Available online at www.sciencedirect.com

SCIENCE DIRECT®

International Journal of Forecasting 20 (2004) 427–434

international journal of forecasting

www.elsevier.com/locate/ijforecast

Forecasting discrete valued low count time series

R.K. Freeland[a], B.P.M. McCabe[b,*]

[a] *University of Waterloo, Waterloo, Canada*

[b] *The School of Management, The University of Liverpool, Chatham Street, Liverpool L69 7ZH, UK*

Abstract

In the past, little emphasis has been placed on producing data coherent forecasts for discrete valued processes. In this paper the conditional median is suggested as a general method for producing coherent forecasts and is in contrast to the conventional conditional mean. When counts are low we suggest that the emphasis of the forecast method be changed from forecasting future values to forecasting the k-step-ahead conditional distribution. In practice, this usually depends on unknown parameters. We modify the distribution to account for estimation error in a coherent way. The ideas are exemplified by an analysis of Poisson Autoregressive model and of wage loss claims data.

Keywords: Birth and death process; Data coherence; Forecasting; Maximum likelihood; Poisson autoregression; Queuing process

1. Introduction

This paper is concerned with forecasting time series which are counts, i.e. data which take values in $\{0,1,2,\ldots\}$. According to Chatfield (2001, p. 81), one of the properties of a 'good' time series model is that it should be unable to predict values which violate known constraints, i.e. we require of a good model that it be forecast coherent. In the count data context we seek a method of forecasting that produces integer values. In the light of this requirement it is clear that traditional *ARIMA* time series methods are inappropriate as they would invariably produce non-integer forecasts. In the absence of a general methodological approach some specific class of models must be entertained and care has to be taken that a particular model is appropriate for the data at hand. The models used here are birth and death processes and their appropriateness for a count data series of burns injuries is discussed in Section 5.

In Section 2 a general method of producing forecasts that are data coherent is presented. This involves calculating the median of the k-step-ahead forecast distribution. Even though coherent forecasts may readily be found by using the median (or even the mode) of the forecast distribution it is argued here that this may be potentially misleading. In cases where the variable is discrete and the cardinality of the support is small, it is suggested that forecasts be provided for each point mass of the distribution. In fact, even in cases where forecasts of continuous variables are required it is becoming increasingly common to use the forecast density itself (see, for example, Wallis (2003), where the Bank of England's Fan charts are discussed).

In almost all practical applications forecast distributions depend on parameters that are unknown. Since these parameters have to be estimated it is important that this source of variation be accounted

* Corresponding author.

E-mail address: brendan.mccabe@liv.ac.uk (B.P.M. McCabe).

0169-2070/$ - see front matter
doi:10.1016/S0169-2070(03)00014-1

for when producing forecasts. Since our forecasts are integers, Section 4 suggests that the probabilities associated with each point mass be modified to reflect the variation in parameter estimation. Point mass forecasting with estimated probabilities is exposited in Sections 2 and 4 in the context of the Poisson Autoregressive model (PAR) model. Finally, Section 5 gives an empirical example of forecasting a count data series using the PAR model. The data are the number of claimants receiving wage loss benefits due to injuries from burns, supplied by the Workers Compensation Board (WCB) of the Province of British Columbia, Canada. The proofs of the propositions in the paper are presented in Appendix A.

2. Coherent forecasting

The most common procedure for constructing forecasts in time series models is to use conditional expectations. The reason is that this technique will yield forecasts with minimum mean square error. But this method lacks data coherency when the time series under consideration has restrictions on its support, e.g. is the set of integers. It is suggested here that the k-step-ahead conditional distribution itself be used to produce coherent forecasts. One obvious idea is to use the median of this distribution. The median always lies in the support and is therefore coherent. It is also has an optimality property. Consider a realisation $X^n = \{X_t\}_{t=0}^{n}$ from a discrete time stochastic process. Then it can be shown that the forecast, $\tilde{X}_{n+k}$, of X_{n+k} that minimises the expected absolute error

$$E[|\ X_{n+k} - \tilde{X}_{n+k}\ |\ |\ X^n]$$

is the median of the k-step-ahead conditional distribution.

However, data coherency notwithstanding, it can be quite misleading to summarise an entire distribution by a single point. That the median forecast may not be very informative is exemplified by the following two situations: in the first case, $P(X=0) = 1-P(X=1) = 0.50$, while in the second $P(X=0) = 1-P(X=5) = 0.90$. In both cases the median of X is 0 (the mean is 0.5), but in the second case, there is almost twice the probability of observing a zero. Since there are only two outcomes in these examples it would be more informative to give the probability distribution for both values in the support.

3. Forecasting with the PAR model

As a prelude to deriving the forecast distribution, we briefly describe the basic PAR model of Al-Osh and Alzaid (1987) and McKenzie (1988). Let $X_0, X_1, \ldots, X_n$ be a series of dependent counts generated according to the following model:

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t,$$

where X_0 has a Poisson distribution with mean λ_0 and $\{\varepsilon_t\}_{t=1}^{\infty}$ is a series of independently distributed Poisson random variables with mean λ. The thinning operator '$\circ$' is defined as follows: given X_{t-1}, $\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} \mathrm{B}_{it}$, where $\mathrm{B}_{1t}, \mathrm{B}_{2t}, \ldots, \mathrm{B}_{X_{t-1}t}$ are iid Bernoulli random variables with $P(\mathrm{B}_{it}=1) = 1-P(\mathrm{B}_{it}=0) = \alpha$. Since $\alpha \circ X_{t-1}$ given X_{t-1} is a sum of iid Bernoulli random variables it follows that it has a binomial distribution with parameters α and X_{t-1}. It is further assumed that B_{jt} and ε_t are independent. Notice that, in this model, X_t is composed of two random components [the complement of the death (i.e. the survivorship) component $\alpha \circ X_{t-1} \mid X_{t-1}$, and the arrivals (birth) component ε_t] and that these two components are not (individually) observed. Thus, the distribution of X_t given X_{t-1} is given by the convolution of the two random components and we denote the conditional probability of X_t given X_{t-1} by

$$p(X_t \mid X_{t-1}) = \sum_{s=0}^{\min(X_t, X_{t-1})} \binom{X_{t-1}}{s} \alpha^s (1-\alpha)^{X_{t-1}-s} \times \frac{e^{-\lambda}\lambda^{X_t - s}}{(X_t - s)!},$$

where ( ) is the standard combinatorial symbol. The model is Markovian and thus the likelihood may be computed from the product of the $p(X_t \mid X_{t-1})$. The stationary distribution of X_t is Poisson with mean $\lambda/(1-\alpha)$. The PAR model may also be interpreted as a birth and death process or as an infinite server queue.

The following result establishes the k-step-ahead conditional distribution and gives its first two moments. The result for $k = 1$ was given by McKenzie (1988), but the current version allows for any finite value of k.

Theorem 1. *The probability mass function of X_{n+k} given X_n is given by*

$$p_k(x \mid X_n) = \sum_{s=0}^{min(x,X_n)} \binom{X_n}{s} (\alpha^k)^s (1-\alpha^k)^{X_n-s} \times \frac{1}{(x-s)!} \exp\left\{-\lambda \frac{1-\alpha^k}{1-\alpha}\right\} \times \left(\lambda \frac{1-\alpha^k}{1-\alpha}\right)^{x-s},$$

and has mean, $\alpha^k X_n + \lambda[(1-\alpha^k)/(1-\alpha)]$, *and variance,* $\alpha^k(1-\alpha^k)X_n + \lambda[(1-\alpha^k)/(1-\alpha)]$.

Thus, $p_k(x \mid X_n)$ is the probability of the value x occurring, according to the k-step-ahead conditional distribution.

4. Forecasting count data when parameters are estimated

Knowledge of the parameters $\boldsymbol{\theta} = (\alpha, \lambda)'$ is required to implement the forecasting ideas of the last section. If the parameters were known it would be easy to calculate $p_k(x \mid X_n)$ as in Theorem 1. Typically, these parameters are estimated; maximum likelihood is often used and this leads to parameter estimates which are asymptotically normal. Now, write the k-step mass function as $p_k(x \mid X_n; \boldsymbol{\theta})$, where we have overtly introduced the vector of parameters $\boldsymbol{\theta}$. In practice, we are only in a position to compute $p_k(x \mid X_n; \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is asymptotically normally distributed around the true parameter value, i.e. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim^a N(\theta, \mathrm{V})$ for some covariance matrix V. Now the δ-method is a technique for finding the asymptotic distribution of a *function* of a random variable, $g(\hat{\boldsymbol{\theta}})$ say, given that the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal. The idea is to apply the δ-method to $g(\hat{\boldsymbol{\theta}}) = p_k(x \mid X_n; \hat{\boldsymbol{\theta}})$.

Let $\hat{\alpha}_n$ and $\hat{\lambda}_n$ be the ML estimators of α and λ in the PAR model based on a sample of size n; it is well know that $(\hat{\alpha}_n, \hat{\lambda}_n)' \sim^a N[(\alpha_0, \lambda_0)', n^{-1}\boldsymbol{V}]$. Computationally efficient methods of computing the matrix **V** may be found in Freeland and McCabe (2001). The ML estimate of the k-step-ahead probability mass is $p_k(x \mid X_n; \hat{\alpha}_n, \hat{\lambda}_n)$. An application of the δ-method gives the asymptotic distribution of this quantity for a fixed value of x. From this we may compute a confidence interval for the probability associated with any value x in the forecast distribution. Obviously, these intervals may be truncated outside [0,1].

Theorem 2. *The quantity $p_k(x \mid X_n; \hat{\alpha}_n, \hat{\lambda}_n)$ has an asymptotically normal distribution with mean $p_k(x \mid X_n; \alpha_0, \lambda_0)$ and variance*

$$\sigma_k^2(x; \alpha_0, \lambda_0) = n^{-1}\left[v_\alpha \left(\left.\frac{\partial p_k}{\partial \alpha}\right|_{\alpha=\alpha_0,\ \lambda=\lambda_0}\right)^2 + 2v_{\alpha\lambda} \frac{\partial p_k}{\partial \alpha}\frac{\partial p_k}{\partial \lambda}\Big|_{\alpha=\alpha_0,\ \lambda=\lambda_0} + v_\lambda \left(\left.\frac{\partial p_k}{\partial \lambda}\right|_{\alpha=\alpha_0,\ \lambda=\lambda_0}\right)^2\right], \tag{1}$$

where v_α and v_λ are the diagonal elements of $\boldsymbol{V}$ and $v_{\alpha\lambda}$ is the off-diagonal element. Here

$$\frac{\partial}{\partial \alpha} p_k(x \mid X_n) = \frac{x_n}{1-\alpha^k}(p_k(x-1 \mid X_n - 1) - p_k(x \mid X_n))k\alpha^{k-1} + (p_k(x-1 \mid X_n) - p_k(x \mid X_n))\lambda \frac{1 - k\alpha^{k-1} - (k-1)\alpha^k}{(1-\alpha)^2},$$

and

$$\frac{\partial}{\partial \lambda} p_k(x \mid X_n) = (p_k(x-1 \mid X_n) - p_k(x \mid X_n)) \frac{1-\alpha^k}{1-\alpha}.$$

Thus we can compute a 95% confidence interval for $p_k(x \mid X_n; \alpha_0, \lambda_0)$, based on its asymptotic distribution, by means of

$$p_k(x \mid X_n; \hat{\alpha}_n, \hat{\lambda}_n) \pm 2\sigma_k(x; \hat{\alpha}_n, \hat{\lambda}_n).$$

In contrast to the treatment of continuous variables, it is the *probability* of the values that is modified to reflect the uncertainty due to estimation of the parameter. In cases where x's form a finite set of discrete values we can treat $\{p_k(x_i \mid X_n; \alpha_0, \lambda_0); i = 1, \ldots, p\}$ as a vector function (one element for each x_i) of a vector argument (α_0, λ_0). In this way it is possible to get joint confidence intervals over the support.

5. Analysis of burns claims data

This section analyses some time series count data obtained from the Workers Compensation Board (WCB) of British Columbia, Canada. The data consist of monthly counts of claimants collecting Short Term Wage Loss Benefit (STWLB) for injuries received in the workplace. In the selected data set all the claimants are male, between the ages of 35 and 54, work in the logging industry and reported their claim to the Richmond, BC service delivery location. The data consist of 120 observations starting in January 1984 and ending in December 1994. The data are claimants whose injuries are burn related. There are two main objectives of the analysis. The first is to estimate the expected number of months during which a claimant will receive benefits. This is an important managerial quantity for the WCB as duration directly affects the total cost of a claim. The second objective is to produce forecasts of the numbers of claimants for the first six months of 1995 for the Richmond delivery area. The number of claimants by location influences the deployment of manpower (and other resources required to service the claim) and impacts on costs and the quality of service delivered.

It is not possible to compute the expected number of months a claimant will remain in the system from the data series of counts. However, if the Poisson AR process can successfully model the counts we may use the queuing interpretation to calculate the expected length of stay. So, let X_t be the number of workers collecting STWLB at time t. This number equals the sum of the number of workers continuing to collect from time $t-1$, $\alpha \circ X_{t-1}$, and the number of new claims at time t, ε_t. It is a standard result in queuing theory that the waiting time is $d = 1/(1-\alpha)$ and this is the mean number of months that a newly disabled worker is expected to collect STWLB.

We may also use the model to provide forecasts of future counts once the parameters have been estimated. While it is perfectly natural to consider the count of claimants process as a queue, it is important that the thinning operator and the arrivals process are correctly specified, especially if ML estimation is to be used or if forecasts are to be made from the k-step-ahead forecast distribution. Now, binomial thinning assumes that the recovery of individuals is independent and this seems plausible, for it is unlikely that one individual's recovery time would affect another's, unless we had limited medical services and recovery of one meant that the next individual could start treatment. The binomial thinning operator also assumes that all individuals recover at the same rate. This is a less realistic assumption, since there is wide variation in individual health due to genetic factors and lifestyle choices, such as diet and exercise. Recovery rates should vary from person to person, since we would expect recovery rates to depend on the person's health immediately prior to injury or illness. However, as the burns data are monthly this source of variation should not be significant. The use of the Poisson distribution for the arrivals process is a standard way to model count data. However, in applications it has often been found lacking because of the fact that variation in real data typically exceeds that of the mean. This sort of overdispersion could be present in the claimant data because of seasonal factors. More workers are employed in the summer months, thereby exposing increased numbers to the risk of injury. However, Table 1 suggests that overdispersion is not a problem for this data set. Finally, if the workers are a small cohort then the arrival and departure processes may be dependent, since, as workers recover, the cohort size increases, which raises the cohort's exposure to injury and increases the number of new injuries (arrivals). In most industries, including forestry, this

Table 1
Descriptive statistics

Data set	Minimum	Maximum	Median	Mode	Mean	Variance
Burns	0	2	0	0	0.34	0.33

is unlikely to be a problem since the number of injured workers is usually a very small fraction of the industry's work force. In summary, therefore, the Poisson AR process is a plausible choice of modelling tool for the WCB data.

5.1. Estimation with burns claims data

In the first instance we conduct some preliminary analysis to get an overall picture of the data at hand. It turns out that the original data contained an 'outlier', in that a single claimant was collecting benefits for a period of 20 months. This is extremely rare and so the observation was removed from the series, as not to do so would imply persistence in the data inconsistent with the short-term memory associated with the Poisson AR model. Table 1 contains a summary of simple descriptive statistics.

It is clear that series of burn related injuries consists of very small counts with a great many zeros. It is also the case that the mean and variance are approximately equal and this suggests that the Poisson AR model may indeed be appropriate for this data set. Figs. 1 and 2 show a plot of the autocorrelation and partial autocorrelation functions of the data. These are also consistent with the data being generated by a Poisson AR(1) and it seems reasonable, therefore, to proceed with estimation of the parameters. The model for burns was estimated by ML and α was calculated to be 0.240 with a 95% confidence interval of (0.007, 0.472). Notice that this interval does not contain zero and we may conclude that there is indeed dependence in the data to be modelled. This value of α implies an average duration per claimant for burns related injuries of $d = 1.316$ months with a 95% confidence interval of (0.881, 1.717). The arrivals parameter, λ, was estimated to be 0.134 with a confidence interval of (0.064, 0.204). The joint information matrix test statistic (see Freeland & McCabe, 2001) for details was 0.240 (P=0.81), which indicates that the Poisson AR(1) model is adequately describing the variation found in the data. These findings suggest that the model (and the estimate of the average duration, d) is satisfactory and that we may proceed to compute the k-step-ahead forecast distribution with reasonable confidence.

5.2. Forecasting burns data

We now calculate forecasts for the first 6 months of 1995; these are shown in Table 2. The table gives

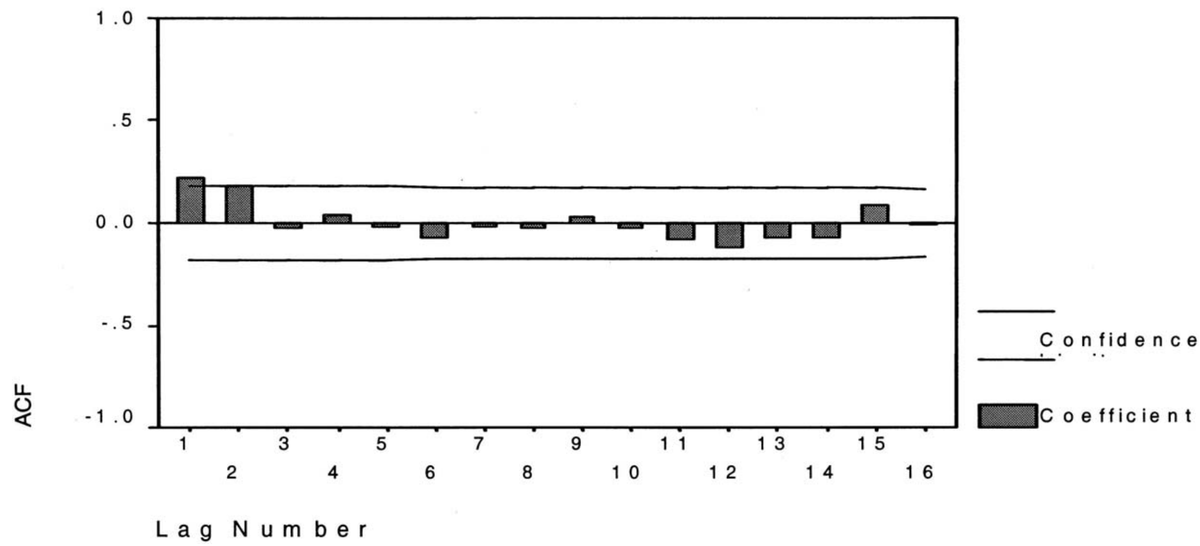

Fig. 1. Autocorrelation function for burns.

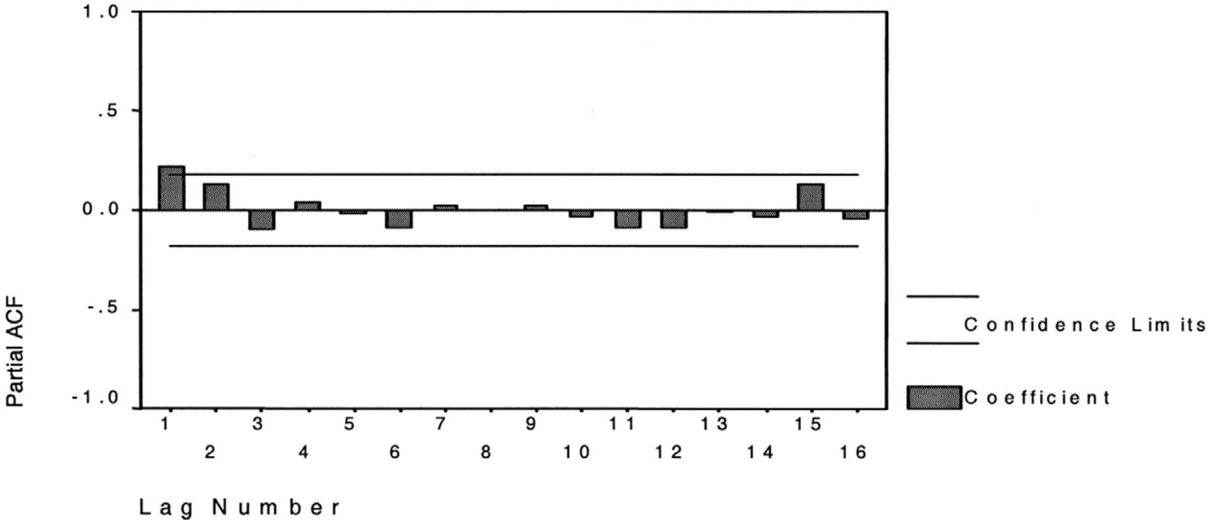

Fig. 2. Partial autocorrelation function for burns.

the k-step-ahead conditional (the current value of the series is 1) mean, median and mode, as well as 95% prediction intervals for the k-step-ahead conditional distribution. Note that the 95% confidence intervals for the six-step-ahead conditional distribution is very close to the 95% confidence intervals for the marginal distribution. Therefore, if we require forecasts beyond 6 months into the future we can simply use the marginal distribution. Also notice that, after five steps ahead, the conditional mean is equal to the marginal mean of 0.176. The k-step-ahead conditional median and mode are both zero for all values of k.

In contrast to the case where the parameters are known, there is now a confidence interval for the probabilities associated with each point mass of the forecast distribution. Thus, for example, in the next period we are 95% confident that the probability of the value 0 occurring lies between 0.812 and 0.937; the point estimate of the probability is 0.875. The point estimates for the values 0, 1, 2 and 3 sum to unity and so the model deems that an observation of 4 or more has zero probability of occurring in the forecast period. Now the median, mode and the mean (if rounded down) all indicate that that no injuries are to be expected in each of the k-step forecast periods. But the table gives more information. It quantifies that the probability of no injury claims in the next period is 88% (point estimate) and

Table 2
Forecasts from burns data

	k						
	1	2	3	4	5	6	∞
Mean	0.374	0.224	0.188	0.179	0.177	0.176	0.176
Median	0	0	0	0	0	0	0
Mode	0	0	0	0	0	0	0
$p_k(0 \mid 1)$	(0.812, 0.937)	(0.773, 0.920)	(0.762, 0.919)	(0.758, 0.920)	(0.757, 0.920)	(0.756, 0.920)	(0.756, 0.920)
$p_k(1 \mid 1)$	(0.063, 0.171)	(0.079, 0.202)	(0.081, 0.211)	(0.081, 0.214)	(0.080, 0.215)	(0.080, 0.215)	(0.080, 0.215)
$p_k(2 \mid 1)$	(0.000, 0.016)	(0.000, 0.023)	(0.000, 0.025)	(0.000, 0.026)	(0.000, 0.026)	(0.000, 0.026)	(0.000, 0.026)
$p_k(3 \mid 1)$	(0.000, 0.001)	(0.000, 0.002)	(0.000, 0.002)	(0.000, 0.002)	(0.000, 0.002)	(0.000, 0.002)	(0.000, 0.002)
$p_k(4 \mid 1)$	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)	(0.000, 0.000)

that a single injury claim has a 12% chance of occurring.

6. Concluding remarks

In this paper we observe that traditional time series *ARMA* methods are inappropriate for forecasting count data. Special models need to be employed to accommodate the integer nature of the data. In addition, the traditional method of forecasting using conditional means will not produce coherent forecasts, i.e. forecasts with integer values in this instance. A method of producing optimal coherent forecasts based on the median of the k-step-ahead conditional distribution is suggested. It is also argued, in cases where the counts are low, that a clearer picture emerges when the median is supplemented by estimates of the probabilities associated with the point masses of the k-step-ahead conditional distribution. A method for calculating confidence intervals for these probabilities is presented. The new techniques are applied to a data set of the number of claimants for wage loss benefit in BC, Canada. The expected duration of a claim cannot be estimated from the numbers of claims themselves. A modelling approach is thus necessary. We argue that the assumptions of the model are a priori plausibly consistent with the structure of the real life data and provide empirical evidence of the appropriateness of the Poisson assumption and the fit of the model. We show that there is a dependence structure to be modelled (i.e. the estimate of the lag-1 correlation, α, is statistically different from zero) and provide an estimate of the expected duration of a claim. Finally, we show that the estimated point mass forecasts are more informative than those supplied by either the mean, median or mode of the forecast distributions.

Acknowledgements

The authors are grateful for support from the National Science and Engineering Research Council of Canada and to Greys Socic for research assistance.

Appendix A

Proof of Theorem 1. Al-Osh and Alzaid (1987) point out the following:

$$(X_t, X_{t-k}) \stackrel{d}{=} \left(\alpha^k \circ X_{t-k} + \sum_{i=0}^{k-1} \alpha^i \circ \varepsilon_{t-i}, X_{t-k} \right),$$

where ε_t is a sequence of uncorrelated non-negative interger-valued random variables with finite mean and variance.

Given X_{t-k}, $\alpha^k \circ X_{t-k}$ has a binomial distribution with parameters (α^k, X_{t-k}). If ε_t is Poisson($\alpha\lambda$) then $\alpha \circ \varepsilon_t$ is Poisson($\alpha\lambda$) (McKenzie, 1988). Further, $\sum_{i=0}^{k-1} \alpha^i \circ \varepsilon_{t-i}$ is Poisson with parameter

$$\sum_{i=0}^{k-1} \alpha^i \lambda = \frac{1-\alpha^k}{1-\alpha} \lambda.$$

Hence the distribution of X_t given X_{t-k} is a convolution of a binomial distribution with parameters (α^k, X_{t-k}) and a Poisson with parameter $[(1-\alpha^k)/(1-\alpha)]\lambda$. The probability mass function and moments follow directly. □

Proof of Theorem 2. This is an application of the δ-method; see Theorem A on p. 122 of Serfling (1980). □

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.

Chatfield, C. (2001). *Time series forecasting*. London: Chapman and Hall/CRC.

Freeland, R. K., & McCabe, B. P. M. (2001). Maximum likelihood estimation of the Poisson Autoregressive Model. Mimeo, University of Liverpool.

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.

Wallis, K. F. (2003). Chi-squared test of interval and density forecasts and the Bank of England's Fan charts. *International Journal of Forecasting 19*, 165–175.

Biographies: Brendan McCABE is Professor of Economics at the University of Liverpool, UK. He has previously taught at the Universities of British Columbia (Canada), Sydney (Australia) and Leeds (UK). His main research interests are in non Gaussian time series.

R. Keith FREELAND is an Assistant Professor in the Department of Statistics and Actuarial Science at the University of Waterloo, Canada. He completed his doctorate at the University of British Columbia. His primary research area is time-series of counts.