Model Selection, Estimation and Forecasting in INAR(p) Models: A Likelihood Based Markov Chain Approach

Ruijun Bu

University of Liverpool, UK

Brendan McCabe*
University of Liverpool, UK

December 4, 2007

*Corresponding author: Management School, Chatham Street, Liverpool, L69 7ZH, UK, Tel: +44-151-795-3705, Fax: +44-151-795-3004, Email: Brendan.Mccabe@liv.ac.uk (Brendan McCabe)

Model Selection, Estimation and Forecasting in INAR(p) Models: A Likelihood Based Markov Chain Approach

Abstract

This paper considers model selection, estimation and forecasting for a class of integer autoregressive models suitable for use when analysing time series count data. Any number of lags may be entertained and estimation may be performed by likelihood methods. Model selection is enhanced by the use of new residual processes that are defined for each of the $p + 1$ unobserved components of the model. Forecasts are produced by treating the model as a Markov Chain and estimation error is accounted for by providing confidence intervals for the probabilities of each member of the support of the count data variable. Confidence intervals are also available for more complicated event forecasts such as functions of the cumulative distribution function e.g. for probabilities that the future count will exceed a given threshold. A data set of Australian counts on medical injuries is analysed in detail.

Keywords: Time Series of Counts; INAR(p) models; Maximum Likelihood Estimation; Markov Chain; Transition Probability; Transition Matrix; Delta Method;

1. Introduction

One of the objectives of modelling time series data is to forecast future values of the variables of interest. The most common procedure for constructing forecasts in time series models is to use conditional expectations as this technique will yield forecasts with minimum mean squared forecast error. However, this method will invariably produce non-integer-valued forecasts, which are thus deemed to lack data coherency in the context of count data models. This paper presents a method of coherent forecasting for count data time series based on the integer autoregressive, $INAR(p)$, class of models. Integer autoregressive models were introduced by Al-Osh and Alzaid (1987) and McKenzie (1988) for models with 1 lag. Both Alzaid and Al-Osh (1990) and Du and Li (1991) considered the $INAR(p)$ class but with differing specifications of the thinning operators. In this paper we use the conditionally independent thinning scheme of Du and Li (1991). Freeland and McCabe (2004b) suggest using the h-step ahead conditional distribution and its median to generate data coherent predictions in the $INAR(1)$ case. They also suggest that the probabilities associated with each point mass be modified to reflect the variation in parameter estimation. McCabe and Martin (2005) explored the issue of coherent forecasting with count data models under the Bayesian framework but they too are only concerned with first-order case. More recently, Jung and Tremayne (2006) proposed a simulation based method for producing coherent forecasts for higher-order $INAR$ models but this too requires considerable computational work and does not use likelihood methods.

The paper makes three contributions. First, we suggest that the model be estimated by Maximum Likelihood (ML) should distributional assumptions warrant it[1]. We may therefore take advantage of the well known asymptotic normality and efficiency properties of the ML method. ML is not difficult computationally and allows for a richer set of tools for model selection and improvement than do other methods of estimation for this class of models. For example, consider testing whether a thinning component should be excluded from the model i.e. testing if the associated parameter $\alpha_k = 0$. Since α_k is a probability, methods of estimation require that $\hat{\alpha}_k$ be restricted to $[0, 1)$ and so tests based on $\hat{\alpha}_k$ will have a non-standard distribution because of the truncation at the boundary point 0. This truncation is not an issue for score based tests in the ML framework. Other tech-

[1] Of course the $INAR$ model with Poisson arrivals could be used as a pseudo-likelihood with the appropriate "sandwich" modification to the usual standard errors. We do not follow up on this suggestion here.

niques like multiple residual analysis and specification testing are also available in the ML framework. Moreover, not only is the model estimated by ML but so too is the *entire h-step ahead probability mass function*. This provides an optimality property for this method of forecasting. Estimation uncertainty can be accommodated by computing confidence intervals for these probabilities. Secondly, we suggest that the forecast mass function be computed by using a Markov Chain (MC) representation of the model. The method, while simple, avoids the need to evaluate complicated convolutions and the same technique may be applied to any arrivals distribution and thinning mechanism. Thirdly, we consider forecasting the *cumulative distribution function* and events based on it. While it is undoubtedly interesting to know what the probability distribution of the size of a queue is, it is often more important to know what the probability that the number will exceed a certain critical threshold is. This requires forecasts of the cumulative distribution function and confidence intervals for the associated probabilities. The paper explains how confidence intervals with the correct coverage may be constructed.

A data set consisting of counts of deaths (by medical injury), monthly from January 1997 to December 2003, is analysed by ML techniques. Lag selection is achieved by means of residuals analysis and specification tests. The selected model is used to forecast up to 8 months ahead. Forecasts are made for both the probability mass and cumulative distribution functions.

The remainder of the paper is organized as follows. Section 2 outlines the $INAR(p)$ model and briefly discusses its properties. In Section 3, we present a method for producing h-step ahead forecasts of the conditional probability distribution of the $INAR(p)$ process. We also show how parameter uncertainty can be reflected in confidence intervals for probability forecasts. The medical injury death count data is analysed in Section 4 while Section 5 concludes.

2. The INAR(p) Model

Du and Li (1991) define the $INAR(p)$ model to be

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots + \alpha_p \circ X_{t-p} + \varepsilon_t, \tag{1}$$

where the innovation process $\{\varepsilon_t\}$ is i.i.d $(\mu_\varepsilon, \sigma_\varepsilon^2)$ and is assumed to be independent of all thinning operations $\alpha_k \circ X_{t-k}$ for $k = 1, 2, \ldots, p$, which are in turn conditionally independent. The "$\circ$" is the thinning operator which, conditional

on X_{t-k}, is defined as

$$\alpha_k \circ X_{t-k} = \sum_{i=1}^{X_{t-k}} \mathrm{B}_{i,k},$$

where each collection $\{\mathrm{B}_{i,k}, i = 1, 2, ..., X_{t-k}\}$ consists of independently distributed Bernoulli random variables with parameter α_k and the collections are mutually independent for $k = 1, 2, \ldots, p$. Intuitively, $\alpha_k \circ X_{t-k}$ is the number of individuals that would independently survive a Binomial experiment in a given period, where each of the X_{t-k} individuals has identical surviving probability α_k. The case where $p = 1$ and $\{\varepsilon_t\}$ is Poisson is known as Poisson autoregression, often denoted as $PoINAR$, since in this case the marginal distribution of X_t is also Poisson. When $p > 1$, it can be shown that the unconditional mean of X_t and the unconditional variance of X_t are generally not equal, so that the marginal distribution of X_t is no longer Poisson even though the innovations are. Dion et al. (1995) show that the $INAR(p)$ process may be generally viewed as a special multitype branching process with immigration. When $\alpha_k \in [0, 1)$, the $INAR(p)$ process is asymptotically stationary as long as $\sum_{k=1}^{p} \alpha_k < 1$ and the correlation properties of this process are identical to the linear Gaussian $AR(p)$ process according to Du and Li (1991).

The conditional moments of X_t are given by

$$\begin{aligned} E\left[X_t | X_{t-1}, \ldots, X_{t-p}\right] &= \mu_\varepsilon + \sum_{k=1}^{p} \alpha_k X_{t-k} \\ Var\left[X_t | X_{t-1}, \ldots, X_{t-p}\right] &= \sigma_\varepsilon^2 + \sum_{k=1}^{p} \alpha_k \left(1 - \alpha_k\right) X_{t-k} \end{aligned}$$

and so while X_t is (unconditionally) stationary it is conditionally heteroscedastic and so the process will exhibit volatility clustering. In contrast to an ARCH model, X_t is serially dependent and the heteroscedastic effect disappears when X_t is uncorrelated. In Bu et al. (2006), a representation of the conditional probability $P(X_t | X_{t-1}, \ldots, X_{t-p})$ is given for the $INAR(p)$ model with Poisson innovations ($INAR(p)$-P) as

$$\begin{aligned} & P(X_t | X_{t-1}, \ldots, X_{t-p}) \\ = & \sum_{i_1=0}^{\min(X_{t-1}, X_t)} \binom{X_{t-1}}{i_1} \alpha_1^{i_1} (1-\alpha_1)^{X_{t-1}-i_1} \sum_{i_2=0}^{\min(X_{t-2}, X_t - i_1)} \binom{X_{t-2}}{i_2} \alpha_2^{i_2} (1-\alpha_2)^{X_{t-2}-i_2} \end{aligned}$$

$$\cdots \sum_{i_p=0}^{\min[X_{t-p},X_t-(i_1+\cdots+i_{p-1})]} \binom{X_{t-p}}{i_p} \alpha_p^{i_p}(1-\alpha_p)^{X_{t-p}-i_p} \frac{e^{-\lambda}\lambda^{X_t-(i_1+\cdots+i_p)}}{[X_t-(i_1+\cdots+i_p)]!}. \quad (2)$$

By multiplying these conditional probabilities we may calculate the likelihood of the data conditional on the initial p observations. By means of the likelihood the parameters may be estimated. Other diagnostics including residuals may also be computed.

A natural way to define residuals in the $INAR(p)$-P model is to define a residual process for each component. So, generalising Freeland and McCabe (2004a), let $\alpha_k \circ X_{t-k} - \alpha_k X_{t-k}$, $t = p+1, ..., T$, be the set of residuals for the kth thinning process and let $\varepsilon_t - \lambda$ be residual set for the arrivals component. These definitions as they stand are not practical, because $\alpha_k \circ X_{t-k}$ and ε_t are not observable but we can replace $\alpha_k \circ X_{t-k}$ and ε_t respectively with $E_t[\alpha_k \circ X_{t-k}]$ and $E_t[\varepsilon_t]$ (their conditional expectations given the observed values of $X_t, X_{t-1}, ..., X_{t-p}$). Thus, we define the computable residuals as

$$r_{kt} = E_t[\alpha_k \circ X_{t-k}] - \alpha_k X_{t-k}$$

and

$$r_{0t} = E_t[\varepsilon_t] - \lambda.$$

It is easy to see that adding the new sets of residuals gives the usual definition of residuals i.e.

$$\sum_{k=0}^{p} r_{kt} = X_t - \sum_{k=1}^{p} \alpha_k X_{t-k} - \lambda.$$

Thus, the usual residuals have been decomposed into sets that reflect each component of the model. However, it should be borne in mind that the decomposition is not orthogonal and the residual sets are correlated. The new residuals may easily be calculated, once the model is estimated, as $E_t[\alpha_k \circ X_{t-k}]$ and $E_t[\varepsilon_t]$ are readily available in terms of the conditional probabilities given in (2) i.e.

$$\begin{aligned} E_t[\alpha_k \circ X_{t-k}] &= \frac{\alpha_k X_{t-k} P(X_t - 1 | X_{t-1}, \ldots, X_{t-k} - 1, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})}, \\ E_t[\varepsilon_t] &= \frac{\lambda P(X_t - 1 | X_{t-1}, \ldots, X_{t-p})}{P(X_t | X_{t-1}, \ldots, X_{t-p})}. \end{aligned}$$

They may also be plotted to assess the adequacy of each component of the model and to possibly suggest improvements.

3. Forecasting Conditional Distribution with the $INAR(p)$-P Model

Coherent forecasting requires the conditional forecast distribution of the count variable at future periods. In the relatively simple case of $PoINAR$ model, the forecast distributions are convolutions of Poisson and Binomial random variables and an explicit expression for $P\left(X_{T+h}|X_T\right)$ is given in Freeland and McCabe (2004b). However, for the higher-order models of principal concern here, analytic solutions are not easily derived. In what follows, we present an efficient procedure for producing h-step ahead distribution forecasts for the $INAR(p)$-P model using the transition probability function of the process.

3.1. Forecasting the Conditional Probability Distribution: A Markov Chain Approach

We may think of any $INAR(p)$ process generated by (1) as a Markov process (chain) X which takes values at time t, X_t. In principle, the number of possible states of the chain, being the values taken by the process, is infinite. But given a data set, there typically exists a sufficiently large positive integer M such that the probability of observing a count larger than M is negligible. Therefore, for a given count series X_t, we can assume that X_t takes values in the finite collection $\{0, 1, \ldots, M\}$. For example, consider the case where $p = 2$. We think of the states of the system as given by pairs of consecutive values of the process. So at time $t-1$, the chain could be in any of the states

$$S = \{(0,0), (0,1), ..., (0,M), (1,0), (1,1), ..., (0,M), (2,0)....\}$$

as (X_{t-2}, X_{t-1}) takes values $(i_2, i_1) \in S$. At time t the process moves to a new pair of values in the same state space and the transition probabilities of going from one state to another are given by

$$P(X_t = j_1, X_{t-1} = j_2 | X_{t-1} = i_1, X_{t-2} = i_2) = P(X_t = j_1 | X_{t-1} = i_1, X_{t-2} = i_2)$$

when $j_2 = i_1$ and zero otherwise. The probability on the right is given by (2) with $p = 2$ for the $INAR(p)$-P model. The zero probability arises when $j_2 \neq i_1$ since both values refer to the process X at the same time period $t-1$. This scheme may be extended to cater for larger values of p. Hence, at any given period t there are $(M+1)^p$ different states in the set S, determined by $\{X_{t-p+1}, X_{t-p+2}, \ldots, X_t\}$. The elements of S are $p \times 1$ vectors and for each of these vectors the first component

refers to X_{t-p+1} while the second refers to X_{t-p+2} and so on. Denote the $(M+1)^p \times 1$ vector, $S(X_t)$, to be the elements of the vectors in the state set S that correspond to X_t. For a Markov system with finite states, the forecast distribution of each state at any time t can be obtained by means of the transition matrix method.

Let $\mathbf{Q}$ denote the $(M+1)^p \times (M+1)^p$ transition probability matrix of an $INAR(p)$ model with maximum possible count M. To get probability forecasts for each state, we let the $(M+1)^p \times 1$ probability vector, $\boldsymbol{\pi}_t$ represent the probabilities of finding the system in each of the different states at a given period t. Also define, for each $i \in \{0, 1, ..., M\}$, a $(M+1)^p \times 1$ selection vector $\mathbf{s}_i$, which has $M+1$ entries equal to 1 in positions that correspond to those in $S(X_t)$ where $X_t = i$; all other entries in $\mathbf{s}_i$ are zero. Thus, the probability of $X_t = i$ can be written as $\boldsymbol{\pi}_t' \mathbf{s}_i$. Hence, for a general $INAR(p)$ process the conditional probability forecasts for X_{T+h} can be obtained from the forecasts of the probability vector $\boldsymbol{\pi}_{T+h}$. That is

$$P\left(X_{T+h} = i | X_T, \ldots, X_{T-p+1}\right) = \boldsymbol{\pi}_{T+h}' \mathbf{s}_i.$$

The following results are well known from the theory of Markov chains (see for example Kemeny and Snell (1976)). Let $\mathbf{Q}$ and $\mathbf{Q}^{(h)}$ denote, respectively, the one-step transition matrix and h-step transition matrix for a homogeneous pth-order Markov system. Then

$$\mathbf{Q}^{(h)} = \mathbf{Q}^{(h-1)} \mathbf{Q} = \mathbf{Q}^h, \tag{3}$$

and

$$\boldsymbol{\pi}_{T+h}' = \boldsymbol{\pi}_{T+h-1}' \mathbf{Q} = \boldsymbol{\pi}_T' \mathbf{Q}^h. \tag{4}$$

Equation (3) says that the h-step transition matrix is equal to the hth power of the one-step transition matrix and Equation (4) says that the h-step ahead forecast of the probability vector $\boldsymbol{\pi}_{T+h}$ is equal to the current probability vector $\boldsymbol{\pi}_T$ times the h-step transition matrix. Thus the current probability vector and the 1-step ahead transition matrix are all that is required to produce forecasts for any number of periods ahead. We may summarise the foregoing developments in the following proposition.

Proposition 3.1. *For a general $INAR(p)$ process with maximum possible count assumed to be M, the h-step ahead forecast of the conditional probability of $X_{T+h|T} = i$, $i = 0, 1, ..., M$ is given by*

$$P\left(X_{T+h} = i | X_T, \ldots, X_{T-p+1}\right) = \boldsymbol{\pi}_T' \mathbf{Q}^h \mathbf{s}_i, \tag{5}$$

where $\boldsymbol{\pi}_T'$ is probability of the current state of the system, $\mathbf{s}_i$ is a selector vector corresponding to the value i and $\mathbf{Q}^h$ is the hth power of the transition matrix $\mathbf{Q}$ of

the process. The transition matrix for the $INAR(p)$-P process may be calculated from (2) which depends on the underlying parameters $\alpha_1, ..., \alpha_p$ and λ.

A method for assessing the uncertainty associated with probability forecasts (5) due to parameter estimation is presented in the next section.

3.2. Forecasting the Conditional Distribution When Parameters are Estimated

If the parameters of the model were known it would be easy to calculate the conditional probability forecasts $P\left(X_{T+h} = i | X_T, \ldots, X_{T-p+1}\right)$ directly using the results of Proposition 3.1. However, in almost all practical applications these parameters are unknown and have to be estimated. Therefore, it is important that this source of variation be accounted for when producing forecasts. In the current context, the values taken by the process in the future are not in doubt; they are low counts and are the elements of $\{0, 1, .., M\}$. The unknown quantities are the *probabilities* $P\left(X_{T+h} = i | X_T, \ldots, X_{T-p+1}\right)$, $i = 0, 1, ..., M$ and it is these that must be estimated. We abbreviate these conditional probabilities (in the manner of conditional expectations) to $P_T\left(X_{T+h} = i\right)$. Estimation uncertainty is the error made in estimating these probabilities. This error is in turn a function of the error made in estimating the unknown parameters implicit in $P_T\left(X_{T+h} = i\right)$ with these probabilities specified in Proposition 3.1 for the $INAR(p)$-P model.

Let $\boldsymbol{\theta} = (\alpha_1, ..., \alpha_p, \lambda)$ be the parameter vector of the $INAR(p)$-P model. The h-step ahead forecast of the conditional probability mass function is written as $P_T\left(X_{T+h} = i; \boldsymbol{\theta}\right)$ to underline dependence on the parameters. Under standard regularity conditions, the ML estimator of $\boldsymbol{\theta}$, denoted $\hat{\boldsymbol{\theta}}$, is asymptotically normally distributed around the true parameter value, i.e. $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N(\mathbf{0}, \mathbf{i^{-1}})$ where $\mathbf{i}^{-1}$ is the inverse of the Fisher information matrix. Let $g_i(\hat{\boldsymbol{\theta}}) = \hat{P}_T\left(X_{T+h} = i\right) = P_T\left(X_{T+h} = i; \hat{\boldsymbol{\theta}}\right)$, $i = 0, ..., M$ and define a *vector* function $\mathbf{g}(\hat{\boldsymbol{\theta}}) = \left(g_0(\hat{\boldsymbol{\theta}}), ..., g_M(\hat{\boldsymbol{\theta}})\right)'$ to deal with all members of the support simultaneously. Since functions of maximum likelihood estimators are themselves also maximum likelihood estimators it follows that $\mathbf{g}(\hat{\boldsymbol{\theta}})$ is the *MLE of the h-step ahead forecast distribution*. Furthermore, an application of the delta method gives the (asymptotic) joint multivariate distribution of the entire estimated forecast mass function. From this we may compute a confidence interval for the estimated probability associated with each value i of X_{T+h} in the forecast distribution as well as for

estimated probabilities of more complex events. The following proposition is a straightforward consequence of Serfling (1980, Section 3.3).

Proposition 3.2. *For the $INAR(p)$-P model, the ML estimator of the h-step ahead forecast, $\mathbf{g}(\hat{\boldsymbol{\theta}})$, has an asymptotically normal distribution with mean vector $\mathbf{g}(\boldsymbol{\theta}_0)$ and variance matrix*

$$\mathbf{V}(\boldsymbol{\theta}_0) = T^{-1}\mathbf{D}\mathbf{i}^{-1}\mathbf{D}', \tag{6}$$

where $\mathbf{i}$ is the Fisher information matrix and $\mathbf{D} = \partial\mathbf{g}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is a $(M+1)\times(p+1)$ matrix of partial derivatives.

Expressions for these derivatives are available in Bu et al (2006). The elements on the diagonal of $\mathbf{V}(\boldsymbol{\theta}_0)$ are the variances of the estimated forecast probabilities for each possible value on its support and the off-diagonal elements represent the covariances between estimated forecast probabilities for different possible values. Accordingly, estimators of the forecast probabilities for different values of i will be correlated. By Proposition 3.2, marginal 95% confidence intervals for the conditional probability $P(X_{T+h} = i|X_T, \ldots, X_{T-p+1}; \boldsymbol{\theta}_0)$ for $i = 0, 1, \ldots, M$, can be computed, using its asymptotic distribution, by means of

$$\hat{P}_T\left(X_{T+h} = i\right) \pm 2\sigma_{i+1}(\hat{\boldsymbol{\theta}}),$$

where $\sigma^2_{i+1}(\hat{\boldsymbol{\theta}})$ is the $(i+1, i+1)$ element of $\mathbf{V}(\hat{\boldsymbol{\theta}})$. While we may compute $\hat{P}_T\left(X_{T+h} = i\right)$ for every $i = 0, 1, \ldots, M$ to obtain pointwise probabilities of the entire mass function, the correlation between $\hat{P}_T\left(X_{T+h} = i\right)$ and $\hat{P}_T\left(X_{T+h} = j\right)$ makes interpretation very difficult when more than a single value of i is involved. This correlation may be extremely large i.e. very close to $+1$ and negative correlation is also possible. Thus, when one is interested in events like $(X_{T+h} \leq i)$ the cumulative distribution function should be used.

Many more complicated events may, in turn, be written as mappings of $\mathbf{g}$; often a linear function $\mathbf{c}'\mathbf{g}(\hat{\boldsymbol{\theta}})$ of $\mathbf{g}(\hat{\boldsymbol{\theta}})$ will suffice. Another application of the delta method gives the asymptotic distribution of these mappings, e.g. $\mathbf{c}'\mathbf{g}(\hat{\boldsymbol{\theta}})$ is distributed as

$$N\left(\mathbf{c}'\mathbf{g}(\boldsymbol{\theta_0}), \mathbf{c}'\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{c}\right). \tag{7}$$

For example, let $\mathbf{c}'\mathbf{g}(\hat{\boldsymbol{\theta}}) = \ell_i'\mathbf{g}(\hat{\boldsymbol{\theta}})$ where ℓ_i is a vector with the first $i+1$ elements equal to 1 and the rest equal to zero. This linear combination is the sum of the

first $i+1$ elements of $\mathbf{g}$ and gives cumulative probabilities i.e.

$$\hat{P}_T\left(X_{T+h} \leq i\right) = \sum_{j=0}^{i} \hat{P}_T\left(X_{T+h} = j\right)$$

and, from (7), the appropriate variance for $\hat{P}_T(X_{T+h} \leq i)$ is the sum of all the elements in the $(i+1) \times (i+1)$ top left submatrix of $\mathbf{V}(\boldsymbol{\theta}_0)$, estimated by $\mathbf{V}(\hat{\boldsymbol{\theta}})$. Thus we may forecast the *cumulative distribution function* and obtain valid confidence intervals.

Another example is the probability of high or low counts. Say we are interested in the estimated probability of the future event, $(u > l)$

$$(X_{T+h} \leq l) \text{ or } (X_{T+h} > u)$$

which may be calculated by adding $\hat{P}_T(X_{T+h} \leq l)$ to $1 - \hat{P}(X_{T+h} \leq u)$. A valid confidence interval is given by

$$1 + \hat{P}_T(X_{T+h} \leq l) - \hat{P}(X_{T+h} \leq u) \pm 2\sqrt{\mathbf{c}'\mathbf{V}(\hat{\boldsymbol{\theta}})\mathbf{c}}$$

where $\mathbf{c}' = (\ell_u - \ell_l)'$ with ℓ_i defined as above. In this manner, we may obtain valid confidence intervals for the probabilities of fairly arbitrary complex events based on the individual outcomes.

4. Analysis of Injury Data

4.1. The Data

In this section, we apply the method developed to monthly counts of medical injury deaths in Australia. This data set, first analysed in Snyder et al (2007), consists of 84 counts recorded from January 1997 to December 2003. In the first instance we conduct some preliminary analysis to get an overall picture of the data at hand. Figure 1 provides the time series plot of the data, which shows neither discernible trends nor clear seasonal patterns. A summary of simple descriptive statistics for the data is given in Table 1. It can be seen that the observed counts vary from 0 to 7 with the sample mean and variance equal to 2.083 and 2.656, respectively. This suggests that there is some slight overdispersion in the data. The marginal distribution of these data is depicted in Figure 2.

[Figure 1]
[Table 1]
[Figure 2]

4.2. Model Selection

For forecasting purposes, we leave out the last 8 observations (10 percent of the whole sample) and use the first 76 data points to select and estimate the model. Figures 3 and 4 plot the sample autocorrelation function (SACFs) and sample partial autocorrelation function (SPACFs) of the process as well as those of the squares. From Figure 3 we conclude that there is correlation in the levels of the process to be modelled though it is not exceptionally large. Interestingly, Figure 4, for the squares process, also reveals significant correlation in the volatility a fact that is also apparent when one closely examines the time series plot in Figure 1.

[Figure 3]
[Figure 4]

Figure 3 shows that the SACFs are significant at both lag 2 and lag 4 at 5% significance level. This is indicative of the presence of serial dependence in the series. However, no significant seasonal patterns are found. Results of the Ljung-Box portmanteau test with various lag lengths (not reported) also confirm the presence of serial dependence in the data. This and the existence of volatility clustering in Figure 4 indicates that the $INAR$-P class is a reasonable set of models to consider. Meanwhile, the SPACFs is significant at lag 2, which suggests that an autoregressive model with dependence of order 2 may be a reasonable starting point.

Based on the above analysis, we decided to proceed by estimating an $INAR(2)$-P model. Estimation of $INAR(p)$-P models can be carried out in several different ways. These include the moments based Yule-Walker (YW) estimation method, the conditional least squares (CLS) estimation method of Klimko and Nelson (1978), and the maximum likelihood (ML) estimation of Bu et al. (2006). Bu (2006) provided detailed accounts of the three estimation methods and examined both the asymptotic efficiency and finite sample performance of the ML estimator (MLE) in relation to both the YW and CLS estimators. It is concluded that even in finite samples it is worth the effort to use MLE for gains in terms of both the bias and the mean squared error (MSE). For this reason, we decided to estimate the model by conditional (on the initial observations) ML (CML). Details of CML estimation of $INAR(p)$-P models are discussed in Bu et al. (2006)[2]. The

[2]Bu (2006) also suggested a procedure for computing the unconditional likelihood but this is not pursued in this study.

CML estimates of the parameters are $\widehat{\alpha}_1 = 0.058(0.094)$, $\widehat{\alpha}_2 = 0.236(0.095)$, and $\widehat{\lambda} = 1.537(0.291)$, respectively. The estimated asymptotic standard errors, which are obtained from the inverse of the observed Hessian, are given in the parenthesis adjacent to each estimate. As the model is estimated by maximum likelihood, it is straightforward to obtain the corresponding AIC and BIC values for the fitted model, which may be used as indications of overall suitability amongst alternative models. The AIC and BIC for the estimated $INAR(2)$-P model are 273.39 and 280.38, respectively.

To assess the adequacy of the fitted model, we examine the residuals for serial dependence. The estimated Pearson residuals of the fitted $INAR(2)$-P model are defined by $\widehat{\varepsilon}_t = X_t - \widehat{\alpha}_1 X_{t-1} - \widehat{\alpha}_2 X_{t-2} - \widehat{\lambda}$. In principle, the existence of any dependence structure in the residuals would suggest that a more general specification is called for. For this reason, we plot the SACFs and SPACFs of $\widehat{\varepsilon}_t$ in Figure 5. Informally, the figure indicates that there is no obvious dependence structure left in the residuals. Results from the Ljung-Box portmanteau tests with various lag lengths (not reported) also do not allow us to reject this hypothesis. However, it should be noted that the residual series $\widehat{\varepsilon}_t$ is an aggregate measure of the residuals from each stochastic component of the model. The results obtained by examining $\widehat{\varepsilon}_t$ do not necessarily reflect the suitability of each individual component.

[Figure 5]

For this reason, we inspect the SACFs and SPACFs for all three residual processes from the estimated $INAR(2)$-P model. While the SACFs and SPACFs for both $\alpha_1 \circ X_{t-1}$ residuals and arrivals residuals (not reported) are all non-significant up to lag 20 at all conventional significance levels, the SACFs and SPACFs for the $\alpha_2 \circ X_{t-2}$ residuals, given in Figure 6, are both significant at lag 4, with the p-values being 0.044 and 0.065, respectively. The presence of serial dependence in $\alpha_2 \circ X_{t-2}$ residuals is also supported by the portmanteau test which yields a test statistic of $Q(4) = 8.170$ (p-value=0.017). These findings suggest that examining only the traditional residuals is not sufficient on its own and that the component residuals are useful tools for detecting the suitability of each component in the model. When combined with the analysis of the usual residuals, they provide a more thorough and robust investigation into the goodness of fit of the estimated model.

[Figure 6]

The results above suggest that a more general $INAR(4)$-P model with Poisson arrivals be investigated. We expect that by doing this we should be able to eliminate the 4th order dependence in the residuals. As before, the model is estimated using conditional maximum likelihood. It is important to note that the thinning parameters are defined in the range $[0, 1)$. This requires restrictions to be imposed on each parameter during estimation. The constrained CML estimation results in $\widehat{\alpha}_3 = 0$. We therefore proceed to estimate an $INAR(4)$-P model without the $\alpha_3 \circ X_{t-3}$ component. The CML estimates of the parameters and the associated standard errors are found to be $\widehat{\alpha}_1 = 0.015(0.095)$, $\widehat{\alpha}_2 = 0.158(0.105)$, $\widehat{\alpha}_4 = 0.137(0.103)$, and $\widehat{\lambda} = 1.550(0.332)$, respectively. From an inspection of the SACFs and SPACFs for both the Pearson residuals and all four component residuals (three thinning processes and one arrival process), we conclude that these residuals correspond to white noise processes and therefore that an $INAR(4)$-P model without $\alpha_3 \circ X_{t-3}$ is adequate in explaining the serial dependence in the data. Meanwhile, the AIC and BIC values for the estimated model reduce to 267.10 and 276.43, respectively, which is also indication of the improved specification.

For forecasting purposes, we tend to prefer a model that is parsimonious to avoid in-sample over fitting. We note from the fitted $INAR(4)$-P model that $\widehat{\alpha}_1$ is fairly small relative to its standard error, suggesting that the first thinning operation process may not be statistically significant. However, a formal test is needed to confirm this hypothesis. Unlike Gaussian $AR(p)$ models, the hypothesis $\alpha_1 = 0$ for $INAR(p)$ models is on the boundary of the parameter space. Therefore, we cannot simply apply the usual test based on the ratio of the parameter estimate to its standard error. Nevertheless, since the model is estimated by ML, the boundary problem in the testing of coefficient significance may be avoided by using a Lagrange Multiplier (LM) test. Unlike the Wald test and the likelihood ratio test, the LM test is valid even when the null hypothesis corresponds to a boundary value of the parameter space. To perform the desired test, we thus estimate the above model by imposing the restriction that $\alpha_1 = 0$ and calculate the LM statistic based on the restricted estimates. The LM statistic is given by $\dot{\ell}'_{\widehat{\boldsymbol{\theta}}_R} i^{-1}_{\widehat{\boldsymbol{\theta}}_R} \dot{\ell}_{\widehat{\boldsymbol{\theta}}_R}$ where $\dot{\ell}'_{\widehat{\boldsymbol{\theta}}_R}$ and $i_{\widehat{\boldsymbol{\theta}}_R}$ are the score vector and information matrix for the unrestricted model evaluated at the restricted estimate $\widehat{\boldsymbol{\theta}}_R$. Explicit expressions for the score function and elements of the information matrix are easily obtained from the results in Section 3 and Appendix B of Bu et al (2006). The resulting test statistic here is given by $LM = 0.022$. Relative to a $\chi^2(1)$ distribution, it corresponds to a p-value of 0.882. Since we can not reject the null hypothesis that $\alpha_1 = 0$, we consider

dropping the $\alpha_1 \circ X_{t-1}$ component and use the restricted model (the $INAR(4)$-P model with only $\alpha_2 \circ X_{t-2}$ and $\alpha_4 \circ X_{t-4}$ as thinning components) for further analysis. The conditional maximum likelihood estimates for the restricted model are $\widehat{\alpha}_2 = 0.158(0.104)$, $\widehat{\alpha}_4 = 0.138(0.103)$, and $\widehat{\lambda} = 1.578(0.276)$, respectively. For the restricted model, neither the SACFs and SPACFs nor the portmanteau Q-tests would allow us to reject hypothesis that Pearson residuals and component residuals are white noise processes. In addition, the AIC and BIC values reduce further to 265.12 and 272.12 respectively, suggesting improvement. The Information Matrix test of the overall adequacy of the model, given in Freeland and McCabe (2004a), has p-value 0.335 and so does not reject the reduced specification. This is the final reduction of the forecasting model.

A further LM test on the joint hypothesis that $\alpha_2 = \alpha_4 = 0$ gives $LM = 5.127$, which corresponds to a p-value of 0.077 relative to the $\chi^2(2)$. This provides evidence against the use of the naive i.i.d. Poisson model. Moreover, the AIC and BIC for the naive i.i.d Poisson model are 284.64 and 286.97, respectively, which are well in excess of those for the chosen $INAR(4)$-P model. Finally, we conducted simulation experiments (not reported) to confirm that the performance of the MLE, in terms of bias and MSE, was satisfactory at $T = 76$ and the LM test had, approximately, the size suggested by the asymptotic approximations.

4.3. Forecasting the Medical Injury Data

This section applies the method developed in Section 3 to produce forecasts for the medical injury data based on the fitted model. For a model with maximum lag length equal to 4, the h-step ahead conditional probability depends on the last 4 observations and can be denoted as $P(X_{T+h}|X_T, X_{T-1}, X_{T-2}, X_{T-3})$; for simplicity, it is henceforth denoted by $P_T(X_{T+h})$. When the parameters of the model are estimated we use the notation $\hat{P}_T(X_{T+h})$. It is observed that the last four observations of the series are $X_T = 6$, $X_{T-1} = 2$, $X_{T-2} = 0$, and $X_{T-3} = 1$, respectively, where $T = 76$.

Table 2 gives the 4-period ahead conditional mean, median and mode forecasts. As expected the conditional mean forecasts are no longer integer values. (In results unreported it can be seen that these conditional mean forecasts converge to the mean of the marginal distribution. This is equal to the unconditional mean of the process implied by the parameter estimates. Similarly, the conditional median and mode forecasts converge to their marginal counterparts.)

[Table 2]

We apply the propositions proposed in Section 3 to compute point estimates and confidence intervals for the probabilities associated with each value of the forecast distribution. These interval forecasts are given in the top panel of Table 3 in the form of the point estimate plus and minus two standard errors. Thus in the $T+1$ period the point estimate of the probability of the value 0 occurring is 0.126 and we are 95% confident that the probability lies between 0.126 ± 0.049; However, it is not easy to interpret multiple rows from Table 3 simultaneously. Figure 7 gives three examples of the contours of the joint probability density functions of the estimated probability forecasts for pairs of possible counts. The probability forecasts are calculated by the delta method and so the joint probability density functions of the probability forecasts for any pair of possible counts are asymptotically bivariate normal. It can be seen from these contours that the estimated probability forecasts for different counts are correlated. For instance, the contour plot in Figure 7(a) suggests that the estimated conditional probabilities $\hat{P}_T(X_{T+1}=0)$ and $\hat{P}_T(X_{T+1}=1)$ have a near perfect positive correlation (0.980) while Figure 7(b) shows a weaker correlation (0.323) between $\hat{P}_T(X_{T+1}=1)$ and $\hat{P}_T(X_{T+1}=2)$. In contrast, Figure 7(c) suggests a negative correlation (-0.157) between $\hat{P}_T(X_{T+1}=2)$ and $\hat{P}_T(X_{T+1}=3)$.

[Table 3]
[Figure 7]

In many circumstances one is often more concerned with the conditional *cumulative* distribution forecasts and here too account must be taken of the correlation between individual forecasts. For example, in the current context, having multiple deaths caused by medical injuries in any single month may be regarded as being very serious and thus the probability of having a count of more than 1, i.e. $P_T(X_{T+h}>1)$, may have particular significance. Although cumulative probabilities can be easily inferred from the top panel of Table 3 by summing over the corresponding point probability estimates, the appropriate standard errors cannot be directly inferred from the individual standard errors as the variance of a cumulative probability depends on off-diagonal elements of $\mathbf{V}(\boldsymbol{\theta}_0)$ in (6) and these covariances are, as we have seen, rarely negligible.

A selection of cumulative probabilities together with their confidence intervals are presented in the bottom panel of Table 3. It can be seen that, for instance in the $T+1$ period, the estimate of the probability of having a count less than or equal to 1 is about 0.392 and the 95% confidence interval of this probability lies

between 0.392 ± 0.101. Equally, the estimate of the probability of having a count greater than 1 is 0.608 with a 95% confidence interval equal to 0.608 ± 0.101.

5. Conclusion

In this paper, we extend the ideas of Freeland and McCabe (2004b) and develop a method for producing data coherent forecasts for higher-order $INAR$ models. We show that the $INAR(p)$ process can be regarded as a Markov system and the forecasts of the distribution of a count series can be obtained by means of a transition matrix of the process. A procedure for calculating confidence intervals for these forecast probabilities is also suggested.

An empirical analysis of Australian medical injury data under a Maximum Likelihood framework is conducted. Estimates of parameters of the $INAR(p)$-P model are obtained by conditional maximum likelihood estimation. Issues of model adequacy are also examined. Our analysis shows that the analysis of traditional residuals alone may ignore serial dependence in the component residuals and that these latter are useful tools for detecting model misspecification. We apply the method developed to produce distribution forecasts for the medical injury data. The results show that the estimated point mass forecasts are more informative than those supplied by either the mean, median or mode of the forecast distributions. In particular, we show that it is also possible to obtain forecasts of the cumulative probabilities as well as their associated confidence intervals. Given the relatively small sample size on which the forecast experiments in this study are based, it is difficult to perform robust statistical tests on the predictive accuracy of competing models as in, for example, Corradi and Swanson (2006). Nevertheless, our analysis does indicate the potential benefit of having constructive model selection tools available for analysing count data and improving forecast performance.

References

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

Alzaid, A.A., & Al-Osh, M.A. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 27, 314-323.

Bu, R. (2006). Essays in financial econometrics and time series analysis. *Ph.D Thesis*, *University of Liverpool*.

Bu, R., Hadri, K., & McCabe, B.P.M. (2006). Maximum likelihood estimation of higher-order integer-valued autoregressive processes. *Working paper*, *University of Liverpool*.

Corradi, V., & Swanson, N.R. (2006). Predictive density and conditional confidence interval accuracy tests. *Journal of Econometrics*, 135, 187-228.

Dion, J.P., Gauthier, G., & Latour, A. (1995). Branching processes with immigration and integer-valued time series. *Serdica,* 21, 123-136.

Du, J.G., & Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis,* 12, 129-142.

Freeland, R.K., & McCabe, B.P.M. (2004a). Analysis of low count time series by Poisson autoregression. *Journal of Time Series Analysis,* 25, 701-722.

Freeland, R.K., & McCabe, B.P.M. (2004b). Forecasting discrete valued low count time series. *International Journal of Forecasting,* 20, 427–434.

Jung, R.C., & Tremayne, A.R. (2006). Coherent forecasting in integer time series models. *International Journal of Forecasting,* 22, 223-238.

Kemeny, J.G., & Snell, J.L. (1976). *Finite Markov Chains*. New York: Springer.

Klimko, L.A., & Nelson, P.I. (1978). On conditional least squares estimation for stochastic processes. *Annals of Statistics,* 6, 629-642.

McCabe, B.P.M., & Martin, G.M. (2005). Bayesian predictions of low count time series. *International Journal of Forecasting,* 21, 315-330.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poison counts. *Advances in Applied Probability,* 20, 822-835.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Snyder, R.D., Martin, G.M., Gould, P., & Feigin, P.D. (2007). An assessment of alternative state space models for count time series. *Working paper*, *Monash University*.

Table 1: Descriptive Statistics of the Medical Injury Data

Minimum	Maximum	Median	Mode	Mean	Variance
0	7	2	2	2.083	2.656

Table 2: Mean, Median and Mode Forecasts and the Observed Values

	Forecasting Period							
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
Observed	2	2	2	3	1	2	2	3
Mean	2.033	2.528	2.177	2.809	2.204	2.373	2.228	2.343
Median	2	2	2	3	2	2	2	2
Mode	2	2	2	2	2	2	2	2

Table 3: Forecasts for the Medical Injury Data

	$\hat{P}_T(X_{T+h}=i)$			
h	$i=0$	$i=1$	$i=2$	$i=3$
1	0.126±0.049	0.266±0.053	0.276±0.010	0.186±0.034
2	0.073±0.095	0.199±0.144	0.261±0.054	0.222±0.061
3	0.111±0.043	0.247±0.052	0.271±0.012	0.197±0.029
4	0.056±0.056	0.166±0.099	0.241±0.055	0.228±0.025
5	0.111±0.049	0.243±0.059	0.267±0.011	0.196±0.031
6	0.094±0.052	0.221±0.071	0.262±0.024	0.207±0.030
7	0.109±0.050	0.240±0.061	0.267±0.013	0.198±0.031
8	0.097±0.053	0.225±0.072	0.263±0.022	0.205±0.031
	$\hat{P}_T(X_{T+h}\leq i)$			
h	$i=0$	$i=1$	$i=2$	$i=3$
1	0.126±0.049	0.392±0.101	0.668±0.104	0.854±0.070
2	0.073±0.095	0.272±0.240	0.533±0.293	0.755±0.232
3	0.111±0.043	0.358±0.095	0.629±0.104	0.826±0.076
4	0.056±0.056	0.223±0.155	0.463±0.209	0.691±0.186
5	0.111±0.049	0.354±0.107	0.622±0.118	0.818±0.087
6	0.094±0.052	0.315±0.123	0.577±0.147	0.784±0.117
7	0.109±0.050	0.349±0.110	0.615±0.123	0.813±0.092
8	0.097±0.053	0.322±0.125	0.585±0.147	0.790±0.116

Figure 1: Time Series Plot of the Medical Injury Data

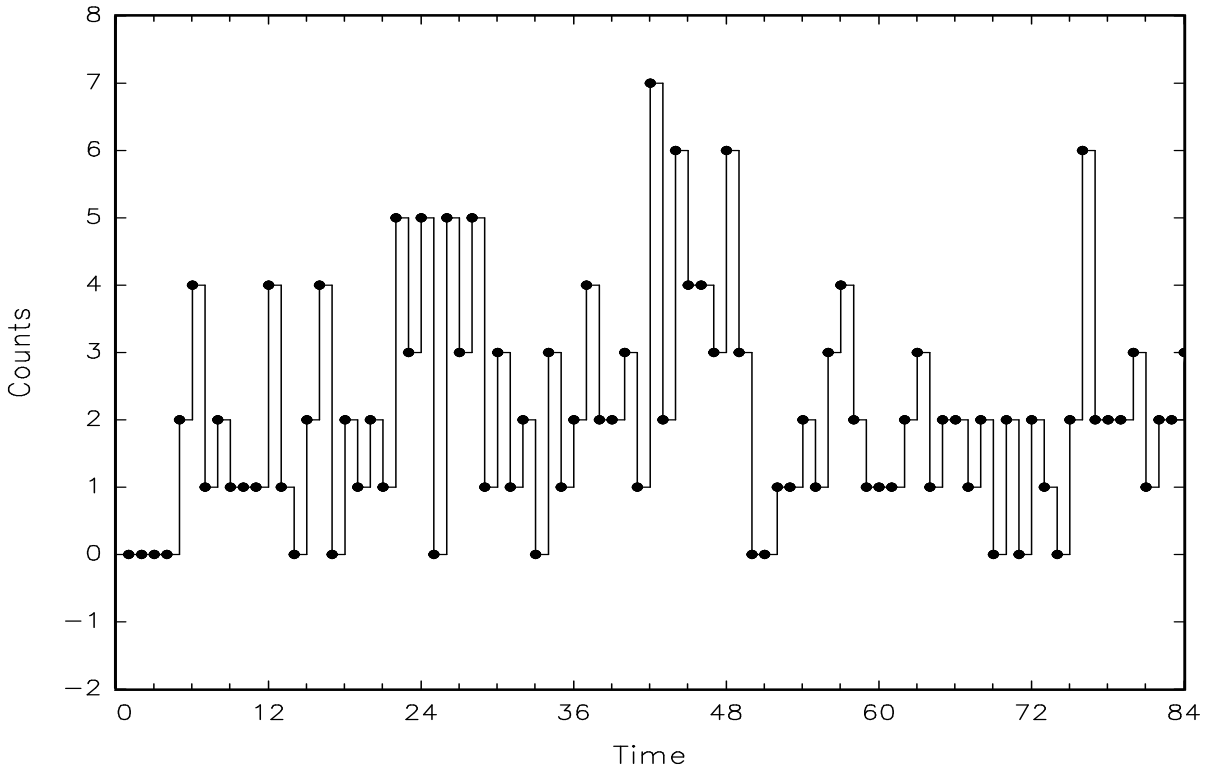


Figure 2: Marginal Distribution of the Medical Injury Data

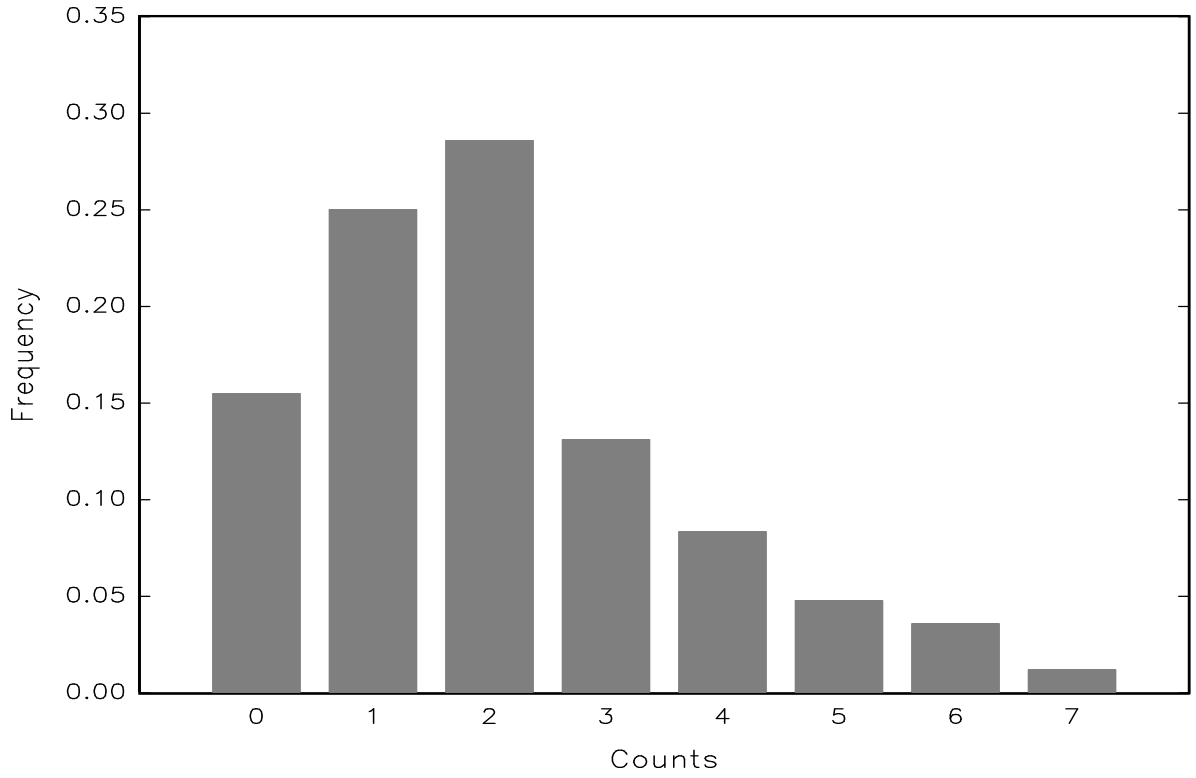


Figure 3: Correlograms of the Medical Injury Data

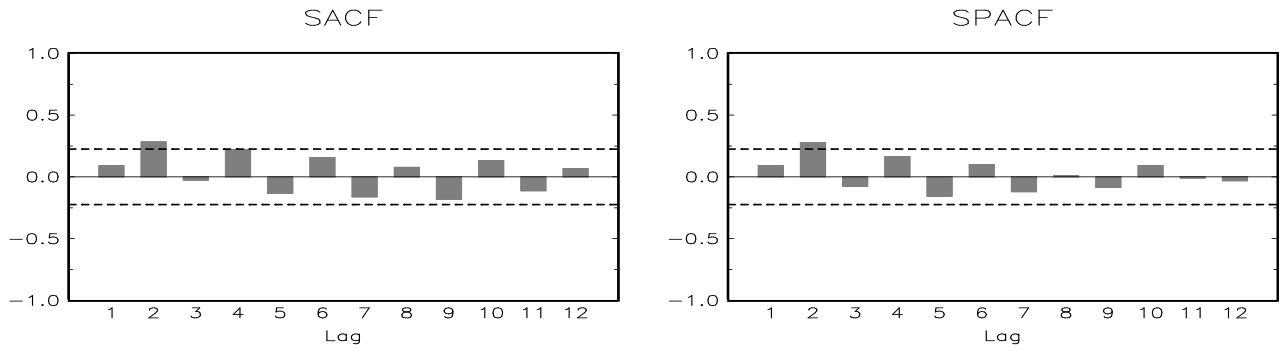


Figure 4: Correlograms of the Squares of the Medical Injury Data

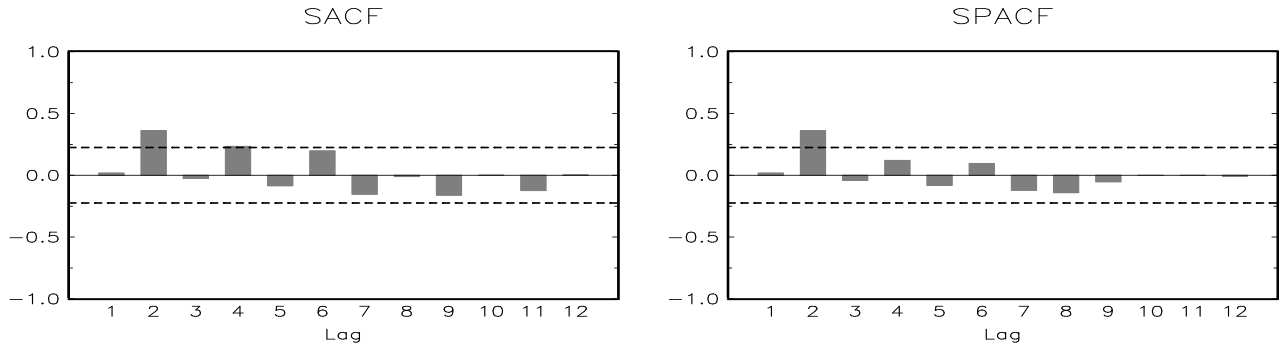


Figure 5: Correlograms of the Residuals $\widehat{\varepsilon}_t$ from the $INAR(2)$-P Model

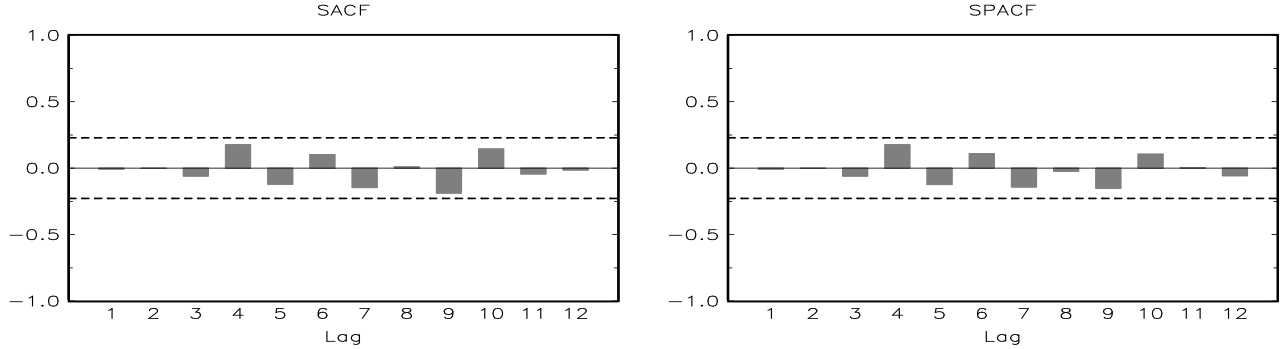
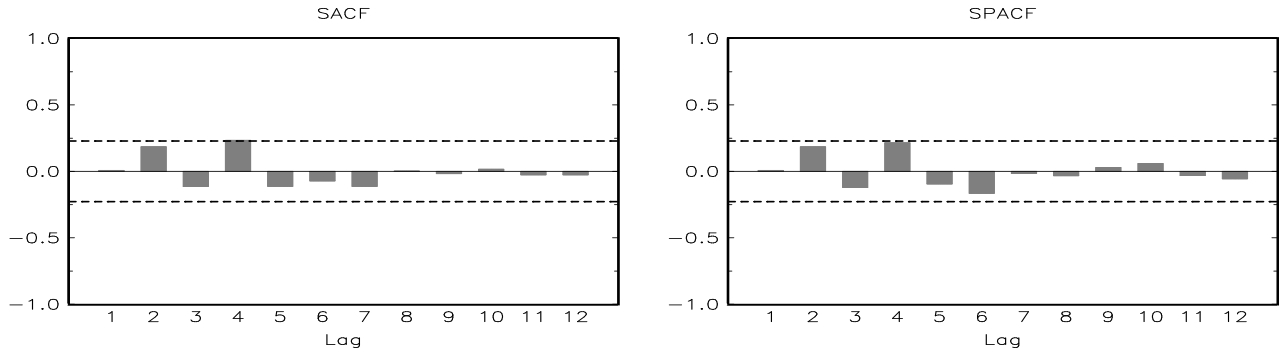


Figure 6: Correlograms of the $\alpha_2 \circ X_{t-2}$ Component Residuals



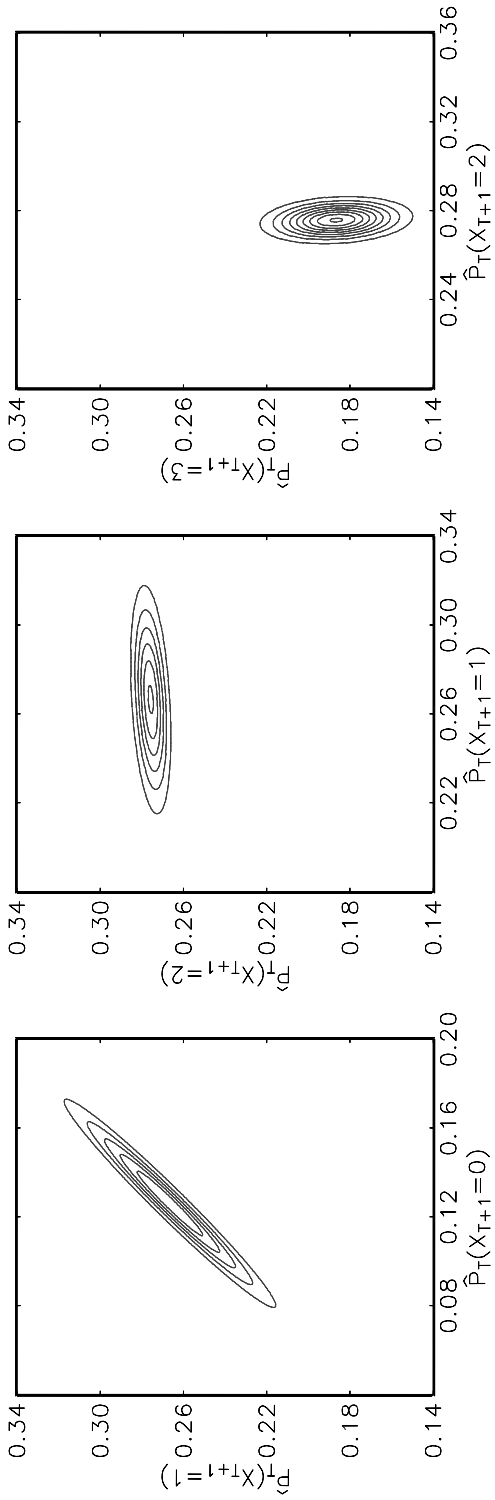


Figure 7: Contour Plots of the Bivariate Densities