (3.5) that, provided $\alpha \neq 0$, the variance is always greater than the

mean. Hence it can be seen that, despite the fact that the ACP model uses an equi-dispersed conditional distribution (3.1), unconditionally the model is over-dispersed.

The autocorrelation function for the ACP(1,1) can be expressed as

$$Corr[y_t, y_{t-s}] = (\alpha+\beta)^{s-1}\frac{\alpha\big(1-\beta(\alpha+\beta)\big)}{1-(\alpha+\beta)^2+\alpha^2} \quad , s = 1, 2, 3, \ldots \tag{3.6}$$

and the derivation is available in Heinen (2003). This autocorrelation is positive for all s. Hence the ACP model only allows for positive serial correlation but, since most time series for count data exhibit positive rather than negative serial correlation, this drawback is considered to be minor.

Maximum likelihood

For simplicity, the ACP(1,1) model will from here on be referred to as the ACP model since higher values for p and q are not being considered in the present study. The parameters of the ACP model can be estimated fairly easily using maximum likelihood estimation and this is in fact one of the strong advantages of the model pointed out by Heinen (2003) and Jung *et al.* (2006). From expression (3.3) it is evident that the parameters that need to be estimated are $\theta = (\omega, \alpha, \beta)'$.

Considering the time series of observations, $y_1, \ldots, y_T$, the likelihood function for a given θ, denoted $L(\theta)$, is constructed by taking the joint pdf of these observations as a product of conditionals on Y_{t-1}, as given by

$$L(\theta) = p(y_1, \ldots, y_T; \theta) = \prod_{t=1}^{T} p(y_t|Y_{t-1}).$$

Since the distribution for $y_t|Y_{t-1}$ is Poisson, as shown in (3.1), the log-likelihood which is used to estimate θ can be expressed as

$$\ln L(\theta) = \sum_{t=1}^{T}\Big\{y_t \ln(\mu_t) - \mu_t - \ln(y_t!)\Big\}, \tag{3.7}$$

where μ_t is written in terms of y_{t-1} and μ_{t-1} as in equation (3.3). In the actual implementation of the model estimation, for a given series of y_t's, the process has to be "kick-started" with initial values for μ_0 and y_0. This can be done by setting y_0 and μ_0 equal to the mean of all the observations, as is done in the applications by Jung *et al.* (2006).

The log-likelihood defined in (3.7) for a given θ can be incorporated into an optimisation routine to find the estimate for θ that maximises this function. This maximum likelihood estimate (MLE) for θ can then be used to compute the MLE for the conditional means μ_t using the mean equation given in (3.3).

Introducing explanatory variables

Thus far the methodology has only described the basic ACP model and this must now be extended to accommodate the effect of explanatory variables. Again only the ACP(1,1) model is considered here. McCullagh and Nelder (1983) show that explanatory variables can be introduced multiplicatively into a static Poisson model using an exponential function, i.e. $\mu_t = \exp(x_t'\delta)$, where x_t is a $k \times 1$ vector of explanatory variables at time t and δ is a $k \times 1$ vector of parameters associated with these variables. An exponential function therefore ensures that μ_t remains positive (McCullagh and Nelder, 1983). For the ACP model, where μ_t is dynamic, Heinen (2003) indicates that μ_t can be combined multiplicatively with the explanatory variables in the original model using an exponential function. Thus

$$\mu_t^* = \mu_t \exp(x_t'\delta), \tag{3.8}$$

where μ_t^* denotes the conditional mean of the Poisson distribution including explanatory variables and μ_t is defined as in the basic model given in expression (3.3).

With the introduction of explanatory variables, the parameters that need to be estimated via maximum likelihood can be assembled as $\phi = (\omega, \alpha, \beta, \delta')'$. Given the observed series and the set of explanatory variables, the log-likelihood for estimating ϕ, denoted $\ln L(\phi)$, can be expressed as

$$\ln L(\phi) = \sum_{t=1}^{T} \Big\{ y_t\big(\ln(\mu_t) + x_t'\delta\big) - \mu_t \exp(x_t'\delta) - \ln(y_t!) \Big\}. \tag{3.9}$$

As in the basic approach, y_0 and μ_0 can be set equal to the mean value of the observations, y_t for $t = 1, \ldots, T$, in order to initialise the process.

Inference

In fitting the ACP model, it is necessary to determine whether the parameters included in the model are significant. In order to do this, approximate standard errors of the maximum likelihood estimates of the parameters can be calculated using the observed Fisher information. These approximate standard errors therefore correspond to the square root of the diagonal elements of the inverse of the Hessian matrix. The Hessian matrix itself can be calculated by taking the second derivative, with respect to μ_t, of the log-likelihood function at each time t, denoted l_t, and then summing over $t = 1, \ldots, T$. Heinen (2003) provides the basic form of the Hessian matrix for the simple case of the ACP model with no explanatory variables and $\theta = (\omega, \alpha, \beta)'$, expressed as

$$\frac{\partial^2 l_t(\theta)}{\partial \theta^2} = -\frac{y_t}{{\mu_t}^2} \frac{\partial \mu_t}{\partial \theta} \Big(\frac{\partial \mu_t}{\partial \theta} \Big)', \tag{3.10}$$

where $l_t(\theta) = y_t \ln(\mu_t) - \mu_t - \ln(y_t!)$ for the simple ACP model and where

$$\frac{\partial \mu_t}{\partial \theta} = z'_t + \beta \frac{\partial \mu_{t-1}}{\partial \theta} \tag{3.11}$$

with

$$z'_t = [1, y_{t-1}, \mu_{t-1}]. \tag{3.12}$$

If explanatory variables are introduced then the Hessian matrix can be obtained numerically.

Diagnostics

The likelihood ratio (LR) test can be used to test for significant autocorrelation in the data using the estimated log-likelihood from the ACP model. Considering that the amount of autocorrelation in the ACP is captured by the parameters α and β in equation (3.3), then the test for autocorrelation involves testing the joint null hypothesis, $H_0 : \alpha = \beta = 0$ against the alternative hypothesis $H_A : (\alpha \neq 0)$ or $(\beta \neq 0)$ or $(\alpha \neq 0$ and $\beta \neq 0)$ (Heinen, 2003). For this test, $\hat{\theta_u}$ is taken as the MLE for the unrestricted likelihood where the log-likelihood, $\ln L(\hat{\theta_u})$, is defined as in equation (3.7) or (3.9) and $\hat{\theta_r}$ is taken as the MLE for the restricted case where the restriction under the null hypothesis is $\alpha = \beta = 0$. The LR test statistic, T_{LR}, can then be computed as

$$T_{LR} = -2\big[\ln L(\hat{\theta_r}) - \ln L(\hat{\theta_u})\big].$$

This test statistic approximately follows a χ^2 distribution with 2 df under the null hypothesis, $H_0 : \alpha = \beta = 0$, and the null hypothesis will therefore be rejected if the test statistic exceeds the χ^2 value for 2 df at a selected significance level. Since most time series data are serially correlated, a high probability of rejection for this null hypothesis is expected in most applications. Cameron and Trivedi (1998) provide details on this LR test and on other tests for likelihood-based models.

In order to test the "goodness of fit" of a model it is necessary to perform diagnostic checking on the models using the standardized or Pearson residuals (Harvey and Fernandes, 1989a). These residuals are found using the conditional mean, derived from equations (3.3) and (3.8), and the conditional variance, which for the ACP model is the same as the conditional mean, as follows:

$$z_t = \frac{y_t - E(y_t|Y_{t-1})}{\sqrt{Var(y_t|Y_{t-1})}}. \tag{3.13}$$

If the model adequately fits the data then these residuals should approximately follow a normal distribution with a mean of zero and a variance of 1. This property can be formally tested but can also be checked using various

graphical techniques, such as a plot of residuals over time or a plot of residuals against the estimated values for μ_t (Harvey and Fernandes, 1989a). It is also necessary to check the residuals for remaining autocorrelations. In order to do this, the autocorrelation functions for the residuals can be plotted and the Ljung-Box statistic for the residuals can be computed, as done by both Heinen (2003) and Jung *et al.* (2006).

Provided the diagnostic checks of the residuals are satisfactory, the "best" model can then be selected based on how well the candidate models fit the data. In order to evaluate the performance of various ACP models that have been fitted to the data using different explanatory variables, a comparison of the log-likelihoods can be made. In time series applications, however, the use of criteria which adjust the log-likelihood for the number of parameters in the model, such as Akaike's Information Criterion (AIC) and Bayesian Information criterion (BIC), are considered to be more appropriate in selecting the best model. When comparing models of different types, as done in Van der Berg *et al.* (2008), the Root Mean Square Error (RMSE), Mean Square Error (MSE) or Mean Absolute Error (MAE) can be of more value (see Armstrong (2001) for definitions of these criteria). These criteria can be particularly useful when a hold out sample is used to test the forecasts from different models (Makridakis *et al.*, 1998). Heinen (2003) uses the RMSE, in combination with other criteria, to compare the ACP model against other models in one of his applications.

Forecasting from the ACP model

Having estimated the parameters via maximum likelihood, values for the dependent series can easily be forecasted using the predicted mean at each time step calculated from equations (3.3) and (3.8). This does, however, require future values for explanatory variables to be sourced or estimated.

3.2.2 Double Autoregressive Conditional Poisson (DACP) model

Description of the DACP model

The Double Autoregressive Conditional Poisson (DACP) model, developed by Heinen (2003), is a generalisation of the ACP framework. It replaces the Poisson distribution with the double Poisson distribution introduced by Efron (1986). As a result, the DACP is an extension of the ACP model which does not restrict the relationship between the conditional variance and mean to that of equality but rather allows the conditional variance to be larger or smaller than the conditional mean. Hence the DACP model

accommodates both under-dispersion or over-dispersion in its conditional distribution. The reasoning, given by Heinen (2003), for this extension is to separate out the over-dispersion in the data that is not caused by serial correlation.

The density for the double Poisson (Efron, 1986) can be taken as the multiplicative combination of two Poisson densities, with an additional parameter γ. The first Poisson density is for the observation y with mean μ and the other Poisson density is for the observation y with mean equal to y, where $y = 0, 1, 2, \ldots$. Writing a Poisson density for observation y with mean μ as $P(y, \mu) = \dfrac{e^{-\mu}\mu^y}{y!}$, the approximate density of the double Poisson can be expressed as

$$\begin{aligned} f(y|\mu,\gamma) &= \gamma^{\frac{1}{2}} P(y,\mu)^{\gamma} P(y,y)^{1-\gamma} \\ &= \gamma^{\frac{1}{2}} \left(\frac{e^{-\mu}\mu^y}{y!}\right)^{\gamma} \left(\frac{e^{-y}y^y}{y!}\right)^{1-\gamma} \\ &= (\gamma^{\frac{1}{2}} e^{-\gamma\mu}) \left(\frac{e^{-y}y^y}{y!}\right) \left(\frac{e\mu}{y}\right)^{\gamma y} \end{aligned} \tag{3.14}$$

for $\mu > 0$ and $\gamma > 0$. This approximate density of the double Poisson can be abbreviated as $DP(\mu, \gamma)$.

The DACP model, therefore, takes the framework of the ACP model and replaces the distributional assumption given in (3.1) with the Double Poisson (DP) distribution as follows

$$y_t | Y_{t-1} \sim DP(\mu_t, \gamma).$$

Efron (1986) shows that the double Poisson distribution has a mean μ and a variance that closely approximates $\dfrac{\mu}{\gamma}$. Using this approximation, Heinen (2003) provides two variations for his double Poisson model. The simpler variation, referred to as the DACP1 model by Heinen (2003), takes γ to be a parameter greater than zero and therefore has an approximate conditional variance as a multiple of the mean, expressed as

$$V[y_t | Y_{t-1}] = \sigma^2 = \frac{\mu_t}{\gamma}. \tag{3.15}$$

Setting $\gamma = 1$ is equivalent to the ACP model. The other variation to the DACP model, referred to by Heinen as the DACP2, takes the variance to be a quadratic function of the mean by setting γ equal to $\dfrac{1}{1 + \delta\mu_t}$. In the current study, however, only the DACP1 model will be considered and for simplicity it will from here on be referred to as the DACP model.

The conditional mean, μ_t, of the DACP(p,q) model is defined as before for the ACP(p,q) model with

$$E[y_t|Y_{t-1}] = \mu_t = \omega + \sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{j=1}^{q} \beta_j \mu_{t-j}$$

and $\omega > 0$ and $\alpha_j, \beta_j \geq 0$ are unknown parameters. Once again the parameters p and q define the number of lags of previous observations and number of lags of previous means respectively that are incorporated into the model. Simplifying the model to $p = q = 1$ gives

$$\mu_t = \omega + \alpha y_{t-1} + \beta \mu_{t-1} \tag{3.16}$$

for the DACP(1,1) model.

The unconditional mean of the DACP(1,1) model is the same as that for the ACP(1,1) model, written as

$$E[y_t] = \mu = \frac{\omega}{1 - (\alpha + \beta)}$$

and the unconditional variance is given by

$$V[y_t] = \sigma^2 = \frac{1}{\gamma} \frac{\mu(1 - (\alpha + \beta)^2 + \alpha^2)}{1 - (\alpha + \beta)^2},$$

which exceeds the unconditional mean μ when $\gamma \leq 1$. Hence, as mentioned previously, the advantage of this model is that it can accommodate both under-dispersion and over-dispersion depending on the value of γ. The unconditional autocorrelation for the DACP(1,1) model is the same as that given in equation (3.6) for the ACP(1,1) model. Full details regarding the derivation of the unconditional variance are given in Heinen (2003).

Maximum likelihood estimation

Efron (1986) observes that $f(y|\mu,\gamma)$ in expression (3.14) should be regarded as an approximate density since the probabilities do not sum to 1. However, he shows that for the exact double density, expressed as

$$\tilde{f}(y|\mu,\gamma) = c(\mu,\gamma) f(y|\mu,\gamma),$$

the normalizing constant, $c(\mu,\gamma)$, can be approximated by

$$\frac{1}{c(\mu,\gamma)} = 1 + \frac{1-\gamma}{12\mu\gamma}\Big(1 + \frac{1}{\mu\gamma}\Big),$$

and is very close to 1 for all values of μ and γ. Taking the sum of the approximate probabilities, $\sum_{y=0}^{\infty} f(y|\mu,\gamma)$, therefore produces a value that is nearly 1. Consequently, Heinen (2003), citing Efron (1986), uses an approximate likelihood derived from the approximate density function and which excludes the multiplicative constant $c(\mu,\gamma)$. This approximate likelihood is then used to estimate the parameters of the model.

The DACP model has parameters $\theta = (\omega,\alpha,\beta)'$, as for the ACP model, and an additional parameter γ. The approximate likelihood function, denoted $L(\theta,\gamma)$, for estimating θ and γ is therefore constructed from the joint pdf of the observations, conditional on Y_{t-1}, and is given as

$$L(\theta,\gamma) = p(y_1,\ldots,y_T;\theta,\gamma) = \prod_{t=1}^{T} p(y_t|Y_{t-1}),$$

where the distribution for $y_t|Y_{t-1}$ is now the approximate double Poisson distribution with the pdf given in (3.14). The approximate log-likelihood for the DACP model, $\ln L(\theta,\gamma)$, can hence be expressed as

$$\begin{aligned}\ln L(\theta,\gamma) &= \sum_{t=1}^{T}\left\{\frac{1}{2}\ln(\gamma) - \gamma\mu_t + y_t(\ln(y_t)-1)\right. \\ &\qquad \left. - \ln(y_t!) + \gamma y_t\left(1+\ln\left(\frac{\mu_t}{y_t}\right)\right)\right\},\end{aligned}$$

where μ_t is defined in equation (3.16).

The maximisation of the approximate log-likelihood needs to be done numerically using an optimisation routine. In order to find the parameter estimates which maximise this approximate log-likelihood function, a choice of initial values for the parameters has to be made. As in the ACP model, the process can be initialised by setting both y_0 and μ_0 equal to the mean of the observed series and starting values for α,β,ω and γ can be experimented with, under the constraints: $\omega,\gamma > 0$, $\alpha,\beta \geq 0$ and $\alpha+\beta < 1$.

Introducing explanatory variables

Similarly to the ACP model, explanatory variables x_t, for $t = 1,\ldots,T$, can be introduced into the DACP model via an exponential function, giving

$$\mu_t^* = \mu_t \exp(x_t'\delta), \tag{3.17}$$

where δ is the corresponding parameter vector for the explanatory variables. The notation μ_t^* defines the mean of the double Poisson including

explanatory variables and μ_t is defined as in equation (3.16). The parameters that need to be estimated via maximum likelihood are both the vector $\phi = (\omega, \alpha, \beta, \delta')'$, as estimated in the ACP model, and the additional γ parameter. The approximate log-likelihood including explanatory variables, which is used to estimate ϕ and γ, is therefore extended to

$$\begin{aligned}\ln L(\phi, \gamma) &= \sum_{t=1}^{T} \left\{ \frac{1}{2}\ln(\gamma) - \gamma\mu_t \exp(x_t'\delta) + y_t(\ln(y_t) - 1) \right. \\ &\qquad \left. - \ln(y_t!) + \gamma y_t \left(1 + \ln\left(\frac{\mu_t}{y_t}\right) + x_t'\delta \right) \right\}.\end{aligned}$$

Inference

As with the ACP model, the standard errors of the maximum likelihood estimates of the parameters can be approximated using the inverse of the observed Fisher information matrix, where the observed Fisher information matrix is equal to minus the Hessian matrix. Considering the log-likelihood of the simple DACP model with no explanatory variables and defining $l_t(\theta, \gamma) = \frac{1}{2}\ln(\gamma) - \gamma\mu_t + y_t(\ln(y_t) - 1) - \ln(y_t!) + \gamma y_t \left(1 + \ln\left(\frac{\mu_t}{y_t}\right) \right)$ to be the component of the log-likelihood obtained from the tth observation where $\theta = (\omega, \alpha, \beta)'$, then the Hessian matrix is given by Heinen (2003) as

$$\begin{aligned}\frac{\partial^2 l_t(\theta,\gamma)}{\partial\theta\partial\theta'} &= -\frac{\gamma y_t}{{\mu_t}^2}\frac{\partial\mu_t}{\partial\theta}\left(\frac{\partial\mu_t}{\partial\theta}\right)' \\ \frac{\partial^2 l_t(\theta,\gamma)}{\partial\theta\partial\gamma} &= \frac{y_t - \mu_t}{\mu_t}\frac{\partial\mu_t}{\partial\theta} \\ \frac{\partial^2 l_t(\theta,\gamma)}{\partial\gamma^2} &= -\frac{1}{2}\frac{1}{\gamma^2}\end{aligned}$$

where $\dfrac{\partial\mu_t}{\partial\theta}$ is the same as in (3.11) for the ACP model. The standard errors of the maximum likelihood estimates can therefore be obtained by using the diagonal from the inverse of this Hessian matrix, after taking the square root of these elements. If explanatory variables are introduced into the DACP model then the Hessian matrix can be obtained numerically.

Diagnostics

In the same manner as the ACP model, the LR test can be carried out to check for significant autocorrelation in the data. Details of this test have been described in the diagnostics section for the ACP model. In addition to this, a new LR test can be used to test for excess over-dispersion in the data by testing the null hypothesis $H_0 : \gamma = 1$ against the alternative $H_A : \gamma \neq 1$.

This LR is computed using the difference between the log-likelihoods of the ACP and DACP models since the ACP model is equivalent to a DACP model with γ set equal to 1. One possible problem, highlighted by Heinen (2003), regarding the LR test for the DACP model is that it is based on approximate likelihoods. The test therefore uses an approximate likelihood ratio and assumes that the approximation error is close to zero.

Checks for goodness of fit can also be carried out by analysing the raw residuals and the Pearson's residuals. Pearson residuals, defined in equation (3.13), are found using the conditional mean and variance (equations (3.16) and (3.15)). The conditional mean and variance for the DACP model including explanatory variables can be obtained by combining equation (3.17) with equations (3.16) and (3.15) respectively. As in the case of the ACP model, these residuals need to be analysed to ensure that there are no autocorrelations or other non-random patterns remaining.

In order to select the "best model" when comparing several DACP models fitted with different combinations of explanatory variables, various criteria based on the likelihood, in particular the AIC and BIC, can be used. In comparing the DACP model against the ACP model or other types of models, general fit statistics may be preferred due to the fact that the DACP uses an approximate likelihood. These fit statistics may include the RMSE, MSE or MAE,

Forecasting from the DACP model

Forecasting values from the DACP model works in the same manner as with the ACP model. The parameter estimates for $\phi = (\omega, \alpha, \beta, \delta')'$ can be obtained via the maximum likelihood procedure and these estimates can be used in equations (3.16) and (3.17) to obtain the predicted mean at each time step. However, forecasts for the explanatory variables are again required in order to generate forecasts for the dependent time series.

3.3 Parameter driven models

3.3.1 State space model and structural model formulation

Before describing the parameter-driven time series models for count data it is appropriate to first look at the general state space model formulation within the context of the normal distribution. State space models, as described in Shumway and Stoffer (2006), are a very general class of time series models whereby the formulation, at each time t, consists of two equations: the observation or measurement equation and the transition or state equation.

At any given time t, for $t = 1, \ldots, T$, the observation equation is expressed as

$$y_t = H_t x_t + \varepsilon_t, \tag{3.18}$$

where y_t is a $q \times 1$ vector of observed variables, x_t is a $p \times 1$ vector of state variables that are usually unobserved, H_t is the $q \times p$ observation matrix and ε_t describes the observation error. It is assumed that ε_t is a $q \times 1$ normal vector with a zero mean and a $q \times q$ covariance matrix R.

At time t, for $t = 1, \ldots, T$, the state vector is determined from the transition equation as

$$x_t = F_t x_{t-1} + \eta_t, \tag{3.19}$$

where F_t is a $p \times p$ transition matrix and η_t is a $p \times 1$ vector that consists of model error components that are assumed to be independently and identically normally distributed with a zero mean and a $p \times p$ covariance matrix Q. As can be seen from this transition equation in (3.19), the state vector x_t can be computed using the previous state x_{t-1}, for all time points $t = 1, \ldots, T$. Equations (3.18) and (3.19) describe the state space formulation for normally distributed data and two assumptions of these state space models, as provided by Shumway and Stoffer (2006), are that ε_t and η_t are uncorrelated and that the process is initialised with the vector x_0 which is normally distributed with mean μ_0 and covariance matrix P_0. The big advantage of these models is that Kalman filter techniques can be used in the estimation process, starting at time $t = 1$ with the initial values and then moving through predict and update steps until time T. This process therefore allows new observations to be incorporated easily.

State space models accommodate the study of many different models in the same mathematical framework (Makridakis *et al.*, 1998) since many time series models and regression models can be expressed in a "state space form". Some examples of these are given in Makridakis *et al.* (1998) and Shumway and Stoffer (2006).

Structural models, developed by Harvey (1989), are a special class of state space models. For the normal distribution, the components of a structural model are linear processes representing trends, cycles and autoregressions and the observed series is the sum of these components. The simplest structural model is the random walk plus noise model, also referred to as the level component plus noise by Harvey and Fernandes (1989a). In this simple structural model, the observation and transition equations can be written as

$$y_t = \mu_t + \varepsilon_t \tag{3.20}$$

and

$$\mu_t = \mu_{t-1} + \eta_t. \tag{3.21}$$

Here μ_t is the trend or level component that can fluctuate based on the white noise disturbance term η_t, which has zero mean, and ε_t is the observation error term, at time t. Note that this follows the same formulation as a univariate state space model, in which H_t and F_t are set equal to one in equations (3.18) and (3.19) respectively.

More complicated structural models could include both the trend component and a seasonal component and may also contain some autoregressive components, incorporated as lagged values of the observed series. Janacek and Swift (1993) and Shumway and Stoffer (2006) describe how to model the seasonality in structural models.

3.3.2 Poisson-gamma model

Overview of the model

One of the parameter-driven time series models for count data selected for this dissertation is the Poisson-gamma model described by Harvey and Fernandes (1989a). This model uses a structural framework similar to that given in equations (3.20) and (3.21). The observation equation models the time series of observations, y_t for $t = 1, \ldots, T$, as a Poisson distribution while the transition equation for the mean, μ_t, is formulated in such a manner that it allows for ease of estimation and prediction. In summary:

- The Poisson-gamma model makes use of Kalman filter techniques to iteratively estimate the parameters with ease, where:
  - the 'predict' step relies on the resulting properties obtained from combining a gamma and beta distribution; and
  - the 'update' step is formulated from the conjugacy of the Poisson and gamma distributions, as used in the Bayesian setting.

- The predictive distribution is negative binomial and therefore the calculation of the likelihood is straightforward.
- The final forecasts take the form of an exponentially weighted moving average (EWMA) and are therefore easy to compute.

The Poisson-gamma model of Harvey and Fernandes (1989a) uses an approach based on that of Smith and Miller (1986) through the use of a multiplicative transition equation. It is a model that has been widely cited and also applied to various types of count data. In the original paper by Harvey and Fernandes (1989a) the model was applied to data involving the number of goals scored in football, purse snatching incidents and the effect of seat belt legislation on accident fatalities. In a later paper in the same year, Harvey and Fernandes (1989b) describe the use of the model in predicting insurance claims. Lambert (1996a) and Lambert (1996b) extended the Poisson-gamma model to handle unequally spaced observations and applied this model to the analyses of drug dose effects on calves' respiratory rates, measured as counts per minute, and population counts of micro-organisms respectively. The model has also been applied to political science data, looking at the effect of global economic and political activities on the number of armed conflicts, by Brandt *et al.* (2000), who used the Poisson-gamma model of Harvey and Fernandes (1989a) with a modified transition equation based on Shephard (1994). Brandt *et al.* (2000) refer to the Poisson-gamma model with Shephard's modification as the Poisson Exponentially Weighted Moving Average (PEWMA) model.

Formulation of the structural model

For the Poisson-gamma model, the observed count y_t taken at time t is assumed to come from a Poisson distribution with unobserved mean μ_t. Therefore, following the state space or structural model framework, our observation distribution is

$$y_t|\mu_t \sim Poisson(\mu_t).$$

The transition equation is given by

$$\mu_t = \frac{\mu_{t-1}\eta_t}{\omega}, \tag{3.22}$$

where $0 < \omega < 1$, $\eta_t \sim Beta(\omega a_{t-1}, (1-\omega)a_{t-1})$ and
$\mu_{t-1}|Y_{t-1} \sim Gamma(a_{t-1}, b_{t-1})$, with Y_{t-1} denoting the information of the observed series up to time $t-1$. The process is initialised with
$\mu_0|Y_0 \sim Gamma(a_0, b_0)$.

The Kalman filter

In practical terms, a method is required to estimate the underlying mean of the process, μ_t, for a given series of observations $y_1, \ldots, y_T$, and to use this estimate to calculate forecasts. An adaption of the Kalman filter given by Harvey and Fernandes (1989a) provides a procedure for predicting and updating filter parameters that can be used in evaluating the likelihood for a given ω. This in turn can be built into an optimisation routine to obtain the MLE for μ_t. The Kalman filter operates using a recursive algorithm to find, for a given ω, the optimal estimates of the model parameters of the transition equation at each time period using all of the available information at that time (Shumway and Stoffer, 2006). The next three subheadings describe the steps of the Kalman filter procedure.

Initialise: The process is initialised with $\mu_0|Y_0 \sim Gamma(a_0, b_0)$. Following the suggestion of Harvey and Fernandes (1989a), initialisation of parameters during implementation of the Kalman filter can be taken as $a_0 = b_0 = 0$ at time $t = 0$ even though it results in $\mu_0 = 0$. This is because the first estimate of a proper distribution for μ_t can simply be taken at time $t = \tau$, where μ_τ is the first non-zero observation, thus discarding all preceding zeros observations.

After initialising the Kalman filter process, the filter moves through each subsequent time period t by first predicting $\mu_t|Y_{t-1}$ from the prior distribution and then using the observation y_t to update the estimate of μ_t using the posterior distribution for $\mu_t|Y_t$. The prior distribution for the 'predict' step of the algorithm and the posterior distribution for the 'update' step therefore need to be introduced.

Predict: In this step the prior distribution, $\mu_t \mid Y_{t-1}$, is considered. The transition equation in (3.22) gives $\mu_t = \dfrac{\mu_{t-1}\eta_t}{\omega}$ but since $\mu_{t-1}|Y_{t-1} \sim Gamma(a_{t-1}, b_{t-1})$ and $\eta_t \sim Beta(\omega a_{t-1}, (1-\omega)a_{t-1})$, the properties of the gamma distribution (given in Appendix A.1: Result 3) can be used to show that the mean μ_t conditional on past observations up to time $t-1$ is from a gamma distribution and can be written as

$$\mu_t|Y_{t-1} \sim Gamma(a_{t|t-1}, b_{t|t-1}),$$

where the parameters $a_{t|t-1}$ and $b_{t|t-1}$ are defined as

$$a_{t|t-1} = \omega a_{t-1}$$
$$b_{t|t-1} = \omega b_{t-1}.$$

Update: In the updating step of the Kalman filter algorithm, the posterior distribution of μ_t needs to be computed once the observation y_t becomes

available. This posterior distribution can be derived using an application of Bayes' theorem (see Section A.2 of the Appendix) and is based on the conjugacy of the Poisson and gamma distributions. Thus the posterior distribution of μ_t given Y_t is a gamma distribution expressed as

$$\mu_t|Y_t \sim Gamma(a_t, b_t),$$

where

$$\begin{aligned} a_t &= a_{t|t-1} + y_t \\ b_t &= b_{t|t-1} + 1. \end{aligned}$$

The Kalman filter process for the Poisson-gamma model can therefore be summarised as follows:

Box 3.1: The Kalman filter process

For a given ω, $0 < \omega < 1$,

1. Initialise: $\mu_0|Y_0 \sim Gamma(a_0, b_0)$

2. Predict: $\mu_t|Y_{t-1} \sim Gamma(\underbrace{\omega a_{t-1}}_{a_{t|t-1}}, \underbrace{\omega b_{t-1}}_{b_{t|t-1}})$

3. Update: $\mu_t|Y_t \sim Gamma(\underbrace{\omega a_{t-1} + y_t}_{a_t}, \underbrace{\omega b_{t-1} + 1}_{b_t})$

Steps 2 and 3 are performed iteratively from $t = \tau, \ldots, T$ and the output of the Kalman filter can be generated as a $T - \tau + 1 \times 2$ matrix containing the values of $a_{t|t-1}$ and $b_{t|t-1}$ for all values of t, for a given ω. This matrix is then used in the calculation of the log-likelihood function which is discussed in the next subsection.

Note that in the Poisson-gamma model, the parameters a_{t-1} and b_{t-1}, in the 'predict' stage of the process, are multiplied by a factor ω which is less than 1, and therefore the conditional mean and variance can be expressed as

$$E[\mu_t|Y_{t-1}] = \frac{a_{t|t-1}}{b_{t|t-1}} = \frac{a_{t-1}}{b_{t-1}} = E[\mu_{t-1} \mid Y_{t-1}]$$

and

$$Var[\mu_t|Y_{t-1}] = \frac{a_{t|t-1}}{b^2{}_{t|t-1}} = \omega^{-1} Var[\mu_{t-1} \mid Y_{t-1}] > Var[\mu_{t-1} \mid Y_{t-1}]$$

respectively. These results are therefore the same as those of the Gaussian state space model where the conditional mean remains the same but the conditional variance increases as t increases (Brandt *et al.*, 2000).

Maximum likelihood estimation

In order to estimate the parameter ω in the Poisson-gamma model, the maximum likelihood approach is used. The advantage of using the natural conjugate of the Poisson in the model formulation is that it gives the predictive distribution as a negative binomial and hence the likelihood can be expressed in an explicit form.

The likelihood, Ł(ω), for the parameter ω, given the observations $y_{\tau+1}, \ldots, y_T$, can be constructed as the joint pdf of these observations, conditional on the information up to time τ, Y_τ. This is expressed as

$$L(\omega) = p(y_{\tau+1}, \ldots, y_T; \omega) = \prod_{t=\tau+1}^{T} p(y_t|Y_{t-1}), \tag{3.23}$$

where $t = \tau$ is the index of the first non-zero observation when initialising $a_0 = b_0 = 0$, as discussed above.

Now the predictive distribution of y_t, conditional on Y_{t-1}, can be evaluated as

$$p(y_t|Y_{t-1}) = \int_0^\infty p(y_t|\mu_t)p(\mu_t|Y_{t-1})d\mu_t.$$

It follows from the fact that $y_t|\mu_t \sim Poisson(\mu_t)$ and $\mu_t|Y_{t-1} \sim Gamma(a, b)$ that this predictive distribution can be written as

$$y_t|Y_{t-1} \sim NegativeBinomial\left(a_{t|t-1}, \frac{b_{t|t-1}}{b_{t|t-1}+1}\right). \tag{3.24}$$

A derivation of this result is given in Section A.2 of the Appendix. By inserting the pdf of this negative binomial distribution (see equation (A.6) in Appendix A.2) into the equation for the likelihood, i.e. combining (3.23) and (3.24), and taking the logarithms, the log-likelihood for ω can be expressed as

$$\begin{aligned} \ln L(\omega) \quad = \quad & \sum_{t=\tau+1}^{T} \Big\{ \ln\Gamma(a_{t|t-1}+y_t) - \ln y_t! - ln\Gamma(a_{t|t-1}) \\ & + a_{t|t-1}\ln b_{t|t-1} - (a_{t|t-1}+y_t)\ln(b_{t|t-1}+1) \Big\}, \end{aligned}$$

where Γ represents the gamma function. For a given ω, the Kalman filter can now be used to calculate $\ln L(\omega)$. By using an optimising routine to

maximise this log likelihood function, the MLE for ω can be obtained and then used in predictions for future values of y_t, that is for $y_{T+1}, \ldots, y_{T+l}$, $l \geq 1$.

Forecasting

In order to compute a one-step ahead forecast, denoted $\tilde{y}_{T+1|T}$, for a given ω and given all past observations Y_T, the expression for $E[y_{T+1}|Y_T]$ is required (Harvey and Fernandes, 1989a). Since $y_{T+1}|Y_T$ follows a negative binomial distribution, as given in (3.24), the properties of that distribution can be used to compute the forecast as

$$\tilde{y}_{T+1|T} = E[y_{T+1}|Y_T] = \frac{a_{T+1|T}}{b_{T+1|T}} = \frac{a_T}{b_T}, \tag{3.25}$$

where a_T and b_T are outputs from the Kalman filter for a given ω. The variance can also be evaluated as

$$\begin{aligned} Var[y_{T+1}|Y_T] &= \frac{a_{T+1|T}(1+b_{T+1|T})}{b^2_{T+1|T}} \\ &= \frac{a_T}{\omega b_T^2} + \frac{a_T}{b_T} \\ &= \frac{1}{\omega} Var[\mu_T|Y_T] + E[\mu_T|Y_T]. \end{aligned} \tag{3.26}$$

Using repeated substitutions (Harvey and Fernandes, 1989a) gives

$$\begin{aligned} a_T &= a_{T|T-1} + y_T \\ &= \omega a_{T-1} + y_T \\ &= \omega(\omega a_{T-2} + y_{T-1}) + y_T \\ &\vdots \\ &= \omega^T a_0 + \sum_{j=0}^{T-1} \omega^j y_{T-j} \end{aligned}$$

and

$$\begin{aligned} b_T &= b_{T|T-1} + 1 \\ &= \omega b_{T-1} + 1 \\ &= \omega(\omega b_{T-2} + 1) + 1 \\ &\vdots \\ &= \omega^T b_0 + \sum_{j=0}^{T-1} \omega^j. \end{aligned}$$

Since initialisation involved taking $a_0 = b_0 = 0$, it follows that the one-step ahead forecast at time T can be expressed as

$$\tilde{y}_{T+1|T} = \frac{a_T}{b_T} = \frac{\sum_{j=0}^{T-1} \omega^j y_{T-j}}{\sum_{j=0}^{T-1} \omega^j}, \tag{3.27}$$

which is an exponentially weighted moving average (EWMA) where the weights of past observations decline exponentially for observations further into the past. It can be seen from (3.27) that small values for ω imply that more recent observations have a larger effect on the forecasted values thus indicating a series with high serial correlation. However, when $\omega = 1$, the forecasted value becomes the same as its previous value, indicating a constant mean and therefore the model reverts back to the static Poisson model.

Although Harvey and Fernandes (1989a) indicate that the distribution of y_{T+l} for lead time $l > 1$ at time T is numerically difficult to evaluate, they show that the multi-step ahead prediction for y_{T+l} at time T, denoted $\tilde{y}_{T+l|T}$, is equal to the one-step ahead forecast and is given as

$$\tilde{y}_{T+l|T} = E(y_{T+l}|Y_T) = \frac{a_T}{b_T}$$

for all lead times l, $l \geq 1$.

Shephard's transition equation

Having described the Poisson-gamma model of Harvey and Fernandes (1989a), an adjustment to the transition equation made by Shephard (1994) is now considered and the reason for this adjustment is examined. The original transition equation of Harvey and Fernandes (1989a) is given as $\mu_t = \frac{\mu_{t-1}\eta_t}{\omega}$. Taking logarithms gives

$$\ln\left(\frac{\mu_t}{\mu_{t-1}}\right) = \ln \eta_t - \ln \omega$$

and taking expectations results in

$$E\Big[\ln\Big(\frac{\mu_t}{\mu_{t-1}}\Big)\Big] = E\Big[\ln \eta_t\Big] - \ln \omega. \tag{3.28}$$

However, since the properties of the beta distribution implies that $E(\eta_t) = \omega$, and Jensen's inequality indicates that $E[\ln \eta_t] \leq \ln E[\eta_t]$, it follows that $E[\ln \eta_t] \leq \ln \omega$. Using this result in (3.28) gives

$$E\Big[\ln\Big(\frac{\mu_t}{\mu_{t-1}}\Big)\Big] \leq 0,$$

which consequently implies that, on average,

$$\mu_t \leq \mu_{t-1},$$

thus indicating a negative trend or growth rate, i.e. μ_t converges to zero as $t \to \infty$ (Nelson, 1990; Shephard, 1994; Brockwell and Davis, 1996; Grunwald *et al.*, 1997).

In order to avoid this problem, Shephard (1994) introduced a transition equation with an expected growth rate of zero. In this transition equation ω^{-1} is replaced with e^{r_t}, thus giving

$$\mu_t = e^{r_t}\mu_{t-1}\eta_t. \tag{3.29}$$

Since $E\Big[\ln\big(\frac{\mu_t}{\mu_{t-1}}\big)\Big] = r_t + E[\ln \eta_t]$, r_t is set equal to $-E[\ln(\eta_t)]$ to get a zero growth rate. Using the properties of the beta distribution, where $E(\eta_t) = \omega$ for all t, Shephard (1994) and Brandt *et al.* (2000) indicate that r_t can be evaluated as

$$r_t = -E[\ln(\eta_t)] = \psi(a_{t-1}) - \psi(\omega a_{t-1}),$$

where ψ denotes the digamma function. Note, however, that although invoking the transition equation in (3.29) avoids the issue of μ_t converging to zero, μ_t still remains nonstationary and hence y_t is also nonstationary (Shephard, 1994).

Introducing explanatory variables

Unlike the static Poisson model, the level component μ_t in the Poisson-gamma model is dynamic. Consequently, when introducing explanatory variables, μ_t can be regarded as independent of these variables (Harvey and Fernandes, 1989a). Therefore, as with the ACP model, the level component μ_t can be combined multiplicatively with the explanatory variables using an exponential function. Thus the distribution of y_t, conditional on μ_t, can be written as

$$y_t|\mu_t \sim Poisson(\underbrace{\mu_t \exp(x_t'\delta)}_{\mu_t^*}), \tag{3.30}$$

where x_t, $t = 1, \ldots, T$, is a sequence of k-dimensional vectors of explanatory variables and δ is the corresponding k-dimensional vector of parameters. Observe that μ_t^* denotes the mean of the Poisson distribution with explanatory variables and hence can be expressed as

$$\mu_t^* = \mu_t \exp(x_t'\delta). \tag{3.31}$$

The transition equation for the separate level component μ_t remains the same as in the basic process, defined in equation (3.22).

The Kalman Filter with explanatory variables

As described for the basic Poisson-gamma model, the Kalman filter requires initialisation at time $t = 0$, setting $a_0 = b_0 = 0$, after which it iterates through each subsequent time period, predicting and updating the filter parameters. This is done for a given set of ω and δ values. The 'predict' and 'update' steps, for the inclusion of explanatory variables, are described as follows:

Predict: For this step the prior distribution of $\mu_t^*|Y_{t-1}$ is required. Considering that $\mu_t^* = \mu_t \exp(x_t'\delta)$ and $\mu_t|Y_{t-1} \sim Gamma(\omega a_{t-1}, \omega b_{t-1})$, it follows from the properties of the gamma distribution (see Appendix A.1: Result 2) that

$$\mu_t^*|Y_{t-1} \sim Gamma(a_{t|t-1}^*, b_{t|t-1}^*),$$

where

$$\begin{aligned} a_{t|t-1}^* &= \omega a_{t-1} \\ b_{t|t-1}^* &= \frac{\omega b_{t-1}}{\exp(x_t'\delta)}. \end{aligned}$$

Update: In order to establish the updating equations for a_t and b_t, as required in the previous step, the distribution of $\mu_t|Y_t$ needs to be determined. To do this, the posterior distribution for the new mean, $\mu_t^*|Y_t$, must first be derived and the exponential function must then be used to find the posterior distribution of $\mu_t|Y_t$ for the model with explanatory variables.

The derivation of $\mu_t^*|Y_t$ follows the same steps as in the basic model (see Section A.2 in the Appendix), using the conjugacy of the Poisson and Gamma distributions to give

$$\mu_t^*|Y_t \sim Gamma(a_{t|t-1}^* + y_t, b_{t|t-1}^* + 1).$$

Substituting for $a_{t|t-1}^*$ and $b_{t|t-1}^*$ and using both $\mu_t = \dfrac{\mu_t^*}{\exp(x_t'\delta)}$ and the properties of the gamma distribution (see Appendix A.1: Result 2), it follows that

$$\mu_t|Y_t \sim Gamma(a_t, b_t),$$

where

$$\begin{aligned} a_t &= \omega a_{t-1} + y_t \\ b_t &= \omega b_{t-1} + \exp(x_t'\delta). \end{aligned}$$

Therefore the updating equations in the Kalman recursions are almost the same as for the basic model with no explanatory variables but with the

multiplier $\exp(x_t'\delta)$ introduced into the scale parameter of the gamma distribution.

The Kalman filter process for the Poisson-gamma model with explanatory variables can therefore be summarised as follows:

Box 3.2: The Kalman filter process including explanatory variables

For a given ω and δ, $0 < \omega < 1$,

1. Initialise: $\mu_0|Y_0 \sim Gamma(a_0, b_0)$, where $a_0 = b_0 = 0$

2. Predict: $\mu_t^*|Y_{t-1} \sim Gamma(\underbrace{\omega a_{t-1}}_{a^*_{t|t-1}}, \underbrace{\omega b_{t-1}\exp(-x_t'\delta)}_{b^*_{t|t-1}})$

3. Update: $\mu_t|Y_t \sim Gamma(\underbrace{\omega a_{t-1} + y_t}_{a_t}, \underbrace{\omega b_{t-1} + \exp(x_t'\delta)}_{b_t})$

As described previously, the Kalman filter iterates through steps 2 and 3 from $t = \tau, \ldots, T$. The values for $a^*_{t|t-1}$ and $b^*_{t|t-1}$, which are required to evaluate the log-likelihood function, are produced as output from the filter for all values of t, for given values of the parameters ω and δ.

Maximum likelihood estimation with explanatory variables

The log-likelihood is the same as for the case with no explanatory variables with $a_{t|t-1}$ and $b_{t|t-1}$ replaced by $a^*_{t|t-1}$ and $b^*_{t|t-1}$ respectively, and can be written as

$$\begin{aligned}\ln L(\omega) \quad = \quad & \sum_{t=\tau+1}^{T} \Big\{ \ln\Gamma(a^*_{t|t-1} + y_t) - \ln y_t! - \ln\Gamma(a^*_{t|t-1}) \\ & + a^*_{t|t-1}\ln(b^*_{t|t-1}) - (a^*_{t|t-1} + y_t)\ln(b^*_{t|t-1} + 1)\Big\}.\end{aligned}$$

The output from the Kalman filter can again be used to evaluate $\ln L(\omega)$ for a given ω and δ. By maximising this log-likelihood function with respect to the unknown parameters ω and δ using a non-linear optimisation routine, the MLEs for these parameters can be obtained and used in the calculation of predictions. The standard errors of the estimates can be computed using the inverse of the Hessian matrix in a similar manner to that used in the observation-driven models.

Diagnostics

Analysis of residuals are again crucial for identifying model fit and for checking whether any autocorrelations still exist. The Pearson residuals can be found using

$$z_t = \frac{y_t - E(y_t|Y_{t-1})}{\sqrt{Var(y_t|Y_{t-1})}},$$

where the conditional mean and variance of y_t are defined in equations (3.25) and (3.26) respectively for the basic model. For models including explanatory variables, the conditional mean and variance can be evaluated as

$$E[y_t|Y_{t-1}] = \frac{a^*_{t|t-1}}{b^*_{t|t-1}} = \frac{a_{t-1}\exp(x'_t\delta)}{b_{t-1}}$$

and

$$\begin{aligned} Var[y_t|Y_{t-1}] &= \frac{a^*_{t|t-1}(1+b^*_{t|t-1})}{(b^*_{t|t-1})^2} \\ &= \frac{a_{t-1}\exp(x'_t\delta)^2\big(1+\omega b_{t-1}\exp(-x'_t\delta)\big)}{\omega b^2_{t-1}} \\ &= \frac{a_{t-1}\exp(x'_t\delta)^2}{\omega b^2_{t-1}} + \frac{a_{t-1}\exp(x'_t\delta)}{b_{t-1}} \end{aligned}$$

respectively, where a_{t-1} and b_{t-1} are outputs from the Kalman filter for a given ω and δ.

Harvey and Fernandes (1989a) mentioned the fact that these residuals should follow a Normal(0,1) distribution if the model adequately fits the data and that plots of residuals against time and against estimates of the mean are also useful in determining "goodness of fit". In addition, there should be no significant autocorrelation left in the residuals. Autocorrelation function (ACF) plots can be produced for these residuals in order to determine whether the model accounts for the autocorrelation in the time series and the Ljung-Box statistic can also be invoked as an additional check for any remaining autocorrelation.

There are various diagnostic checks that can be used in selecting the best model. The same fit statistics as used in the ACP and DACP models apply here for the Poisson-gamma model. Thus the criteria AIC and BIC can be used when comparing several Poisson-gamma models, each containing a different selection of variables. The AIC and BIC can also be used to find the "best fit" amongst different types of likelihood-based models but alternative measures for such comparisons include RMSE, MSE and MAE.

Forecasting from the Poisson-gamma model with explanatory variables

Harvey and Fernandes (1989a) show that, using the same derivation as for the Poisson-gamma without explanatory variables, the l-step ahead prediction at time T for the model including explanatory variables can be written as

$$\tilde{y}_{T+l|T} = \frac{\exp(x'_{T+l}\delta)\sum_{j=0}^{T-1}\omega^j y_{T-j}}{\sum_{j=0}^{T-1}\omega^j \exp(x'_{T-j}\delta)}. \tag{3.32}$$

Introducing

$$\text{EWMA}[y] = \frac{\sum_{j=0}^{T-1}\omega^j y_{T-j}}{\sum_{j=0}^{T-1}\omega^j} \quad \text{and} \quad \text{EWMA}[exp(x'\delta)] = \frac{\sum_{j=0}^{T-1}\omega^j exp(x'_{T-j}\delta)}{\sum_{j=0}^{T-1}\omega^j},$$

the equation for the l-step ahead prediction at time T in (3.32), for given estimates of ω and δ, can be expressed as

$$\tilde{y}_{T+l|T} = \frac{\exp(x'_{T+l}\delta)\text{EWMA}[y]}{\text{EWMA}[\exp(x'\delta)]}.$$

3.3.3 Stochastic Autoregressive Mean (SAM) model

Description of the SAM model

Another model for time series of count data considered in the class of parameter-driven models is the Poisson model with a stochastic autoregressive mean, termed the SAM model (Jung *et al.*, 2006). This formulation stems from the model by Zeger (1988), who introduced a separate latent variable into the mean of the Poisson regression model in order to accommodate both over-dispersion and autocorrelation of the time series of counts. The SAM model, described by Jung *et al.* (2006), is closely related to the Poisson-gamma model in terms of its specification as it can also be described in terms of a state space formulation. With the SAM model, however, the conditional mean function contains a latent autoregressive process which

changes independently of the observed counts due to the inclusion of a separate dynamic error term.

In describing the model, the series of observed counts, y_t, the time series of k explanatory variables, x_t, and the latent non-negative stochastic process, u_t, where $t = 1, \ldots, T$, are considered. Denoting the mean as μ_t, the conditional distribution of $y_t|\mu_t$ is then assumed to follow a Poisson distribution expressed as

$$y_t|\mu_t \sim Poisson(\mu_t)$$

with mean

$$\mu_t = \exp(x_t'\delta)u_t, \tag{3.33}$$

where δ is a k-dimensional vector of regression coefficients. Following the assumption used by several authors, in particular Jung *et al.* (2006), $\lambda_t = \ln(u_t)$ can be taken to be a Gaussian first-order autoregressive process such that

$$\ln(u_t) = \lambda_t = \gamma\lambda_{t-1} + \nu\epsilon_t \tag{3.34}$$

where $\epsilon_t \sim NID(0, 1)$. Hence the parameters that need to be estimated are δ, γ and ν and these can be summarised as $\theta = (\delta', \gamma, \nu)'$. The mean of the Poisson regression can now be expressed as $\mu_t = \exp(x_t'\delta + \lambda_t)$, or equivalently, $\ln(\mu_t) = x_t'\delta + \lambda_t$.

To achieve stationarity in the Poisson process, the condition $E(u_t) = 1$ needs to be met thus requiring $E(\exp(\lambda_t)) = 1$ which implies $E(\lambda_t) = 0$. Also note that in order to ensure stationarity of the process described by λ_t, $|\gamma| < 1$ is required.

Issues regarding the estimation of the SAM model

Although the formulation of the SAM model is straightforward, the implementation is extremely challenging. The implementation firstly involves approximating the likelihood for fixed values of the parameters, θ, using Monte Carlo methods, and secondly, estimating these parameters as MLE's by maximising the approximated likelihood using an optimisation routine. The need for an approximate likelihood is explained in this subsection.

Considering that the conditional density of y_t given λ_t, denoted $g_t(y_t|\lambda_t, \theta)$, is Poisson($\mu_t$) with $\mu_t = \exp(x_t'\delta + \lambda_t)$ and the conditional density of λ_t given λ_{t-1}, denoted $p_t(\lambda_t|\lambda_{t-1}, \theta)$, is $N(\gamma\lambda_{t-1}, \nu^2)$, it follows that

$$g_t(y_t|\lambda_t, \theta) = \frac{e^{-\mu_t}\mu_t^{y_t}}{y_t!} \tag{3.35}$$

and

$$p_t(\lambda_t|\lambda_{t-1},\theta) = \frac{1}{\nu\sqrt{2\pi}}\exp\left\{-\frac{1}{2\nu^2}(\lambda_t-\gamma\lambda_{t-1})^2\right\}. \quad (3.36)$$

The joint conditional density of y_t and λ_t given all past information up to time $t-1$ and given θ, denoted $f_t(y_t,\lambda_t|\lambda_{t-1},y_1,\ldots,y_{t-1},\theta)$, is therefore the product of the conditional densities and is written as

$$f_t(y_t,\lambda_t|\lambda_{t-1},y_1,\ldots,y_{t-1},\theta) = g_t(y_t|\lambda_t,\theta)\times p_t(\lambda_t|\lambda_{t-1},\theta). \quad (3.37)$$

Hence the joint density of all the observations, $y_1,\ldots,y_T$, and of the latent variables $\lambda_1,\ldots,\lambda_T$, is the product of the conditional densities specified in (3.37) for $t=1,\ldots,T$. The term at $t=1$ in this joint density requires a value for λ_0. For convenience Jung *et al.* (2006) simply take $\lambda_0 = E(\lambda_t) = 0$.

In order to evaluate the likelihood, $L(\theta)$, of the parameters θ for the data $y_1,\ldots,y_T$, it is necessary to integrate the product in (3.37) over λ_t as

$$L(\theta) = \int\cdots\int\prod_{t=1}^{T} f_t(y_t,\lambda_t|\lambda_{t-1},y_1,\ldots,y_{t-1},\theta)d\lambda_1,\ldots,d\lambda_T. \quad (3.38)$$

However, since the likelihood function expressed in equation (3.38) is a high-dimensional integral, it cannot be computed directly and approximations are therefore required. Zeger (1988), who originally proposed the model, uses a quasi-likelihood approach while Durbin and Koopman (1997, 2000) use a partial importance sampling approach. Specifically, Durbin and Koopman (1997, 2000) only estimate part of the likelihood with importance sampling and then use Kalman filter techniques to complete the estimation. Jung *et al.* (2006) estimate the likelihood using efficient importance sampling (EIS). EIS, developed by Richard and Zhang (2006), is a Monte Carlo (MC) integration technique used for such instances where high-dimensional integrals need to be evaluated. Jung *et al.* (2006) use the EIS implementation in two approaches, namely; the maximum likelihood (ML) approach and the Bayesian approach whereby EIS is incorporated into MCMC analysis. For the purpose of this dissertation, the maximum likelihood efficient importance sampling (ML-EIS) technique is selected for the estimation of the SAM model.

Approximating the likelihood in the ML-EIS procedure

The estimation of the SAM model uses Monte Carlo integration in order to address the issue of evaluating the high-dimensional integral in (3.38). Given the formulation for $f_t(y_t,\lambda_t|\lambda_{t-1},y_1,\ldots,y_{t-1},\theta)$ in (3.37), the simple simulation approach would be to generate, for a given value of θ, N independent trajectories of $\lambda_1^{(i)},\ldots,\lambda_T^{(i)}$, for $i=1,\ldots,N$, from the distribution

of $\lambda_t|(\lambda_{t-1}, \theta)$ and to use these trajectories in the Monte Carlo integration framework to approximate the likelihood as

$$\begin{aligned} L(\theta) &= \int \cdots \int \prod_{t=1}^{T} \left(\frac{f_t(y_t, \lambda_t|\lambda_{t-1}, y_1, \ldots, y_{t-1}, \theta)}{p_t(\lambda_t|\lambda_{t-1}, \theta)} \right) p_t(\lambda_t|\lambda_{t-1}, \theta) d\lambda_1, \ldots, d\lambda_T \\ &\simeq \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} \left(\frac{f_t(y_t, \lambda_t^{(i)}|\lambda_{t-1}^{(i)}, y_1, \ldots, y_{t-1}, \theta)}{p_t(\lambda_t^{(i)}|\lambda_{t-1}^{(i)}, \theta)} \right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \prod_{t=1}^{T} g_t(y_t|\lambda_t^{(i)}, \theta), \end{aligned}$$

where $g_t(y_t|\lambda_t^{(i)}, \theta)$ is defined in equation (3.35). However, Liesenfeld and Richard (2005) and Jung *et al.* (2006) point out that this method is highly inefficient since the simulated trajectories do not follow the latent process. They solve this problem by using efficient importance sampling. Instead of drawing samples from the sequence of $p_t(\lambda_t|\lambda_{t-1}, \theta)$ densities, referred to as the natural sampler by Liesenfeld and Richard (2005), they use a sequence of auxiliary importance samplers, written as $m_t(\lambda_t|\lambda_{t-1}, a_t)$, and indexed by auxiliary parameters a_t. The choice of $m_t(\lambda_t|\lambda_{t-1}, a_t)$ usually includes a parametric extension to the original natural sampler $p_t(\lambda_t|\lambda_{t-1}, \theta)$. Considering then the likelihood, N independent importance trajectories $\lambda_1^{(i)}, \ldots, \lambda_T^{(i)}$, for $i = 1, \ldots, N$, can be simulated from the sequence of sampling densities $m_t(\lambda_t|\lambda_{t-1}, a_t)$ and the approximation to the likelihood can be introduced as

$$\begin{aligned} L(\theta) &= \int \cdots \int \prod_{t=1}^{T} \left(\frac{f_t(y_t, \lambda_t|\lambda_{t-1}, y_1, \ldots, y_{t-1}, \theta)}{m_t(\lambda_t|\lambda_{t-1}, a_t)} \right) m_t(\lambda_t|\lambda_{t-1}, a_t) d\lambda_1, \ldots, d\lambda_T \\ &\simeq \frac{1}{N} \sum_{i=1}^{N} \left[\prod_{t=1}^{T} \frac{f_t(y_t, \lambda_t^{(i)}|\lambda_{t-1}^{(i)}, y_1, \ldots, y_{t-1}, \theta)}{m_t(\lambda_t^{(i)}|\lambda_{t-1}^{(i)}, a_t)} \right]. \end{aligned} \tag{3.39}$$

(3.40)

In order to estimate the likelihood for a given θ, the following three steps are therefore necessary:

Box 3.3: Estimating the likelihood for the SAM model

Step 1. Construct an importance sampler $m_t(\lambda_t|\lambda_{t-1}, a_t)$ so that the ratio

$$\left(\frac{f_t(y_t, \lambda_t^{(i)}|\lambda_{t-1}^{(i)}, y_1, \ldots, y_{t-1}, \theta)}{m_t(\lambda_t^{(i)}|\lambda_{t-1}^{(i)}, a_t)} \right)$$

remains approximately constant over all the trajectories for each time step. This is done in order to minimize the error in the Monte Carlo estimation of $L(\theta)$.

Step 2. Find the auxiliary parameters, a_t for each t, which minimise the Monte Carlo error. This step utilises a sequential process working backwards from $t = T$ to $t = 1$ and this process is repeated through several sets of trajectories until the values of the improved trajectories converge.

Step 3. Use the results from step 2 to estimate the likelihood, $L(\theta)$, or the log-likelihood.

The likelihood, $L(\theta)$, is now approximated as the objective function in an optimisation procedure. The three steps specified in Box 3.3 are explained under subheadings as follows:

Step 1: Constructing an importance sampler

This step describes the construction of an importance sampler for a time point t. Note that the importance sampler has the same form but different auxiliary parameters for each t, $t = 1, \ldots, T$. Liesenfeld and Richard (2005) indicate that, in the construction of the importance sampler $m_t(\lambda_t|\lambda_{t-1}, a_t)$, the use of density kernels instead of densities provides a better approximation to $f_t(y_t, \lambda_t|\lambda_{t-1}, y_1, \ldots, y_{t-1}, \theta)$. Therefore, taking $k_t(\lambda_t, \lambda_{t-1}, a_t)$ as the density kernel of $m_t(\lambda_t|\lambda_{t-1}, a_t)$, the importance sampler can be expressed as

$$m_t(\lambda_t|\lambda_{t-1}, a_t) = \frac{k_t(\lambda_t, \lambda_{t-1}, a_t)}{\chi_t(\lambda_{t-1}, a_t)},$$

where the integrating constant $\chi_t(\lambda_{t-1}, a_t)$ is given by $\int k_t(\lambda_t, \lambda_{t-1}, a_t) d\lambda_t$.

In the implementation of the EIS algorithm for the SAM model, Jung *et al.* (2006) take the density kernel $k_t(\lambda_t, \lambda_{t-1}, a_t)$ to be an extension of the original density $p_t(\lambda_t|\lambda_{t-1}, \theta)$ given in (3.36). By doing this they simplify the estimation procedure since it results in a normal distribution for the impor-

tance sampler. Their density kernel is expressed as

$$k_t(\lambda_t, \lambda_{t-1}, a_t) \propto p_t(\lambda_t|\lambda_{t-1}, \theta) \times \exp\Big\{ -\frac{1}{2}\big[\alpha_t\lambda_t^2 - 2\beta_t\lambda_t\big]\Big\},$$

with $a_t = (\alpha_t, \beta_t)'$ being the auxiliary parameters that need to be estimated. Since $\lambda_t|\lambda_{t-1}, \theta \sim N(\gamma\lambda_{t-1}, \nu^2)$, and the conditional density $m_t(\lambda_t|\lambda_{t-1}, a_t)$ is proportional to its density kernel $k_t(\lambda_t, \lambda_{t-1}, a_t)$, it follows that

$$\begin{aligned} m_t(\lambda_t|\lambda_{t-1}, a_t) &\propto p_t(\lambda_t|\lambda_{t-1}, \theta) \times \exp\Big\{ -\frac{1}{2}\big[\alpha_t\lambda_t^2 - 2\beta_t\lambda_t\big]\Big\} \\ &\propto \exp\Big\{ -\frac{1}{2\nu^2}\big[\lambda_t - \gamma\lambda_{t-1}\big]^2\Big\} \times \exp\Big\{ -\frac{1}{2}\big[\alpha_t\lambda_t^2 - 2\beta_t\lambda_t\big]\Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[\lambda_t^2\Big(\underbrace{\frac{1}{\nu^2} + \alpha_t}_{b}\Big) - 2\lambda_t\Big(\underbrace{\frac{\gamma\lambda_{t-1}}{\nu^2} + \beta_t}_{c}\Big)\Big]\Big\}. \quad (3.41) \end{aligned}$$

Using the result from the normal distribution given in Section A.3 of the Appendix, with b and c as specified in (3.41), it can be deduced that the importance sampler $m_t(\lambda_t|\lambda_{t-1}, a_t)$ follows a normal distribution with variance, σ_t^2 given by

$$\sigma_t^2 = \frac{1}{\frac{1}{\nu^2} + \alpha_t} = \frac{\nu^2}{(1 + \nu^2\alpha_t)} \quad (3.42)$$

and mean, κ_t, by

$$\kappa_t = \sigma_t^2\Big(\frac{\gamma\lambda_{t-1}}{\nu^2} + \beta_t\Big). \quad (3.43)$$

Integrating the density kernel $k_t(\lambda_t, \lambda_{t-1}, a_t)$ with respect to λ_t gives the expression for the integrating constant, $\chi_t(\lambda_{t-1}, a_t)$, as follows:

$$\begin{aligned} \chi_t(\lambda_{t-1}, a_t) &= \int k_t(\lambda_t, \lambda_{t-1}, a_t)d\lambda_t \\ &\propto \int \exp\Big\{ -\frac{1}{2}\Big[\frac{(\lambda_t - \gamma\lambda_{t-1})^2}{\nu^2} + \alpha_t\lambda_t^2 - 2\beta_t\lambda_t\Big]\Big\}d\lambda_t \\ &\propto \int \exp\Big\{ -\frac{1}{2}\Big[\lambda_t^2\underbrace{\Big(\frac{1}{\nu^2} + \alpha_t\Big)}_{1/\sigma_t^2} - 2\lambda_t\underbrace{\Big(\frac{\gamma\lambda_{t-1}}{\nu^2} + \beta_t\Big)}_{\kappa_t/\sigma_t^2} + \frac{(\gamma\lambda_{t-1})^2}{\nu^2}\Big]\Big\}d\lambda_t \\ &\propto \int \underbrace{\exp\Big\{ -\frac{1}{2}\Big[\frac{1}{\sigma_t^2}(\lambda_t - \kappa_t)^2\Big]\Big\}}_{\text{normal kernel}} \times \underbrace{\exp\Big\{ -\frac{1}{2}\Big[\frac{(\gamma\lambda_{t-1})^2}{\nu^2} - \frac{\kappa_t^2}{\sigma_t^2}\Big]\Big\}}_{\text{constant w.r.t. } \lambda_t} d\lambda_t \\ &\propto \exp\Big\{ -\frac{1}{2}\Big[\frac{(\gamma\lambda_{t-1})^2}{\nu^2} - \frac{\kappa_t^2}{\sigma_t^2}\Big]\Big\}. \quad (3.44) \end{aligned}$$

Step 2: Minimising the Monte Carlo error
Having constructed the importance sampler, $m_t(\lambda_t|\lambda_{t-1}, a_t)$, those parameters, a_t, which minimise the Monte Carlo error in the estimation of the likelihood $L(\theta)$ need to be found. This process can be summarised using the following steps:

Box 3.4: Step 2 - Minimising the MC error for the ML-EIS procedure

Step 2.1. Generate N independent trajectories from the initial natural sampler $p_t(\lambda_t|\lambda_{t-1}, \theta)$ using (3.34) and a common set of random N(0,1) variates, $\epsilon_t^{(i)}, t = 1, \ldots, T$ and $i = 1, \ldots, N$. These are regarded as inefficient samples but are used to "kick start" the process.

Step 2.2. Minimise the Monte Carlo error for the generated set of trajectories using a sequential process over t, working backwards from $t = T$ to $t = 1$. This is also referred to as the EIS regression.

Step 2.3. Use the estimated values for a_t, that is $\hat{a}_t = (\hat{\alpha}_t, \hat{\beta}_t)$, from the sequential process in Step 2.2 to calculate the means κ_t, from (3.43), and variances σ_t^2, from (3.42), of the importance sampler $m_t(\lambda_t|\lambda_{t-1}, a_t)$.

Step 2.4. Since $m_t(\lambda_t|\lambda_{t-1}, \hat{a}_t)$ is distributed as a $N(\kappa_t, \sigma_t^2)$, use this to generate a set of N improved trajectories by taking $\lambda_t^{(i)} = \kappa_t^{(i)} + \sigma_t \epsilon_t^{(i)}$, for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, where $\epsilon_t^{(i)}$ are the common set of N(0,1) variates used in step 2.1.

Step 2.5. Iterate through Steps 2.2 to 2.4 several times, stopping when the improved trajectories converge. Typically, no more than five iterations are required.

Note that a common set of random N(0,1) variates are used throughout the process whenever a set of trajectories is generated, that is in Steps 2.1 and 2.4, in order to achieve convergence in Step 2.5. The final trajectories, $\lambda_t^{(i)}$, for $t = 1, \ldots, T$ and $i = 1, \ldots, N$, and estimates of $\hat{a}_t$ obtained at the end of the iterations in Step 2.5 are then used to estimate the log-likelihood.

It is necessary to provide more detail on Step 2.2 in Box 3.4 which is referred to as sequential minimisation or EIS regression. This step involves a sequential process, for a given set of trajectories, stepping through each

time unit, where for each time t, the ratio given in Step 1 of Box 3.3, or the log of this ratio, must be as close as possible to a constant, c_t. This can be formulated as a least squares regression problem, that of finding $\hat{a}_t$ and $\hat{c}_t$, for each t, which minimise the following expression:

$$(\hat{c}_t, \hat{a}_t) = \arg\min_{c_t, a_t} \sum_{i=1}^{N} \left[\ln \left(\frac{f_t(y_t, \lambda_t^{(i)} | \lambda_{t-1}^{(i)}, y_1, \ldots, y_{t-1}, \theta)}{m_t(\lambda_t^{(i)} | \lambda_{t-1}^{(i)}, a_t)} \right) - c_t \right]^2.$$

Substituting for $f_t(y_t, \lambda_t^{(i)} | \lambda_{t-1}^{(i)}, y_1, \ldots, y_{t-1}, \theta)$ and $m_t(\lambda_t^{(i)} | \lambda_{t-1}^{(i)}, a_t)$, and collecting the terms that are not dependent on λ_t into the constant, c_t, this expression reduces to

$$\begin{aligned} (\hat{c}_t, \hat{a}_t) \quad = \quad & \arg\min_{c_t, a_t} \sum_{i=1}^{N} \Bigg[- \exp(\lambda_t^{(i)}) + y_t \lambda_t^{(i)} + \ln\left(\chi_t(\lambda_{t-1}^{(i)}, a_t)\right) \\ & - \frac{1}{2}\alpha_t \left(\lambda_t^{(i)}\right)^2 - \beta_t \lambda_t^{(i)} - c_t \Bigg]^2. \end{aligned} \tag{3.45}$$

Note that since the value of $\chi_t(\lambda_{t-1}, a_t)$ does not depend on λ_t at time t, Liesenfeld and Richard (2005) use $\chi_{t+1}(\lambda_t, a_{t+1})$ to do the minimsation at period t and for convenience introduce the additional constant $\chi_{T+1}(\lambda_T, a_{T+1})$ to be equal to 1. As a result, the start of the process, at time T, can be initiated with $\chi_{T+1}(\lambda_T, a_{T+1}) = 1$, and the process works backwards through t to $t = 1$.

The minimisation of the expression (3.45) at time t is equivalent to finding LS estimators of the parameters b_{0t}, b_{1t} and b_{2t} in the regression equation

$$z_t^{(i)} = b_{0t} + b_{1t}\lambda_t^{(i)} + b_{2t}\left(\lambda_t^{(i)}\right)^2 \tag{3.46}$$

where $i = 1, \ldots, N$ and the dependent variable is defined as

$$z_t^{(i)} = - \exp(\lambda_t^{(i)}) + y_t \lambda_t^{(i)} + \ln\left(\chi_{t+1}(\lambda_t^{(i)}, \hat{a}_{t+1})\right) \tag{3.47}$$

and

$$\begin{aligned} b_{0t} &= c_t \\ b_{1t} &= -\beta_t \\ b_{2t} &= -\frac{1}{2}\alpha_t. \end{aligned} \tag{3.48}$$

In other words, the dependent variable z_t is computed using (3.47) and the design matrix of regressors X_t has an ith row containing $\left(1, \lambda_t^{(i)}, (\lambda_t^{(i)})^2\right)$

where $i = 1, \ldots, N$ and N is the number of trajectories. Setting $\hat{b}_t = (\hat{b}_{0t}, \hat{b}_{1t}, \hat{b}_{2t})$ and evaluating

$$\hat{b_t} = (X_t' X_t)^{-1} X_t' z_t \tag{3.49}$$

provides values for the $\hat{b_t}$ parameters and subsequently produces the parameters of interest $\hat{a}_t = (\hat{\alpha}_t, \hat{\beta}_t)$ via substitution of the expressions given in (3.48).

The overall sequential minimisation or EIS regression process, which forms step 2.2 in Box 3.4, therefore operates as follows:

Box 3.5: Step 2.2 - EIS regression

Step 2.2.1. Set $t = T$. Taking the value $\chi_{T+1}(\lambda_T^{(i)}, \hat{a}_{T+1}) = 1$, find the estimates, $\hat{a}_T = (\hat{\alpha}_T, \hat{\beta}_T)$ and $\hat{c}_T$, which minimise the Monte Carlo error at time T by means of regression, using equations (3.46) - (3.49).

Step 2.2.2. Set $t = t - 1$ and then calculate $\chi_{t+1}(\lambda_t, \hat{a}_{t+1})$ using (3.42), (3.43), (3.44) and the value for $\hat{a}_{t+1}$ which was estimated from the previous time step. Then minimise the Monte Carlo error for the current time step t using the regression process.

Step 2.2.3. Repeat step 2.2.2. until $t = 1$.

Step 3: Approximating the log-likelihood

To approximate the log-likelihood, $\ln L(\theta)$, logarithms of the equation for the likelihood (3.39) are first taken to obtain

$$\begin{aligned} \ln L(\theta) &= \ln \left[\frac{1}{N} \sum_{i=1}^{N} \left\{ \prod_{t=1}^{T} \frac{f_t(y_t, \lambda_t^{(i)} | \lambda_{t-1}^{(i)}, Y_{t-1}, \theta)}{m_t(\lambda_t^{(i)} | \lambda_{t-1}^{(i)}, a_t)} \right\} \right] \\ &= \ln \left[\frac{1}{N} \sum_{i=1}^{N} \exp \left\{ \sum_{t=1}^{T} \underbrace{\ln \left(\frac{f_t(y_t, \lambda_t^{(i)} | \lambda_{t-1}^{(i)}, Y_{t-1}, \theta)}{m_t(\lambda_t^{(i)} | \lambda_{t-1}^{(i)}, a_t)} \right)}_{r_t^{(i)}} \right\} \right]. \end{aligned} \tag{3.50}$$

Now taking $r_t^{(i)} = \ln \left(\frac{f_t(y_t, \lambda_t^{(i)} | \lambda_{t-1}^{(i)}, Y_{t-1}, \theta)}{m_t(\lambda_t^{(i)} | \lambda_{t-1}^{(i)}, a_t)} \right)$ in (3.50), $r_t^{(i)}$ can be calcu-

lated as

$$\begin{aligned} r_t^{(i)} &= \ln\left[\frac{g_t(y_t|\lambda_t^{(i)},\theta).\frac{1}{\nu\sqrt{2\pi}}\exp\big[-\frac{1}{2\nu^2}(\lambda_t^{(i)}-\gamma\lambda_{t-1}^{(i)})^2\big]}{\frac{1}{\sigma_t\sqrt{2\pi}}\exp\big[-\frac{1}{2\sigma_t^2}(\lambda_t^{(i)}-\kappa_t^{(i)})^2\big]}\right] \\ &= \ln\left[\left(\frac{\exp\big(-\exp(x_t'\delta+\lambda_t^{(i)})\big)\exp\big(x_t'\delta+\lambda_t^{(i)}\big)^{y_t}}{y_t!}\right)\times \right. \\ &\qquad \left.\left(\frac{\sigma_t}{\nu}\exp\Big[-\frac{1}{2\nu^2}\big(\lambda_t^{(i)}-\gamma\lambda_{t-1}^{(i)}\big)^2+\frac{1}{2\sigma_t^2}\big(\lambda_t^{(i)}-\kappa_t^{(i)}\big)^2\Big]\right)\right] \\ &= -\exp(x_t'\delta+\lambda_t^{(i)})+y_t(x_t'\delta+\lambda_t^{(i)})-\ln(y_t!)+\ln\big(\frac{\sigma_t}{\nu}\big) \\ &\qquad -\frac{1}{2\nu^2}\big(\lambda_t^{(i)}-\gamma\lambda_{t-1}^{(i)}\big)^2+\frac{1}{2\sigma_t^2}\big(\lambda_t^{(i)}-\kappa_t^{(i)}\big)^2. \end{aligned}$$

In this manner, an approximation to the log-likelihood, $\ln L(\theta)$, can be calculated using the final trajectories, $\lambda_t^{(i)}$ for $i = 1,\ldots,N$, and using the values for $\kappa_t^{(i)}$ and σ_t^2 calculated from the final estimates for $\hat{\alpha}_t$ and $\hat{\beta}_t$, for all $t = 1,\ldots,T$.

Note that for large T, the sum of the exponential ratios $r_t^{(i)}$ in (3.50) becomes too large to be evaluated. This problem can be addressed by subtracting a constant equivalent to the mean of all the $r_t^{(i)}$ values, $t = 1,\ldots,T$ and $i = 1,\ldots,N$. The product of this mean and T is then added to the end result to get the final log-likelihood. This is computed for fixed values of $\theta = (\delta',\gamma,\nu)'$. In order to find the MLE's for θ, this whole process is nested into an optimiser so as to find the values for θ which maximise the approximate log-likelihood.

Diagnostics

As with the Poisson-gamma model, an analysis of Pearson residuals is required to assess the "goodness of fit" of the model. Pearson residuals are found using the conditional mean and variance as follows:

$$z_t = \frac{y_t - E(y_t|Y_{t-1},x_t)}{\sqrt{Var(y_t|Y_{t-1},x_t)}}.$$

For the SAM model, defining Y_{t-1} as the information on the observed series up to time $t-1$, the conditional mean and variance can be expressed as

$$E[y_t|Y_{t-1},x_t] = \exp(x_t'\beta)E[\exp(\lambda_t)|Y_{t-1},x_{t-1}] \tag{3.51}$$

and

$$\begin{aligned} Var[y_t|Y_{t-1},x_t] &= \exp(x_t'\beta)\Big(E[\exp(\lambda_t)|Y_{t-1},x_{t-1}]+ \\ &\qquad \exp(x_t'\beta)Var[\exp(\lambda_t)|Y_{t-1},x_{t-1}]\Big) \end{aligned} \tag{3.52}$$

The equations (3.51) and (3.52) require the evaluation of the conditional mean and variance of $\exp(\lambda_t)$, which, as indicated by Jung *et al.* (2006), can be done via EIS.

The Pearson residuals should follow an N(0,1) distribution and in addition, there should be no significant autocorrelation left in the residuals. The ACF plots and Ljung-Box statistic can be computed for the residuals in order to check for any remaining autocorrelation in the time series.

Similarly, to the ACP, DACP, and Poisson-gamma models discussed, various fit statistics, such as the log-likelihood, AIC, BIC, RMSE and MAE, can be used for model comparison and for selecting the best model. The log-likelihoods, AIC or BIC for the SAM model can be useful when comparing one SAM model to other SAM models fitted to different combinations of explanatory variables. The RMSE and MAE can be used for comparing the overall fit of different types of models and may be more appropriate when comparing models such as the DACP and SAM which use approximations of the likelihoods since Heinen (2003) warns that problems may potentially arise when tests are used with the approximate maximum likelihood of the DACP model.

Forecasting from the SAM model

Forecasting from the SAM model can be achieved using the parameter estimates for $\theta = (\delta', \gamma, \nu)'$ obtained by maximising the approximate log-likelihood, $\ln L(\theta)$, in (3.50). These estimates can be used in equations (3.33) and (3.34) where $\lambda_t = \ln(u_t)$ to obtain the predicted mean at each time step being forecasted. However, in order to calculate future values for the dependent time series, forecasts for the explanatory variables are required.

Chapter 4

Cholera case study

4.1 Introduction

The background to the cholera study has been discussed in the introductory chapter. This chapter therefore deals with the application of the selected observation-driven and parameter-driven count data time series models, which were discussed in Chapter 3, to the cholera data. Details of the data and a basic exploratory analysis performed on the data are described first, after which some results for the Poisson and negative binomial regression models, similar to those documented in Van der Berg *et al.* (2008), are provided. The results for the count data time series models, separated into sections for observation-driven and parameter-driven models, are then presented. Fit statistics for the various models are listed and a comparison of the fits of the models is given at the end of the chapter.

4.2 Data

The data that were available for this study were weekly data comprising the number of cholera cases in Beira, as well as certain climatic data. The cholera counts were recorded as the number of patients treated for cholera on a weekly basis. The data comprising the cholera counts were provided by the Centre for Environmental Hygiene and Medical Examinations (CHAEM) in Mozambique and based on data collected at the Cholera Treatment Centre (CTC) in Beira. The climatic data available included daily rainfall (mm), air temperature (^{o}C) and humidity (%), all of which were recorded at the Beira airport and were obtained from official sources. These data were converted to weekly values, namely, total weekly rainfall, average weekly air temperature and average weekly humidity, in order to correspond to the weekly cholera counts. Although water temperature is regarded as an important variable in modelling the incidence of cholera due to the increased breeding of the cholera bacteria in warmer water, the average weekly air tempera-

ture was considered to be a good proxy for this variable. In earlier work done at the CSIR, humidity was found to have no correlation with cholera counts and was therefore excluded from any further analyses. As a result, only air temperature and rainfall were considered as explanatory variables in this study. Seasonal terms were also introduced in order to model seasonal cycles not implicitly captured through the use of the climatic variables.

Six years of weekly cholera counts were used spanning the period from January 1999 to December 2004 and comprising a total of 313 counts. The data are the property of the CSIR and therefore for confidentiality reasons the entire dataset cannot be provided. However, an extract of this data is given in Table 4.1.

Table 4.1: Extract of the cholera dataset.

Year	**Week**	**Cholera counts**	**Rainfall (mm)**	**Air Temp. (^{o}C)**
1999	1	122	99.9	27.1
1999	2	123	219.2	26.3
1999	3	135	5.2	28.0
1999	4	144	118.7	26.6
1999	5	168	71.9	28.2
1999	6	267	177.7	26.4
1999	7	246	131.2	26.4
1999	8	381	144.6	26.0
1999	9	424	20	26.9
1999	10	267	36.2	27.2
1999	11	208	10.8	27.7
1999	12	130	23.6	27.3
1999	13	71	40.1	25.9
1999	14	85	3.6	25.4
1999	15	55	20	24.6
1999	16	16	0	27.0
1999	17	26	84.2	24.6
1999	18	6	19	23.3
1999	19	6	0.3	22.8
1999	20	0	15.9	22.6
1999	21	0	0	23.2
1999	22	0	4.1	21.9
1999	23	0	4.5	21.3
1999	24	0	0	22.5
1999	25	0	12.2	22.1
1999	26	0	16	20.6

4.3 Exploratory analysis

A plot of weekly cholera counts over time is presented in Figure 4.1. It is clear from the graph that the cholera cases have a strong seasonal component with regular outbreaks occurring almost every year, with 2001 being the exception. There does not appear to be any evidence of an increasing or decreasing trend over time and therefore it is the seasonality that is of interest.

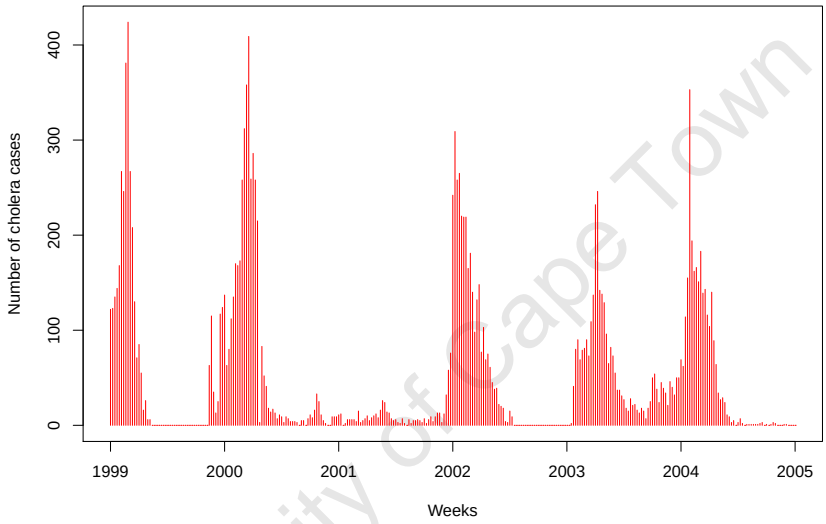


Figure 4.1: *Time series plot of weekly cholera counts in Beira: Jan 1999 - Dec 2004.*

A histogram of the cholera counts is displayed in Figure 4.2. It can be observed from the histogram that the data are highly skewed with a high frequency of small counts and only a few large counts, thus suggesting that the data follow a Poisson distribution.

Table 4.2 includes the mean, median and variance of the weekly cholera counts. The large difference between the mean and the median again confirms the skewness in the data. However, these statistics show that the cholera data also exhibit a large amount of over-dispersion, since the variance of 6487 is far greater than the mean of 50. Since the underlying condition of a Poisson distribution is equi-dispersion with the mean equal to the variance, the statistics in Table 4.2 suggest that a Poisson-based model which accommodates over-dispersion or a negative binomial distribution may be more appropriate. The choice of distribution is important when selecting

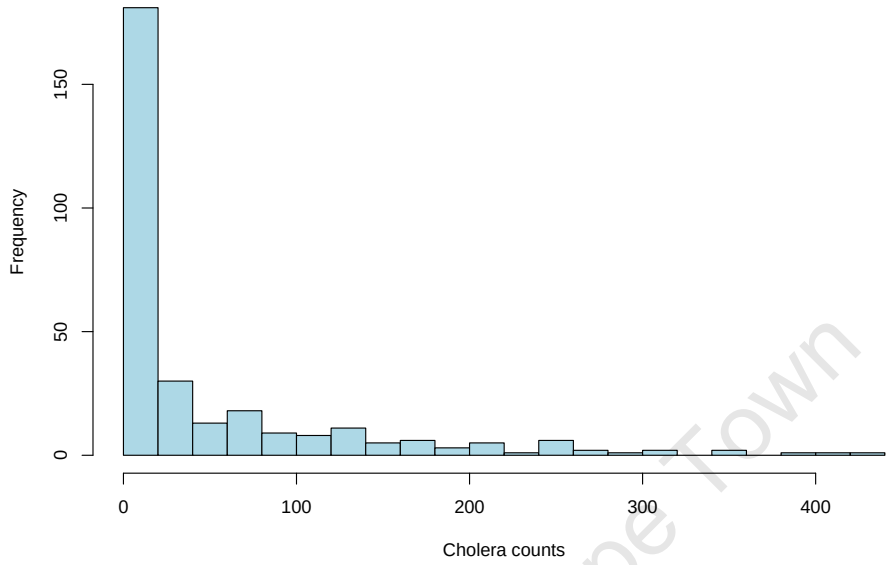


Figure 4.2: *Histogram of cholera cases.*

which models should be fitted to the data. Examples exist of both Poisson regression and negative binomial regression being fitted to cholera case data (Huq *et al.*, 2005; Masahiro *et al.*, 2008; Constantin de Magny *et al.*, 2008; Van der Berg *et al.*, 2008; Emch *et al.*, 2008; Fernández *et al.*, 2009).

Table 4.2: Mean, median and variance of the cholera counts.

Mean	**Median**	**Variance**
50.0	11	6487.5

As with most time series data, another property of the cholera count data is that of autocorrelation. This autocorrelation is evident from Figure 4.3 but appears to be masked by the seasonality. The autocorrelation function (ACF) plot in this figure displays a slow decay in autocorrelations with a definite seasonal cycle - first positive autocorrelations and then negative autocorrelations.

Allowing for the fact that the seasonality may be masking the autocorrelation in the underlying series, the data were seasonally differenced and the ACF plot of the seasonally differenced data is given in Figure 4.4. This figure indicates that even on the "deseasonalised" data, the underlying autocorrelation is very strong. A measure for assessing the models fitted to

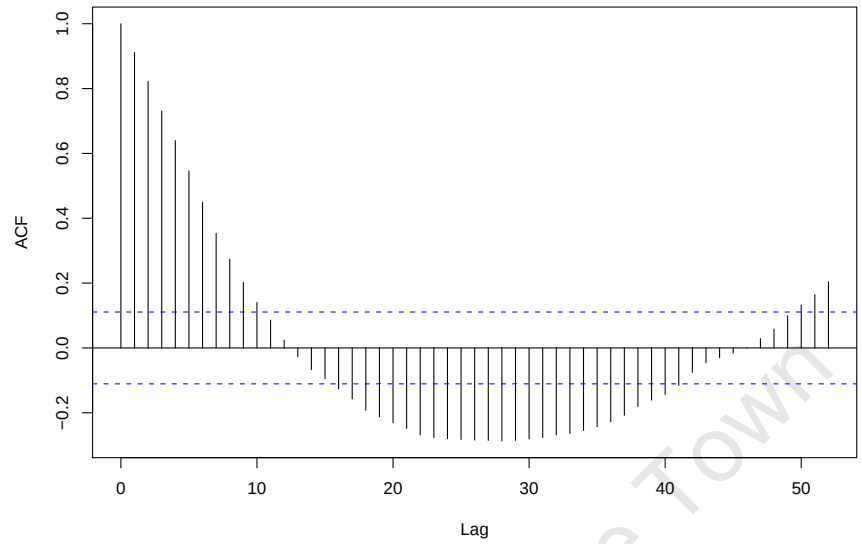


Figure 4.3: *Autocorrelation function plot for numbers of cholera cases.*

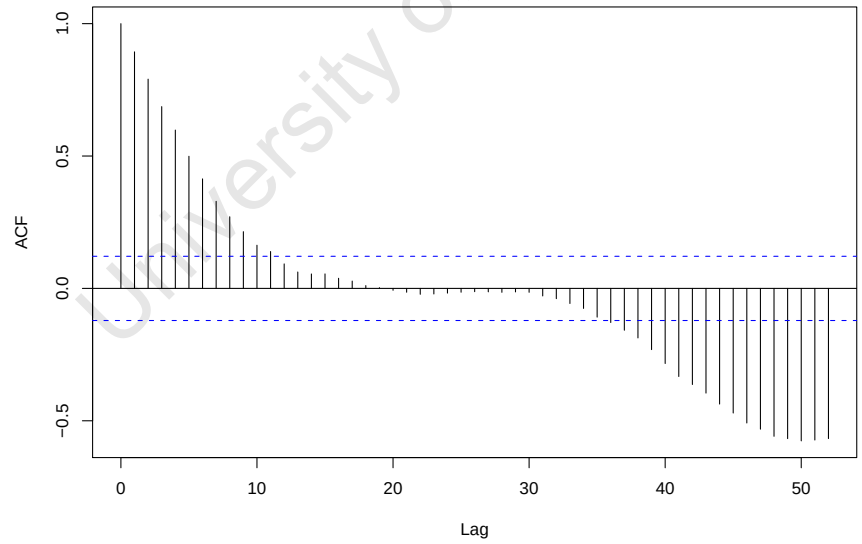


Figure 4.4: *Autocorrelation function plot for numbers of cholera cases with seasonality removed.*

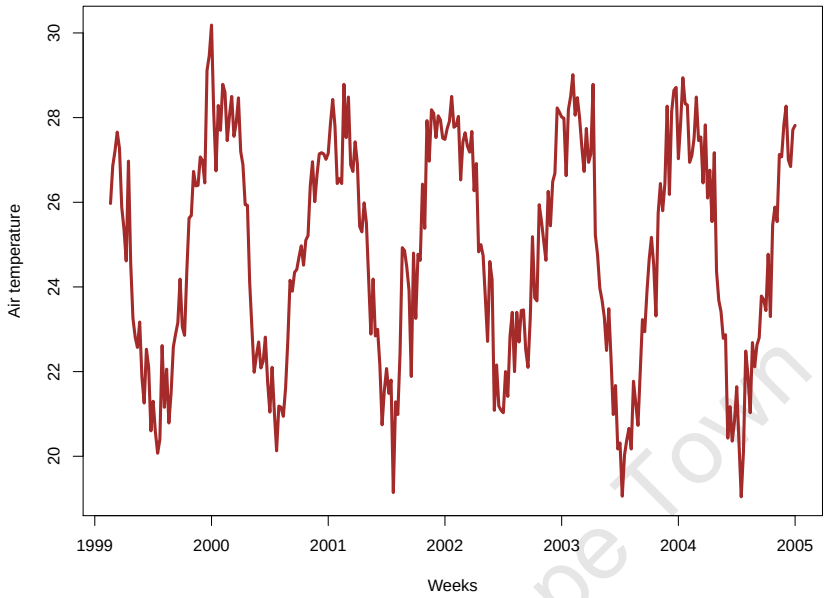
these data is thus required in order to check that the residuals have no significant autocorrelation after a particular model has been fitted.

As mentioned previously, the explanatory variables considered in this study are those of average weekly air temperature and weekly rainfall. These climatic variables are now examined and time series plots of the individual series are shown in Figure 4.5.

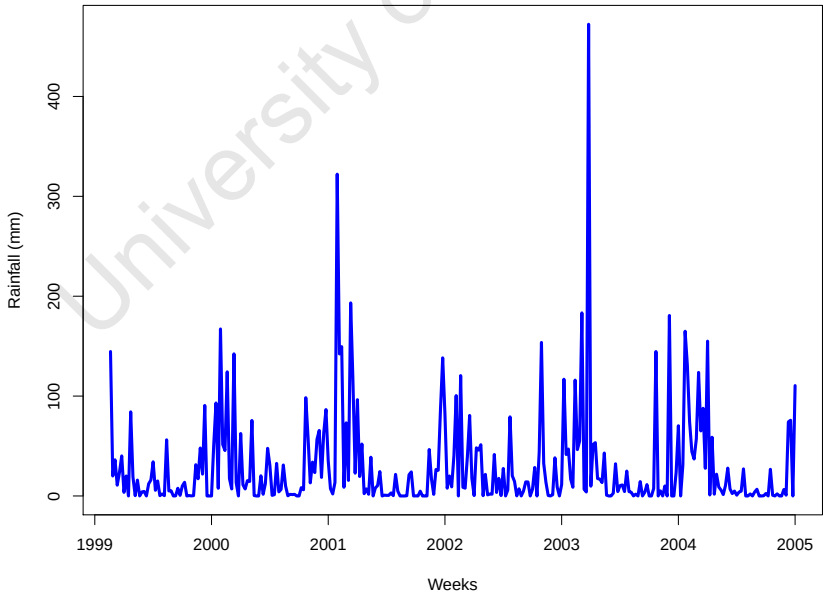
The graphs in Figure 4.5 indicate a similar seasonal pattern to that of the cholera counts, with peaks over the summer season. In order to compare these patterns in more detail, the plots in Figure 4.6 (a) and (b) are considered where standardised values of air temperature and rainfall, respectively, together with standardised values of cholera counts, are plotted against time. Each series was standardised by subtracting the mean from each observation and dividing this difference by the standard deviation.

From Figure 4.6 (a), a lagged relationship between air temperature and cholera counts can be observed with increases in air temperature preceding the increase in cholera counts by a few weeks. A visual comparison of cholera counts and rainfall does not reveal such a pattern. There was, however, evidence from an earlier study that the cumulative effect of rainfall has a stronger impact on cholera outbreaks (Van der Berg *et al.*, 2008). This is specifically due to the fact that a build up of rain results in flooding which inevitably leads to the spread of the disease. Consequently, an additional variable was created representing the cumulative rainfall over a two-week period. Longer periods of accumulation were also tested previously in statistical models (Van der Berg *et al.*, 2008) but these were found to be not as significant in the modelling of cholera counts as the two-week accumulation period.

What is apparent from the visual inspection of the cholera case data is the clear seasonality in the data. Although some seasonality is evident in the climatic variables which are used in the modelling of the cholera counts, it was necessary to determine whether additional seasonal terms were required. Consequently, additional seasonal variables were created using Fourier series terms, also referred to as harmonic terms. These variables, which capture seasonal cycles for weekly data, were constructed as $cos(2\pi kt/52)$ and $sin(2\pi kt/52)$, for $k = 1, 2, 3, 4$ and $t = 1, \ldots, 52$. The terms created using $k = 1$ represent one complete annual cycle, while the use of $k = 2, 3$ and 4 create 6-monthly, 4-monthly and quarterly cycles respectively. The use of these harmonic terms follows that of Davis *et al.* (2000) and Jung *et al.* (2006) in the analysis of asthma data.

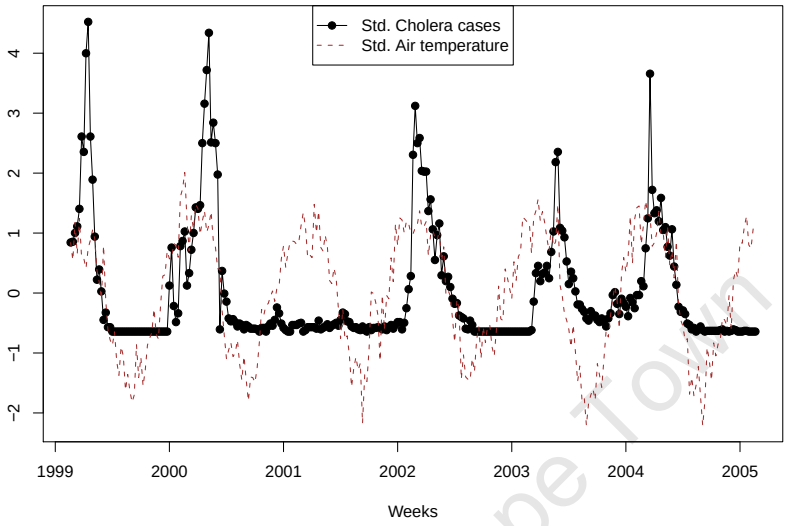


(a) Air temperature

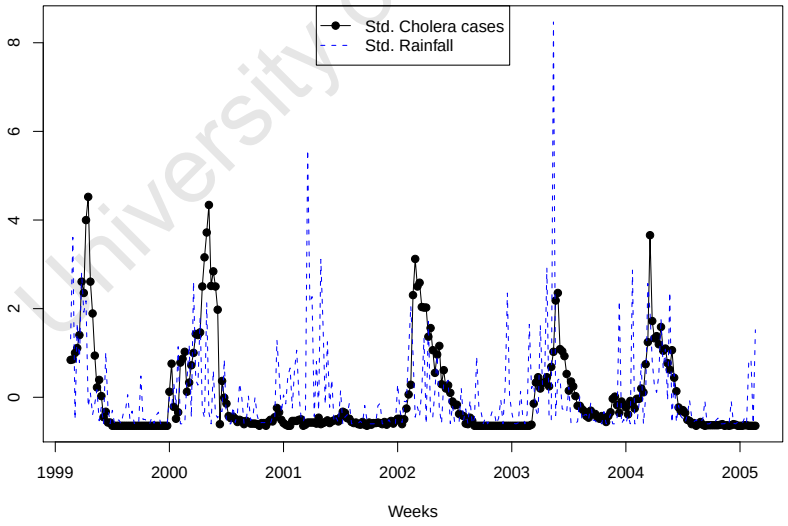


(b) Rainfall

Figure 4.5: *Plots showing air temperature and rainfall in Beira: Jan 1999 - Dec 2004.*



(a) Standardised cholera and air temperature



(b) Standardised cholera and rainfall

Figure 4.6: *Standardised values of cholera counts plotted against standardised air temperature and rainfall.*

4.4 Results for the Poisson and negative binomial regression models

For the purpose of comparing and evaluating the time series count models, namely ACP, DACP, Poisson-gamma and SAM, the cholera data were first modelled using static Poisson and negative binomial regressions, thus ignoring serial correlation. The exploratory analysis revealed evidence of high over-dispersion thus indicating a preference for the negative binomial regression model. Cameron and Trivedi (1998), however, point out that this over-dispersion is typical of most real-life data and that Poisson regression can still be used in such cases since it gives consistent estimates of the coefficients for the explanatory variables. They also note that the standard error estimates have to be adjusted as they are always under-estimated. The simplest means of adjusting the standard errors is to use a variance function in which the variance is a multiple of the mean and this multiple, or scale parameter, is used to scale the log likelihood, the standard errors and the t-statistics. The fitting of a Poisson distribution to such data is referred to as Poisson pseudo-MLE (PMLE) and further details are provided in Cameron and Trivedi (1998). In the present study, the results obtained in this manner are referred to as results from the "scaled Poisson" model. In the study by Van der Berg *et al.* (2008) only results from certain negative binomial regression models are presented and the use of Poisson regression is said to be avoided due to the evidence of over-dispersion in the cholera data.

Poisson and negative binomial regression fall into the class of generalised linear models (GLM). Since these models are well documented in many statistical texts, for example McCullagh and Nelder (1983) and Cameron and Trivedi (1998), the theory is not given here. Software required for fitting such models is widely available and for the purposes of this dissertation, the GLM procedure (PROC GENMOD) in the SAS/STAT software package (SAS Institute Inc., 2003) was utilised for fitting these models. PROC GENMOD also allows for the specification of the scale parameter when using Poisson regression and here the 'SCALE=PEARSON' option accommodates the estimation of the scale as the square root of Pearson's chi-square divided by the degrees of freedom. For the Poisson regression model, Pearson's chi-square is computed as $\chi^2 = \sum_{t=1}^{T} \frac{(y_t - \mu)^2}{\mu}$, where y_t represents the observation at time t, for $t = 1, \ldots, T$, and μ represents the static mean of the Poisson regression model.

When fitting these models, there were various combinations of lagged air temperature and lagged, cumulated rainfall variables that were tested for inclusion in the models. Seasonal variables were also included, as defined

in the previous exploratory analysis section. During the process of variable selection it was evident that only the annual harmonic terms needed to be used in the models, namely those constructed using the terms $cos(2\pi t/52)$ and $sin(2\pi t/52)$, where t represents the corresponding week of the year. Variables were selected using a combination of chi-square significance tests of the variable coefficients and a comparison of the AIC values of the various models. The "best" combination of explanatory variables was therefore selected as the one which included only significant variables and which resulted in the highest log-likelihood and the lowest AIC and produced residuals that were satisfactory.

When fitting the Poisson regression model using all possible combinations of available variables, the annual seasonal terms were found to be strong predictors of cholera counts. In addition, air temperature at lag 6, i.e. the average weekly air temperature value from 6 weeks before, was also found to be a significant explanatory variable for cholera counts. The effect of rainfall in the Poisson regression, however, was not significant in any model in which seasonal variables were included. The AIC values for some of the key models which had the best fits together with satisfactory residual patterns and for which all estimates of coefficients of the variables remaining in the model were significant at the 5% significance level, are listed in Table 4.3.

Table 4.3: AIC values for key Poisson regression models.

Variables fitted	**AIC**
Lag 6 air temperature	15311.6
Lag 5, lag 6 and lag 7 air temperature	13968.7
Lag 6 air temperature and lag 5 2-week cumulative rainfall	14940.4
Lag 6 air temperature and lag 6 2-week cumulative rainfall	14515.1
Lag 5 air temperature and annual seasonal terms	13465.5
Lag 6 air temperature and annual seasonal terms	13334.8
Lag 7 air temperature and annual seasonal terms	13428.4

More detailed output for the "best" fitting Poisson regression model, selected as the model from Table 4.3 with the lowest AIC, is presented in Table 4.4. This latter table presents the estimates of the coefficients, as well as the standard errors and chi-square values for determining the significance of these coefficients. The mean and variance of the Pearson residuals, as well as the value for the log-likelihood, are also computed. The scaled values for the Poisson standard errors, their corresponding chi-square values and the mean and variance of the scaled Pearson residuals are provided in square brackets and the value of the scale parameter, derived from the Pearson chi-square statistic, is also given.

Table 4.4: Maximum likelihood estimates of the parameters of the "best" fitting Poisson regression model, together with details of residuals and fit statistics. Values from the scaled Poisson model are included in square brackets and the p-value of the Ljung-Box statistic is given in round brackets.

	Poisson [+ scaled Poisson]		
Parameters/ Coefficients	**Estimates**	**s.e.**	**Chi-square**
Intercept	-1.2435	0.2317 [1.5327]	
$cos(2\pi t/52)$	0.6634	0.0178 [0.1177]	1390.4 [31.8]
$sin(2\pi t/52)$	1.2527	0.0311 [0.2056]	1623.9 [37.1]
Lag 6 temperature	0.1750	0.0092 [0.0606]	365.5 [8.4]
Mean(residuals)	0.185 [0.028]		
Var(residuals)	43.3 [0.996]		
Ljung-Box (52)	1191.5 (<0.0001)		
Log likelihood	-6659.4		
AIC	13334.8		
Scale	6.62		

The "best" fitting Poisson regression model, as presented in Table 4.4, includes both annual seasonal terms together with lag 6 air temperature. This model indicates that a positive relationship exists between the lagged temperature and cholera counts, i.e. an increase in temperature six weeks earlier is linked to an increase in cholera cases in the current week.

The Pearson residuals are standardised residuals and for a well specified model should have a mean of zero, a variance close to 1 and have no evidence of significant autocorrelation. Given these criteria it can be seen that the mean and variance of these residuals for the scaled Poisson are acceptable. However, since the Poisson regression model does not model the serial correlation in the data it is expected that the residuals would still include elements of autocorrelation. An autocorrelation function (ACF) plot of Pearson residuals is therefore considered, as given in Figure 4.7. This ACF plot shows that for the static Poisson regression model, the residuals still exhibit autocorrelation, with a definite seasonal cycle. The Ljung-Box statistic in Table 4.4 is computed on the Pearson residuals up to lag 52 and also indicates highly significant serial correlation.

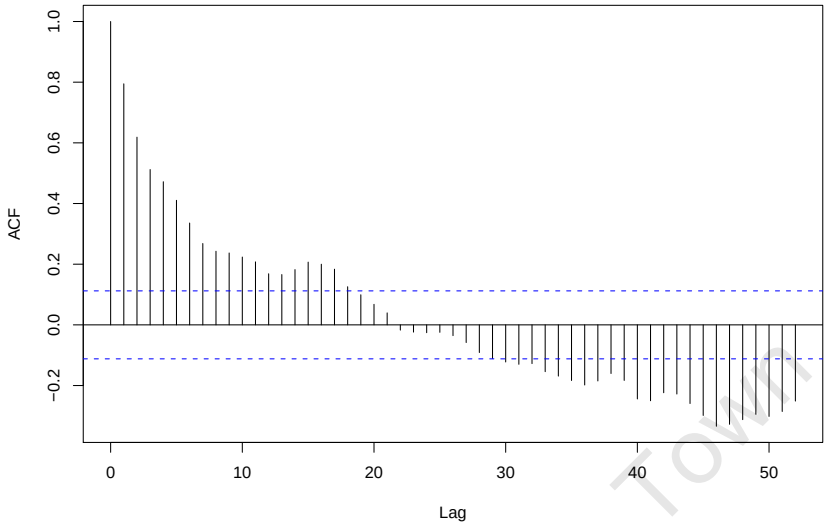


Figure 4.7: *Autocorrelation function plots of Pearson residuals for the Poisson regression model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

Figure 4.8 presents a plot of the scaled Pearson residuals against the predicted counts while Figure 4.9 shows the actual and predicted values for the model, plotted over the 6 year period. The former graph indicates a random pattern in the residuals but it can be seen from the latter graph that the Poisson regression model does not adequately capture the peaks and the prolonged lows in the data.

When fitting the negative binomial regression, it was found that the seasonal terms were highly significant and dominated the regression model, thus resulting in the rainfall and temperature variables having no significant influence on the predicted values. However, although the model including only the two annual seasonal terms produced the best AIC, the residuals patterns were not satisfactory and it was therefore excluded from consideration. The other key negative binomial regression models for which the residual patterns were acceptable and which produced the best fits, as measured with the AIC statistic, are listed in Table 4.5. All coefficients associated with the variables included in these models were significant at the 5% significance level.

From Table 4.5 it can be seen that there was virtually no difference between the third, fourth and fifth models in terms of AIC statistics and any of these

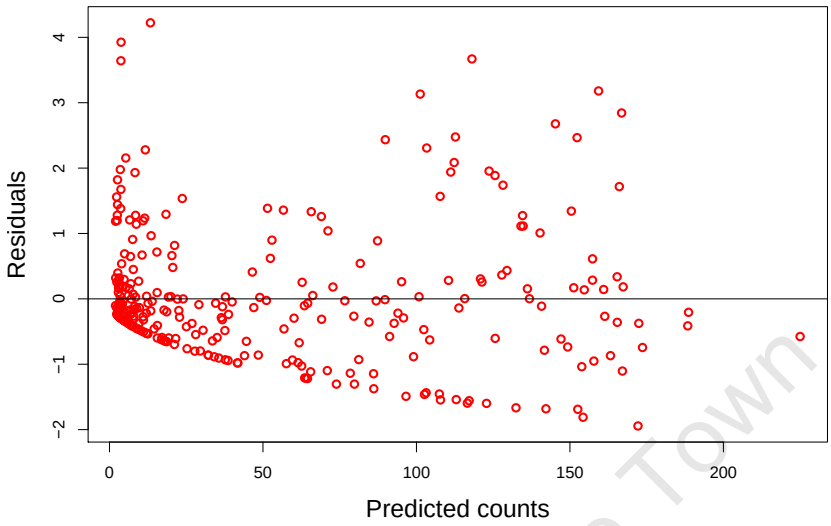


Figure 4.8: *Plot of scaled Pearson residuals vs predicted cholera counts from the Poisson regression model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

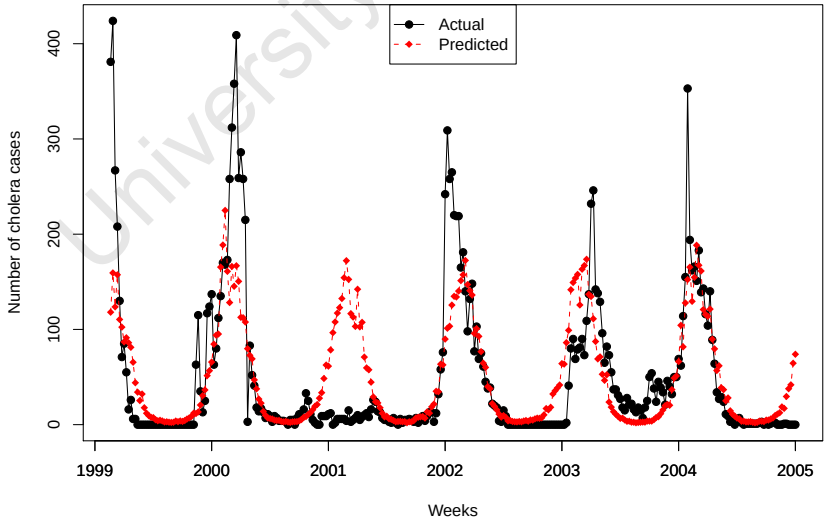


Figure 4.9: *Time series plots of actual and predicted cholera counts for the Poisson regression model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

Table 4.5: AIC values for key negative binomial regression models.

Variables fitted	**AIC**
Lag 6 air temperature	2568.7
Lag 7 air temperature	2572.1
Lag 6 and lag 7 air temperature	2563.5
Lag 6 air temperature and lag 5 2-week cumulative rainfall	2563.3
Lag 6 air temperature and lag 6 2-week cumulative rainfall	2561.9
Lag 7 air temperature and lag 5 2-week cumulative rainfall	2570.3
Lag 7 air temperature and lag 6 2-week cumulative rainfall	2567.9

could be selected as the "best" negative binomial model. It was decided that the results of the negative binomial regression model using lag 6 air temperature and lag 5 cumulative rainfall would be presented here for comparison purposes. These results are summarised in Table 4.6, together with the fit statistics, while the corresponding graphs of ACFs, standardised residuals versus predicted counts and the time series plots of predicted and actual values, are displayed in Figures 4.10, 4.11 and 4.12 respectively.

Table 4.6: Maximum likelihood estimates of the parameters of the selected negative binomial regression model, together with details of residuals and fit statistics.

	Negative Binomial		
Parameters/ Coefficients	**Estimates**	**s.e.**	**Chi-square**
Intercept	-6.291	0.7959	
Lag 5 cumulative rain	0.003	0.0014	4.93
Lag 6 temperature	0.3786	0.0333	129.22
Mean(residuals)	0.0015		
Var(residuals)	0.938		
Ljung-Box (52)	1011.1 (<0.0001)		
Log likelihood	-1282.14		
AIC	2570.28		

Similarly to the Poisson regression model, it is clear from the autocorrelation function plot in Figure 4.10 that the negative binomial regression model is not an adequate model for the analysis of the cholera counts due to the remaining autocorrelation in the residuals. This is also confirmed by the high value for the Ljung-Box statistic. The residual patterns for the negative binomial model, however, appear to be satisfactory, as evaluated from the residual versus predicted plot in Figure 4.11. The time series plots of actual

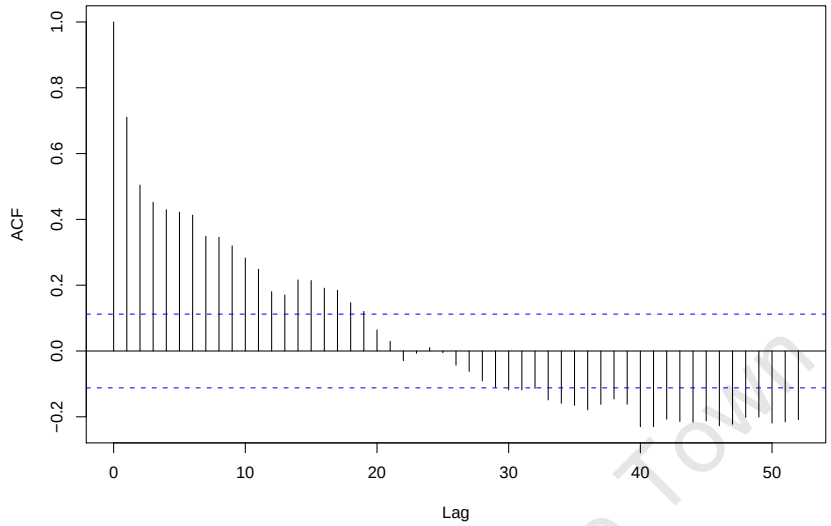


Figure 4.10: *Autocorrelation function plot of Pearson residuals for the negative binomial regression model, which includes lag 5 cumulative rainfall and lag 6 air temperature as explanatory variables.*

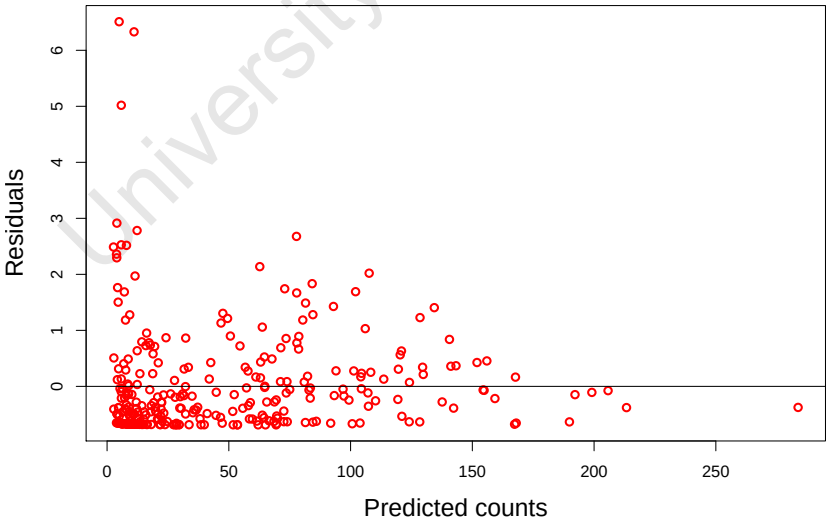


Figure 4.11: *Plot of Pearson residuals vs predicted cholera counts from the negative binomial regression model, which includes lag 5 cumulative rainfall and lag 6 air temperature as explanatory variables.*

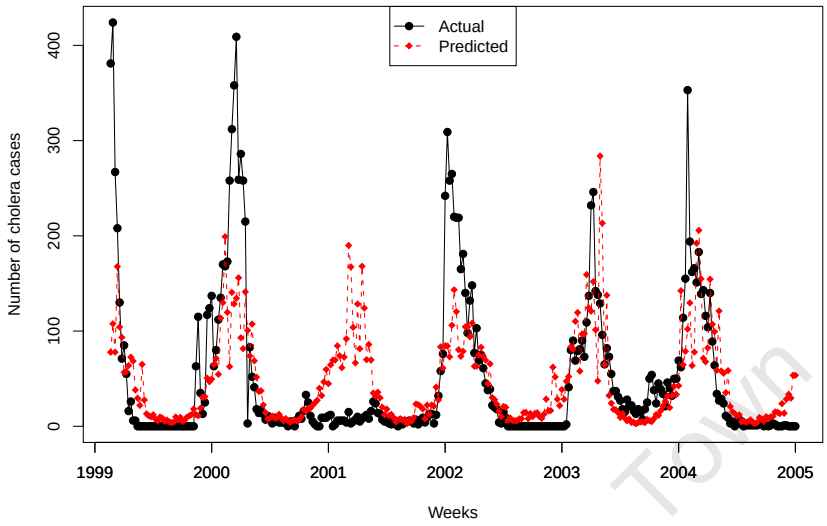


Figure 4.12: *Time series plots of actual and predicted cholera counts for the negative binomial regression model, which includes lag 5 cumulative rainfall and lag 6 air temperature as explanatory variables.*

and predicted values in Figure 4.12 display a strong correlation between the predicted counts and the cholera epidemics but also show that the model is not able to adequately predict the peaks.

4.5 Results for the observation-driven models

In this section, the performance of the selected observation-driven models, namely the ACP and DACP, when fitted to the cholera data is observed. The ACP and DACP models were fairly straightforward to program and the programming was done using the R software package (R Development Core Team, 2009). The R code for both models is provided in Sections B.1 and B.2 in Appendix B. The programs were tested using the polio and asthma data and then compared with published results. It should be noted that both of these models were very easy to implement, although it was sometimes found necessary to experiment with starting values as the optimiser did not always provide the best solution on the first set of starting values.

As with the static Poisson and negative binomial regression models, a variable selection process based on significance or non-significance of the parameters and on AIC values, was used to find the "best" fitting model for both the ACP and DACP. The results of the ACP models are presented first.

Table 4.7 lists the AIC statistics of the ACP models which fitted the cholera counts best and for which the coefficients associated with the variables were significant.

Table 4.7: AIC values for key ACP models.

Variables fitted	**AIC**
Lag 6 air temperature	4893.8
Lag 6 air temperature and lag 5 2-week cumulative rainfall	4848.3
Lag 6 air temperature and annual seasonal terms	4399.9
Lag 7 air temperature and annual seasonal terms	4629.2
Annual seasonal terms	4454.1
Lag 6 air temperature and lag 5 2-week cumulative rainfall and annual seasonal terms	4396.4
Lag 7 air temperature and lag 5 2-week cumulative rainfall and annual seasonal terms	4625.1

All the models presented in Table 4.7 fit the data well and have no remaining patterns in the residuals, except for the model including only air temperature which has some autocorrelation in the residuals with a slight seasonal pattern. Of the remaining models, the model including annual seasonal variables, lag 5 cumulative rainfall accumulated over two weeks, and lag 6 air temperature is the "best" model. The results of this model are summarised in Table 4.8.

Table 4.8 indicates that all of the explanatory variables are significant in the model and that there is a positive relationship between the lagged air temperature and cholera counts thus linking an increase in lagged temperature to an increase in cholera counts. The negative coefficient for the lag 5 2-week cumulative rainfall variable is not expected but is probably due to the combined effect of all the variables included in the model.

A likelihood ratio (LR) test for the ACP model to test for autocorrelation is a test of the null hypothesis that $\alpha = \beta = 0$, where α and β are the parameters associated with lagged values of the observations and lagged values of the conditional mean respectively (see Chapter 3). This LR test is found to be highly significant. When the parameters α and β are taken to be zero, the ACP model is reduced to the normal unscaled Poisson regression model with a constant mean and hence this LR test is rejecting the static Poisson regression in favour of the Poisson model with an autoregressive conditional mean (ACP).

The predicted (fitted) values refer to the values of μ_t^* which are updated

Table 4.8: Maximum likelihood estimates of the parameters of the "best" fitting ACP model, together with details of residuals and fit statistics.

	ACP model		
Parameters/ Coefficients	**Estimates**	**s.e.**	**Z-score**
Intercept	0.00991	0.2563	
ω	0.03061	0.0035	8.83
α	0.02058	0.0014	14.31
β	0.22997	0.0157	14.66
$cos(2\pi t/52)$	0.12998	0.0195	6.65
$sin(2\pi t/52)$	-0.40010	0.0365	-10.95
lag 5 cumulative rain	-0.00042	0.0001	-3.99
Lag 6 temperature	0.14013	0.0098	14.25
Mean(residuals)	0.18		
Var(residuals)	18.7		
Ljung-Box (52)	54.0 (0.326)		
Log likelihood	-2190.2		
AIC	4396.4		

once observations become known and these were computed using equations (3.3) and (3.8) from pages 18 and 20 respectively. The Ljung-Box statistic, computed on the Pearson residuals for the ACP model, suggests that there is no remaining autocorrelation in the residuals since the statistic is not significant at the 10% level of significance. This is confirmed by the ACF plot of the Pearson residuals, as given in Figure 4.13. The high value for the variance of the residuals in Table 4.8, however, shows that although the ACP model has accounted for most of the autocorrelation, it has not captured all of the dispersion in the data and the residuals are still highly over-dispersed.

The plot of residuals against predicted (fitted) values in Figure 4.14 does not indicate any major patterns in the residuals and the plot of predicted (fitted) and actual counts over time in Figure 4.15 exhibits a much closer fit to the actual data than was previously obtained when fitting the static Poisson and negative binomial regression models.

The fit of the DACP model to the cholera data is now considered. The AIC values of the key DACP models which fitted the cholera counts well are listed in Table 4.9. The "best" model for the DACP included the two annual seasonal terms, $cos(2\pi t/52)$ and $sin(2\pi t/52)$, and air temperature at lag 6. The different lags of the 2-week accumulated rainfall variable were found to be non-significant for the DACP model in which annual seasonal

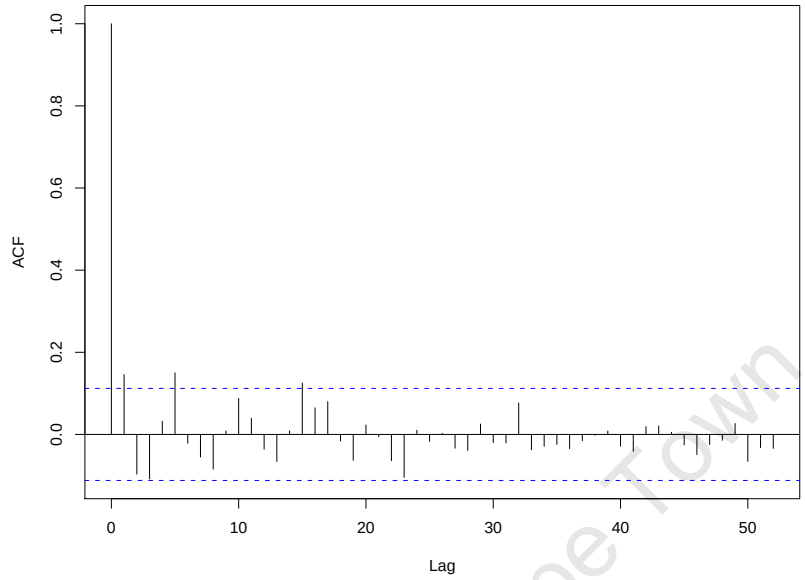


Figure 4.13: *Autocorrelation function plot of Pearson residuals from the ACP model, which includes annual seasonal variables, lag 5 2-week cumulative rainfall and lag 6 air temperature as explanatory variables.*

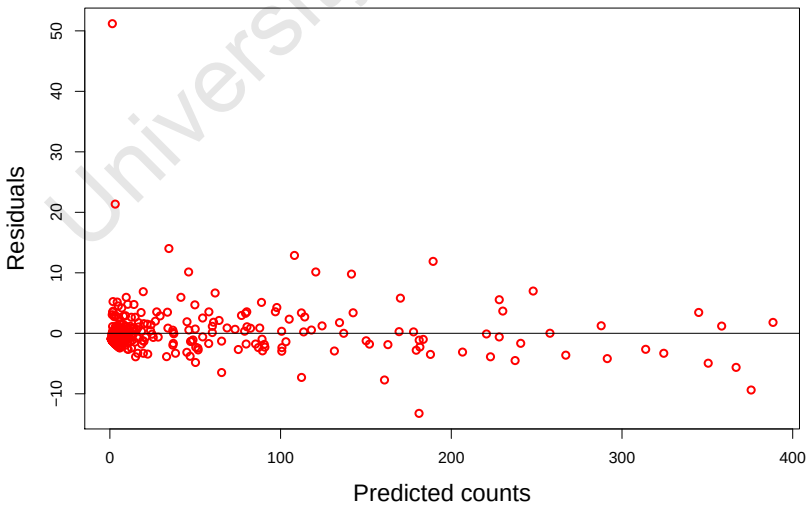


Figure 4.14: *Plot of Pearson residuals vs predicted (fitted) cholera counts from the ACP model, which includes annual seasonal variables, lag 5 2-week cumulative rainfall and lag 6 air temperature as explanatory variables.*

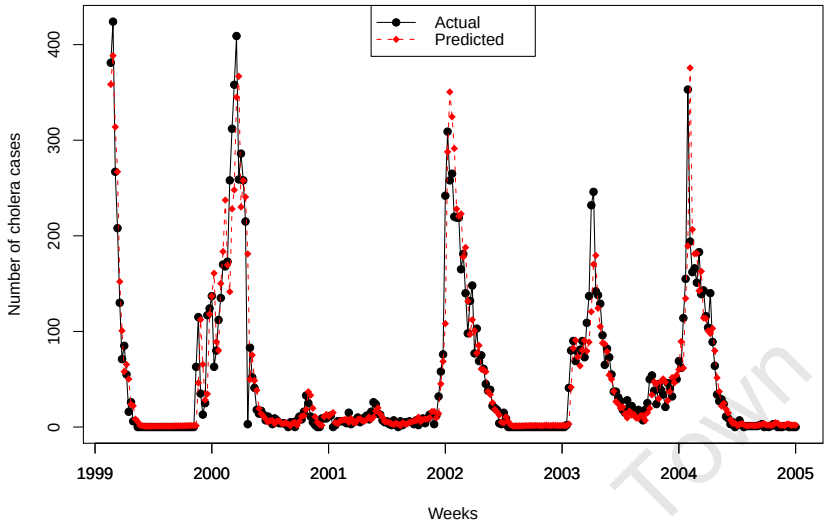


Figure 4.15: *Time series plots of actual and predicted (fitted) cholera counts from the ACP model, which includes annual seasonal variables, lag 5 2-week cumulative rainfall and lag 6 air temperature as explanatory variables.*

terms and lag 6 temperature were used. The results of the "best" DACP model are given in Table 4.10.

Table 4.9: AIC values for key DACP models.

Variables fitted	**AIC**
Lag 6 air temperature	2269.3
Lag 6 air temperature and lag 5 2-week cumulative rainfall	2263.7
Lag 7 air temperature and lag 5 2-week cumulative rainfall	2284.2
Lag 7 air temperature and lag 5 2-week cumulative rainfall and seasonal terms	2251.6
Lag 6 air temperature and annual seasonal terms	2230.5
Lag 7 air temperature and annual seasonal terms	2251.0

Similarly to the ACP model, the DACP exhibits a positive relationship between lagged air temperature and cholera counts. Comparing the estimates of the explanatory variable coefficients between the ACP and DACP models, as presented in Tables 4.8 and 4.10 respectively, there appears to be, as expected, a similarity between these two models. The cumulative rainfall variable, however, is only present in the "best" ACP model. The DACP has a considerably better variance of residuals than the ACP (i.e. 1.54 as

Table 4.10: Maximum likelihood estimates of the parameters of the "best" fitting DACP model, together with details of residuals and fit statistics.

	DACP model		
Parameters/ Coefficients	**Estimates**	**s.e.**	**Z-score**
Intercept	0.0118	0.8196	
ω	0.0516	0.0194	2.66
α	0.0250	0.0056	4.50
β	0.2600	0.0538	4.83
γ	0.0953	0.0079	12.16
$cos(2\pi t/52)$	0.1900	0.0645	2.95
$sin(2\pi t/52)$	-0.3223	0.1113	-2.89
lag 6 temperature	0.1271	0.0317	4.01
Mean(residuals)	0.050		
Var(residuals)	1.58		
Ljung-Box (52)	59.67 (0.2170)		
Log likelihood	-1107.3		
AIC	2230.5		

apposed to 18.7) but it is not as close to 1 as the negative binomial regression model and therefore there is still some measure of over-dispersion in the residuals.

In contrast to the ACP model, the DACP model makes use of an additional parameter γ which models the over-dispersion in the data. From Table 4.10 it can be seen that the estimate of this parameter of 0.095 is very different from 1. An LR test which tests for equi-dispersion in the data, i.e. tests the null hypothesis that $\gamma = 1$, is a test which compares the log-likelihoods of the DACP and ACP models since the DACP model with $\gamma = 1$ is equivalent to the ACP model. This test is highly significant in the present case and therefore rejects the ACP model in favour of the DACP model. The fact that the parameter estimates between the ACP and DACP models are so similar yet the ACP standard errors are so small indicates that although the ACP is capturing the parameter estimates adequately, it is not capturing all the dispersion and hence the over-dispersion is creating an under-estimation in the standard errors. This is analogous to the problems faced by the Poisson regression model fitted to over-dispersed data.

The predicted (fitted) values from the DACP model were computed using equations (3.16) and (3.17) from pages 24 and 25 respectively. The Ljung-Box statistic given in Table 4.10, which is computed on the Pearson residuals

of the DACP model, is not significant thus indicating that there is no remaining autocorrelation. The ACF plot of Pearson residuals given in Figure 4.16 indicates the same result. A plot of Pearson residuals against predicted (fitted) cholera counts for the DACP model is displayed in Figure 4.17 and here no visible patterns in the residuals can be detected.

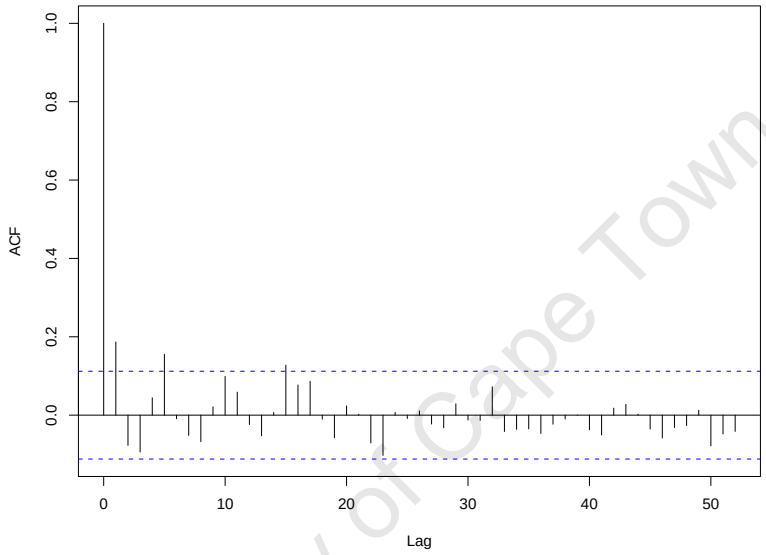


Figure 4.16: *Autocorrelation function plot of Pearson residuals from the DACP model, which includes annual seasonal variables and lag6 air temperature as explanatory variables.*

Figure 4.18 comprises time series plots of the actual and predicted (fitted) counts of the DACP model represented in Table 4.10. It is important to observe from this graph, as with the same graph for the ACP model in Figure 4.15, that the predicted always lag the actual counts by one week thus indicating the strong dependence on the number of cholera counts from the previous week. This is due to the strong autocorrelation in the data which is inherently accommodated in the model. This property is discussed further in Section 4.8 where concluding remarks on the case study are given.

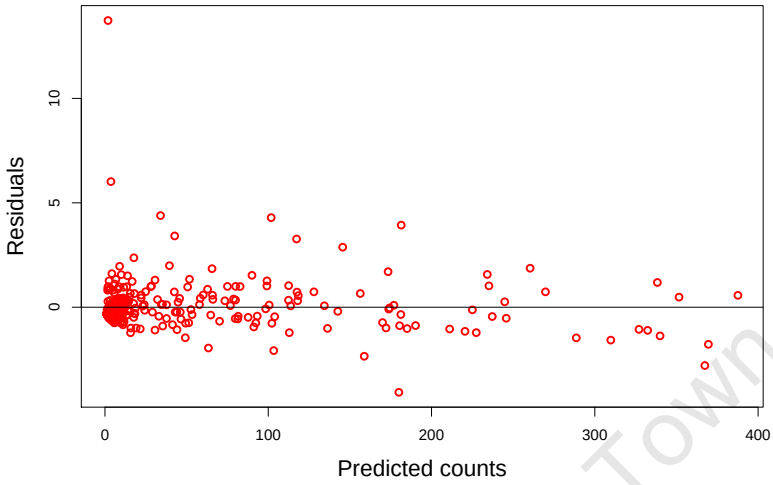


Figure 4.17: *Plot of Pearson residuals vs predicted (fitted) cholera counts from the DACP model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

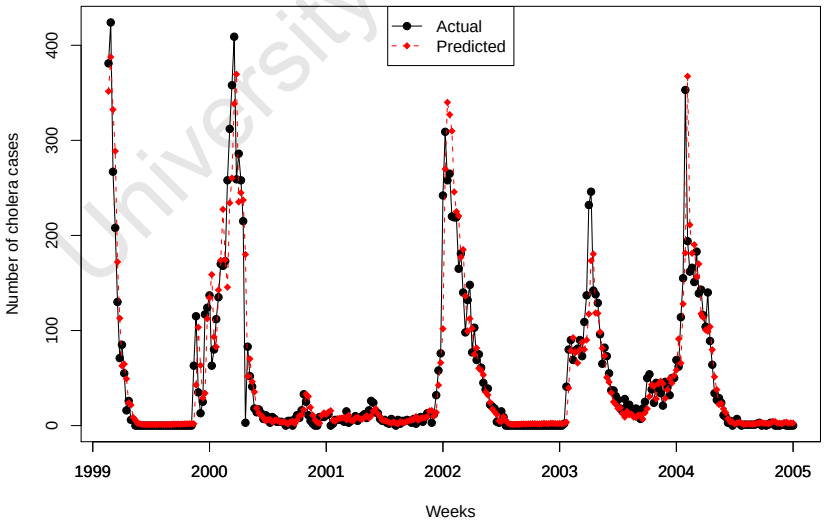


Figure 4.18: *Time series plots of actual and predicted (fitted) cholera counts for the DACP model, which includes annual seasonal variables and lag6 air temperature as explanatory variables.*

4.6 Results for the parameter-driven models

In contrast to the observation-driven models, the Poisson-gamma and SAM models were considerably more difficult to program although in essence the Poisson-gamma model only required the addition of a Kalman filter procedure. The Poisson-gamma method used by Harvey and Fernandes (1989a) was coded as the R program provided in Section B.3 of Appendix B. Attempts were made to use the transition equation proposed by Shephard (1994). However, in fitting Shephard's corrected transition equation, as defined in Chapter 3, to the cholera counts, the log-likelihoods became unstable for certain selections of parameter estimates while the Harvey and Fernandes method performed better. With the Harvey and Fernandes method though, it was found to be still important to test different starting values for the optimiser when applying it to the cholera data since a successful solution was not always obtained for certain starting values. Using the coefficients from a static Poisson regression model as starting values for the explanatory variables proved to be useful for both the Poisson-gamma and SAM models. Note that the program for the Poisson-gamma model was initially validated on the goal data used by Harvey and Fernandes (1989a).

The estimation of the SAM model proved to be the most challenging of the four models to program in R as the whole method is extremely complex. There were also certain scaling issues that needed to be considered and addressed in the program. The R program was tested on the asthma data and was found to produce the same results as those published in Jung *et al.* (2006). The code for this program is provided in Section B.4 of Appendix B. However, when trying to fit the SAM model to the cholera count data, the value for χ_t, as defined in Chapter 3, frequently went to infinity for a number of selected parameter values. The same result was observed when running the GAUSS program (Aptech Systems Inc., 2011), provided by Professor Robert Jung, on the cholera count data (Haines, personal communication). In the R program, the values of infinity resulted in a failure in the optimiser routine whenever such parameter estimates were initiated. The code was subsequently adjusted to bypass such occurrences in the optimser but, in order to obtain the best possible parameter estimates that maximised the approximate log-likelihood, numerous trial and error attempts involving different starting values were required. In addition, it was found that the final results achieved in this manner did not always produce sensible standard errors. Although the SAM model itself is straightforward and the approach to the estimation is intriguing, the implementation of the model involves too many computational challenges to make it a recommended approach for data with the same characteristics as the cholera case counts.

The Poisson-gamma and SAM models were fitted to the cholera data us-

ing various combinations of explanatory variables, as was done with the observation-driven models. The key Poisson-gamma models, together with their corresponding AIC values, are listed in Table 4.11. The "best" of these Poisson-gamma models is the model which includes annual seasonal variables and lag 6 air temperature. As with the DACP model, cumulative rainfall is not included in this "best" fitting model. The resulting parameter estimates, standard errors and various statistics for the Poisson-gamma model are presented in Table 4.12.

Table 4.11: AIC values for key Poisson-gamma models.

Variables fitted	**AIC**
Lag 6 air temperature	2406.2
Lag 5 and lag 6 air temperature	2398.4
Annual seasonal terms	2378.5
Lag 6 air temperature and annual seasonal terms	2373.3

Table 4.12: Maximum likelihood estimates of the parameters of the "best" fitting Poisson-gamma model, together with details of residuals and fit statistics.

	Poisson-gamma		
Parameters/ Coefficients	**Estimates**	**s.e.**	**Z-score**
Intercept	0.0100	0.0033	
ω	0.1333	0.0095	13.96
$cos(2\pi t/52)$	0.9612	0.2059	4.67
$sin(2\pi t/52)$	1.4243	0.2738	5.20
Lag 6 temperature	0.0616	0.0230	2.68
Ljung-Box (52)	40.30 (0.881)		
Log likelihood	-1181.7		
AIC	2373.3		

From the results of the Poisson-gamma model in Table 4.12 it can be seen that, similarly to the observation-driven models fitted to the cholera counts, a positive relationship exists between the air temperature at lag 6 and the cholera counts. It is observed, however, that the dependence on temperature as an explanatory variable is not as strong as in the other models. The parameter estimate for ω in the Poisson-gamma model is highly significant and very different from 1, that is 0.13, thus suggesting strong autocorrelation with recent observations having a much stronger influence than other observations. A value of $\omega = 1$ would imply that a constant mean describes the cholera count process and the model reverts to the static Poisson regres-

sion model.

The predicted (fitted) values make use of the outputs obtained from the 'Predict' step of the Kalman filter process including explanatory variables (Step 2 of Box 3.2 on page 38), as also indicated in the diagnostics section on page 39. Note that the variance of the residuals for the Poisson-gamma model have been omitted from Table 4.12 due to two abnormal Pearson residual values which skewed the results. These two large residuals appeared in all instances where the Poisson-gamma model was fitted to the cholera data and therefore only the raw residuals were used when assessing the "goodness of fit". In analysing these two residuals further, it was found that they both occurred at the start of an epidemic, one in 1999 and one in 2003, and in particular that these two epidemics were sudden, going from constant zero counts over a number of weeks to 63 and 41 counts respectively. From the time series plots of the cholera data, as displayed in Figure 4.1, it was seen that the start of the epidemics in other years was more gradual and did not have consistent zero incidences of cholera prior to the epidemic. The resulting effect in the Poisson-gamma model was such that the conditional variance became so small, after predicting close to zero counts for several weeks, that when the epidemic actually started in these years, the model was still predicting the incidences of cholera to be close to zero. This resulted in an abnormally large Pearson residual since the difference between the actual count and the conditional mean was very large while the conditional variance was extremely small. Although this may explain the occurrence of these spurious residuals it also highlights the fact that the explanatory variables in this instance, and in all instances concerning the Poisson-gamma model, cannot actually predict the start of the epidemic well. Only once an outbreak of cholera has started does the strong autocorrelation with cholera cases from time $t-1$ enable the model to adequately predict the cases at time t.

The ACF plot of the raw residuals from the Poisson-gamma model, displayed in Figure 4.19, indicates little remaining serial correlation and the Ljung-Box statistic given in Table 4.12 is non-significant thus confirming the absence of autocorrelation in the residuals. Note that the Ljung-Box statistic for the Poisson-gamma model was also computed using the raw residuals.

From the plot of residual versus predicted (fitted) counts for the Poisson-gamma model, given in Figure 4.20, there appears to be no obvious unexplained pattern in the residuals. The actual and predicted (fitted) counts plotted over the six years of the data are shown in Figure 4.21 and in these graphs the 1 week lag between actual and predicted values is evident.

Using a variable selection process with the SAM model, in which diagnostics

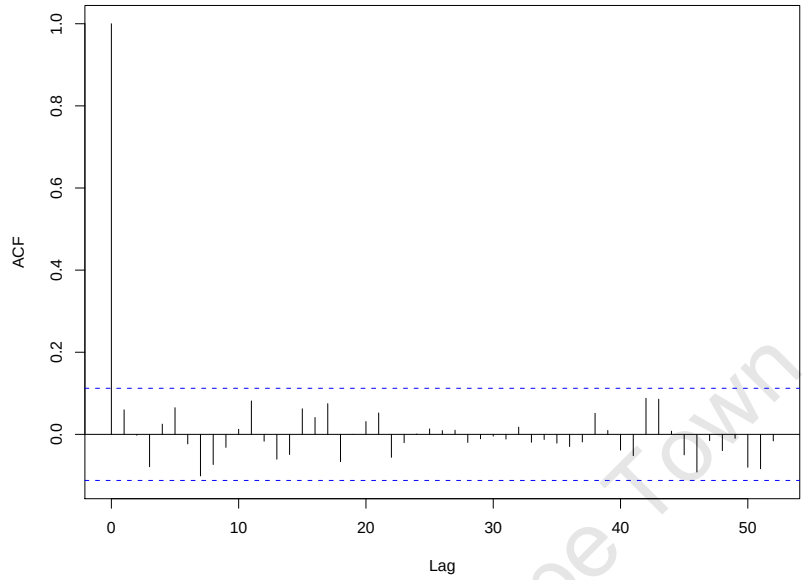


Figure 4.19: *Autocorrelation function plot of raw (non-standardised) residuals from the Poisson-gamma model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

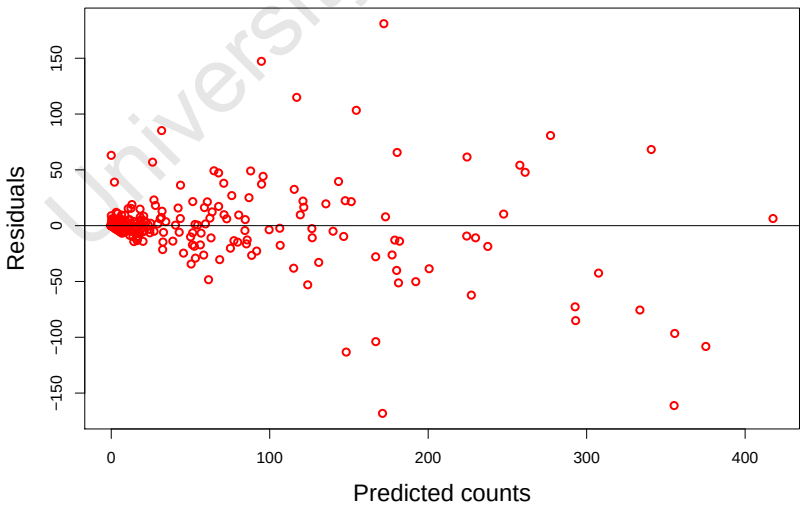


Figure 4.20: *Plot of raw (non-standardised) residuals vs predicted (fitted) cholera counts from the Poisson-gamma model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

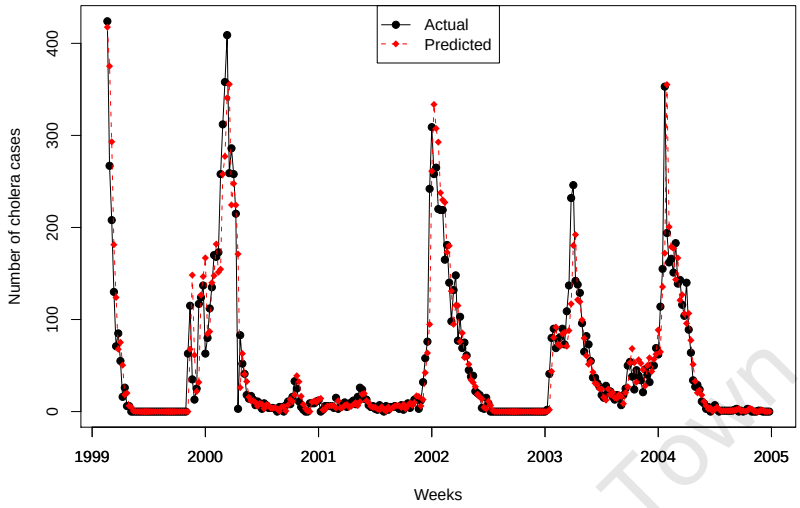


Figure 4.21: *Time series plots of actual and predicted (fitted) cholera counts for the Poisson-gamma model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

of residual patterns, significance of variables and "goodness of fit" statistics were analysed, certain models which produced a good fit to the cholera counts were identified and these are given in Table 4.13.

Table 4.13: AIC values for key SAM models.

Variables fitted	**AIC**
Lag 6 air temperature	2233.7
Lag 6 air temperature and lag 5 2-week cumulative rainfall	2271.0
Lag 6 air temperature and annual seasonal terms	2221.9
Lag 7 air temperature and annual seasonal terms	2240.6

Table 4.14 gives the results of the "best" SAM model that could be fitted to the cholera counts using the available explanatory variables. As with the DACP and Poisson-gamma model the final SAM model includes the annual seasonal terms and the lag 6 air temperature variable.

The coefficients for both of the seasonal variables and for lag 6 air temperature in Table 4.14 are all positive, thus indicating a positive relationship between these variables and cholera cases. There is a stronger dependence on air temperature in the SAM model when compared to the Poisson-gamma model but the coefficients for the seasonal variables are similar. Observe

Table 4.14: Maximum likelihood estimates of the parameters of the "best" fitting SAM model, together with details of residuals and fit statistics.

	SAM		
Parameters/ Coefficients	**Estimates**	**s.e.**	**Z-score**
Intercept	0.0050	0.4278	
γ	0.9838	0.0080	122.39
ν	0.3334	0.00003	10404.57
$cos(2\pi t/52)$	0.7189	0.2278	3.16
$sin(2\pi t/52)$	1.3322	0.2326	5.73
Lag 6 temperature	0.1392	0.0080	17.31
Mean(residuals)	-0.2927		
Var(residuals)	1.3879		
Ljung-Box (52)	56.34 (0.316)		
Log likelihood	-1104.9		
AIC	2221.9		

that the standard errors are somewhat unusual, as noted previously.

The predicted (fitted) values from the SAM model are computed using equation (3.51) given on page 49. The ACF plot of Pearson residuals for the SAM model is given in Figure 4.22. This plot indicates that the model has captured most of the serial correlation in the data, which is confirmed by the non-significant Ljung-Box statistic presented in Table 4.14.

For comparison purposes, the raw residuals are also used in the plot of residuals versus predicted values for the SAM model, as displayed in Figure 4.23 and there is no evidence of an obvious pattern in the residuals. The graph of actual and predicted counts over time, shown in Figure 4.24, displays a good fit of predicted (fitted) values to the actual data with the similar 1 week lag that was previously observed in the Poisson-gamma model.

The results for the "best" parameter-driven models, as presented in Tables 4.12 and 4.14, and the various graphs of the residuals, indicate that both the Poisson-gamma and SAM models appear to fit the data well and therefore on fit alone there is little difference between them. However, the challenges encountered when attempting to fit the SAM model to the cholera counts suggest that the Poisson-gamma model would be the preferred parameter-driven model for the cholera case data. This choice is in spite of the difficulties encountered with the Pearson residuals for the Poisson-gamma model and the issues with regard to starting values.

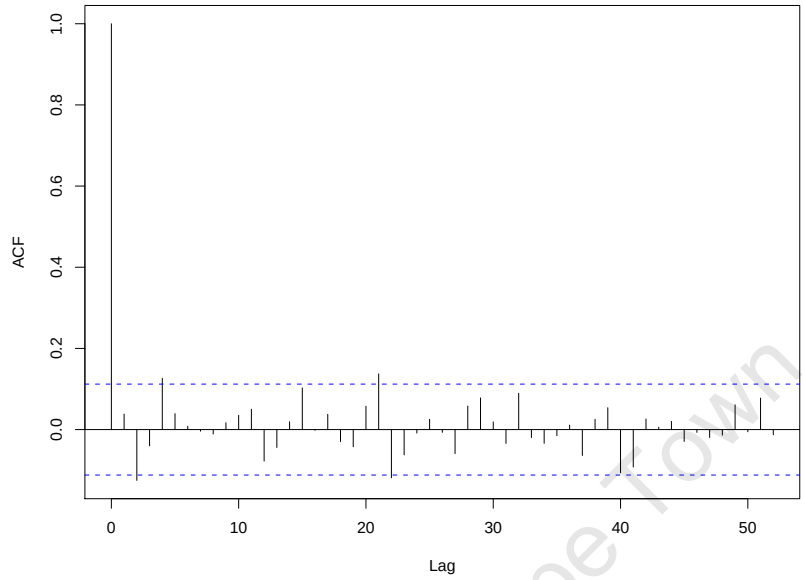


Figure 4.22: *Autocorrelation function plot of Pearson residuals from the SAM model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

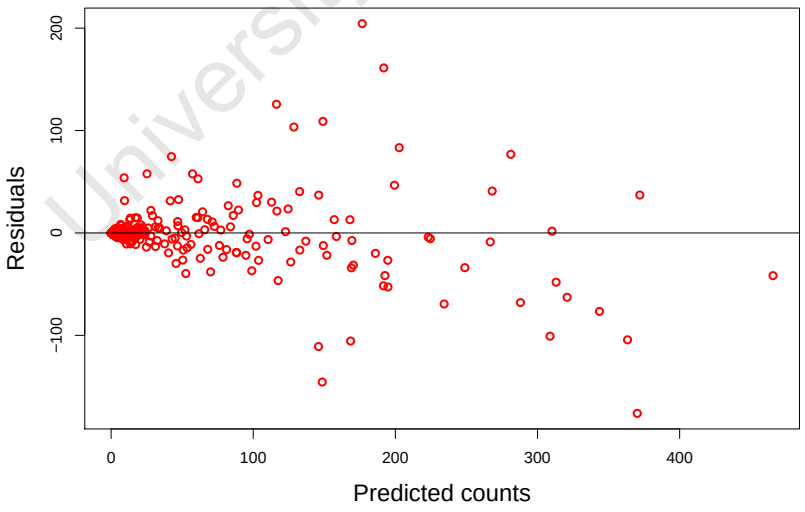


Figure 4.23: *Plot of raw residuals (non-standardised) vs predicted (fitted) cholera counts from the SAM model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

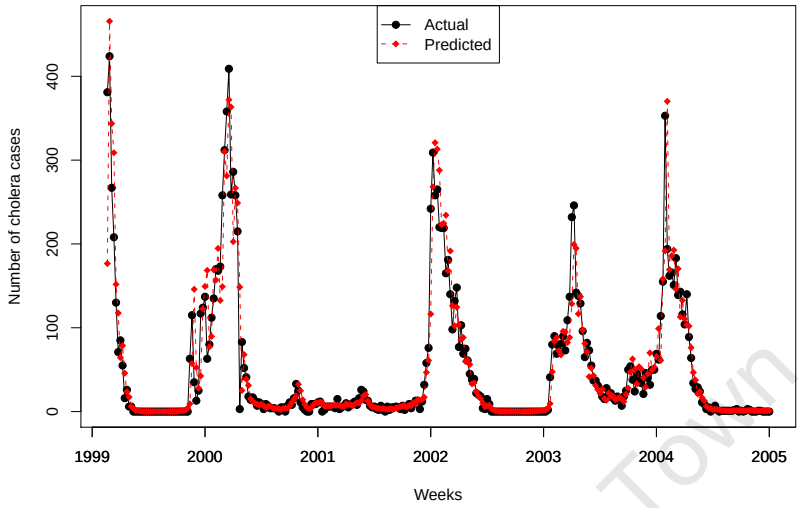


Figure 4.24: *Time series plots of actual and predicted (fitted) cholera counts for the SAM model, which includes annual seasonal variables and lag 6 air temperature as explanatory variables.*

4.7 Comparison of models

Having presented the results for all the models of interest, a comparison of these models is now provided. An overall comparison is given looking at the differences in coefficients for the explanatory variables, comparing the ACFs and other plots of residuals, as well as comparing the fit statistics.

It is interesting to note that all of the chosen observation-driven and parameter-driven models, except for the ACP model, found the rainfall variable to be non-significant in predicting cholera counts, while all the models included annual seasonal terms and lag 6 air temperature as explanatory variables. However, the ACP model had the same limitations as the static Poisson regression model when modelling highly over-dispersed data and a method of rectifying this problem could be scaling the standard errors in some way. In so doing, the scaled standard errors would potentially indicate a non-significant coefficient for lag 5 cumulative rainfall in the ACP model, as observed for the other models. The negative binomial regression model was the only model in which seasonal terms and air temperature could not be used together in the same model and instead reflected the need for lagged air temperature and lagged 2-week cumulative rainfall for predicting cholera counts. The difference in the results for the negative binomial regression model from those for the Poisson regression model was probably due to the

difference in the underlying distribution.

The ACF plots for the six models discussed in the previous sections show the expected result that the static Poisson and negative binomial regression models do not account for the serial correlation in the data, as shown in Figures 4.7 and 4.10, while the count data time series models, with ACFs presented in Figures 4.13, 4.16, 4.19 and 4.22, manage to accommodate most of the autocorrelation in the cholera count series. This is also emphasized by their respective Ljung-Box statistics, which are highly significant for the static models and non-significant for the time series models.

The plots of actual and predicted counts over time for each of the six models, given in Figures 4.9, 4.12, 4.15, 4.18, 4.21 and 4.24, indicate a similar performance between the observation-driven and parameter-driven models. The static negative binomial and Poisson regression models, however, do not perform as well as the other models, particularly with regard to capturing the peaks of the epidemics, and this is due to the fact that they are influenced by the explanatory variables alone and not by past observations.

When comparing the static Poisson regression, DACP, Poisson-gamma and SAM models all fitted using annual seasonal variables and lag 6 air temperature, it can be seen from Tables 4.4 and 4.14 that the static Poisson regression models and SAM models have similar coefficients for the three predictor variables. The Poisson-gamma model summarised in Table 4.12 has coefficients for the annual seasonal variables that are not too dissimilar from those of the static Poisson regression and SAM models but the coefficient for the air temperature predictor is considerably lower, namely 0.062 as apposed to 0.175 and 0.139 for the Poisson regression and SAM models respectively. The DACP model, presented in Table 4.10, has a similar coefficient for the air temperature predictor as that of the static Poisson regression and SAM models, namely 0.127, but the coefficients for the annual seasonal terms of this DACP model are very different, with the annual sine term having a negative coefficient. The large discrepancies between the estimated coefficients of the explanatory variables for the different models may be somewhat concerning but the differences could be due to the fact that these temperature and seasonal drivers, although having a significant influence, play a much smaller role in the overall model fit than the parameters which capture the serial correlation in the dependent series. The value of γ for the SAM model of 0.984 is very close to 1 indicating the strong dependence on previous values of cholera counts, as is similarly implied by the low value for ω of 0.133 in the Poisson-gamma model.

Although likelihood based models can typically be compared using their log-likelihoods and AIC statistics, due to scaling issues in the Poisson regression

and the approximation of the log-likelihoods for the DACP and SAM models, the RMSE and MAE statistics are used instead. The RMSE and MAE statistics both evaluate the fit of the model based on the raw residuals and these statistics are listed in Table 4.15 for all of the fitted models.

Table 4.15: Fit statistics from all the models.

Model	RMSE	MAE
Poisson	57.04	32.81
Negative binomial	63.23	37.04
ACP	31.00	14.95
DACP	31.23	15.27
Poisson-gamma	31.17	14.80
SAM	31.78	14.55

A lower RMSE or MAE indicates a better fit. However, the values of both statistics for the observation-driven and parameter-driven time series models in Table 4.15 indicate that there is very little difference in terms of the overall fits of the models to the cholera counts. The static Poisson and negative binomial regression models, however, have RMSE and MAE values that are roughly double those of the times series models, thus indicating the inadequacy of such models when applied to serially correlated time series data.

4.8 Concluding remarks on the case study

This case study has confirmed the existence of a relationship between cholera counts and both season and lagged air temperature. It has also indicated that the effects of accumulated rainfall are not significant once seasonal patterns have been accounted for. The best fit from each of the count data time series models indicates that all of the models have managed to capture most of the serial correlation in the data and in so doing exhibit a strong dependence on the most recently observed cholera counts. By capturing this autocorrelation, however, the models are placing less emphasis on the effect of climatic or seasonal variables as drivers of cholera and in fact indicate the failure of all these models to actually predict the onset until after the cholera epidemic has already commenced. This was emphasized by the two abnormally large Pearson residuals in the Poisson-gamma model and by the lagged effect shown in the plots of actual and predicted values for these time series models. Despite these practical issues however, the application of the selected time series models for count data has clearly shown the benefit gained in terms of model fit when compared to the static Poisson and negative binomial regression models, which do not take the time series properties of the data into account.

Chapter 5

Summary and conclusions

5.1 Summary

The aim of this dissertation has been to investigate models that are applicable to time series of count data and to apply these models to weekly cholera counts recorded in Beira over a six year period. Two classes of models were used, namely observation-driven and parameter-driven, and two models from each of these classes were explored. An overview of other count data time series models has been given as well as an overview of models previously applied in cholera studies. An in-depth case study of the cholera counts has been presented and in the process, the effect of environmental drivers on the outbreaks of cholera has been observed and discussed. Final conclusions on the cholera study are now provided in this chapter. The selected models have been compared in terms of their fit to cholera counts and in the present chapter, a general comparison is provided and computational aspects are discussed.

5.2 General comparison of models

The results from the analysis of cholera count data presented in Chapter 4 clearly showed that the static Poisson and negative binomial regression models were not suitable for data which are serially correlated. The ACP and DACP models both performed well when fitted to the cholera data, with similar parameter estimates for the models including lag 6 air temperature and seasonal terms. The DACP model, however, had the added advantage of capturing the over-dispersion in the data which was not adequately achieved by the ACP model. The ACP model could be useful in cases with small to moderate amounts of over-dispersion but in the case of the cholera counts this over-dispersion was excessive and thus affected the estimation of the log-likelihood and the standard errors of the parameters. The DACP model, on the other hand, provided a formulation where the standard er-

rors of the parameters, the log-likelihood and the Pearson residuals could be better estimated and compared to that of other models.

The Poisson-gamma and SAM models fitted using lag 6 air temperature and annual seasonal variables had similar fit statistics to the DACP model. Generally speaking, the overall fit of these different observation-driven and parameter-driven models, as measured by the RMSE and MAE, differed negligibly. The observation-driven ACP and DACP models, however, have the advantage over the parameter-driven Poisson-gamma and SAM models in that they are far easier to implement and estimate. Both the Poisson-gamma and SAM models were found to be rather challenging to program and implement and in particular, the program for the SAM model was not only extremely complex but also very unstable when applied to the cholera data. A gain in model simplicity can sometimes outweigh small gains in model fit and therefore the DACP model would be the preferred choice for data that exhibit similar characteristics to those of the cholera counts considered in the present study.

5.3 Conclusions on the cholera study

In terms of understanding the cholera data, and the environmental variables that drive the cholera epidemics in Beira, all the models fitted to the cholera case data indicated a strong relationship to a lagged effect of air temperature, specifically with a lag of 6 weeks. However, the benefit gained from including rainfall in a predictive model appeared to be limited. The use of harmonic terms to describe the additional annual seasonal effect in the cholera counts, in other words a seasonal effect which is over and above what is already implicit in the temperature data, added value to the predictions. However, as is typical of time series data, the cholera counts are strongly autocorrelated and in capturing this autocorrelation, all four time series models indicated the strong dependence of weekly cholera counts on previous levels of the disease. Note that the dependence on previous counts was modelled directly through lagged values of the observations, y_{t-1}, for the observation-driven ACP and DACP models, while for the parameter-driven Poisson-gamma and SAM models the dependence was modelled through a latent process.

Although many studies have shown that rainfall, temperature and season all have an effect on cholera epidemics and that these variables are indeed correlated with the cholera epidemics as recorded in the Beira data, determining a trigger from climatic variables at a weekly scale has not been achieved. The very nature of the cholera disease, namely the fact that it is contagious,

makes it much more difficult to predict from external variables alone, unlike conditions such as asthma to which Jung *et al.* (2006) fitted both the ACP and SAM models and found the condition to be strongly dependent on seasonal effects. The static Poisson and negative binomial regression models, which only relied on environmental drivers, showed that cholera cases are linked to air temperature and season but that the exact start of cholera outbreaks cannot be predicted from these explanatory variables. The time series models described here have in fact captured the characteristic of such a contagious disease by showing that the number of cholera cases from the previous week is the main predictor of the number of cholera cases in the current week, particularly since the spread of the disease is predominantly caused by the number of infected people rather than by the original primary source in the environment. This re-emphasizes the remarks by Fernández *et al.* (2009), citing Pascual *et al.* (2002), that climatic factors on their own are not sufficient to determine cholera epidemics and that further information, specifically with regard to immunity levels, is necessary for a better understanding of such outbreaks.

5.4 Final remarks

This study has highlighted the benefits of using models specifically developed for time series of counts over and above the standard techniques used in modelling count data, and has demonstrated the improved fit which was obtained for the cholera data case study. In addition, this dissertation has explored the cholera count data and the relationship between cholera epidemics and external environmental drivers in more detail. Although this study has emphasized the existence of a relationship of cholera counts to both season and lagged air temperature values, it has also shown the strong dependence that the number of cases of the disease in a particular week has on the cholera counts from the previous week. In essence it has shown that these relationships with climatic factors are not sufficient to predict the exact occurrence of a cholera outbreak and further information with regard to the susceptible population would probably add value to such studies, should data on these factors be available. However, given the available data for the incidences of cholera in Beira, and the absence of corresponding data on immunity levels or the susceptible population, there does not appear to be any scope to study this particular dataset further.

In terms of the actual count data time series models that have been studied in this dissertation, the conclusions drawn from these models have been based on only one case study, namely that of weekly cholera counts. It would therefore be of interest to repeat such comparisons by applying the models to other types of time series of counts and not necessarily restricting

them to the area of disease counts. In addition, the scope of the dissertation has been restricted to the study of only four different models for time series of counts and therefore expanding the study to include additional models would be helpful in gaining a broader understanding of this field of study.

Bibliography

Al-Osh, M. and Alzaid, A. (1987). First order integer valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8:261–275.

Aptech Systems Inc. (2011). *GAUSS, Version 11.* WA, USA. URL http://www.aptech.com.

Armstrong, J. S. (2001). *Principles of Forecasting.* Kluwer Academic Publishers, Dordrecht.

Brandt, P. T., Williams, J. T., Fordham, B. O., and Pollins, B. (2000). Dynamic modeling for persistent event-count time series. *American Journal of Political Science*, 44:823–843.

Brännäs, K. (1995). Explanatory variables in the AR(1) model. Umea Economic Studies No. 381, University of Umea.

Brännäs, K. and Johansson, P. (1994). Time series count data regression. *Communication in Statistics: Theory and Methods*, 23:2907–2925.

Brockwell, P. J. and Davis, R. A. (1996). *Introduction to Time Series and Forecasting.* Springer-Verlag, New York.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data.* Cambridge University Press, Cambridge.

Campbell, M. (1994). Time series regression for counts: an investigation into the relationship between sudden infant death syndrome and environmental temperature. *Journal of the Royal Statistical Society, Series A*, 157:191–208.

Cazelles, B., Chavez, M., Constantin de Magny, G., Guégan, J., and Hales, S. (2007). Time-dependent spectral analysis of epidemiological time-series with wavelets. *Journal of the Royal Society Interface*, 4:625–636.

Constantin de Magny, G., Cazelles, B., and Guégan, J. (2006). Cholera threat to humans in Ghana is influenced by both global and regional climatic variability. *EcoHealth*, 3:223–231.

Constantin de Magny, G., Murtugudde, R., Sapiano, M., Nizam, A., Brown, C., Busalacchi, A., Yunus, M., Nair, G., Gil, A., Lanata, C., Calkins, J., Manna, B., Rajendran, K., Bhattacharya, M., Huq, A., Sack, R., and Colwell, R. (2008). Environmental signatures associated with colera epidemics. *Proceedings of the National Academy of Sciences*, 105:17676–17681.

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

Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2003). Observation-driven models for Poisson counts. *Biometrika*, 90:777–790.

Davis, R. A., Dunsmuir, W. T. M., and Wang, Y. (1999). *Asymptotics, Nonparametrics, and Time Series*, Chapter 3: Modelling time series of count data. Marcel Dekker, New York.

Davis, R. A., Dunsmuir, W. T. M., and Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika*, 87:491–505.

Davis, R. A. and Wu, R. (2009). A negative binomial model for time series of counts. *Biometrika*, 96:735–749.

Durbin, J. and Koopman, S. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84:669–684.

Durbin, J. and Koopman, S. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Series B*, 62:3–56.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, 81:709–721.

Emch, M., Feldacker, C., Islam, M. S., and Ali, M. (2008). Seasonality of cholera from 1974 to 2005: a review of global patterns. *International Journal of Health Geographics*, 7:31.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalised Linear Models*. Springer-Verlag, New York.

Fernández, M. A. L., Bauernfeind, A., Jiménez, J. D., Gil, C. L., Omeiri, N. E., and Guibert, D. H. (2009). Influence of temperature and rainfall on the evolution of cholera epidemics in Lusaka, Zambia, 2003-2006: analysis of a time series. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103:137–143.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2000). *Bayesian Data Analysis*. 2nd edition. Chapman and Hall, London.

Gil, A. I., Louis, V. R., Rivera, I. N., Lipp, E., Huq, A., Lanata, C. F., Taylor, D. N., Russek-Cohen, E., Choopun, N., Sack, R. B., and Colwell, R. R. (2004). Occurrence and distribution of *Vibrio cholerae* in the coastal environment of Peru. *Environmental Microbiology*, 6:699–706.

Grunwald, G., Hamza, K., and Hyndman, R. (1997). Some properties and generalizations of non-negative Bayesian time series models. *Journal of the Royal Statistical Society, Series B*, 59:615–626.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Harvey, A. C. and Fernandes, C. (1989a). Time series models for count or qualitative observations. *Journal of Business and Economic Statistics*, 7:407–417.

Harvey, A. C. and Fernandes, C. (1989b). Time series models for insurance claims. *Journal of the Institute of Actuaries*, 116:515–528.

Heinen, A. (2003). Modelling time series count data: An autoregressive conditional Poisson model. Core discussion paper No. 2003-63, Catholic University of Louvain, Belgium.

Huq, A., Sack, R. B., Nizam, A., Longini, I. M., Nair, G. B., Ali, A., Morris, J. G., Khan, M. H., Siddique, A. K., Yunus, M., Albert, M. J., Sack, D. A., and Colwell, R. R. (2005). Critical factors influencing the occurrence of *Vibrio Cholerae* in the environment of Bangladesh. *Applied and Environmental Microbiology*, 71:4645–4654.

Jacobs, P. and Lewis, P. (1978a). Discrete time series generated by mixtures I: Correlation and runs properties. *Journal of the Royal Statistical Society, Series B*, 40:94–105.

Jacobs, P. and Lewis, P. (1978b). Discrete time series generated by mixtures II: Asymptotic properties. *Journal of the Royal Statistical Society, Series B*, 40:222–228.

Jacobs, P. and Lewis, P. (1983). Stationary discrete autoregressive moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4:19–36.

Janacek, G. and Swift, L. (1993). *Time Series: Forecasting, Simulation, Applications*. Ellis Horwood, New York.

Jung, R. C., Kukuk, M., and Liesenfeld, R. (2006). Time series of count data: modeling, estimation and diagnostics. *Computational Statistics and Data Analysis*, 51:2350–2364.

Koelle, K. and Pascual, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: A nonlinear time series approach with an application to cholera. *The American Naturalist*, 163:901–913.

Lambert, P. (1996a). Modeling of nonlinear growth curve on series of correlated count data measured at unequally spaced times: A full likelihood based approach. *Biometrics*, 52:50–55.

Lambert, P. (1996b). Modelling of repeated series of count data measured at unequally spaced times. *Journal of the Royal Statistical Society, Series C*, 45:31–38.

Lewis, P., McKenzie, E., and Hugus, D. (1989). Gamma processes. *Communications in Statistics - Stochastic Models*, 5:1–30.

Liesenfeld, R. and Richard, J. F. (2005). Classical and Bayesian analysis of univariate and multivariate stochastic volatility models. *Econometric Reviews*, 25:335–360.

Lipp, E. K., Huq, A., and Colwell, R. R. (2002). Effects of global climate on infectious disease: the cholera model. *Clinical Microbiology Reviews*, 15:757–770.

Lobitz, B., Beck, L., Huq, A., Wood, B., Fuchs, G., Faruque, A., and Colwell, R. (2004). Climate and infectious disease: Use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Environmental Microbiology*, 6:699–706.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman and Hall, London.

Makridakis, S., Wheelwright, S., and Hyndman, R. (1998). *Forecasting Methods and Applications*, Chapter: Advanced Forecasting Models. 3rd edition. Wiley, New York.

Masahiro, H., Armstrong, B., Hajat, S., Wagatsuma, Y., Faruque, A. S. G., Hayashi, T., and Sack, D. A. (2008). The effect of rainfall on the incidence of cholera in Bangladesh. *Epidemiology*, 19:103–110.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*. Chapman and Hall, London.

McKenzie, E. (2003). *Handbook of Statistics*, Volume 21, Chapter: Discrete Variate Time Series. Elsevier Science Publishers, Amsterdam.

Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6:318–334.

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models*. Wiley, New York.

Pascual, M., Bouma, M. J., and Dobson, A. P. (2002). Cholera and climate: revisiting the quantitative evidence. *Microbes and Infection*, 4:237–245.

Pascual, M. and Ellner, S. P. (2000). Linking ecological patterns to environmental forcing via nonlinear time series models. *Ecology*, 81:2767–2780.

Pascual, M., Rodó, X., Ellner, S., Colwell, R., and Bouma, M. (2000). Cholera dynamics and El Niño–Southern Oscillation. *Science*, 289:1766–1769.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria. URL http://www.R-project.org.

Richard, J. F. and Zhang, W. (2006). Efficient high-dimensional Monte Carlo importance sampling. Working Paper, University of Pittsburgh.

Rodó, X., Pascual, M., Fuchs, G., and Faruque, A. S. G. (2002). ENSO and cholera: A nonstationary link related to climate change? *Proceedings of the National Academy of Sciences*, 99:12901–12906.

SAS Institute Inc. (2002-2003). *SAS/STAT software, Version 9.1 of the SAS system for Windows*. Cary, NC, USA.

Shephard, N. (1994). Local scale models. *Journal of Econometrics*, 60:181–202.

Shephard, N. (1995). Generalized linear autoregressions. Working paper, Nuffield College, Oxford.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–667.

Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications: With R Examples*, Chapter 6: State-space models. 2nd edition. Springer, New York.

Smith, R. and Miller, J. (1986). A non-Gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society, Series B*, 48:79–88.

Steutel, F. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability*, 7:893–899.

Van der Berg, F., Pienaar, M., Holloway, J., Koen, R., and Elphinstone, C. (2008). A comparison of various modelling approaches applied to cholera case data. *Orion*, 24:17–36.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, 75:621–629.

Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44:1019–1031.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall, London.

Appendix A

Appendix: Definitions and theoretical results

A.1 Theoretical results regarding the gamma distribution

In the formulation of the Poisson-gamma model there are some useful theoretical results that are required with regard to the gamma distribution. These results are described here.

The first result can be taken from the beta-gamma transformation given in the paper by Lewis *et al.* (1989) in which they state that multiplying a $Gamma(m+n,\beta)$ random variable by an independent $Beta(m,n)$ random variable results in a $Gamma(m,\beta)$, thus allowing for the reduction of the shape parameter of a gamma distribution using this transformation. Using a slightly different notation with $b=\beta$ and a equivalent to the $m+n$ used in Lewis *et al.* (1989) and introducing a parameter ω, with $0<\omega<1$, such that $a=m+n=\omega a+(1-\omega)a$, the following result is obtained:

Result 1:
Taking x and z such that

$$x\sim Gamma(a,b), \qquad z\sim Beta(\omega a,(1-\omega)a)$$

with $0<\omega<1$, and taking $y=xz$, then

$$y=xz\sim Gamma(\omega a,b). \tag{A.1}$$

The second result follows from the scaling properties of a gamma distribution, as used by Smith and Miller (1986), and concerns the multiplication of a $Gamma(a,b)$ variate with a constant c and the resulting effect on the scale parameter b of the distribution.

Result 2:
Taking x such that

$$x \sim Gamma(a, b)$$

and taking $y = cx$, where c is a constant, it follows that

$$y = cx \sim Gamma(a, \frac{b}{c}) \tag{A.2}$$

Combining results 1 and 2 gives the following:
Result 3:
Taking c as a constant, $x \sim Gamma(a, b)$ and $z \sim Beta(\omega a, (1-\omega)a)$, then

$$y = cxz \sim Gamma(\omega a, \frac{b}{c}) \tag{A.3}$$

A.2 Conjugate priors and the Poisson-gamma conjugacy

The Poisson-gamma model makes use of a conjugate prior in the formulation of the model. In the Bayesian paradigm, the term conjugacy relates to the property whereby the posterior distribution is from the same family of distributions as the prior distribution (Gelman *et al.*, 2000). The use of conjugate families can be algebraically convenient since the form of the posterior distribution is then known. A full description of different conjugate priors associated with the various families of distributions, as well as applications of this property, can be found in Gelman *et al.* (2000). The observation distribution for the Poisson-gamma model is Poisson and for this distribution, the conjugate prior is the gamma distribution.

Posterior distribution for the Poisson-gamma model
Using Bayes' theorem and the conjugacy of the Poisson and gamma distributions, where $y_t|\mu_t \sim Poisson(\mu_t)$ and $\mu_t|Y_{t-1} \sim Gamma(a, b)$, results in the following posterior distribution:

$$\begin{aligned} f(\mu_t|y_t) &= \frac{Pr(y_t|\mu_t) \times f(\mu_t)}{Pr(Y_t)} \\ &\propto Pr(y_t|\mu_t) \times f(\mu_t) \\ &\propto \frac{e^{-\mu_t}{\mu_t}^{y_t}}{y_t!} \times \frac{e^{-b_{t|t-1}\mu_t}{\mu_t}^{a_{t|t-1}-1}}{\Gamma(a_{t|t-1}){b_{t|t-1}}^{a_{t|t-1}}} \\ &\propto e^{-\mu_t(b_{t|t-1}+1)}{\mu_t}^{(a_{t|t-1}+y_t)-1}, \end{aligned}$$

where Γ represents the gamma function. Thus

$$\mu_t|Y_t \sim Gamma(a_{t|t-1} + y_t, b_{t|t-1} + 1) \tag{A.4}$$

or, in other words, (A.4) shows that due to the properties of conjugate families, the posterior, $\mu_t|Y_t$, is from a gamma distribution.

Predictive distribution for the Poisson-gamma model
The conditional or predictive distribution of y_t has a pdf given by

$$p(y_t|Y_{t-1}) = \int_0^\infty p(y_t|\mu_t)p(\mu_t|Y_{t-1})d\mu_t$$

and since these are Poisson observations with a gamma prior, it follows that

$$\begin{aligned} p(y_t|Y_{t-1}) &= \int_0^\infty \frac{e^{-\mu_t}{\mu_t}^{y_t}}{y_t!}\frac{b^a}{\Gamma(a)}\mu_t^{a-1}e^{-\mu_t b}d\mu_t \\ &= \frac{b^a}{\Gamma(a)y_t!}\int_0^\infty \mu_t^{(y_t+a)-1}e^{-\mu_t(b+1)}d\mu_t, \end{aligned} \tag{A.5}$$

where $a = a_{t|t-1}$ and $b = b_{t|t-1}$.

Now considering the known integral result that $\int_0^\infty \mu^{d-1}e^{-c\mu}\frac{c^d}{\Gamma(d)}d\mu = 1$, it follows that $\int_0^\infty \mu^{d-1}e^{-c\mu}d\mu = \frac{\Gamma(d)}{c^d}$. Using this result in equation (A.5) gives

$$\begin{aligned} p(y_t|Y_{t-1}) &= \frac{b^a}{\Gamma(a)y_t!}\frac{\Gamma(y_t+a)}{(b+1)^{y_t+a}} \\ &= \frac{\Gamma(a+y_t)}{y_t!\Gamma(a)}(\frac{b}{b+1})^a(\frac{1}{b+1})^{y_t}. \end{aligned} \tag{A.6}$$

This is the pmf of a negative binomial distribution and by substituting back for a and b, it can be written as

$$y_t|Y_{t-1} \sim NegBin(a_{t|t-1}, \frac{b_{t|t-1}}{b_{t|t-1}+1}).$$

As a result of the gamma distribution being the natural conjugate of the Poisson, the end result is a predictive distribution from the same family of distributions as the Poisson, namely the negative binomial distribution.

A.3 Result used in the derivation of the SAM model

Taking a random variable x where the pdf for x has a kernel equal to $\exp\{-\frac{1}{2}(bx^2-2cx)\}$, then the result that $\exp\{-\frac{1}{2}(bx^2-2cx)\} \propto \exp\{-\frac{1}{2}[b(x-\frac{c}{b})^2]\}$ can be considered when all constants not involving x are removed. Hence it can be deduced that $x \sim N(\frac{c}{b}, \frac{1}{b})$.

Appendix B

Appendix: R Code

B.1 ACP model

```
# File "ACP_expl.r" containing R code for fitting Heinen's ACP(1,1) model
# with explanatory variables

# y = observed series
# x = matrix of explanatory variables
# n = number of observations in generated series
# a = value of alpha
# b = value of beta
# w = value of omega
# d = value of delta (parameters associated with explanatory variables)
# p = (a, b, w, d)
# m0 = inital value of poisson mean (mu)
# y0 = initial value of y counts


# Log-likelihood function -
# This function is called from the optimiser in function maxloglikeexpl_acp.
# It returns the value of the log-likelihood for the given data and parameters.
loglikeexpl_acp<-function(p,y,x)
  {
  #initialise the parameters
   a<-p[1]; b<-p[2]; w<-p[3]; d<-p[4:length(p)]
   y<-as.ts(y);
   x<-as.ts(as.matrix(x))
   y0<-mean(y)
   m0=y0
   n <- length(y)
   mu<-matrix(0,nrow=n);
```

```
   mu2<-matrix(0,nrow=n)
   logl<-matrix(0,nrow=n)
  #Compute the log-likelihood function
     for (t in 1:n)
     { if (t==1)
        { mu[t] <- w+a*y0+b*m0
          mu2[t] <- mu[t]*exp(x[t,]%*%d)
          logl[t] <-(y[t]*log(mu2[t]) - mu2[t] - lgamma(y[t]+1))
        }
       else
        { mu[t] <- w+a*y[t-1]+b*mu[t-1]
          mu2[t] <- mu[t]*exp(x[t,]%*%d)
          logl[t] <-(y[t]*log(mu2[t]) - mu2[t] - lgamma(y[t]+1))
        }
     }
    llabwd<-sum(logl)
    llabwd  #return the loglikeilhood for the given values of a, b, w and d
  }

# This funtion maximises the log-likelihood using the "optim" function in R
# a.init = starting value for alpha
# b.init = starting value for beta
# w.init = starting value for omega
# d.init = vector of starting values for delta
maxloglikeexpl_acp<-function(y,a.init,b.init,w.init,d.init,x)
  {
    #Define the lower and upper bounds to use a bounded optimising routine
    d.lower <- matrix(-1,ncol=1,nrow=length(d.init));
    lower.bound <- c(0,0,0,d.lower)
    d.upper <- matrix(10,ncol=1,nrow=length(d.init));
    upper.bound <- c(1,1,10,d.upper)

    #Define x as a matrix
    x <- as.matrix(x)
    #Define no. of parameters in model
    k <- ncol(x) + 3

    # Fit the model
    fitted.param<-optim(c(a.init,b.init,w.init,d.init),loglikeexpl_acp,
                        method="L-BFGS-B",
                        lower=lower.bound,upper=upper.bound,
                        hessian=TRUE,
                        control=list(trace=2,REPORT=5,fnscale=-(length(y))),
                        y=as.matrix(y),x=as.matrix(x))
```

```
 # Retrieve the parameters and the log-likelihood function (llf) value
 param<-fitted.param$par
 llf <- fitted.param$value

# Estimate the standard errors and compute the z scores
 covar <- -solve(fitted.param$hessian)
 se<-sqrt(diag(covar))
 z<-param/se
 coefs <- t(matrix(rbind(param,se,z),nrow=3))
 colnames(coefs) <- c("Parameters","Std. Errors","Z-score")

# Get names of parameters
if (k>4) {x.names <- colnames(x)}
#If only one explanatory variable then can't retrieve column name
else {x.names <- "x.variable"}
rownames(coefs) <- c("Alpha","Beta","Omega",x.names)

#Compute the predicted and residual values
 n <- length(y)
 y0<-mean(y)
 m0<-y0
 pnum <- length(param)
 a<-param[1]; b<-param[2]; w<-param[3]; d<-param[4:pnum]
 mu<-matrix(0,nrow=n);
 pred<-matrix(0,nrow=n);
 res<-matrix(0,nrow=n);
 rawres<-matrix(0,nrow=n);
 for (tt in 1:n)
 {
    {if (tt==1)
      { mu[tt] <- w+a*y0+b*m0 }
     else
        { mu[tt] <- w+a*y[tt-1]+b*mu[tt-1]}
    }
    pred[tt] <- mu[tt]*exp(x[tt,]%*%d)
    res[tt] <- (y[tt]-pred[tt])/sqrt(pred[tt])
    rawres[tt] <- y[tt]-pred[tt]
 }

# Write out the output
 cat(" ","\n")
 cat("-------------------------------------","\n")
 cat("ACP(1,1) regression output","\n")
```

```
    cat("---------------------------------------","\n")
    printCoefmat(coefs)
    cat("---------------------------------------","\n")
    cat("Log-likelihood value  : ", llf, "\n")
    aic <- -2*llf+2*k
    cat("AIC                   : ", aic, "\n")
    dof <- length(y)-k
    cat("Degrees of Freedom    : ", dof, "\n")
    cat("---------------------------------------","\n")
   # Return all fit statistics with function call
    fit <- list(coefs=coefs,
                optim.param=fitted.param,
                param=param,
                covar=covar,
                std.err=se,
                z=z,
                llf=llf,
                aic=aic,
                dof=dof,
                k=k,
                residuals=res,
                rawresid=rawres,
                predicted=pred)
    return(fit)
  }


# Restricted log-likelihood where hypothesis is a=b=0
# i.e. no auto-correlation
loglikeexpl_acp_restricted<-function(p,y,x)
  {
  #initialise the parameters
   w<-p[1]; d<-p[2:length(p)]
   y<-as.ts(y); x<-as.ts(as.matrix(x))
   n <- length(y)
   mu<-matrix(0,nrow=n); mu2<-matrix(0,nrow=n)
   logl<-matrix(0,nrow=n)
   #Compute the log-likelihood function
     for (t in 1:n)
     { mu[t] <- w
       mu2[t] <- mu[t]*exp(x[t,]%*%d)
       logl[t] <-(y[t]*log(mu2[t]) - mu2[t] - lgamma(y[t]+1))
     }
    llabw<-sum(logl)
```

```
    llabw
  }

# This funtion maximises the restricted log-likelihood
# where hypothesis is a=b=0 i.e. no auto-correlation
maxloglikeexpl_acp_restricted<-function(y,w.init,d.init,x)
  { d.lower <- matrix(-1,ncol=1,nrow=length(d.init));
    lower.bound <- c(0,d.lower)
    d.upper <- matrix(10,ncol=1,nrow=length(d.init));
    upper.bound <- c(100,d.upper)
    x <- as.matrix(x)
    # Fit the model
    fitted.param<-optim(c(w.init,d.init),loglikeexpl_acp_restricted,
                        method="L-BFGS-B",
                        lower=lower.bound,upper=upper.bound,
                        hessian=TRUE,
                        control=list(trace=2,REPORT=5,fnscale=-(length(y))),
                        y=as.matrix(y),x=as.matrix(x))
    param<-fitted.param$par
    llf <- fitted.param$value
   #Write out the value of the restricted log-likelihood as output
    cat("-------------------------------------","\n")
    cat("Log-likelihood value : ", llf, "\n")
    cat("Parameters(omega,x)  : ", param, "\n")
   #Return only the value of the restricted log-likelihood from function call
    fit <- list(param=param,
                llf=llf)
   return(fit)
  }

# Compute the likelihood ratio between the restricted and unrestricted llfs
# unres.llf = value of unrestricted log-likelihood function
# res.llf = value of restricted log-likelihoods function
computeLR<-function(unres.llf,res.llf)
  { lr <- 2*(unres.llf - res.llf)
    cat("-------------------------------------","\n")
    cat("Log-likelihood ratio test   : ", lr, "\n")
    cat("-------------------------------------","\n")
    return(lr)
  }
```

B.2 DACP model

```
# File "DACP_expl.r" containing R code for fitting Heinen's DACP(1,1) model
# with explanatory variables

# y = observed series
# x = matrix of explanatory variables
# n = number of observations in generated series
# a = value of alpha
# b = value of beta
# w = value of omega
# g = value of gamma
# d = value of delta (parameters associated with explanatory variables)
# p = (a, b, w, d)
# m0 = inital value of poisson mean (mu)
# y0 = initial value of y counts


# Log-likelihood function -
# This function is called from the optimiser in function maxloglikeexpl_dacp.
# It returns the value of the log-likelihood for the given data and parameters.
loglikeexpl_dacp<-function(p,y,x)
  {
  #initialise the parameters
   a<-p[1]; b<-p[2]; w<-p[3]; g<-p[4]; d<-p[5:length(p)]
   y<-as.ts(y); x<-as.ts(as.matrix(x))
   y0<-mean(y)
   m0<-y0
   n <- length(y)
   mu<-matrix(0,nrow=n);
   mu2<-matrix(0,nrow=n)
   logl<-matrix(0,nrow=n)
   #Compute the log-likelihood function
     for (t in 1:n)
     { if (t==1)
        { mu[t] <- w+a*y0+b*m0
          mu2[t] <- mu[t]*exp(x[t,]%*%d)
          if (y[t]==0)
            {logl[t] <-(log(g)*0.5 - lamda*mu2[t] - lgamma(1))}
          else
            {logl[t] <-(log(g)*0.5 - g*mu2[t] + y[t]*(log(y[t])-1) -
                        lgamma(y[t]+1) + g*y[t]*(1+log(mu2[t]/y[t])))}
        }
       else
```

```
        { mu[t] <- w+a*y[t-1]+b*mu[t-1]
          mu2[t] <- mu[t]*exp(x[t,]%*%d)
          if (y[t]==0)
            {logl[t] <-(log(g)*0.5 - g*mu2[t] - lgamma(1))}
          else
            {logl[t] <-(log(g)*0.5 - g*mu2[t] + y[t]*(log(y[t])-1) -
                        lgamma(y[t]+1) + g*y[t]*(1+log(mu2[t]/y[t])))}
        }
     }
    llabw<-sum(logl)
    llabw  #return the loglikeilhood for the given values of a, b, w, g and d
  }

# This funtion maximises the log-likelihood using the "optim" function in R
# a.init = starting value for alpha
# b.init = starting value for beta
# w.init = starting value for omega
# g.init = starting value for gamma
# d.init = vector of starting values for delta
maxloglikeexpl_dacp<-function(y,a.init,b.init,w.init,g.init,d.init,x)
  {
   #Define the lower and upper bounds to use a bounded optimising routine
    d.lower <- matrix(-1,ncol=1,nrow=length(d.init));
    lower.bound <- c(0,0,0,0,d.lower)
    d.upper <- matrix(1,ncol=1,nrow=length(d.init));
    upper.bound <- c(1,1,1,5,d.upper)

    #Define x as a matrix
    x <- as.matrix(x)
    #Define no. of parameters in model
    k <- ncol(x) + 4

    # Fit the model
    fitted.param<-optim(c(a.init,b.init,w.init,g.init,d.init),loglikeexpl_dacp,
                        method="L-BFGS-B",
                        lower=lower.bound,upper=upper.bound,
                        hessian=TRUE,
                        control=list(trace=2,REPORT=5,fnscale=-(length(y))),
                        y=as.matrix(y),x=as.matrix(x))

    # Retrieve the parameters and the log-likelihood function (llf) value
    param<-fitted.param$par
    llf <- fitted.param$value
   # Estimate the standard errors and z scores
```

```
   covar <- -solve(fitted.param$hessian)
   se<-sqrt(diag(covar))
   z<-param/se
   coefs <- t(matrix(rbind(param,se,z),nrow=3))
   colnames(coefs) <- c("Parameters","Std. Errors","Z-score")

  # Get names of parameters
  if (k>5) {x.names <- colnames(x)}
  #If only one explanatory variable then can't retrieve column name
  else {x.names <- "x.variable"}
  rownames(coefs) <- c("Alpha","Beta","Omega","Gamma",x.names)

#Compute the predicted and residual values
   n <- length(y)
   y0<-mean(y)
   m0<-y0
   pnum <- length(param)
   a<-param[1]; b<-param[2]; w<-param[3]; g<-param[4]; d<-param[5:pnum]
   mu<-matrix(0,nrow=n);
   pred<-matrix(0,nrow=n);
   stdres<-matrix(0,nrow=n);
   rawres<-matrix(0,nrow=n);
   for (tt in 1:n)
   {
      {if (tt==1)
         { mu[tt] <- w+a*y0+b*m0 }
       else
          { mu[tt] <- w+a*y[tt-1]+b*mu[tt-1]}
      }
      pred[tt] <- mu[tt]*exp(x[tt,]%*%d)
      stdres[tt] <- (y[tt]-pred[tt])/sqrt(pred[tt]/lam)
      rawres[tt] <- y[tt]-pred[tt]
   }

  # Write out the output
   cat(" ","\n")
   cat("------------------------------------","\n")
   cat("DACP(1,1) regression output","\n")
   cat("------------------------------------","\n")
   printCoefmat(coefs)
   cat("------------------------------------","\n")
   cat("Log-likelihood value  : ", llf, "\n")
   aic <- -2*llf+2*k
   cat("AIC                   : ", aic, "\n")
```

```
    dof <- length(y)-k
    cat("Degrees of Freedom    : ", dof, "\n")
    cat("-------------------------------------","\n")
   # Return all fit statistics with function call
    fit <- list(coefs=coefs,
                optim.param=fitted.param,
                param=param,
                covar=covar,
                std.err=se,
                z=z,
                llf=llf,
                aic=aic,
                dof=dof,
                k=k,
                residuals=stdres,
                rawresid=rawres,
                predicted=pred)
   return(fit)
  }
```

B.3 Poisson-gamma model

```
# File "Poisson_gamma_expl_hf.r" containing R code for fitting the
# Poisson-gamma model with explanatory variables and using Harvey &
# Fernandes method

# y = observed series
# x = matrix of explanatory variables
# n = number of observations in generated series
# a = value of a
# b = value of b
# w = value of omega
# d = value of delta (parameters associated with explanatory variables)

#Kalman filter for the model. Used to compute the filter and
#parameters for the log-likelihood function.
#Calculates for a given w and d.
 kalmanfilterexpl.hf <- function(y,x,w,d)
   {
     #Create matrices a and b for predict step, a_u and b_u for update step
     a<-matrix(0,nrow=length(y));
     b<-a;
```

```
    a_u<-a;
    b_u<-a;
    #Define x as a matrix and time series object
    x<-as.ts(as.matrix(x))
   #Create matrices aw and bw to use later in calulating predictions
    aw<-a;
    bw<-a;

    #Compute the filter parameters
    for (t in 1:length(y))
      { if (t==1)
           { a_u[t]<-y[t];
             b_u[t]<-exp(x[t,]%*%d);
           }
        else
           { a[t]<-a_u[t-1];
             b[t]<-b_u[t-1];
             aw[t]<-w*a[t]
             bw[t]<-w*b[t]*exp(-x[t,]%*%d);
             a_u[t]<-w*a[t]+y[t];
             b_u[t]<-w*b[t]+exp(x[t,]%*%d)
           }
      }

    #Return the filter parameters in a time series object
    kf_data<-matrix(cbind(a, aw, b, bw),ncol=4)
    kf_data<-ts(kf_data,start=1,names=c("a","aw","b","bw"))
    kf_data
  }

#loglikeexpl.hf  - log likelihood function using H&F method.
#Calculates the log-likelihood function for given values of w and d.
#This function is called from the optimiser in function maxloglikeexpl.hf.
#Returns the value for the log-likelihood
loglikeexpl.hf<-function(p,y,x)
  {
  #initialise the parameters
   w<-p[1]
   d<-as.matrix(p[2:length(p)])
   #Define y and x as a matrices
   y<-as.ts(as.matrix(y));
   x<-as.ts(as.matrix(x));

   #Call the filter to get values for a and b
```

```
   kf<-kalmanfilterexpl.hf(y,x,w,d);
   i<-1
   #Count how many rows until the first non-zero "a"
   while (kf[i,1]==0){i<-i+1}
   #Store values of filter parameters,
   #excluding initial zeros for "a" and "b" in first few rows
   kf<-kf[i:nrow(kf),]
   a<-kf[,"a"]; b<-kf[,"b"];
   #Take same rows for x and y
   y<-y[i:nrow(y),]
   x<-x[i:nrow(x),]

   #Calculate the log-likelihood function using the filter parameters
   #and given values for w and d
    llw<-(lgamma((w*a) + y) - lgamma(y+1) - lgamma(w*a) +
         ((w*a)*log(w*b*exp(-x%*%d))) -
         (((w*a) + y)*log(1+(w*b*exp(-x%*%d)))));
    llw<-sum(llw)
    #Return the value for the loglikelihood function
    llw
  }

# This funtion maximises the log-likelihood using the "optim" function in R
# w.init = starting value for omega
# d.init = vector of starting values for delta
maxloglikeexpl.hf<-function(y,w.init,d.init,x)
  {
    #Define the lower and upper bounds to use a bounded optimising routine
    d.lower <- matrix(-1,ncol=1,nrow=length(d.init));
    lower.bound <- c(0.05,d.lower)
    d.upper <- matrix(5,ncol=1,nrow=length(d.init));
    upper.bound <- c(1,d.upper)

    #Define x as a matrix
    x <- as.matrix(x)
    #Define no. of parameters in model
    k <- ncol(x) + 1

    # Fit the model
    fitted.param<-optim(c(w.init,d.init),loglikeexpl.hf,method="L-BFGS-B",
                        lower=lower.bound,upper=upper.bound,
                        hessian=TRUE,
                        control=list(trace=2,REPORT=5,fnscale=-(length(y))),
                        y=as.matrix(y),x=as.matrix(x))
```

```
   # Get the parameters and the log-likelihood function (llf) value
   param<-fitted.param$par
   llf <- fitted.param$value

  # Estimate the standard errors and z scores
   covar <- -solve(fitted.param$hessian)
   se<-sqrt(diag(covar))
   z<-param/se
   coefs <- t(matrix(rbind(param,se,z),nrow=3))
   colnames(coefs) <- c("Parameters","Std. Errors","Z-score")

  # Get names of parameters
  if (k>2) {x.names <- colnames(x)}
  #If only one explanatory variable then can't retrieve column name
  else {x.names <- "x.variable"}
  rownames(coefs) <- c("Omega",x.names)

#Initialise values for computing the prediction and residual values
   n <- length(y)
   y0<-mean(y)
   m0<-y0
   pnum <- length(param)
   w<-param[1]; d<-param[2:pnum]
   pred<-matrix(0,nrow=n);
   vary<-matrix(0,nrow=n)
   stdres<-matrix(0,nrow=n);
   rawres<-matrix(0,nrow=n);

   #Call the filter to get values for a and b
   kf<-kalmanfilterexpl.hf(y,x,w,d);
   i<-1
   #Count how many rows until the first non-zero "a"
   while (kf[i,1]==0){i<-i+1}
   #Store values of filter parameters aa and bb,
  #excluding initial zeros for "aw" and "bw" in first few rows
   aw<-kf[,"aw"]; bw<-kf[,"bw"];
   #Calculated predicted values and Pearson residuals
   for (tt in 1:n)
   {
      {if (tt < i)
         { pred[tt] <- m0
           vary[tt] <- m0}
       else
```

```
          { pred[tt] <- aw[tt]/bw[tt]
            vary[tt] <- aw[tt]*(1+bw[tt])/(bw[tt]^2)}
       }
       stdres[tt] <- (y[tt]-pred[tt])/sqrt(vary[tt])
       rawres[tt] <- y[tt]-pred[tt]
    }

   # Write out the output
    cat(" ","\n")
    cat("--------------------------------------","\n")
    cat("Poisson-gamma regression output - H&F method","\n")
    cat("--------------------------------------","\n")
    printCoefmat(coefs)
    cat("--------------------------------------","\n")
    cat("Log-likelihood value  : ", llf, "\n")
    aic <- -2*llf+2*k
    cat("AIC                   : ", aic, "\n")
    dof <- length(y)-k
    cat("Degrees of Freedom    : ", dof, "\n")
    cat("--------------------------------------","\n")
   # Return all fit statistics with function call
    fit <- list(coefs=coefs,
                optim.param=fitted.param,
                param=param,
                covar=covar,
                std.err=se,
                z=z,
                llf=llf,
                aic=aic,
                dof=dof,
                k=k,
                residuals=stdres,
                rawresid=rawres,
                predicted=pred)
    return(fit)
  }

computeLR<-function(unres.llf,res.llf)
  { lr <- 2*(unres.llf - res.llf)
    cat("--------------------------------------","\n")
    cat("Log-likelihood ratio test   : ", lr, "\n")
    cat("--------------------------------------","\n")
    return(lr)
  }
```

B.4 SAM model

```
# File "SAM_expl.r" containing R code for fitting the SAM model
# with explanatory variables, using ML-EIS estimation

library(MASS)

# y = observed series
# x = matrix of explanatory variables
# tt = number of observations in generated series
# n = number of trajectories
# e = matrix of random N(0,1) variates with tt+1 rows and n columns
# v = value of nu
# g = value of gamma
# d = value of delta (parameters associated with explanatory variables)


#Function to do the EIS regression
eis_regression_expl <- function(y.data,x.data,e,tt,n,g,v,d)
{
  # Generate n trajectories of lambdas and put in matrix
  lmd<-matrix(0,nrow=(tt+1),ncol=n);
    for (t in 2:(tt+1))
    {
     lmd[t,] <- g*lmd[t-1,] + v*e[t,]
    }

  # Remove first row of zeros from the matrix, the lambda0
  lam<-lmd[2:(tt+1),]

  flag <- 0     #set flag if values become unstable

  for (i in 1:20)   #iterations of regression procedure
  {
  #Starting values for regression process
  mean.m <- matrix(0,nrow=tt,ncol=n)  #mean of importance sample m_t
  var.m <- matrix(0,nrow=tt)   #variance of importance sample m_t
  chi <- matrix(1,nrow=tt+1,ncol=n)    #integrating constant
  betat <- matrix(0,nrow=tt)  #matrix for beta values
  alpha <- matrix(0,nrow=tt)  #matrix for alpha values

  #create matrices for weighted regression although not currently using
  ll0<-matrix(0,nrow=tt,ncol=n);  #matrix of Log-likelihoods from Poisson part
  weight<-matrix(1,nrow=tt,ncol=n);
```

```
  #Calc Poisson log-likelihood to use for weighting in regression
#  for (t in 1:tt)
#  {
     #Calc Poisson log likelihood
#    ll0[t,] <- (y.data[t])*t(x.data[t,]%*%d+lam[t,]) -
#                  exp(x.data[t,]%*%d+lam[t,]) - lfactorial(y.data[t])
#  }

#  if (i>1)
#  {
#    if (i<10)
#    {
#      weight <- exp(0.5*ll0)  #Weights used in regression
#    }
#  }


  #Do the regression - looping through time values in reverse order
  #because chi(t+1) used in equation
  for (l in tt:1)  #Use l as time index
    {
  if (flag==0)
    {
     x1 <- matrix(1,ncol=n)*weight[l,]
     x2 <- t(lam[l,])*weight[l,]
     x3 <- t(lam[l,]^2)*weight[l,]

     #chi taken to be 1 at l=tt+1
     y <- (-exp(x.data[l,]%*%d + (lam[l,])) + y.data[l] %*% t(x.data[l,]%*%d +
          (lam[l,])) + log(chi[l+1,]))*weight[l,]

     #Transpose to get x's as columns
     x <- t(rbind(x1,x2,x3))
     B <- ginv(t(x)%*%x)%*%(t(x)%*%t(y))   #Transpose y to get in column

     alpha[l] <- -2*B[3]
     betat[l] <- B[2]

     #Compute variance
     var.m[l] <- v^2/(1 + (v^2)*(alpha[l]))

     if (l==1)
     {
```

```
      #compute mean where lambda0 is zero
      mean.m[l,] <- var.m[l]*(betat[l])
     }
     else
     {
      #compute mean
      mean.m[l,] <- var.m[l]*(betat[l] + lam[l-1,]*g/(v^2))
      #compute chi for l=2:tt
      chi[l,] <- exp((mean.m[l,])^2/(2*var.m[l])-(g*lam[l-1,])^2/(2*v^2))
     }

   #check if chi matrix has missing values -
   #Value of chi goes to inifinty for some starting values
    if (is.na(chi[l,1]))
       { flag <- 1}        #Set flag to 1 if missing values
    }  #end of if statement checking for flag
    }  #end of for loop


  #Generate improved trajectories using mean and variance of m_t
    for (t in 1:(tt))
      {
       lam[t,] <- mean.m[t,] + t(sqrt(var.m[t]))%*%(e[t+1,])
      }
  }

  #Return values for lambda, mean and variance and flag
  out <- list(lam=lam,
              mean.m=mean.m,
              var.m=var.m,
              flag=flag)
  return(out)
}

#loglikeexpl_sam  - approximate log likelihood function for the SAM model.
#Calculates the log-likelihood function for given values of g, v and d.
#This function is called from the optimiser in function maxloglikeexpl_sam.
#Returns the value for the log-likelihood
loglikeexpl_sam <- function(p,y,x,e)
  {
   #initialise the parameters
    g<-p[1]; v<-p[2]; d<-p[3:length(p)]
    tt <- length(y)
    y <- as.matrix(y)
```

```
x <- as.matrix(x)
n <- 10  #No. of trajectories
llp<-matrix(0,nrow=tt,ncol=n);
llt<-matrix(0,nrow=tt,ncol=n);

#Call the EIS regression and retrieve output values
reg<-eis_regression_expl(y,x,e,tt,n,g,v,d);
flag=reg$flag
if (flag==0)
  {
lam=reg$lam
var.m=reg$var.m
mean.m=reg$mean.m

#Compute the approximate log-likelihood function
for (t in 1:(tt))
  {
   llp[t,] <- y[t]*(x[t,]%*%d + lam[t,]) - exp(x[t,]%*%d + lam[t,]) -
              lfactorial(y[t])    #Calc poisson log likelihood

   if (t==1)
   {
    llt[t,] <- llp[t,] + log(sqrt(var.m[t])/v) - 1/(2*v^2)*(lam[t,])^2 +
               1/(2*var.m[t])*(lam[t,]-mean.m[t,])^2
   }
   else
   {
    llt[t,] <- llp[t,] + log(sqrt(var.m[t])/v) -
               1/(2*v^2)*(lam[t,] - g*lam[t-1,])^2 +
               1/(2*var.m[t])*(lam[t,]-mean.m[t,])^2
   }
  }

#The sum of the llts becomes to big for large tt which makes exp value 0.
#Therefore scale down the llt values and adjust again at the end.
scale2 <- mean(llt)
scaled.llt <- llt-scale2
llf <- tt*scale2 + log(mean(exp(colSums(scaled.llt[1:tt,]))))
}
else
#if missing values in regression then give penalty value to log-likelihood
{llf <- -5000}

#Return value of log-likelihood function
```

```
    llf
  }


#Find parameters which maximise the log-likelihood -
#adjust bounds to handle cholera data
maxloglikeexpl_sam<-function(y,x,e,g.init,v.init,d.init)
  {
    #Define the lower and upper bounds to use a bounded optimising routine
    d.lower <- matrix(-0.9,ncol=1,nrow=length(d.init));
    lower.bound <- c(-0.999,0.00001,d.lower)
    d.upper <- matrix(2,ncol=1,nrow=length(d.init));
    upper.bound <- c(0.999,0.999,d.upper)

    #Define y, x and e as matrices
    x <- as.matrix(x)
    y <- as.matrix(y)
    e <- as.matrix(e)
    #Define no. of parameters in model
    k <- ncol(x) + 2

    fitted.param<-optim(c(g.init,v.init,d.init),loglikeexpl_sam,
                        method="L-BFGS-B",
                        lower=lower.bound,upper=upper.bound,
                        hessian=TRUE,
                        control=list(trace=2,REPORT=5,fnscale=-(length(y))),
                        y=as.matrix(y),x=as.matrix(x),e=as.matrix(e))

    # Retrieve the parameters and the log-likelihood function (llf) value
    param<-fitted.param$par
    llf <- fitted.param$value

   # Estimate the standard errors and z scores
    covar <- -solve(fitted.param$hessian)
    se<-sqrt(diag(covar))
    z<-param/se
    coefs <- t(matrix(rbind(param,se,z),nrow=3))
    colnames(coefs) <- c("Parameters","Std. Errors","Z-score")
   # Get names of parameters
   if (k>3) {x.names <- colnames(x)}
   #If only one explanatory variable then can't retrieve column name
   else {x.names <- "x.variable"}
   rownames(coefs) <- c("Gamma","Nu",x.names)
```

```
#Compute the predicted and residual values
   tt <- length(y)
   n <- 10  #No. of trajectories
   pnum <- length(param)
   g<-param[1]; v<-param[2]; d<-param[3:pnum]
   predmat<-matrix(0,nrow=tt,ncol=n);
   pred<-matrix(0,nrow=tt);
   varmat<-matrix(0,nrow=tt,ncol=n);
   vary<-matrix(0,nrow=tt);
   res<-matrix(0,nrow=tt);
   rawres<-matrix(0,nrow=tt);
#Call the eis regression to get values for lambda
   reg<-eis_regression_expl(y,x,e,tt,n,g,v,d);
   lam <- reg$lam
   flag <- reg$flag

  #Calculate predicted mean and variance using delta method
   for (t in 1:tt)
     { if (t==1)
      {predmat[t,] <- exp(x[t,]%*%d)
       varmat[t,] <- exp(x[t,]%*%d) *
                    (1+ exp(x[t,]%*%d)*((exp(lam[t,]))^2*v^2))
      }
      else
      {predmat[t,] <- exp(x[t,]%*%d+g*lam[t-1,])
       varmat[t,] <- exp(x[t,]%*%d) *
                   (exp(g*lam[t-1,])+ exp(x[t,]%*%d)*((exp(lam[t,]))^2*v^2))
      }
     pred[t] <- mean(predmat[t,])
     vary[t] <- mean(varmat[t,])
     res[t] <- (y[t]-pred[t])/sqrt(vary[t])
     rawres[t] <- y[t]-pred[t]
   }

   rmse <- sqrt(mean(rawres^2))
   mse <- mean(rawres^2)
   mae <- mean(abs(rawres))

  # Write out the output
   cat(" ","\n")
   cat("-------------------------------------","\n")
   cat("SAM regression output","\n")
   cat("-------------------------------------","\n")
   printCoefmat(coefs)
```

```
 cat("--------------------------------------","\n")
 cat("Log-likelihood value  : ", llf, "\n")
 aic <- -2*llf+2*k
 cat("AIC                   : ", aic, "\n")
 dof <- length(y)-k
 cat("Degrees of Freedom    : ", dof, "\n")
 cat("--------------------------------------","\n")
# Return all fit statistics with function call
 fit <- list(coefs=coefs,
             optim.param=fitted.param,
             param=param,
             covar=covar,
             std.err=se,
             z=z,
             llf=llf,
             aic=aic,
             dof=dof,
             k=k,
             rmse=rmse,
             mse=mse,
             mae=mae,
             residuals=res,
             rawresid=rawres,
             predicted=pred)
 return(fit)
}
```