NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF FINANCE AND MANAGEMENT SCIENCE

FOR 14 2008

ISSN: 1500-4066

JULY 2008

Discussion paper

Treating missing values in INAR(1) models

BY

JONAS ANDERSSON AND **DIMITRIS KARLIS**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS AND BUSINESS ADMINISTRATION

Treating missing values in INAR(1) models

Jonas Andersson*and Dimitris Karlis†

Abstract

Time series models for count data have found increased interest in recent days. The existing literature refers to the case of data that have been fully observed. In the present paper, methods for estimating the parameters of the first-order integer-valued autoregressive model in the presence of missing data are proposed. The first method maximizes a conditional likelihood constructed via the observed data based on the k-step-ahead conditional distributions to account for the gaps in the data. The second approach is based on an iterative scheme where missing values are imputed in order to update the estimated parameters. The first method is useful when the predictive distributions have simple forms. We derive in full details this approach when the innovations are assumed to follow a finite mixture of Poisson distributions. The second method is applicable when there are not closed form expressions for the conditional likelihood or they are hard to derive. Simulation results and comparisons of the methods are reported. The proposed methods are applied to a data set concerning syndromic surveillance during the Athens 2004 Olympic Games.

Keywords: imputation; Markov Chain EM algorithm; mixed Poisson; discrete valued time series;

1 Introduction

1.1 Motivation

In recent years the need to create and maintain health surveillance systems has been recognized. A definition of health surveillance refers to the systematic collection, collation, analysis and interpretation of health related data and dissemination of information to those who can decide about actions to be taken to prevent and control a disease (usually one of an infectious nature). Among various aspects of health surveillance, "syndromic surveillance" has been considered as an important tool since it is based on symptoms rather than diagnosis and hence it can create alerts faster. For example, syndromic surveillance systems for detection of biologic terrorism after the September 2001 terrorist attacks have been launched in New York city (Das et al. (2003)). In addition, during large athletic events such surveillance systems can be useful to quickly detect threats for the public health, and they have been used in winter Olympic Games in Salt Lake City 2002 and Athens 2004 Olympic Games. (see Dafni et al. (2004))

Syndromic surveillance data refer to the number of incidences of a particular syndrome in a hospital; a sudden increase on the number of incidences of this symptom may be related to an outbreak and measures must be taken as soon as possible. The data usually are collected at the emergency departments (EDs) of major hospitals. This kind of data are by nature low counts data series, in particular if they refer to symptoms that are not so common and we do not observe a large number so as to be able to assume normal approximations. Hence the data must be treated as discrete valued time series. Most of the existing method either ignore the discrete nature of the data or their time series nature. Moreover a common element of such data is the existence of missing values because for example the hospital does not accept patients for some days, either because of maintenance reasons or because each day only some of the hospitals are on duty in a wider area. This creates missing values. Moreover in certain cases the data are themselves irregularly spaced because data are not collected on a daily basis.

The data refer to different hospitals in Athens, for the period from March 2004 until end of September 2004. The period covers the period of the Olympic Games, so we have used data from hospitals not in the main Olympic places so as not to have problems with the changes in the potential population of the

*Department of Finance and Management, Norwegian School of Economics and Business Administration, Helleveien 30, 5045 Bergen, Norway, email : jonas.andersson@nhh.no

†Corresponding author. Department of Statistics, Athens University of Economics and Business, 76 Patission str, Athens 10434, Greece, email : karlis@aueb.gr

areas. We have plotted 4 cases with different proportion of missing values in Figure 1. For the first case (a)the proportion of missing values is 5% while for the other cases the proportion is 28%, 11.5% and 3.3% respectively. One can also see some patterns in the missing values. These data sets are very typical to the problem at hand. The aim is to be able to fit a suitable model in order to allow for correct prediction of future values and thus able to see if the next observations are acceptable (in the sense that they do not indicate a sudden increase of this syndrome) given the historical data or it is too high which perhaps imply the start of an outbreak.

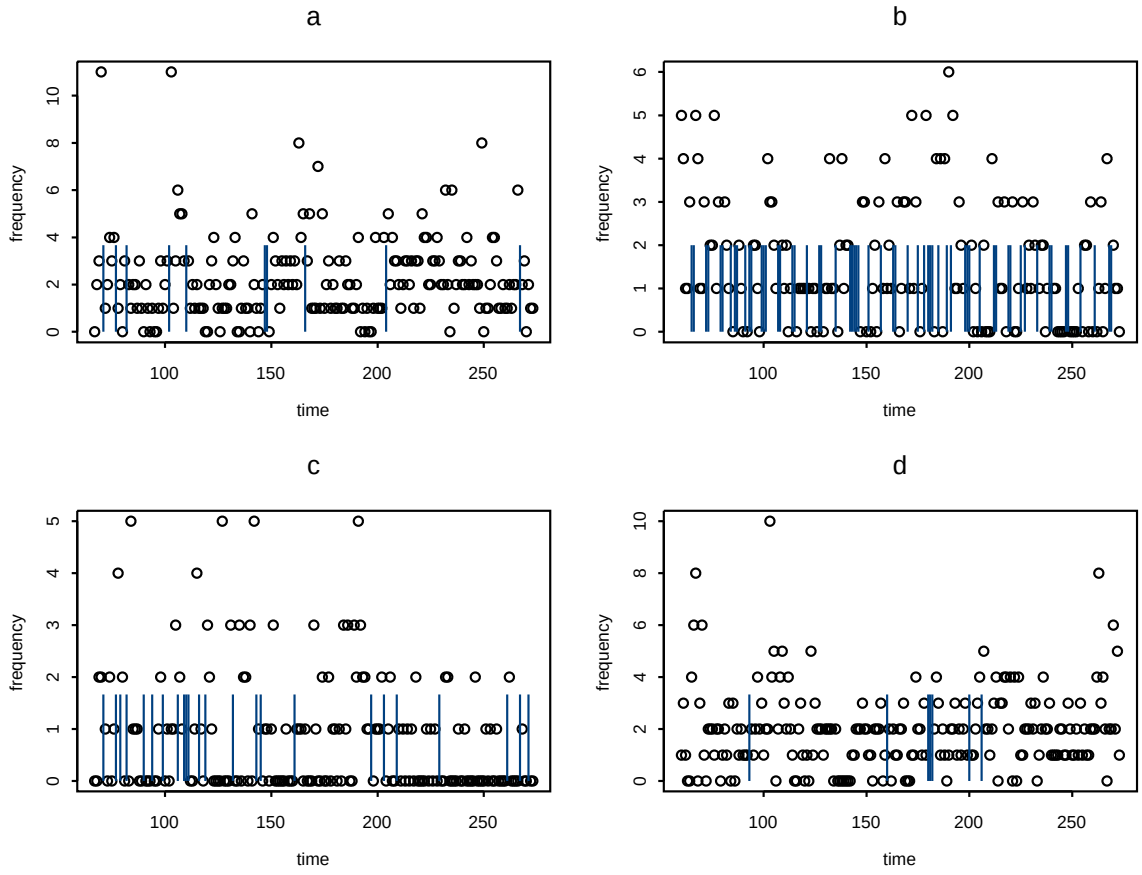

Figure 1: Datasets related to syndromic surveillance. The vertical lines imply missing values. One can see different patterns on the data with respect to the missing data and their proportion.

Note also that integer valued time series occur in several other circumstances where the response is a discrete random variable, like the number of purchases of a particular product, the occurrence of a certain type of crime in a province, the number of cars passing a certain point etc. While issues like estimation for such models have been worked out, the literature lacks of methodologies to handle time series containing missing values.

1.2 Summary of the paper

In this paper we start from the general Integer Autoregressive Model of order 1 (INAR(1) hereafter) proposed by McKenzie (1985) and Al-Osh and Al-Zaid (1987) and we discuss how the missing values in such time series models can be treated so as to allow for estimating the parameters and making inference about the data. We consider the general setting by allowing a variety of different distributions for the innovations so as to be able to model several different time series.

In the present paper we propose two different approaches for dealing with missing values. The first one uses the conditional likelihood created via the conditional distributions of the missing data given the observed ones. This approach is suitable when these distributions can be written in closed form expressions and thus computations are relatively easy. For certain models we derive the conditional distribution based on the results of Freeland and McCabe (2004). For this approach we consider a general model by assuming that the innovations follow a finite mixture of Poisson distribution (see Pavlopoulos and Karlis (2007)) and we fully derive the conditional distributions needed.

The second approach uses an iterative algorithm, where the missing data are generated from the conditional distribution given the observed data and the current parameters (instead of evaluating the distribution we need only to simulate from it, which is quite easier). This second approach is useful when the conditional distributions are cumbersome. Variants of this approach are discussed. A comparison of the two methods is made using the simplest case, i.e. the case based on Poisson innovations.

The remainder of the paper proceeds as follows: Section 2 gives a brief overview of the INAR model and the existing variants of this model. Section 3 describes the predictive distribution if mixed Poisson innovations are considered. In Section 4 the conditional maximum likelihood approach is given while in Section 5 the methods based on imputation are discussed. In Section 6 simulations results are reported to show the efficiency of the methods proposed and a comparison between them. Application of the methods to the syndromic surveillance data can be found in section 7 while concluding remarks and discussion in section 8.

2 The INAR model

2.1 Description of the model

McKenzie (1985) and Al-Osh and Al-Zaid (1987) defined a process for discrete data analogous to the standard autoregressive process for continuous data, called the Integer-valued autoregressive ($INAR$) process as follows:

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

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

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

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

where Y_t are independent Bernoulli random variables with $P(Y_t = 1) = \alpha = 1 - P(Y_t = 0)$, $\alpha \in [0, 1]$ and is called the binomial thinning operator. The binomial thinning operator $\circ$ implies that, conditional on X, $\alpha \circ X$ is a binomial random variable, where X denotes the number of trials and α the probability of success in every trial. The term ϵ_t is referred to as the *innovation term* and must be independent of $\alpha \circ X_{t-1}$ and follows any discrete distribution F.

It is straightforward to show that for $0 \leq \alpha < 1$ the stationary mean and variance of the $INAR(1)$ process $\{X_t\}$ are $E(X_t) = \mu/(1 - \alpha)$ and $Var(X_t) = (\alpha\mu + \sigma^2)/(1 - \alpha^2)$ and that the autocorrelation function is given by $\rho(k) = \alpha^k$, for any non-negative integer k. This implies that the autocorrelation decays exponentially.

Note that according to the choice of the distribution F for the innovations certain models can be constructed. Most of the literature assumes that the innovations follow a Poisson distribution. For non Poisson innovations one can see the papers of Böckenholt (1999); Pavlopoulos and Karlis (2007); McKenzie (1986); Brännäs and Hellstrom (2001); Mc Cabe and Martin (2005)

Al-Osh and Al-Zaid (1987) were concerned with estimation of the two parameters of the INAR(1) model, where innovations follow a Poisson law. Parameter estimation is also treated in Freeland and McCabe (2005); Jung et al. (2005); Böckenholt (1999); Pavlopoulos and Karlis (2007) among others. In all the above papers the data are considered as no having missing values.

Our aim lies at examining the case where the data series contains missing values. More formally, consider time series data X_i, $i \in \mathcal{T} = \{1, 2, \ldots, T\}$. However for some reasons we have not observed all the data points but only the data points X_s, $s \in \mathcal{S} \subseteq \mathcal{T}$, while the points X_r, $r \in \mathcal{R} = \mathcal{T} \setminus \mathcal{S}$ are missing. Let also Θ denote the parameters of interest. We aim at estimating Θ by using all the available information from the observed data.

The first approach is a conditional likelihood approach. Denote as $\mathcal{S} = \{s_1, s_2, \ldots, s_{T_o}\}$, where $s_1 < s_2 \ldots < s_{T_o}$. Then based on the Markovian property of the INAR model the conditional likelihood of the observed data will be of the form

$$CL(\Theta) = \prod_{i=1}^{T_o} Pr(X_{s_i} \mid X_{s_{i-1}}, \Theta)$$

Note that since the distance between s_i and s_{i-1} will not necessarily be equal to 1, i.e. we have successive points, in certain case we need to calculate the p-step ahead predictive distribution, where $p = s_i - s_{i-1}$.

The second approach is based on imputing the missing values and then estimating the parameters from the complete data. Iterating between imputation and updating can provide estimates. The two approaches are described in detail in later sections.

A common element in both methods is the need to find the conditional distribution in a certain time point given the values in some time points earlier. This corresponds to the $p-$step predictive distributions of the model.

2.2 The predictive distribution

Following Mc Cabe and Martin (2005) the p-step ahead predictive distribution for a general INAR model will be of the form

$$P(X_t = x_t \mid X_{t-p} = x_{t-p}, \Theta) = \sum_{s=0}^{min(x_t, x_{t-p})} P(B = s) P(W = x_t - s) \tag{2}$$

where B is a binomial random variable, following a $Bin(\alpha^p, x_{t-p})$ distribution and $W = \sum_{j=0}^{p-1} W_j$ where $W_j = \alpha^j \circ \epsilon_{t-j}$. So, the predictive distribution will be the convolution of a binomial random variable with a discrete random variable defined as the sum of independent random variables as given by W. Note that the exact form of the distribution of W depends on the choice of the distribution F for the innovation terms.

Consider for example the case of Poisson innovations. Since, if $X \sim Poisson(\lambda)$ the unconditional distribution of $\beta \circ X$ is again a $Poisson(\beta\lambda)$ one can see that the distribution of W is the sum of independent Poisson variables and hence also follows a Poisson distribution. This implies that the p-step ahead predictive distribution is the convolution of a $Bin(\alpha^p, x_{t-p})$ variate and a $Poisson(\sum_{j=0}^{p-1} \alpha^j \lambda)$ variate and hence it has a neat and easy to compute formula.

Moving away from the simple Poisson innovations, let us assume negative binomial innovations. This implies that the data are overdispersed relative to the Poisson distributions, an assumption which is realistic for real data. Now, consider the case where $X \sim NBIN(p, r)$, then the unconditional distribution of $\beta \circ X$ is again a $NBIN(\frac{p}{1-(1-\alpha)(1-p)}, r)$ distribution but the convolution of negative binomial random variables is not a negative binomial This implies that the distribution of W becomes quite complicated and computationally demanding to be used in practice.

Similar problems occur when other innovations distributions are considered, like the binomial distribution or the generalized Poisson distribution. For all these cases the predictive distributions are somewhat difficult to be obtained in simple forms and numerical calculations must be employed which perhaps complicates the computations involved.

To disentangle the problem we will consider finite Poisson mixtures innovations as in Pavlopoulos and Karlis (2007) and we will show that in this case the predictive distributions belong to a specific family of distributions that allow for relatively easy computations (see Section 3).

Finally, note that, even if the distribution is complicated to be written in closed form, simulation of random variables from it is relatively straightforward since generation of W is obvious. This will be used later on for the imputation methods.

3 Poisson mixture innovations

3.1 Predictive distribution

Consider the INAR model defined in (1) where the innovations follow a distribution F. Since for the INAR model the innovations ϵ_t, $t = 1, \ldots, T$ are independent and identically distributed the distribution of W is determined as the convolution of the W_j's. For the latter we have that for W_j the unconditional distribution is derived as

$$\begin{aligned} P(W_j = w) &= \sum_y P(W_j = w \mid \epsilon_{t-j} = y) P(\epsilon_{t-j} = y) \\ &= \sum_y P(\alpha^j \circ \epsilon_{t-j} = w \mid \epsilon_{t-j} = y) P(\epsilon_{t-j} = y) \end{aligned}$$

One can see that the conditional distribution of W_j given ϵ_{t-j} is a $Bin(\alpha^j, \epsilon_{t-j})$ distribution. The unconditional distribution depends on the choice made for F.

Consider the case where F is a finite mixture of Poisson distributions, i.e.

$$P(\epsilon_t = y) = \sum_{j=1}^{k} p_j \frac{\exp(-\lambda_j)\lambda_j^y}{y!}, \quad y = 0, 1, \ldots$$

where $\sum p_j = 1$, $0 < p_j \leq 1$ and $0 \leq \lambda_1, < \lambda_2, < \ldots < \lambda_k$. We will denote this distribution as $FP(k, \theta_k)$, where k is the number of components and θ_k the vector of parameters for this mixture, i.e. $\theta_k = (p_1, \ldots, p_{k-1}, \lambda_1, \ldots, \lambda_k)$. Clearly for $k = 1$ we have the simple Poisson distribution.

In order to derive the predictive distribution we need to see what is the distribution of the random variable $\beta \circ \epsilon$ when ϵ follows a $FP(k, \theta_k)$ distribution. It is relatively easy to see that if ϵ follows a simple Poisson distribution with parameter λ then $\beta \circ \epsilon$ follows also a Poisson distribution with parameter $\beta\lambda$.

For the case of a finite mixture of Poisson distributions one can see that

$$\begin{aligned} P(W = w) &= \sum P(W = w \mid \epsilon = y) P(\epsilon = y) \\ &= \sum P(\beta \circ \epsilon = w \mid \epsilon = y) P(\epsilon = y) \\ &= \sum_{j=1}^{k} \sum_{y-w=0}^{\infty} \frac{y!}{w!(y-w)!} \beta^w (1-\beta)^{(y-w)} p_j \frac{\exp(-\lambda_j)\lambda_j^y}{y!} \end{aligned}$$

Changing the order of summation one relatively easy obtains that the distribution of $W = \beta \circ \epsilon$ is a $FP(k, \theta_k^*)$ distribution where $\theta_k^* = (p_1, \ldots, p_{k-1}, \beta\lambda_1, \ldots, \beta\lambda_k)$. i.e. a finite mixture of Poisson distributions with the same number of components, the same mixing proportions but with different mixing parameters.

So, going back, since the ϵ_{t-j}'s all follow a $FP(\cdot)$ distribution all the W_j's will be also FP distributions and hence the density of W will be the sum of independent finite Poisson mixture random variables, each one having, however, a different parameter vector (but the same number of components). Hence the final step is to derive the distribution for the convolution of two random variables each one following a finite Poisson mixture distribution.

Corollary: Suppose that X follows a $FP(k, \theta_k)$ distribution with $\theta_k = (p_1, ..., p_{k-1}, \lambda_1, ..., \lambda_k)$ and Y follows a $FP(m, \theta_m)$ distribution with $\theta_m = (\pi_1, ..., \pi_{m-1}, \mu_1, ..., \mu_m)$ and it is independent of X. Then, $Z = X + Y$ follows a $FP(km, \theta_M)$ distribution where $\theta_M = (\phi_1, ..., \phi_{mk-1}, t_1, ..., t_{mk})$ where the ϕ_i's are the km products $p_i\pi_j$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m$ and the t_i's are the km sums $\lambda_i + \mu_j$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m$.

Proof: The probability generating function of the random variable Z, denoted as $g_z(t)$ is the product of the probability generating functions of X and Y. Thus we have that

$$\begin{aligned} g_z(t) &= \left[\sum_{i=1}^{k} p_i \exp\left[\lambda_i(t-1)\right]\right]\left[\sum_{i=1}^{m} \pi_i \exp\left[\mu_i(t-1)\right]\right] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{m} p_i\pi_j \exp\left[\left(\lambda_i+\mu_j\right)(t-1)\right] \end{aligned}$$

which is the probability generating function of a $FP(mk, (\phi_s, t_s)$ distribution with $\phi_s = p_i\pi_j$ and $t_s = \lambda_i + \mu_j$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m$ and $s = 1, 2, \ldots, mk$.

Remark: Note that for $k = 1$ and $m = 1$ this is the well known additive property of the Poisson variates. We can generalize this result for the sum of more than two rv with each one having a finite Poisson mixture distribution. If we let a Poisson distribution with $\lambda = 0$ to represent a degenerate distribution with all its mass at 0, then the $FP(2, (p_1, 0, \lambda_2))$ distribution is a zero-inflated Poisson distribution.

Using the above results we can see that W_j follows a $FP(k, \theta_k^{(j)})$ distribution where $\theta_k^{(j)} = (p_1, \ldots, p_{k-1}, \alpha^j\lambda_1, \ldots, \alpha^j\lambda_k)$. Hence by the above corollary we can see that $W_0 + W_1$ will follow also a $FP(k^2, \theta_{k^2})$ distribution with $\theta_{k^2} = (\pi_1, \ldots, \pi_{k^2}, \mu_1, \ldots, \mu_{k^2})$ with $\pi_{(i-1)j+i} = p_ip_j, \quad i, j = 1, \ldots, k$ and $\mu_{(i-1)j+i} = \lambda_i + \alpha\lambda_j, \quad i, j = 1, \ldots, k$. As one can see $W_0 + W_1 + W_2$ will have k^3 components and in general $\sum_{j=0}^{p-1} W_j$ will have k^p components. One can find the corresponding parameters by a recursive scheme, but for even moderate k the number of components can be prohibitively large after a few steps.

Finally in order to get the p-step ahead predictive distribution for the case of the finite Poisson mixture innovation distribution, one can see that this will be the convolution of a finite Poisson mixture with a binomial random variable, which results to a mixture of Poisson-Binomials, the order is the same as the order of the mixture for the finite Poisson mixture. This can be seen easily using the following representation for the finite Poisson mixture random variable. Consider random variables Z_j each one following a Poisson distribution with parameter λ_j, $j = 1, \ldots, k$. Consider also multinomial random variable $\Omega = (\Omega_1, \ldots, \Omega_k)$ with corresponding probabilities $(p_1, \ldots, p_k)$. Then $\sum_{j=1}^{k} \Omega_j Z_j$ follows an $FP(k, (p_1, \ldots, p_{k-1}, \lambda_1, \ldots, \lambda_k))$ distribution. Consider now a binomial random variable X. We need the distribution of $X + \sum_{j=1}^{k} \Omega_j Z_j$ or equivalently the distribution for the random variable $\sum_{j=1}^{k} \Omega_j X + \sum_{j=1}^{k} \Omega_j Z_j = \sum_{j=1}^{k} \Omega_j (Z_j + X)$ which implies a finite mixture of Poisson binomial random variables.

3.2 Approximation of high-order Poisson mixtures

An issue that arises when a finite Poisson mixture is assumed for the innovation terms is the fact that the order of the mixture increased rapidly. A solution to this problem could be to approximate a high-order Poisson mixture with one of lower order. We are therefore interested in how well such an approximation works.

Assume that the data are independent draws from a $FP(k+1, (\mathbf{p}, \lambda))$-distribution where $\mathbf{p} = (p_1, ..., p_k, p_{k+1})'$ and $\lambda = (\lambda_1, ..., \lambda_k, \lambda_{k+1})'$. The sums of the elements in $\mathbf{p}$ is one. We will approximate this distribution by a $FP(k, (\tilde{\mathbf{p}}, \tilde{\lambda}))$-distribution where the elements of $\tilde{\mathbf{p}}$ are reweighted in order to sum to one and given by $\tilde{p}_i = p_i/(1 - p_{k+1})$ and $\tilde{\lambda} = (\lambda_1, ..., \lambda_k)'$.

Under the data generating process we define the $(k+1)$-class probability function as

$$\pi_{k+1}(x) = \sum_{i=1}^{k+1} p_i \exp\left(-\lambda_i\right)\frac{\lambda_i^x}{x!}.$$

This will be approximated by the function

$$\tilde{\pi}_k(x) = \sum_{i=1}^{k} \tilde{p}_i \exp(-\lambda_i) \frac{\lambda_i^x}{x!},$$

where the latter can be rewritten as

$$\tilde{\pi}_k(x) = \frac{1}{1-p_{k+1}} \sum_{i=1}^{k} p_i \exp(-\lambda_i) \frac{\lambda_i^x}{x!}.$$

We will now analyse the effect of removing the $k+1$'th class. This will be done if $p_{k+1} < \epsilon$ for some small positive ϵ. In practice we will remove any class having a small probability or a lambda close to a lambda of another class. However, there is no loss of generality by assuming the order of the states, as is done here.

We can formulate the approximation error as

$$\begin{aligned} e(x) &= \tilde{\pi}_k(x) - \pi_{k+1}(x) \\ &= \frac{p_{k+1}}{1-p_{k+1}} \sum_{i=1}^{k} p_i \exp(-\lambda_i) \tfrac{\lambda_i^x}{x!} - p_{k+1} \exp(-\lambda_{k+1}) \tfrac{\lambda_{k+1}^x}{x!} \end{aligned} \tag{3}$$

By considering the two summands of (3) separately, we see that

$$e(x) \leq \frac{p_{k+1}}{1-p_{k+1}} \sum_{i=1}^{k} p_i \exp(-\lambda_i) \frac{\lambda_i^x}{x!} \leq \frac{p_{k+1}}{1-p_{k+1}} \tag{4}$$

and

$$e(x) \geq -p_{k+1} \exp(-\lambda_{k+1}) \frac{\lambda_{k+1}^x}{x!} \geq -p_{k+1} \tag{5}$$

The inequalities (4) and (5) together implies

$$|e(x)| \leq \max\left(p_{k+1}, \frac{p_{k+1}}{1-p_{k+1}}\right) = \frac{p_{k+1}}{1-p_{k+1}} \tag{6}$$

A numerical example of is that if we remove a class with probability less than 0.005 ($\epsilon < 0.005$) an upper bound of the absolute error we can get in calculating a probability $\pi_{k+1}(x) = P(X = x)$ is 0.005025.

4 Conditional Maximum Likelihood Estimation

The literature on linear models of continuously valued time series with missing values is very mature. An extensive review of this literature here would therefore be beyond the scope of this paper. Consider the simple AR(1) model of the form $x_t = \phi x_{t-1} + \epsilon_t$ were $\{\epsilon_t\}$ now is a normally distributed white noise process with standard deviation σ. One way to handle missing variables in this case is to use the conditional distributions

$$(x_t | \mathcal{F}_{t-p}) \sim N\left(\phi^p x_{t-p}, \sigma^2 \frac{1-\phi^{2p}}{1-\phi^2}\right) \tag{7}$$

where $\mathcal{F}_{t-p} = \{\text{all observations until time } t-p\}$ to calculate the conditional log-likelihood. This conditional likelihood can then be maximized numerically.

The conditional distribution is not always as easy to express as in the simple case given above. However, for other linear models, the problem of missing data can be solved by formulating the model of interest in terms of a state space model and apply the Kalman filter to obtain the likelihood (see e.g. Hamilton, 1994; Harvey et al.). The special, but nevertheless rather including, case of vector autoregressive- moving average (VARMA) processes is treated extensively in Luceno (1989).

Analogously one may treat the case of the INAR model. For the cases where we have an explicit expression of the conditional probabilities for X_{s_i} given $X_{s_{i-1}}$ we also have, given the Markovian nature of the process, an explicit expression for the conditional log-likelihood function of the model

$$\log CL(\Theta) = \sum_{i=1}^{T_o} \log Pr(X_{s_i} \mid X_{s_{i-1}}, \Theta). \tag{8}$$

Each conditional probability is calculated under the p-step ahead predictive distribution with $p = s_i - s_{i-1}$. Equation (8) is then maximized numerically.

5 Imputation

5.1 The general idea

Imputation in the time series context is not so simple since we need to find a way to impute the values without destroying the underlying autocorrelation structure. Consider, for example, the simple case where the missing values are imputed by merely considering the conditional distribution, conditioned on the previously observed data point. To help the illustration consider a series where X_t and X_{t+3} are observed but the values X_{t+1} and X_{t+2} need to be imputed. A straightforward imputation would be based on the previously observed value and, hence, to generate the data by simulating X_{t+1} conditional on X_t and X_{t+2} conditional on the imputed X_{t+1}. While this would reproduce a large part of the autocorrelation and the marginal properties of the series, the imputed values will not have any connection with X_{t+3} destroying the structure at this point.

The key to overcome this problem is to impute the missing values by taking into account not only the information at the previous observed point but also the information of the next observed point. To do so we propose two alternative methods.

The first method imputes data based only on the conditional distribution where the conditioning set consists of previous observations. This is repeated a number of times and the imputation that provides the larger log-likelihood using the current values of the parameters is kept. In some sense we keep the imputed values that given the current parameters are more probable.

The second method imputes the values from the conditional distribution where the conditioning set consists of values both before and after the missing value. A value for the next observed value is also generated and we keep the imputed value only if it coincides with the observed one. This implies that we have bridged the observed values using the imputed ones.

5.2 Imputation based on the likelihood

The steps for this approach are the following:

Step A: Impute data X_r, $r \in \mathcal{R}$ by using the conditional distribution $Pr(X_r \mid X_{s_r}, \Theta)$ where $s_r \in \mathcal{S}$ such as $s_r < r$ and $s_r - r$ is minimum.

Step B: Combining the observed and the imputed data, consider the entire series of observations. For this new series and using the current values of the parameters we calculate the conditional log-likelihood of the series as the one when all the data were observed. Denote this value as L_i.

Step C: Repeat steps A and B, a number of times, say M. This will result in M values of the log-likelihood, $L_1, \ldots, L_M$. L_i indicates how possible is the imputation given the parameter values.

Step D: Select the imputation that provides the best likelihood.

Step E: Using the selected series maximize the conditional likelihood with respect to the parameters and update their values.

Step F: If a stopping criterion is satisfied then stop otherwise go back to step A.

The key idea behind this is that we expect that at least one of the many simulated datasets will resemble the true one or at least be close, the closeness is measured by the log-likelihood. At least for M large this method will succeed in generating data close to reality. However large M can be computationally demanding.

5.3 Bridge imputation

This imputation tries to generate the missing data taking into consideration all the available data. The missing values are generated by conditioning on the previous values but in order not to destroy the structure we force the missing data to tie with the observed one by simulating also the first available point after the missing datapoint(s) and rejecting all the paths where this point differs from the previous one.

So this imputation proceeds as follows. Consider that we have observed points X_t and X_{t+k}, for some $k > 1$ but all the points between are missing. The algorithm proceeds by simulating X_{t+r+1} conditional on the previous point X_{t+r} for $r = 0, \ldots, k-1$ and then if the simulated value $\tilde{X}_{t+k}$ does not coincide with the true value to discard this path and regenerate a new one until the condition is satisfied. This procedure is used to fill all the missing points. In Figure 2 one can see the approach. While several paths were generated, we kept only the one that bridged the two edges. For simulating the missing points the current values of the parameters are used. Once an entire series have been simulated, then the INAR model is fit in order to obtain new parameters.

In practice a few iterations suffice for the parameters to reach the area where the likelihood is maximized, however, due to the Monte Carlo nature then they fluctuate. This method can be problematic

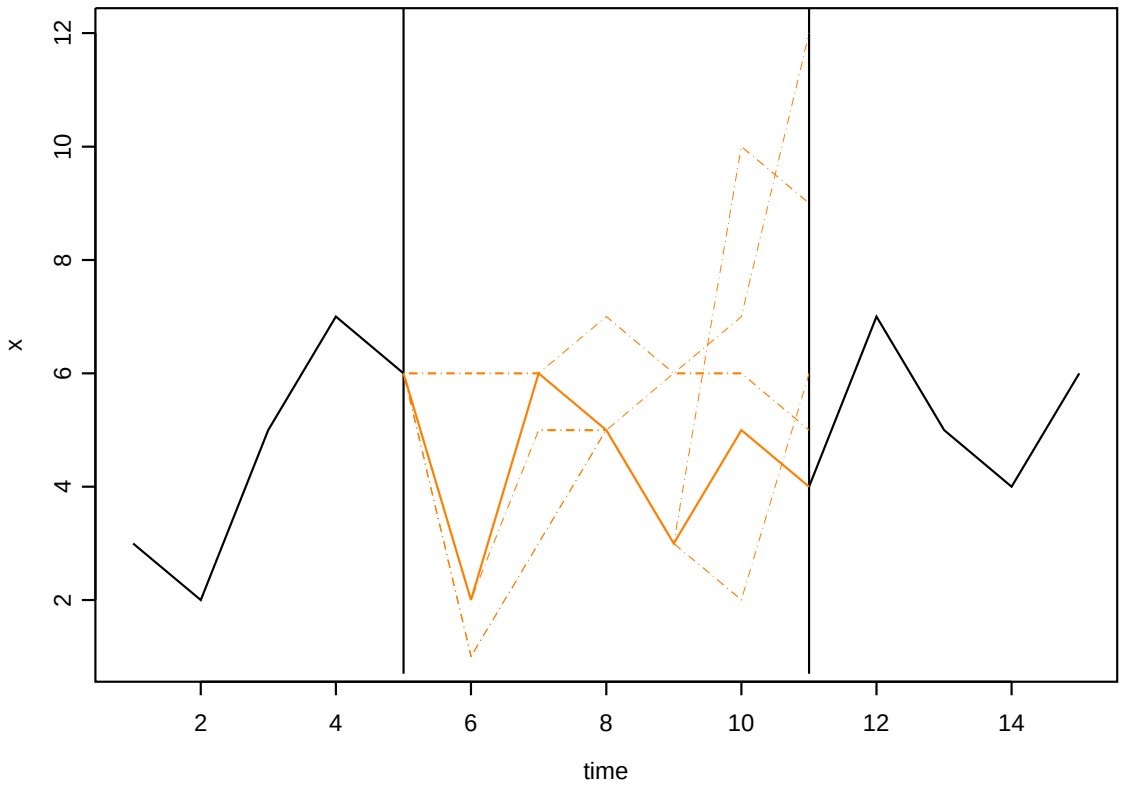

Figure 2: Representation of the bridge method. Some imputed series (dotted lines) and the one that is accepted (solid line). the missing values are those between the vertical lines.

if the assumed model is purely specified as the generated values can be far away from the observed ones and hence one may need a lot of attempts in order to accept a trajectory. However such problems can be solved by considering reasonable starting values and/or the assumed model is not very contradicting

with the observed data.

Having discussed imputation method we end up this section by considering a combination of the conditional likelihood and the imputations approaches. This algorithm can be considered as an MCEM algorithm

5.4 A Monte Carlo EM-type algorithm

The idea is to create the complete data set, i.e. to augment the observed data with the missing data by imputing the missing data from the appropriate conditional distributions. To do this we use the bridge imputation described above. Then the expectation of the complete data log-likelihood is approximated using Monte Carlo at the E-step while the M-step involves maximization of the expected complete log-likelihood. A description of the algorithm follows:

Let $T = s_{T_o}$.

1. Choose some starting values Θ_0 for the parameters Θ in the model.
2. Generate $x_1^g, x_2^g, ..., x_T^g$ where x_t^g is equal to observed values if available and through the following simulation procedure when not. If $s_i - s_j = p$ we generate data from

$$x_t^g = \alpha \circ x_{t-1}^g + \epsilon_t, \quad t = s_{i-1}, \ldots, s_i$$

   with parameter values obtained from the last iteration and conditional on $x_{s_{i-1}}$. If $x_t^g \neq x_{s_i}$ we discard it and try with a new trajectory. This is done until we obtain $x_t^g = x_{s_i}$.
3. Step 2 is repeated M times. We now have M trajectories consisting of a combination of observed and generated data.
4. For each trajectory, we calculate $\log CL_i = \log CL(\Theta|\text{trajectory } i)$, $i = 1, 2, ..., M$.
5. Monte Carlo E-step: Approximate the expected conditional log-likelihood for the parameter value by

$$\widehat{\log CL}(\Theta) = \frac{1}{M} \sum_{i=1}^{M} \log CL_i$$

6. M-step: Maximize (the approximation of) the expected conditional log-likelihood to obtain a new Θ.
7. Repeat steps 2 to 6 with the parameter values obtained in the last iteration until some convergence criteria is reached.

Looking carefully on the procedure above one sees that, for large M, the behavior will be that of an EM-type algorithm.

When the proportion of missing data is small this algorithm works well, but if the missingness increases it can be slow.

6 Simulation and Comparisons

A simulation experiment was conducted in order to show the performance of the proposed methodology. We worked by assuming Poisson innovations as this implies the simple Poisson INAR model and hence a formal comparison of several methods is applicable. Data of two different sample sizes, namely $n = 100, 500$ were generated. For all cases parameter λ of the Poisson innovations was set equal to 3 while the autocorrelation parameter α was set equal to 0.1 and 0.5 respectively to indicate small and strong autocorrelation respectively. Time series were generated by mimicking the INAR process. Then, a number of random data points were considered as missing. We used two different scenarios of missingness. Namely the first one assumes that only 5% of the data are missing while the second that 30% of the data points are missing, implying large proportion of missing values. All computations were run in `R`, (R Development Core Team, 2005)

For each data set we used 6 different methods to estimate the parameters. The first two just ignore the missing data either by creating a time series with only the observed data and fitting the model (Ign) or by using for the calculations only the available pairs (Ign2). Such methods while very simple in nature ignore either part of the autocorrelation structure as the former or decrease the sample size the latter and hence they are suspect to produce estimates with bias and/or large variances.

The third method is the conditional ML (CML) method described in section 4. Depending on the gaps between the observations the adequate conditional distribution was employed. The rest methods are in fact imputation methods described in section 5. The interesting feature is to examine the effect of using one imputation as in the bridge method (BR), or many as in the two other methods namely the method of section 5.2 (denoted as LI) and the MCEM approach (denoted as SML).

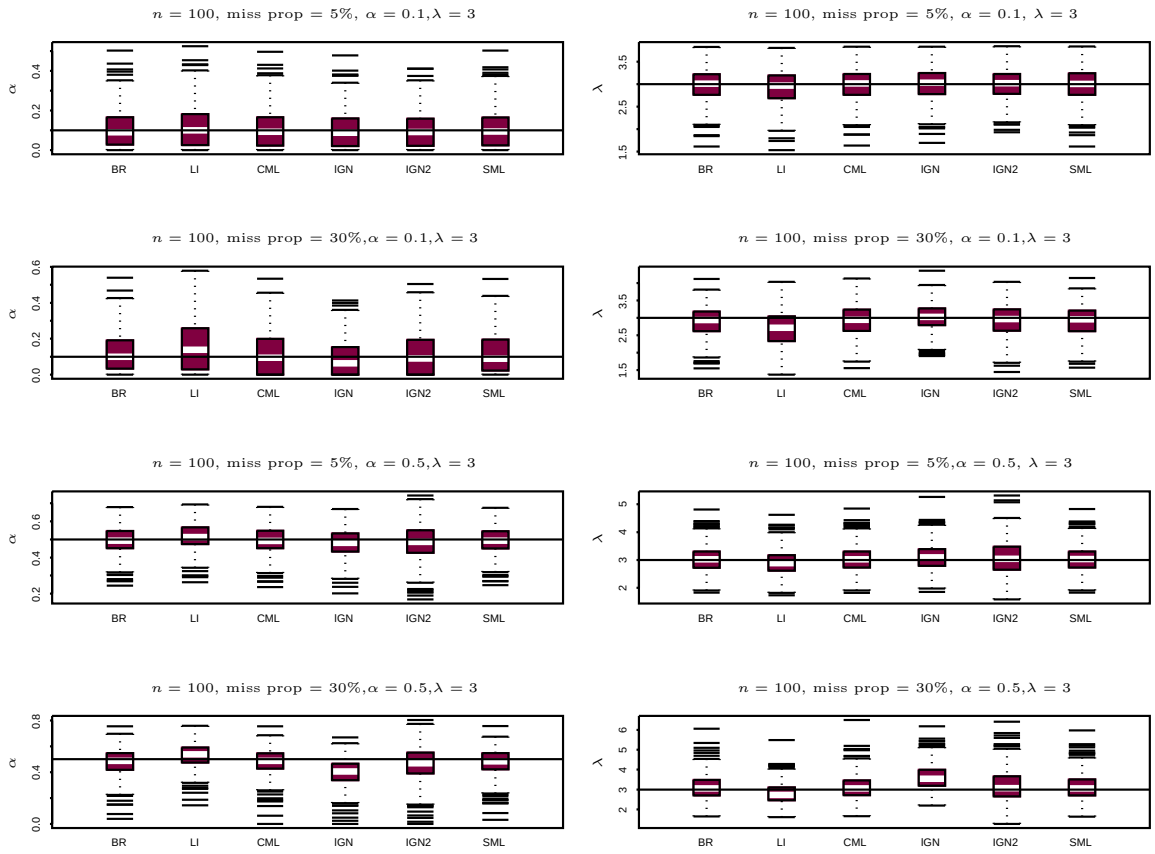

Figure 3: Boxplot with results from the simulation, $n = 100$

Figures 3 and 4 show the boxplots based on 500 replications for all the methods and the configurations used. For the iterative methods we have used 50 iterations and report as estimate the average of the last 20 iterations. From figures, one can see some interesting findings. First of all the imputation methods, especially BR, perform quite well producing good estimates with relatively small variance comparable to the CML method. This implies that in cases where the CML method is not computationally feasible the method can be a standard approach for estimation. On the other hand the methods that ignore observations lead to either bias or large variance as they ignore the structure and decrease the number of observations respectively. The LI method is the worst imputation method, however its performance can be improved by simulating more series at the cost of increasing the computing time.

Concluding the simulations verify the good properties of the CML method while in cases where this

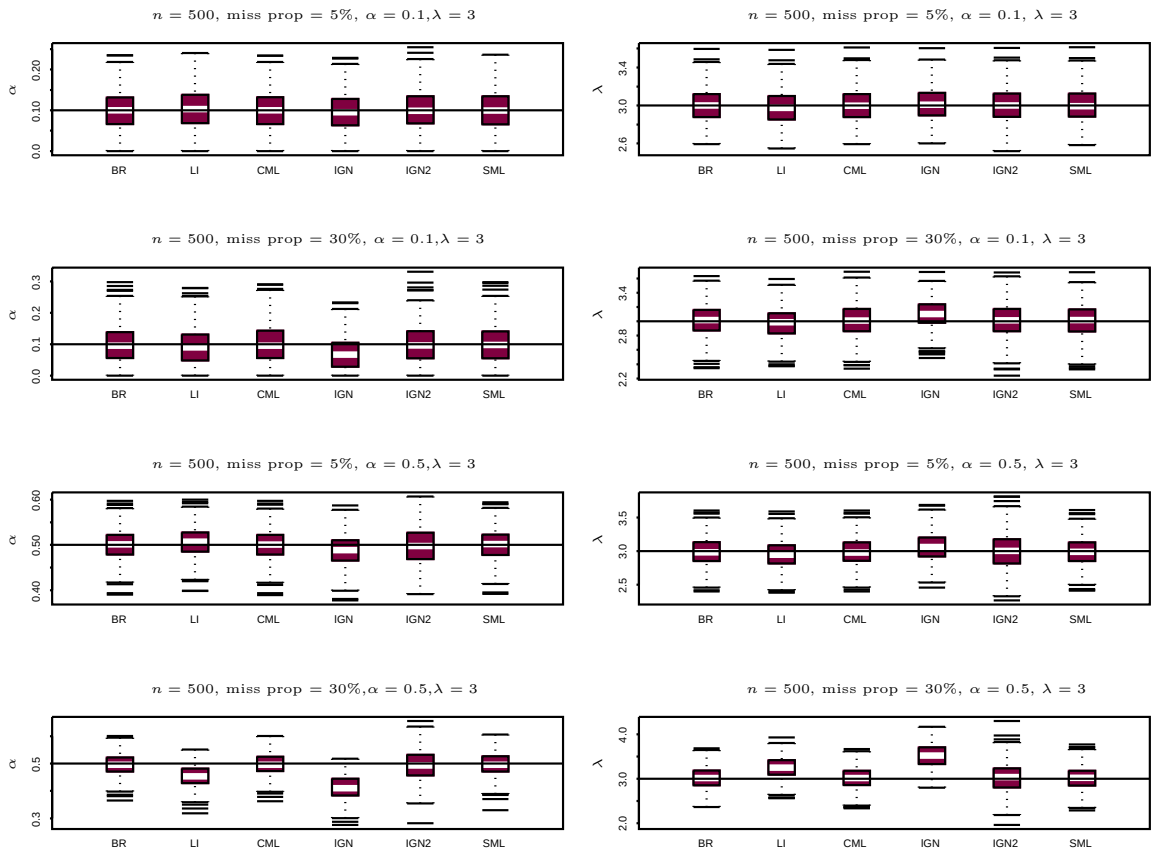

Figure 4: Boxplot with results from the simulation, $n = 500$

method is computationally problematic, the imputation methods can be used.

7 Syndromic data from Athens 2004 Olympic Games

We will use the data from the syndromic surveillance of Athens Olympic Games in 2004. Plot of the data can be seen in Figure 1. We focus on the series (a). The syndrome considered is "febrile illness with rash". Since the data show some overdispersion (variance = 3.194, mean= 2.222, variance to the mean ratio is equal to 1.43) the simple Poisson INAR case is not adequate. The sample autocorrelation at lag 1 is 0.10 but this underestimates the true one since there are some missing values that destroy the underlying structure. We have fitted two models, the INAR(1) with negative binomial innovation term and one with finite mixture of Poisson distributions as innovation distribution. Both models allow for overdispersion. For the negative binomial innovations we use the parametrization were λ denotes the mean of the distribution and n^* the size parameter.

For the negative binomial model the CML is not applicable as the predictive distribution is cumbersome. The form of the distribution can be computed only recursively, which implies that it is time consuming and any errors are accumulated. This may lead to problems during the optimization. For this reason we used the imputation methods described. The conditional distribution is the convolution of a binomial and a negative binomial distributions and hence simulation is relatively straightforward. These computations were also run in `R`.

Results from fitting this model to the data are reported at Table 1 based on different methods. The

Negative Binomial Innovation			
Method	α	λ	n^*
IGN	0.1389	1.9247	5.075
IGN2	0.1594	1.854	6.512
Bridge	0.1733	1.8640	4.7216
Lik, $M = 100$	0.1707	.18461	4.9810
$M = 500$	0.1704	1.8447	5.0059
$M = 1000$	0.1710	1.8421	4.9886
Mixed Poisson innovation			
	α	mix. prop.	mix. parameters
CML	0.16751	0.9580	$\lambda_1 = 1.6557$
			$\lambda_2 = 6.9271$

Table 1: Results from fitting a model with negative binomial innovations based on different methods

first two correspond to the cases where the missing values were ignored. As one can see the autocorrelation parameter is underestimated. The imputation methods provide larger estimates of the autocorrelation and give quite similar results. Moreover the LI method when increasing the number of imputed series approaches better the bridge method as expected. We have also fitted mixed Poisson innovations. As mentioned above CML is applicable and thus we report such estimates in the second part of Table 1. The conditional likelihood stopped increasing for $k > 2$ components. Hence two components suffice to represent the data. The conditional likelihood was -353.1975, while its value for $k = 1$, i.e. simple Poisson innovations, was -363.6542.

Figure 5 shows the history of the 500 iterations used for the bridge imputation method (negative binomial innovations assumed). Each plot corresponds to one parameter estimate. It is interesting that the algorithm converged very quickly, 5 iterations were sufficient to approach the estimate and then it just fluctuated, as expected, around this value. The plot also depict the ergodic mean (discarding the first 100 iterations) and the horizontal line refers to the reported estimate based on the mean of the last 500 iterations. Note that for such stochastic algorithms other functionals could have been reported instead of the mean.

Figure 6 shows the predictive one step ahead distribution $P(X_{t+1} = y \mid X_t = x)$ based on both models, the one with negative binomial innovations and the one with mixed Poisson innovations with 2 components with parameters as estimated from the data. One can see that there are some differences. The mixed Poisson leads to larger tails than the negative binomial one. We have selected certain values x of the previous observation. Such predictive distributions can be used to recognize an unexpectedly large observation and for surveillance reason to alarm so as to examine the situation more thoroughly.

One of the issues arising is which of the two models we must select. On a theoretical ground, since the mixed Poisson model with $k = 2$ gives the largest possible likelihood over all models of mixed Poisson family we know that this is the best possible model in likelihood terms. On the other hand such model lacks parsimony as it has a relatively large number of parameters, especially if k is large. So a trade off is needed. The problems becomes more complicated since we do not have a likelihood score available from each model so as to be able to penalize it and decide on these grounds. To overcome the problem we followed a simulation based procedure, where the log-likelihood of the complete data was estimated based on 500 replications from each model, where missing values were imputed based on the conditional distributions as described in a previous section. Such a procedure led to an estimate of the complete data log-likelihood and using these estimates we can have an objective comparison of the two models penalizing for the parameters as well. The complete log-likelihood of the mixed Poisson model is larger. Taking the average over the 500 replications we found -387.5171 for the negative binomial model and -382.1978 for the mixed Poisson. The difference is statistically significant. Penalizing for the difference in the number of parameters the mixed Poisson model is statistically better using either the AIC or the BIC criteria and taking into account the Monte Carlo error due to simulation approach. This leads to the conclusion that for this dataset the mixed Poisson innovations are preferable.

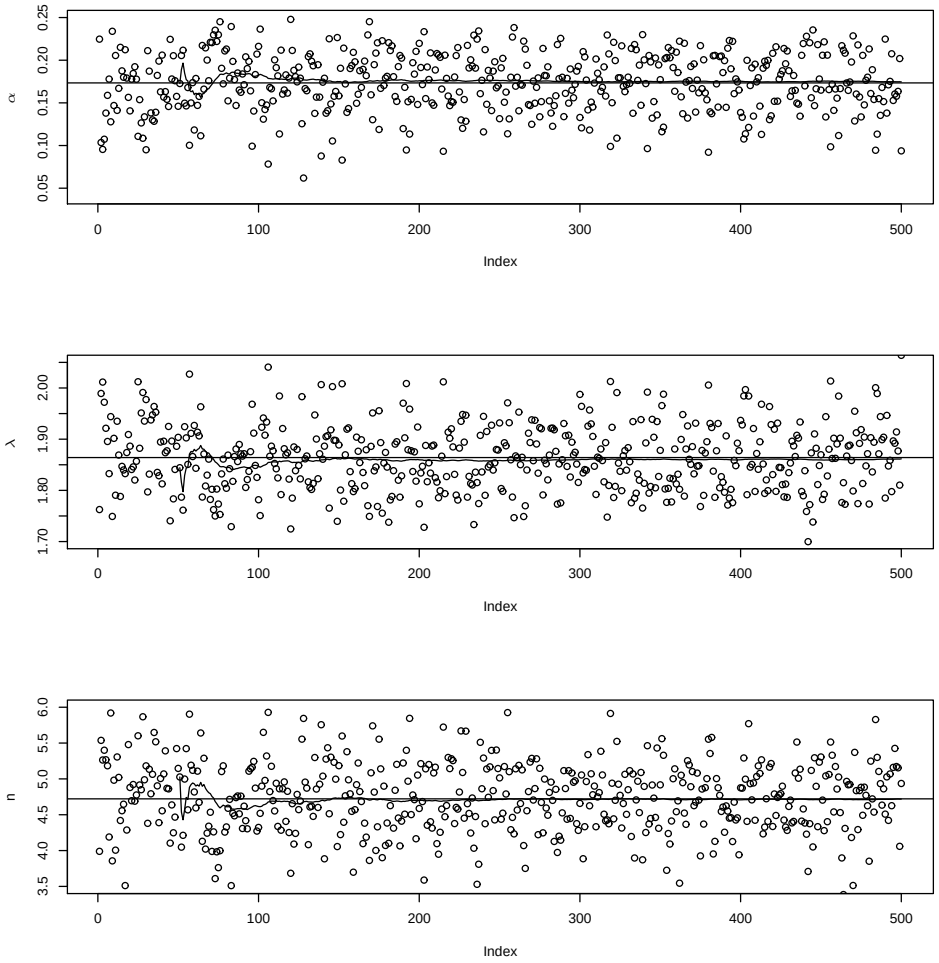

Figure 5: The estimates for the 500 iterations of the algorithm using the bridge imputation. The ergodic mean is depicted after the first 100 iterations. the horizontal line refers to the reported estimate, namely the last 250 iterations.

8 Conclusions

In the present paper we examine the problem of making statistical inference for time series of discrete valued data that contain missing data. Missing data can destroy the underlying structure to a certain extent and hence, by simply removing or ignoring them, we may end up with inaccurate estimates and thus distort the inferential procedures. To disentangle the problem we propose two approaches. The first one is based on the conditional likelihood where each observation is conditioned on the previously observed one. This approach, can be very time consuming since the predictive p step ahead distributions are convolutions of discrete random variables and in general recursive algorithms are needed. We saw that for a broad class of models, that of finite Poisson mixture distributions for the innovations, the conditional likelihood is tractable. The second approach is based on an iterative algorithm where the missing data are imputed and then we use the complete data to update our estimates. The performance of both methods are quite satisfactorily based on the simulation experiments presented.

There are some more points that we did not pursue in this paper. First of all, the main focus in the paper is on the INAR(1) model, but extensions to general INAR(p) can be obtained. The conditional distributions are much more cumbersome but the imputation approach based on simulation is easily

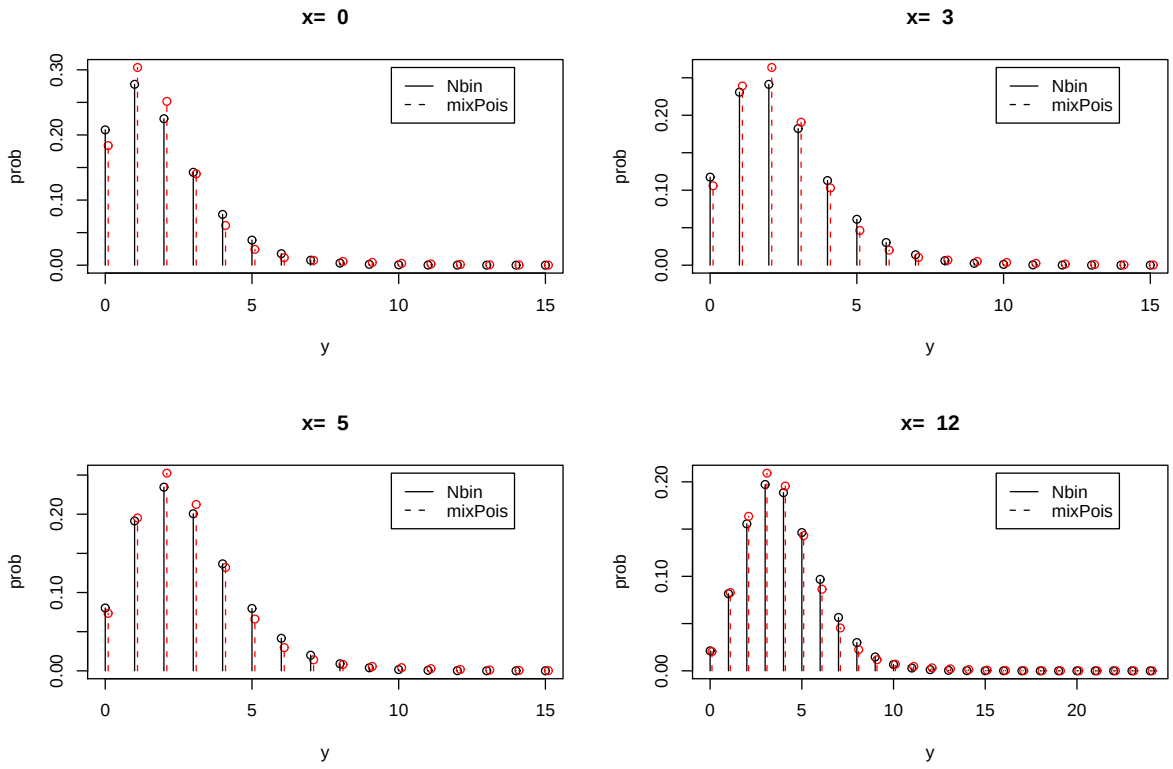

Figure 6: The predictive distributions $P(X_{t+1} = y \mid X_t = x)$ conditional on certain values x of the previous observation. The dotted lines correspond to the model with mixed Poisson innovations and the solid lines to the model with negative binomial innovations. The value of the parameters used are those estimated by the two models respectively. One can see that the mixed Poisson model results to larger tails

extendable and applicable as it involves simply simulation from standard discrete distributions.

Another interesting point relates to the application discussed in the paper. Since the interest relies mostly on being able to promptly recognize an unexpectedly large value, it would be better to separate the data in two parts. The model is fitted to the first part and it is used for the second part to recognize suspect values. On the other case, when all the data are used to fit the model, a problem of over-fitting may occur, especially while in finite mixture models outliers tend to create themselves a component in the mixture. So, we advice careful use of different data sets to obtain the estimates and apply the model for future prediction.

Acknowledgements

The authors would like to thank the Hellenic Center for Infectious Diseases Control and Prof. Urania Dafni for making the data used in the paper available to us.

References

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

U. Böckenholt. Mixed INAR(1) Poisson regression models: Analyzing heterogeneity and serial dependencies in longitudinal count data. *Journal of Econometrics*, 89:317–338, 1999.

K. Brännäs and J. Hellstrom. Generalized integer valued autoregression. *Econometric Reviews*, 20: 425–443, 2001.

U.G. Dafni, S. Tsiodras, D. Panagiotakos, K. Gkolfinopoulou, G. Kouvatseas, Z. Tsourti, and G. Saroglou. Algorithm for statistical detection of peaks — syndromic surveillance system for the Athens 2004 Olympic Games. *Morbidity and Mortality Weekly Report*, 53 (supplement):86–94, 2004.

D. Das, D. Weiss, and F. et al. Mostashari. Enhanced drop-in syndromic surveillance in new york city following September 11, 2001. *Journal of Urban Health*, 80 (supplement 1):i76–i88, 2003.

R.K. Freeland and B.P.M. McCabe. Forecasting discrete valued low count time series. 20:427–434, 2004.

R.K. Freeland and B.P.M. McCabe. Asymptotic properties of CLS estimators in the Poisson AR(1) model. *Statistics and Probability Letters*, 73:147–153, 2005.

J.D Hamilton. *Time Series Analysis*. Princeton, 1994.

A. Harvey, Koopman S.J., and J. Penzer. *Advances in Econometrics*, volume 13, chapter Messy time series: A unified approach, pages 103–143.

R.C. Jung, G. Ronning, and A.R. Tremayne. Estimation in conditional first order autoregression with discrete support. *Statistical Papers*, 46:195–224, 2005.

A. Luceno. Estimation of missing values in possibly partially nonstationary vector time series. *Biometrika*, 84:495–499, 1989.

B.M.P. Mc Cabe and G.M. Martin. Bayesian prediction of low count time series. *International Journal of Forecasting*, 21:315–330, 2005.

E. McKenzie. Some simple models for discrete variable time series. *Water Resources Bulletin*, 21:645–650, 1985.

E. McKenzie. Autoregressive moving-average processes with negative binomial and geometric marginal distributions. *Advances in Applied Probability*, 18:679–695, 1986.

H. Pavlopoulos and D. Karlis. INAR(1) modelling of overdispersed count series with an environmental application. *Environmetrics*, forthcoming, 2007.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL `http://www.R-project.org`. ISBN 3-900051-07-0.