Biostatistics (2001), **2**, 4, *pp.* 433–444
Printed in Great Britain

Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease

J. L. HAY*
Department of Social and Preventive Medicine, University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland, 4102, Australia
J.Hay@spmed.uq.edu.au

A. N. PETTITT
Centre in Statistical Science and Industrial Mathematics, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, 4001, Australia

SUMMARY

This paper presents a Bayesian analysis of a time series of counts to assess its dependence on an explanatory variable. The time series represented is the incidence of the infectious disease ESBL-producing *Klebsiella pneumoniae* in an Australian hospital and the explanatory variable is the number of grams of antibiotic (third generation) cephalosporin used during that time. We demonstrate that there is a statistically significant relationship between disease occurrence and use of the antibiotic, lagged by three months. The model used is a parameter-driven model in the form of a generalized linear mixed model. Comparison of models is made in terms of mean square error.

Keywords: Bayesian hierarchical model; Count data; Klebsiella pneumoniae; Markov chain Monte Carlo; Parameter driven model; Time series effects.

1. INTRODUCTION

In epidemiological studies, the reported occurrences of a disease are often expressed as daily, weekly or monthly counts. Studies that model these counts and their associations with other variables provide important information leading to better understanding of the disease and, hopefully, better control measures. When the counts are relatively large, methods based on transformation and consequent use of standard Gaussian techniques are mostly adequate. However, when counts are small such approximate techniques are less accurate and there is a need for better techniques based on discrete distributions for counts.

This paper investigates the fully Bayesian analysis of a time series of counts when the goal is to assess its association with an explanatory variable. The time series represented is the incidence of an infectious disease ESBL-producing (ESBL: extended-spectrum beta-lactamase) *Klebsiella pneumoniae* in an Australian hospital and the explanatory variable is the number of grams of antibiotic third generation cephalosporins used over that time period. ESBL-producing *K.pneumoniae* was first isolated in Australia in 1988 and is of serious concern because it responds only to the more expensive aminoglycoside agents

*To whom correspondence should be addressed

such as Amikacin and because the plasmid which encodes for the enzyme is transferable to the other members of the enterobacteriaceae.

In Australia, monotherapy with a third-generation cephalosporin is the preferred regime for severe community-acquired pneumoniae such as *Klebsiella*. However, it has been noted that a restriction in the use of antibiotics (especially third-generation cephalosporins) coincided with a reduction in the rate of infection. In this paper, we analyse the number of infections of ESBL-producing *K.pneumoniae* to allow for both a time dependence for the incidence of infection and the direct regression relationship between antibiotic use and infection.

By far the most popular model for analysing this type of count regression data is the Poisson regression model. However, it has been recognized by most researchers that alternative methods are necessary to account for the possibility of overdispersion and, particularly, serial correlation. Cox (1981) suggested that there were two classes of models for time-dependent data: observation-driven models and parameter-driven models. In observation-driven models, the conditional distribution of the response y_t is specified as a function of past responses $y_{t-1}, \ldots, y_1$. Observation-driven models have been discussed by Zeger and Qaqish (1988) and Li (1994). In parameter-driven models, autocorrelation is introduced via a latent process. Parameter-driven models in regression analysis have been discussed by Zeger (1988); Azzalini (1982); Stiratelli *et al.* (1984); Anderson and Aitkin (1985).

Another type of model which could be used in this context is the dynamic generalized linear model (DGLM). West *et al.* (1985) consider this model and, more recently, Shephard and Pitt (1997); Gamerman (1998) consider fully Bayesian analyses based on Markov chain Monte Carlo (MCMC) approaches. Although the DGLM is very general in formulation, the examples in Shephard and Pitt (1997); Gamerman (1998) are restricted to simple order-one autoregressive or random walk time series models. Additionally, for a scientific interpretation of the effect of covariates or explanatory variables, it is simpler, in our view, to use models with constant regression parameters rather than stochastically varying parameters, although the DGLMs are particularly useful to estimate the dynamic nature of relationships. Durbin and Koopman (1997) consider maximum likelihood estimation for the DGLM or state space class of models with non-Gaussian errors. They estimate the likelihood using simulation and give an analysis for an example involving low counts of road traffic accident deaths involving stochastically varying trend and seasonal parameters, and a time constant regression effect for an intervention variable. This model agrees with our view that, for ease of scientific explanation, time constant regression parameters are more appropriate.

As far as applied studies are concerned, the most popular time series regression model is the parameter-driven model of Zeger (1988). This model is commonly used as a model for daily health outcomes. Anderson *et al.* (1996); Pope *et al.* (1992); Schwartz and Dockery (1992a,b); Dockery *et al.* (1992); Schwartz (1993); Spix *et al.* (1993) all use this model to examine the effect of different types of air pollution on daily mortality. In addition to these studies, further applications are found in Campbell (1994); Brännäs and Johansson (1994) where the model is applied to data on daily incidences of sudden infant death syndrome and monthly Swedish traffic accident data, respectively.

To estimate the parameters of the parameter-driven model, Zeger (1988) employs a generalized estimating equation (GEE) approach. This is by far the most popular estimation method (see references cited in the previous paragraph). More recently, Chan and Ledolter (1995) suggest a Monte Carlo EM algorithm (MCEM) approach for the estimation of the parameters. In this paper, we provide the framework for a fully Bayesian analysis of a time series of counts. The general specification of the model is similar to the hierarchical model of Chan and Ledolter (1995); Diggle *et al.* (1998). To conduct the analysis, a MCMC technique (see for example Gilks *et al.*, 1995) is used. Unlike the GEE and MCEM approaches, the MCMC algorithm provides posterior distributions for both regression and time series parameters. Section 2 introduces the parameter-driven model for this type of data and examines alternative specifications of the model. Section 3 applies the methodology to the *Klebsiella* data set. Section 4 offers a

comparison of the models introduced in Section 3 using the mean square error. Some concluding remarks are made in the final section.

2. THE MODEL

The model that is used for this Bayesian analysis is similar to that used by Chan and Ledolter (1995). This model is a parameter-driven model with a random effect having a time series correlation structure. The model is formulated here as a Bayesian hierarchical model. In the first level of the model, the $(y_t|\mu_t, t = 1, \ldots, n)$ are assumed to be independently distributed as Poisson random variables with mean μ_t, giving

$$p(\mathbf{y}|\mu) = \prod_{t=1}^{n} \frac{\mathrm{e}^{-\mu_t}\mu_t^{y_t}}{y_t!}.$$

The second level of the model relates the conditional means $\mu = (\mu_1, \ldots, \mu_n)$ to the other regression effects and the time series random effects such that

$$\log(\mu) = \mathbf{X}\beta + \mathbf{w},$$

where $\mathbf{X}$ is the $n \times r$ model matrix, $\beta = (\beta_1, \ldots, \beta_r)$ is the vector of regression effects and $\mathbf{w} = (w_1, \ldots, w_n)$ forms an appropriately chosen time series process. If the time series process forms a stationary and invertible Gaussian ARMA(p, q) process (see Box and Jenkins, 1976) the process may be written in terms of a multivariate normal distribution:

$$\mathbf{w} \sim \mathbf{MVN}(\mathbf{0}, \Sigma)$$

where $\mathbf{0}$ is the $n \times 1$ mean vector of zeroes and Σ is the covariance matrix chosen to correspond to the time series process as required. Alternatively, one may write the process

$$\Phi_p(B)\mathbf{w} = \Psi_q(B)\mathbf{u}$$

where Φ_p is the $p \times 1$ vector of autoregressive coefficients, Ψ_q is the $q \times 1$ vector of moving average coefficients and $\mathbf{u} = (u_1, \ldots, u_n)$ consists of Gaussian white noise with variance σ^2. The form of this transformation is derived by using a Cholesky decomposition (see Graybill, 1983) or equivalently the Gram–Schmidt orthogonalization (see Fuller, 1995).

In the Bayesian framework the posterior for all unknowns is proportional to the product of the likelihood and prior distributions. The specification of this model is as follows:

$$\begin{aligned} p(\beta, \Phi_p, \Psi_q, \sigma, \mathbf{w}|y) \propto\ & p(y|\beta, \mathbf{w}, \Phi_p, \Psi_q, \sigma) \times p(\mathbf{w}|\Phi_p, \Psi_q, \sigma) \times p(\beta) \\ & \times p(\Phi_p, \Psi_q) \times p(\sigma) \end{aligned} \tag{1}$$

where we have assumed unknown parameters β, Φ_p, Ψ_q, and σ are *a priori* independent. MCMC methods (see, for example, Gilks *et al.*, 1995) are used to sample from the posterior distributions of the unknown parameters.

We note that this specialization to the Poisson distribution could be generalized by taking the distribution of y conditional on μ to be any distribution in the exponential family, such as binomial or gamma. We should also compare our approach with that of Gamerman (1998) who gives a fully Bayesian approach for Poisson time series. He considers

$$\log(\mu_t) = \beta_{1t} + x_t\beta_{2t}$$

and allows $\beta_t = (\beta_{1t}, \beta_{2t})$ to evolve according to a Gaussian random walk

$$\beta_t = \beta_{t-1} + a_t,$$

where $\{a_t\}$ is an independent Gaussian process.

The approach we take here does allow for stationary models so that marginal expectations of cross products of y can be related to the time series parameters. Shephard and Pitt (1997) consider a multiplicative Gaussian model with a simple lag-one autoregressive error in the multiplicative term. We note that Durbin and Koopman (1997), whose modelling approach is similar to Shephard and Pitt (1997), emphasize maximum likelihood estimates rather than finding posterior distributions. They give an example involving dynamical trend and seasonal effects for modelling a time series of counts.

3. CONTROL OF KLEBSIELLA PNEUMONIAE

In this section we examine a data set that was obtained from Morton *et al.* (1999). The *Klebsiella* data set consists of the monthly number of cases of ESBL-producing *K.pneumoniae* and the number of grams of third-generation cephalosporins used in the hospital from January 1992 to March 1998. ESBL-producing *K.pneumoniae* was first isolated in Germany in 1983 (see Knothe *et al.*, 1983) and since then there have been outbreaks in all parts of the world. The organism was first recognized in Australia in 1988. ESBL-producing *K.pneumoniae* is of serious concern because it responds only to the more expensive aminoglycoside agents such as Amikacin and because the plasmid which encodes for the enzyme is transferable to the other members of the enterobacteriaceae. At the Princess Alexandra Hospital ESBL-producing *Enterobacter, Citrobacter* and *Escherischia* have been isolated in small numbers. ESBL-producing *K.pneumoniae* was first isolated at the Princess Alexandra Hospital in December 1991.

Paterson and Playford (1998) suggest that monotherapy with a third-generation cephalosporin was the preferred regime for severe community-acquired pneumoniae such as *Klebsiella* in Australia. Morton *et al.* (1999) suggested that a restriction in the use of antibiotics (especially third-generation cephalosporins) coincided with a reduction in the rate of infection. This hypothesis has also been noted by Rahal *et al.* (1998). However, both of these investigations had provided high-quality data to investigate the relationship but had not quantified it using statistical methods. Here we analyse the data of Morton *et al.* (1999) to provide both a time dependence for the incidence of infection and the direct regression relationship between antibiotic use and infection. Both of these relationships are important to the management of hospitals to improve health of patients.

3.1 *Initial analysis*

In an initial analysis of the data we employed an independent log–linear model which regressed the monthly number of cases of ESBL-producing *K.pneumoniae* (hereafter referred to as *Klebsiella*) on the number of grams of third-generation cephalosporins (hereafter referred to as cephalosporin) at various time lags. We found that the most significant log–linear model was one that included the number of grams of cephalosporin lagged by 3 months. Estimates of standard errors of a subset of the examined models are given in Table 1. A plot of the number of *Klebsiella* infections and the number of grams of cephalosporin lagged by 3 months is given in Figure 1.

Although this model suggests that the number of grams of cephalosporin lagged by 3 months is statistically significant ($p \approx 0.036$), the residual mean deviance is still very high (210.65 on 70 degrees of freedom), suggesting that the model could be substantially improved. Morton *et al.* (1999) noted that *Klebsiella* is a contagious organism and therefore infections with it are not independent events. An examination of the sample partial autocorrelation function of the residuals from the log–linear model with

Table 1. *Parameter estimates and standard errors for a subset of models initially examined for the* Klebsiella *Data found from MLE for an independent log–linear model*

Parameter	Model A	Model B	Model C	Model D	Model E
Intercept	1.204 (0.311)	1.072 (0.311)	0.955 (0.395)	0.708 (0.414)	0.666 (0.327)
Ceph. $\times 10^{-4}$	0.806 (2.909)	−0.105 (3.285)	−1.017 (3.266)	−2.573 (3.600)	
Lag 1 Ceph. $\times 10^{-4}$		2.300 (3.233)	1.574 (3.444)	1.383 (3.459)	
Lag 2 Ceph. $\times 10^{-4}$			2.771 (3.426)	1.487 (3.556)	
Lag 3 Ceph. $\times 10^{-4}$				6.385 (3.387)	5.939 (2.829)

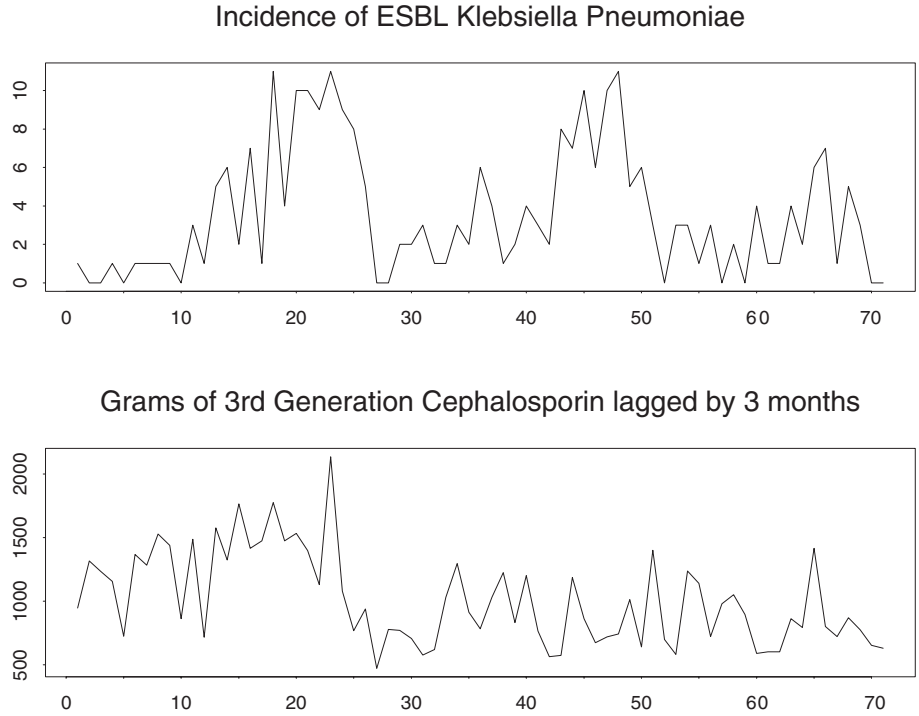

Fig. 1. Comparison of number of *Klebsiella* cases (y-axis uppermost figure) with the grams of 3rd generations cephalosporins (y-axis lowermost figure) used. The x-axis for both graphs is time in months from January 1992—March 1998.

the number of grams of cephalosporin lagged by 3 months as its covariate (see Figure 2) supports this assertion.

3.2 *AR(1) approach*

In the light of Figure 2, we used the Bayesian hierarchical model of Section 2 with the residuals modelled as a first-order autoregressive model. The autoregressive model of order one is one of the most commonly utilized time series models. Both Zeger (1988); Chan and Ledolter (1995) examine an $AR(1)$ process for their time series random effect, $\mathbf{w}$. Shephard and Pitt (1997) use this process in their stochastic volatility

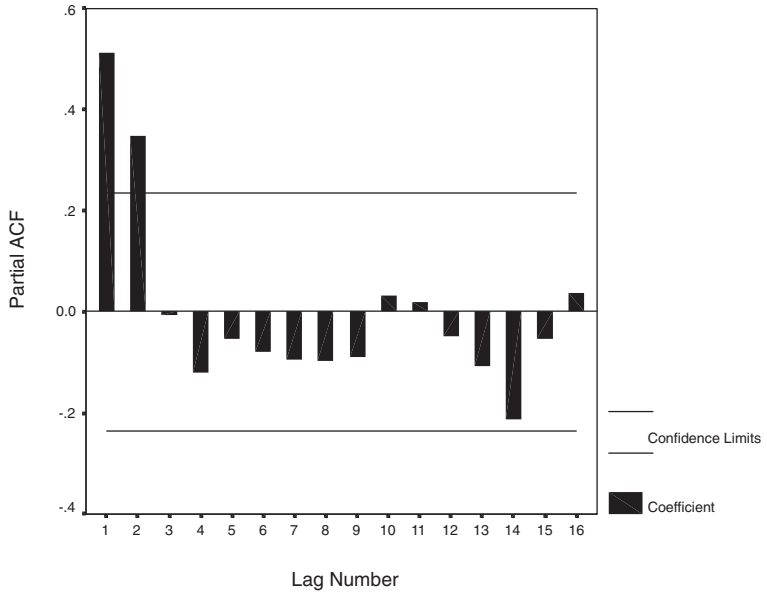

Fig. 2. Partial autocorrelation function of the residuals from the independent log–linear model with an intercept and the number of grams of cephalosporin lagged by 3 months as covariates. Confidence limits are ±2 standard errors.

model. The model can be written (see Judge *et al.*, 1980) as

$$w_1 = u_1 \Big/ \sqrt{1-\phi_1^2}$$
$$w_t = \phi_1 w_{t-1} + u_t, \, t > 2,$$

where $u_t \sim N(0, \sigma^2)$ for $t = 1, \ldots, n$ and ϕ_1 satisfies $|\phi_1| < 1$. In an initial run of the MCMC algorithm it was noticed that the intercept term was very unstable due to the aliasing effect between the intercept term and the autocorrelation parameter. This behaviour is accentuated when the autocorrelation parameter approaches 1 as is the case here. Since the priority of this analysis was to track the number of *Klebsiella* cases, it was decided to adopt an autoregressive smoothing approach in which the intercept term was omitted from the model. Thus the model can be written

$$y_t \sim \text{Poisson}(\mu_t)$$
$$\log(\mu_t) = w_t + x_t \beta,$$

and it can be shown straightforwardly that

$$w_1 \sim \text{Normal}(0, \sigma^2/(1-\phi_1^2))$$
$$w_t | w_{1,\ldots,t-1} \sim \text{Normal}(\phi_1 w_{t-1}, \sigma^2), \, t > 2.$$

Non-informative prior distributions were given for both β ($\beta \sim \text{Normal}(0, 10^6)$) and the precision parameter ($1/\sigma^2 \sim \text{Gamma}(0.001, 0.001)$) and a Uniform$(-1, 1)$ prior distribution was given for the autocorrelation parameter ϕ_1. It is clear from the results of the analysis that since the posterior variances of the parameters are so much smaller than the prior variances, the prior distributions have effectively no influence on the posterior distributions.

Table 2. *Parameter estimates, empirical standard errors and* 2.5 *and* 97.5% *posterior quantiles for the* Klebsiella *data*

Parameter	Posterior mean	s.e.	2.5%	97.5%
Lag 3 Ceph.$\times 10^{-4}$	6.899	0.123	1.729	11.060
ϕ_1	0.819	3.780×10^{-3}	0.584	0.965
σ	0.469	4.261×10^{-3}	0.297	0.709

To assess convergence of the algorithm, we used the Gelman and Rubin (1992) diagnostic. We ran four parallel chains started from four sets of overdispersed starting values. We ran the chain for 20 000 iterations, discarding the first 1000 iterations as burn-in. Four sets of chains are approximately 1, consistent with convergence. The estimated posterior means, Monte Carlo standard errors of simulation estimation (empirical standard errors) and the 2.5 and 97.5% quantiles are reported in Table 2.

We note that the regression coefficient for Lag 3 cephalosporin is estimated as $5.94(2.83) \times 10^{-4}$ (standard error in parentheses) using the Poisson log–linear model results of Table 1, whereas the posterior mean of this parameter from the Bayesian time series model of Table 2 is $6.90(2.36) \times 10^{-4}$ (posterior standard deviation in parentheses). Thus, the effect of Lag 3 cephalosporin is estimated to be larger and estimated more precisely when modelling the time series error explicitly.

3.3 *First- and second-difference approach*

An alternative approach to the analysis of the *Klebsiella* data retains a Poisson likelihood for the data, but employs a first- or second-difference smoother for the intercept, w_t in the model

$$y_t \sim \text{Poisson}(\mu_t)$$
$$\log(\mu_t) = w_t + \beta x$$

by assuming an autoregressive model for $\mathbf{w}$ such that either the first- or second-differences are independent normal variates. These models offer alternative prior specification for the random effect $\mathbf{w}$ in light of the high autocorrelation evident in both Figure 2 and Table 2.

For the first-difference model, a non-informative proper prior distribution is chosen for w_1 such that

$$w_1 \sim \text{Normal}(0, 10^6\sigma^2)$$

where σ is the standard deviation. For $t > 1$, prior distributions are chosen such that

$$w_t | w_{1,\ldots,t-1} \sim \text{Normal}(w_{t-1}, \sigma^2).$$

For the second-difference model, non-informative proper prior distributions are chosen for w_1 and w_2 such that

$$w_1 \sim \text{Normal}(0, 10^6\sigma^2)$$
$$w_2 | w_1 \sim \text{Normal}(0, 10^6\sigma^2),$$

where σ is the standard deviation. For $t > 2$, prior distributions are chosen such that

$$w_t | w_{1,\ldots,t-1} \sim \text{Normal}(2w_{t-1} - w_{t-2}, \sigma^2).$$

The MCMC algorithms were run for 100 000 iterates with the first 1000 iterates discarded as burn-in. The results of these analyses are given in Table 3.

Table 3. *Parameter estimates, empirical standard errors and* 2.5 *and* 97.5% *posterior quantiles using the first- and second-difference models. The symbol* † *denotes the results of the first-difference model*

Parameter	Posterior mean	s.e.	2.5%	97.5%
†Lag 3 Ceph.$\times 10^{-4}$	4.738	0.083	−0.656	0.001
$^{\dagger}\sigma$	0.387	1.233×10^{-3}	0.240	0.585
Lag 3 Ceph.$\times 10^{-4}$	3.843	0.115	−0.591	8.200
σ	0.133	1.215×10^{-3}	0.072	0.225

It can be seen from Table 3, that the $AR(1)$ approach and the first-difference approach yield similar results. Unlike the $AR(1)$ model, both the first- and second-difference models estimate the effect of Lag 3 cephalosporin to be slightly smaller, $4.74(2.67) \times 10^{-4}$ and $3.84(2.27) \times 10^{-4}$ respectively (posterior standard deviation in parentheses), than the independent log–linear model approach $5.94(2.83) \times 10^{-4}$ (standard error in parentheses). However, as for the $AR(1)$ approach, both difference models estimate the effect of Lag 3 cephalosporin more precisely than the independent log–linear model approach.

4. COMPARISON OF THE MODELLING APPROACHES

There are several ways to approach model comparison within a practical Bayesian analysis. Some approaches involve computation of a Bayes factor, cross-validatory statistics and deviance measures: see, for example, Spiegelhalter *et al.* (1995) or Key *et al.* (1999). Here we consider Bayesian variants on the mean square error (M) where

$$M = \sum_{t=1}^{n} (y_t - \mu_t)^2,$$

and μ_t is the posterior mean of $\exp(x_t\beta + w_t)$, which is the conditional mean of y_t given the random effect w_t and β. There are two points to consider here: the mean square error criterion and which posterior distribution to use. First, the mean square error criterion, rather than some weighted criterion, appears appropriate to use here since any loss in predicting numbers of cases would most likely be defined in terms of absolute numbers. Second, the posterior distribution of w_t can be defined in a number of ways, some taking into account the sequential time series nature of the data. For example, the simplest distribution is the full marginal distribution obtained from equation (1). We consider implementation of this below. On the other hand, in order to obtain a one-step-ahead predictive distribution for the model, we can consider the posterior distribution of $w_1, \ldots, w_{t-1}$ given data $y_1, \ldots, y_{t-1}$ and the parameter values. From this a predictive distribution of w_t can be determined using the time series structure and μ_t found either by using the full posterior of the parameter β based on all the data or one based only on the data up to time $t - 1$. We also consider implementation of this latter method below. For general ideas concerning these cross-validatory leave-out-k-cases methods, see Key *et al.* (1999), for example.

For the first method above, we evaluate $\exp(x_t\beta + w_t)$ at every iteration of the MCMC algorithm giving a sample from the posterior distribution of M, $(M^{(i)}, i = 1, \ldots, Z)$, where Z is the number of MCMC iterates used, from which we can derive the posterior mean and 2.5 and 97.5% quantiles of M.

For the second method we evaluate $\exp(x_t\beta + w_t)$ for t equal to $61, \ldots, 72$ in twelve separate MCMC runs. This is carried out by omitting in turn the last $12, \ldots, 1$ observations or observation, carrying out a separate posterior analysis using the remaining $(t - 1)$ observations and then generating the value of w_t from the resulting posterior. There are certainly much more efficient means of carrying this out but

Table 4. *Posterior means (p.m.) and* 2.5 *and* 97.5% *posterior quantiles for the mean square error (M) and and the predicted mean number of* Klebsiella *cases* ($\hat{y}_t$) *under the AR(1) approach and the first- and second-difference approaches.* M_P *gives the predictive mean square error for the three modelling approaches*

	$AR(1)$			First difference			Second difference		
Mean square error	p.m.	2.5%	97.5%	p.m.	2.5%	97.5%	p.m.	2.5%	97.5%
M	266.0	184.1	368.5	275.5	197.9	369.9	299.1	245.2	374.2
Observed value	$\hat{y}_t$	2.5%	97.5%	$\hat{y}_t$	2.5%	97.5%	$\hat{y}_t$	2.5%	97.5%
$y_{72} = 0$	1.789	0.414	4.766	2.103	0.512	5.359	1.907	0.398	4.853
$y_{71} = 0$	2.927	0.808	7.408	3.595	1.102	8.607	4.040	1.238	9.086
$y_{70} = 3$	3.618	1.004	9.292	4.304	1.343	10.27	5.221	1.730	11.47
$y_{69} = 5$	2.966	0.758	7.520	3.539	1.005	8.522	4.591	1.356	10.46
$y_{68} = 1$	4.623	1.291	12.07	5.605	1.818	13.64	8.195	2.931	17.91
$y_{67} = 7$	3.192	0.909	7.993	3.801	1.182	9.412	5.271	1.683	12.33
$y_{66} = 6$	3.526	0.840	9.604	3.099	0.788	8.213	3.500	0.784	9.386
$y_{65} = 2$	2.689	0.707	7.178	2.674	0.748	7.253	3.181	0.753	8.611
$y_{64} = 4$	1.843	0.395	5.170	1.679	0.376	4.616	1.511	0.256	4.469
$y_{63} = 1$	1.768	0.420	4.799	1.815	0.425	4.913	1.654	0.312	4.881
$y_{62} = 1$	2.235	0.534	6.415	2.314	0.567	6.306	2.199	0.439	6.320
$y_{61} = 4$	1.108	0.221	3.178	1.078	0.197	3.092	0.607	0.075	1.977
M_P	65.6			77.8			106.8		

we do not mention these here. We denote this one-step-ahead prediction of the mean of y_t based on data $y_1, \ldots, y_{t-1}$ by $\hat{y}_t$.

The predictive one-step-ahead mean square error M_P for each modelling approach is evaluated as the posterior mean of $\sum_{t=n-11}^{n} (y_t - \hat{y}_t)^2$. Table 4 gives the posterior means and 2.5 and 97.5% quantiles of the estimated posterior distributions of M and $(\hat{y}_t, t = 61, \ldots, 72)$, and the value of M_P for the three modelling approaches.

We note that the measure M_p automatically allows for both the model complexity and predictive error. It is clear from Table 4 that in terms of fitting and predicting the number of *Klebsiella* cases, the $AR(1)$ model is the most appropriate choice with the smallest M and M_P values of 266.0 and 65.6, respectively. However, the non-stationary first-difference model also fits the data quite well with M and M_P values of 275.5 and 77.8 respectively. The second-difference model seems inappropriate for these data. In terms of one-step-ahead prediction quantiles, we note that the 2.5–97.5% interval is generally more precise for the $AR(1)$ model.

To further demonstrate the fit of these models to the data, Figure 3 gives a plot of the posterior means of the fitted values, $E(y_t|x_t, w_t)$, together with the observed number of *Klebsiella* cases. All models appear to smooth the data well.

5. Discussion

We have demonstrated the scientific usefulness of various complex statistical modelling approaches by identifying a lagged relationship between disease occurrence and antibiotic use. This information should lead to improved health care provision in hospitals.

The Bayesian approach to the analysis of a time series of counts is a useful addition to the methods currently available. We have used WINBUGS for all applications of the MCMC algorithms (see `http:`

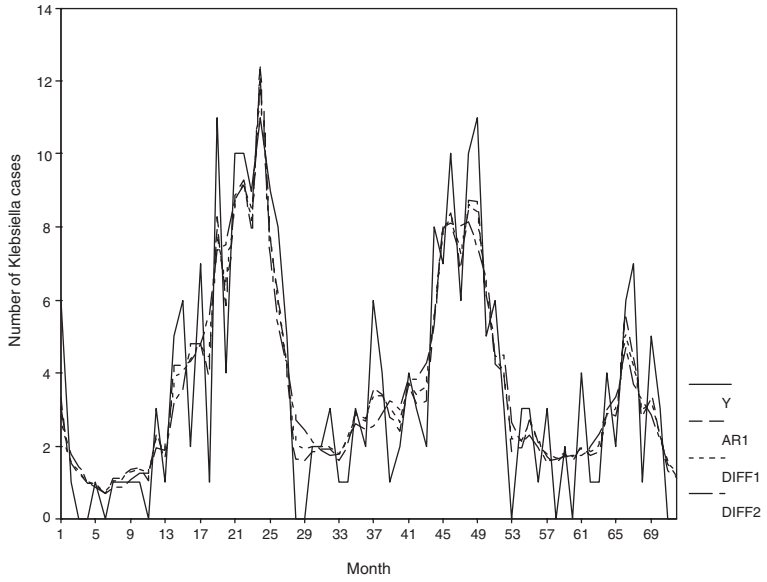

Fig. 3. Observed *Klebsiella* cases and posterior means of the fitted values from the $AR(1)$ approach and the first- and second-difference approach (DIFF1 and DIFF2 respectively).

`//www.mrc-bsu.cam.ac.uk/bugs` for details). For the *Klebsiella* data set, the popular Gelman and Rubin convergence diagnostic suggests that convergence of the Metropolis algorithm may be diagnosed after 20 000 iterations. In addition, the MCMC approach provides the analyst with posterior distributions for all parameters, rather than just maximum likelihood estimates for each parameter, and predictions taking into account the uncertainty.

Another advantage of the Bayesian approach is its flexibility. One may specify any time series model for the effects, $\mathbf{w}$. In addition to the $AR(1)$ and first- and second-difference approaches, seasonality may be modelled parsimoniously using using a seasonal time series effect. This seasonal approach provides the analyst with a useful alternative to the more traditional sinusoid or fixed effects approach and opens the way for these models to be extended to other modelling situations. In general, one could replace the time series structure by a more general cubic smoothing spline: see, for example, Hastie and Tibshirani (1990). This would allow quite a general but structural covariance pattern. The flexibility of this model is again demonstrated by the choice of the time series distribution. In this paper we have assumed that the time series process is Gaussian. However, one may assume other distributions for the time series process such as the Student-t distribution which might provide more robustness against observational (as against process) outliers.

In conclusion we suggest that in terms of implementation, flexibility and interpretation, an MCMC algorithm approach is an extremely good addition to the current approaches.

ACKNOWLEDGEMENTS

The work of J.L.H. was supported by an ARC Large Grant and an Australian Postgraduate Award. The authors are indebted to Dr Rodney Wolff and Dr Dawei Huang for references on time series and to Dr Kerrie Mengersen for references on MCMC diagnostics. We are particularly grateful to Dr Anthony Morton for making available to us the *Klebsiella* data set and discussions about the health care aspects of the disease infections.

REFERENCES

ANDERSON, D. A. AND AITKIN, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society* B **47**, 203–210.

ANDERSON, H. R., PONCE DE LEON, A., BLAND, J. M., BOWER, J. S. AND STRACHA, D. P. (1996). Air pollution and daily mortality in London: 1987–92. *British Medical Journal* **312**, 665–669.

AZZALINI, A. (1982). Approximate filtering of parameter driven processes. *Journal of Time Series Analysis* **3**, 219–223.

BOX, G. E. P. AND JENKINS, G. M. (1976). *Time Series Analysis Forecasting and Control*, Rev. edn. San Fransisco: Holden Day, pp. 73–82.

BRÄNNÄS, K. AND JOHANSSON, P. (1994). Time series count data regression. *Communications in Statistics - Theory and Methods* **23**, 2907–2925.

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

CHAN, K. AND LEDOLTER, J. (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association* **90**, 242–252.

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

DIGGLE, P. J., TAWN, J. A. AND MOYEED, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47**, 299–350.

DOCKERY, D. W., SCHWARTZ, J. AND SPENGER, J. D. (1992). Air pollution and daily mortality: association with particulates and acid aerosols. *Environmental Research* **59**, 362–373.

DURBIN, J. AND KOOPMAN, S. J. (1997). Monte Carl maximum likelihood estimation for non-Gaussian state space models. *Biometrika* **84**, 669–684.

FULLER, W. A. (1995). *Introduction to Statistical Time Series*, 2nd edn. New York: Wiley.

GAMERMAN, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* **85**, 215–227.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

GILKS, W. R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1995). Introducing Markov chain Monte Carlo. In Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds), *Markov Chain Monte Carlo in Practice*. pp. 1–19. London: Chapman and Hall.

GRAYBILL, F. A. (1983). *Matrices with Applications in Statistics*, 2nd edn. Belmont, CA: Wadsworth.

HASTIE, T. J. AND TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

JUDGE, G. G., GRIFFITHS, W. E., HILL, R. C., LÜTKEPOHL, H. AND LEE, T. C. (1980). *The Theory and Practice of Econometrics*. New York: Wiley, p. 441.

KEY, J. T., PERICCI, L. R. AND SMITH, A. F. M. (1999). Bayesian model choice: what and why? In Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Bayesian Statistics 6*, Oxford: University Press, pp. 343–370 (with discussion).

KNOTHE, H., SHAH, P., KREMERY, V., ANTAL, M. AND MITSUHASHI, S. (1983). Transferable resistance to cefotaxime, cefoxitin, cefamandole and cefuroxime in clinical isolates of *Klebsiella pneumoniae* and Serratia marcensens. *Infection* **11**, 315–317.

LI, W. K. (1994). Time series models based on generalized linear models: some further results. *Biometrics* **50**, 506–511.

MORTON, A. P., MCELWAIN, D. L. S., DOBSON, A., WHITBY, M., STACKELROTH, J. AND WILLS, C. (1999). Nosocomial infection surveillance: monitoring an outbreak of ESBL-producing *Klebsiella pneumoniae*. Report for the Department of Infectious Diseases, The Princess Alexandra Hospital.

PATERSON, D. L. AND PLAYFORD, E. G. (1998). Should third-generation cephalosporins be the empirical treatment of choice for severe community-acquired pneumonia in adults? *Medical Journal of Australia* **168**, 344–348.

POPE, C. A., SCHWARTZ, J. AND RANSOM, M. R. (1992). Daily mortality and PM10 pollution in Utah Valley. *Archives of Environmental Health* **47**, 211–217.

RAHAL, J. J., URBAN, C., HORN, D., FREEMAN, K., SEGAL-MAURER, S., MAURER, J., MARIANO, N., MARKS, S., BURNS, J. M., DOMINICK, D. AND LIM, M. (1998). Class restriction of cephalosporin use to control total cephalosporin resistance in nosocomial *Klebsiella*. *Journal of the American Medical Association* **280**, 1233–1237.

SCHWARTZ, J. (1993). Air pollution and daily mortality in Birmingham, Alabama. *American Journal of Epidemiology* **137**, 1136–1147.

SCHWARTZ, J. AND DOCKERY, D. W. (1992a). Increased mortality in Philadelphia associated with daily air pollution concentrations. *American Review of Respiratory Disease* **145**, 600–604.

SCHWARTZ, J. AND DOCKERY, D. W. (1992b). Particulate air pollution and daily mortality in Steubenville, Ohio. *American Journal of Epidemiology* **135**, 12–19.

SHEPHARD, N. AND PITT, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84**, 653–657.

SPIEGELHALTER, D., THOMAS, A., BEST, N. AND GILKS, W. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling*. Cambridge: Medical Research Council Biostatistics Unit, p. 42.

SPIX, C., HEINRICH, J., DOCKERY, D. W., SCHWARTZ, J., VOLKSCH, G., SCHWINKOWSKI, K. *et al.* (1993). Air pollution and daily mortality in Erfurt, East Germany from 1980–1989. *Environmental Health Perspectives* **101**, 518–526.

STIRATELLI, R., LAIRD, N. AND WARE, J. H. (1984). Random effects models for serial observations with binary response. *Biometrics* **40**, 961–971.

WEST, M., HARRISON, J. P. AND MIGON, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association* **80**, 73–96.

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

ZEGER, S. L. AND QAQISH, B. (1988). Markov regression models for time series: a quasilikelihood approach. *Biometrics* **44**, 1019–1031.

[*Received 28 October, 1999; revised 28 February, 2000; accepted for publication 28 April, 2000*]