Biostatistics (2004), **5**, 2, *pp.* 223–237
Printed in Great Britain

The analysis of hospital infection data using hidden Markov models

BEN COOPER[†]
Department of Epidemiology, Kresge Building, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA
ben.cooper@hpa.org.uk

MARC LIPSITCH
Department of Epidemiology, Kresge Building, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA

SUMMARY

Surveillance data for communicable nosocomial pathogens usually consist of short time series of low-numbered counts of infected patients. These often show overdispersion and autocorrelation. To date, almost all analyses of such data have ignored the communicable nature of the organisms and have used methods appropriate only for independent outcomes. Inferences that depend on such analyses cannot be considered reliable when patient-to-patient transmission is important.

We propose a new method for analysing these data based on a mechanistic model of the epidemic process. Since important nosocomial pathogens are often carried asymptomatically with overt infection developing in only a proportion of patients, the epidemic process is usually only partially observed by routine surveillance data. We therefore develop a 'structured' hidden Markov model where the underlying Markov chain is generated by a simple transmission model.

We apply both structured and standard (unstructured) hidden Markov models to time series for three important pathogens. We find that both methods can offer marked improvements over currently used approaches when nosocomial spread is important. Compared to the standard hidden Markov model, the new approach is more parsimonious, is more biologically plausible, and allows key epidemiological parameters to be estimated.

Keywords: Count data; Hidden Markov models; Hospital epidemiology; Interrupted time series; SIS epidemic model; Time series.

1. INTRODUCTION

Hospital-acquired infections caused by transmissible nosocomial pathogens can severely detriment patient welfare and place large burdens on health-care resources (Plowman *et al.*, 1999). They may also present a headache for the analyst. Data usually consist of short time series of low-numbered counts. Often these time series display overdispersion and autocorrelation.

Molecular typing shows that for many important nosocomial pathogens, such as methicillin-resistant *Staphylococcus aureus* (MRSA) and vancomycin-resistant enterococci (VRE), new cases arise through

[†]To whom correspondence should be addressed. Current address: Statistics, Modelling and Economics Unit, 61 Colindale Ave, London NW9 5EQ, UK.

patient-to-patient transmission (thought to be largely mediated by contacts from transiently-colonized health-care workers) rather than by the acquisition of resistance during antibiotic treatment via spontaneous mutation. The majority of patients harbouring such organisms carry them asymptomatically, with overt infections developing in only a proportion of patients. Consequently, the underlying transmission process is usually only partially observed.

Analyses of such time series data may serve a number of purposes. Often the main interest lies in the regression problem, for example relating infection incidence to staffing levels or antibiotic usage data. Other applications include forecasting and the development of alert systems to detect periods or places where transmission exceeds some threshold (Brown *et al.*, 2002). At a more basic level, the object may be to estimate important epidemiological parameters from the time series data. This latter application is the main focus of our paper, though the method we propose has potential applications in the other domains.

To date, most analyses of hospital infection time series data have been rudimentary. For example, the most common application of formal statistical methods has been in the analysis of interrupted time series (ITS) studies, the principal tool for evaluating interventions intended to control hospital infections. A recent systematic review of the use of isolation policies for controlling MRSA found that of 24 ITS studies presenting any statistical analysis, all but one used methods that assumed outcomes to be independent (Cooper *et al.*, 2003).

The assumption that outcomes are independent is untenable for a communicable disease under most circumstances, and more appropriate time series models are clearly required. Recently, traditional autoregressive integrated moving average (ARIMA) time series models have been applied to hospital infection data (Lopez-Lozano *et al.*, 2000). However, ARIMA models, which assume continuous outcomes, will be of limited value when outcome data are in the form of low-numbered counts (Cardinal *et al.*, 1999). Only when the counts are large is the continuous approximation likely to be justified.

As pointed out by Zeger (1988), with a single time series valid inferences depend on the correct specification of the time series model. A model based on a mechanistic understanding of the transmission process would therefore seem a natural choice. However, while appropriate methods for analysing infectious disease spread in small populations have been developed (Becker, 1989), these are not applicable to most hospital infections. The rapid turnover of hospitalized patients and the asymptomatic nature of carriage render methods designed for typical community pathogens inappropriate.

A number of transmission models for hospital pathogens have been suggested (Sébille *et al.*, 1997; Austin *et al.*, 1999; Cooper *et al.*, 1999; Lipsitch *et al.*, 2000). Recently Pelupessy *et al.* (2002) proposed using a simplified version of these for the analysis of transmission data in a hospital ward. However, this Markov model assumed a sequence of whole-ward surveillance swabs capable of detecting carriage with certainty. Such data are rarely available in practice.

A number of generic approaches have also been used or suggested for modelling time series of counts. Cameron and Trivedi (1998, chapter 7) and MacDonald and Zucchini (1997, chapter 1) provide recent reviews. Following Cox (1981), a distinction is often made between observation-driven models, where the model for the response at each time point explicitly conditions on responses at earlier times, and parameter-driven models, where serial dependence arises through a latent (unobserved) process. In the present context, serial dependence is thought to arise through a largely unobserved transmission process (since both asymptomatic and symptomatic/infected patients may transmit the organism) and parameter-driven models would seem the more appropriate. One possible approach that has been applied to infectious disease data in a regression context is to use estimating equations, treating autocorrelation terms as nuisance parameters (Zeger, 1988). In contrast, we are particularly interested in making inferences about the unobserved process generating the autocorrelation. We show how this can be done using a hidden Markov model to explicitly describe the latent process.

Section 2 of this paper first describes how the usual (unstructured) hidden Markov model can be used to model a time series of counts of infected patients. We then present the structured hidden Markov

model in Section 3 where the underlying Markov chain models the unobserved epidemic process and is constructed using a few standard epidemiological parameters. We then compare the fits of these hidden Markov models (together with a simple Poisson model) to three time series for major nosocomial pathogens in Section 4. We consider limitations and possible extensions of the current approach in the discussion.

2. Hidden Markov models

We consider a sequence of observations, y_t, made at times $t = 1, 2, \ldots, n$, where y_t represents the observed number of new infections occurring during the interval $(t-1, t]$, and Y_t is the corresponding discrete random variable. At this stage we ignore covariates and assume that denominators remain constant throughout and can be neglected.

In a hidden Markov model there is an underlying unobserved state of the system that changes in time according to a Markov process. The distribution of observations at a given time is determined by the system's state at that time (MacDonald and Zucchini, 1997).

Let C_t $(t = 1 \ldots n)$ represent an irreducible Markov chain on state space $1, \ldots, m$, where the Markov chain is defined by the transition matrix $\Gamma = (\gamma_{ij})$. Then we associate an observation model with each state. Thus, for example, with a Poisson observation model we have

$$\Pr(Y_t = s|C_t = i) = \mathrm{e}^{-\lambda_i}\lambda_i^s/s!. \tag{2.1}$$

Fitting a hidden Markov model involves both estimating the elements of the transition matrix, $\gamma_{ij} = \Pr(C_t = j|C_{t-1} = i)$, for the Markov chain, and the parameters, λ_i, of the observation model. The naïve approach to evaluating the likelihood involves summing over all possible sequences of states. This has complexity $\mathrm{O}(m^n)$, and will be computationally infeasible for all but the simplest problems. Instead the likelihood can be rewritten and evaluated far more efficiently. Thus, putting

$$\pi_t(s) = \mathrm{diag}(\Pr(Y_t = s|C_t = 1), \Pr(Y_t = s|C_t = 2), \ldots, \Pr(Y_t = s|C_t = m)), \tag{2.2}$$

the likelihood of the observed data, L, can be written as

$$L = \delta\pi_1(y_1)\Gamma\pi_2(y_2)\Gamma \ldots \Gamma\pi_n(y_n)\mathbf{1}' \tag{2.3}$$

where δ is the probability distribution of C_1 (MacDonald and Zucchini, 1997). Throughout this paper δ is taken to be the stationary distribution of the Markov chain. This expression can be evaluated in $\mathrm{O}(nm^2)$ time.

By log and logit transforming parameters taking values on the positive real line and the (0,1) interval respectively maximum likelihood estimation can be accomplished using an algorithm for unconstrained numerical maximization. An alternative approach is to use an expectation-maximization (EM) algorithm (Baum *et al.*, 1970; Le Strat and Carrat, 1999; Bureau *et al.*, 2003).

If C_t is stationary (the existence of a unique strictly positive stationary distribution is guaranteed by the chain's irreducibility), with the stationary distribution defined by state probabilities δ_i, then for a Poisson observation model we have

$$\mathrm{E}[Y_t] = \sum_{i=1}^{m} \delta_i\lambda_i \tag{2.4}$$

and for a two-state model

$$\mathrm{var}[Y_t] = \mathrm{E}[Y_t] + \delta_1\delta_2(\lambda_2 - \lambda_1)^2, \tag{2.5}$$

showing that provided $\lambda_1 \neq \lambda_2$ the y_t will be overdispersed. In general, this model will produce overdispersion provided that conditional means depend on the state; for most transition matrices there will also be autocorrelation (MacDonald and Zucchini, 1997).

Advantages of this approach include the flexibility of hidden Markov models, and their ability to cope with missing observations and unequally spaced time intervals.

Hidden Markov models have previously been used for the study of infectious diseases when individuals can be in a number of imperfectly observed states (Satten and Longini, 1996; Bureau *et al.*, 2003; Smith and Vounatsou, 2003). Of more relevance to the current problem is the use of hidden Markov models in the analysis of infectious disease surveillance data. In this context, Le Strat and Carrat (1999) proposed that in a two-state hidden Markov model the different states or regimes could correspond to either 'epidemic' or 'non-epidemic' periods, though no mechanistic interpretation was attached to the underlying Markov chain. In the higher-dimension models (which gave much better fits to their influenza-like illness data) no meaningful interpretations were attached to different states. Applied to hospital infection data, such non-epidemic periods could perhaps be interpreted as periods when *fade-out* of the pathogen in a given setting had occurred (i.e. when ward-level prevalence has fallen to zero, breaking the chain of transmission).

3. A STRUCTURED HIDDEN MARKOV MODEL

This approach represents a modification of the standard hidden Markov model, motivated by mechanistic considerations. In this case a continuous time Markov chain is derived from a dynamic transmission model for the epidemic process in a single hospital ward or unit with N patients. This Markov chain is constructed in terms of biologically meaningful parameters: β/N gives the rate of transmission to each susceptible patient per colonized or infected patient; ν, the probability that each patient is already carrying the organism when admitted to the ward; μ, the rate at which patients are discharged from the ward. The state of the system, $C_t \in \{0, 1, 2, \ldots, N\}$, is interpreted as the number of infected or colonized patients on the ward at time t.

Explicitly, if N gives the total patient population, then for a small time increment, h, we use a continuous time Markov chain with transition probabilities:

$$\Pr[C_{t+h} = i + 1 | C_t = i] = \begin{cases} \frac{\beta}{N} i(N - i)h + \nu(N - i)\mu h + \mathrm{o}(h) & \text{if } i < N \\ 0 & \text{if } i = N \end{cases} \tag{3.1}$$

$$\Pr[C_{t+h} = i - 1 | C_t = i] = \begin{cases} (1 - \nu) i \mu h + \mathrm{o}(h) & \text{if } i > 0 \\ 0 & \text{if } i = 0. \end{cases} \tag{3.2}$$

The balance of probabilities is given to there being no change of state between t and $t + h$. All other transitions have probability $\mathrm{o}(h)$.

The model corresponds to a reparametrization of the stochastic susceptible–infectious–susceptible (SIS) epidemic model with immigration; see, for example, Bailey (1975, chapter 7). In this model, new cases arise due to cross-infection at a rate proportional to the product of the number of infected or colonized hosts, C_t, and the number of susceptible hosts, $N - C_t$. This is known as the mass action assumption, and would be true for a homogeneous randomly mixing population. New cases can also arise due to susceptible patients being discharged (which occurs with rate $(N - C_t)\mu$) and immediately replaced by patients who are infected or colonized on admission. Decreases in the number of colonized patients occur only through the discharge of colonized or infected patients. Explicit consideration of the rate of loss of carriage can be neglected as the duration of carriage for most important nosocomial bacterial pathogens is long compared to typical lengths of stay. The $1/N$ term appears in equation (3.1) since for a

contact-transmitted pathogen we do not expect the total instantaneous transmission rate to increase with N, unless patient contact rates also increase with N (and this seems unlikely).

If P_h is the tridiagonal matrix of probabilities constructed from the expressions above with ijth element given by $\Pr(C_{t+h} = j|C_t = i)$, then the generator matrix, G, for the Markov process is given by

$$G = \lim_{h \downarrow 0} \frac{1}{h}(P_h - I). \tag{3.3}$$

It can be shown that the matrix Γ_t with elements $\gamma_{ij}^t = \Pr(C_{t+a} = j|C_a = i)$, (where a is arbitrary as the Markov process is time homogeneous), is given by $\Gamma_t = \exp(tG)$ (Grimmet and Stirzaker, 1992). If all observations are separated by a time interval, t, the likelihood will be given by equation (2.3), replacing Γ with Γ_t. The extension to data with unequal intervals between observations is straightforward.

Effectively, the underlying Markov model is just an alternative parametrization of the model proposed by Pelupessy *et al.* (2002). If a sequence of whole-ward surveillance swabs were available and assumed to be capable of detecting carriage with certainty then the standard Markov model could be used. In our case, since only infection data are used, the states, C_t, are not observed, and the hidden Markov model is required.

The state of the system (the number of patients harbouring the organism) then determines the conditional distribution of Y_t (the number of observed infections), where $\mathrm{E}[Y_t|C_t]$ should increase with C_t.

A natural choice, and the one adopted throughout this paper, is to assume a Poisson distribution with mean λC_t, i.e.

$$\Pr[Y_t = y|C_t = i] = \frac{\mathrm{e}^{-\lambda i}(\lambda i)^y}{y!}, \tag{3.4}$$

where λ is to be estimated but is the same for all states. Mechanistically, this can be derived if we consider the prevalence to be approximately constant during the interval over which infection counts are aggregated, and consider there to be some small chance of an infection developing for each colonized patient day. Unequal intervals of aggregation can be accommodated by appropriate scaling of the parameter λ.

This approach has all the advantages of the standard hidden Markov model; in particular, overdispersion and autocorrelation are produced as before. In addition it is more parsimonious, particularly compared to the higher-dimension hidden Markov models. Thus, we may have a 21-state Markov process (for a 20-bed ward) with only four parameters. Of these, the hospital discharge rate, μ, can usually be observed with certainty (as assumed here). It may also be possible to estimate the probability patients are colonized on admission, ν, from other data sources in many applications. The other major advantage of this approach is that it is based on a mechanistic understanding of the data generation process and estimated parameters are of interest in themselves, and often able to provide insight into the population processes underlying the observed data.

Another difference between this approach and conventional hidden Markov models relates to identifiability. The standard hidden Markov model is not affected by a permutation of the hidden states, so for an n-state model there will will be $n!$ equivalent solutions. In the structured model, because the observation model depends explicitly on the labelling of the states, this will not be the case.

4. Model evaluation

Data used to evaluate the competing models were taken from the ICARE (Intensive Care Antimicrobial Resistance Epidemiology) project, a surveillance study of intensive care units (ICUs) at 41 U.S. hospitals

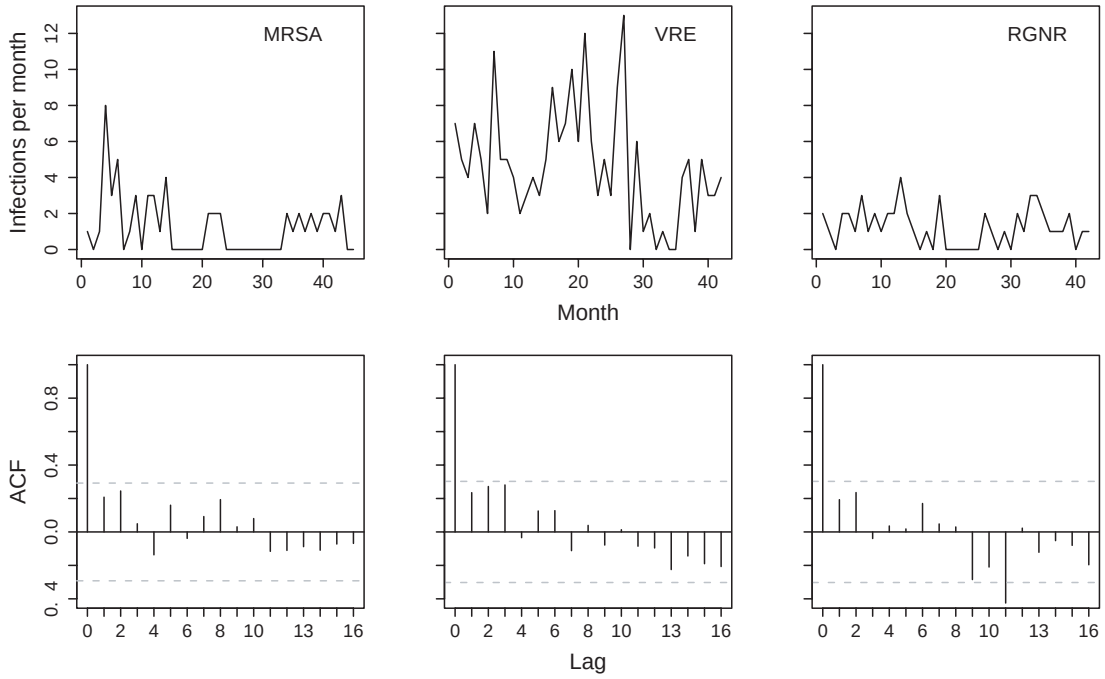

Fig. 1. Monthly infection data used for initial model assessment.

(Fridkin *et al.*, 1999). Participating centres recorded the monthly counts of non-duplicate clinical isolates of major nosocomial pathogens. We restricted our attention to time series for three classes of pathogens: methicillin-resistant *Staphylococcus aureus* (MRSA); vancomycin-resistant enterococci (VRE); and third generation cephalosporin-resistant Gram-negative rods (R-GNR). Most of these time series were very short, had gaps, or reported very low numbers of infections. We considered only those time series with at least 40 contiguous months of data and an average of at least one case per month for the given pathogen. To simplify the analysis we also excluded time series showing clear trends (assessed by regressing counts on time in an autoregressive negative binomial model). Three time series met these criteria, one for each pathogen. The MRSA data came from a ten-bed ward over 45 months, and the VRE and R-GNR data from the same 16-bed ward over 42 months. Mean (SD) patient days per month were 257.2 (23.5) and 338.0 (43.4) respectively. These time series are shown in Figure 1 together with their correlograms.

For each of the three time series of counts we fit three different models: a simple Poisson model (corresponding to the current practice in the hospital infection literature of assuming outcomes to be independent), a standard (unstructured) hidden Markov model using a Poisson observation model (Poisson HMM); and the structured hidden Markov model based on the SIS epidemic model (SIS HMM) assuming a mean ICU stay of 8 days.

All analyses were performed using *R 1.5.0*, a free open-source statistical package (Ihaka and Gentleman, 1996). In particular, the structured hidden Markov model was implemented with Fortran and *R* routines which built on existing code due to MacDonald and Zucchini (1997) and Lindsey (1999). Likelihood maximization was accomplished using the Newton-type algorithm implemented in *R*'s *nlm* function (Schnabel *et al.*, 1985). Variance estimates for the fitted parameters were obtained by inverting the Hessian.

We assessed model fit using two approaches. First, we report the Akaike information criterion (AIC),

Table 1. *AICs and parametric bootstrap results for models fit to the three time series*

				Bootstrap results	
				var/mean	mean run length
Data	Model	df	AIC	2.5–97.5%	2.5–97.5%
MRSA (*n*: 45)	Poisson	44	155.70	0.63–1.46	1.18–1.67
var/mean: 2.20	Poisson HMM	41	135.51	1.09–2.63	1.29–3.01
mean run length: 1.73	SIS HMM	42	132.03	1.14–3.11	1.22–3.75
VRE (*n*: 42)	Poisson	41	229.47	0.62–1.46	1.02–1.31
var/mean: 2.20	Poisson HMM	38	212.36	1.39–3.65	1.02–1.36
mean run length: 1.08	SIS HMM	39	210.59	1.24–3.27	1.02–1.36
R-GNR (*n*: 42)	Poisson	41	119.73	0.64–1.55	1.17–1.75
var/mean: 0.90	Poisson HMM	38	123.37	0.77–2.06	1.20–2.80
mean run length: 1.35	SIS HMM	39	122.40	0.75–1.90	1.17–2.00

which we take as minus twice the log likelihood plus twice the number of estimated parameters (Akaike, 1973). Second, we used a parametric bootstrap method, as proposed by Tsay (1992). Following Grunwald *et al.* (2000), we used the following diagnostics: the positive/negative ratio, which is used to assess time-reversibility; the variance to mean ratio, to assess overdispersion; and skewness. We also used a fourth diagnostic, the mean run length (i.e. the mean length of runs of equal counts). The choice of this statistic was motivated in part by a consideration of the importance of stochastic fade-out in epidemics in small populations (Bailey, 1975). For each fitted model we simulated 1000 data-sets of monthly infections from that model, calculated 2.5th and 97.5th percentiles for each statistic, and compared these with values observed for the actual time series.

4.1 *Results of model evaluation*

Table 1 shows AICs and bootstrap results for the models applied to the three time series. For the unstructured hidden Markov model the dimension, m, of the state space of reported models was chosen to be the smallest integer to satisfy $m \geqslant 2$ and for which no reduction in the AIC was seen for the corresponding model with dimension $m + 1$. For all three time series this criterion was met by a state space of dimension 2.

From a comparison of the AIC results, a simple Poisson model is clearly inadequate for both the MRSA and VRE time series. The structured hidden Markov model gives the best fit to the MRSA data. For the VRE data this model is only slightly better than the standard hidden Markov model. The R-GNR data are different in that there is no overdispersion; the simple Poisson model appears to provides an adequate fit to the data. Compared to the Poisson model, no reductions in AIC are seen for other models. It is worth noting that while MRSA and VRE are known to be highly contagious in hospital settings, the role of patient-to-patient spread for resistant Gram-negative rods is more ambiguous. While molecular typing has demonstrated nosocomial transmission in some settings (Almuneef *et al.*, 2001) a number of other studies where strains have been typed have found little evidence of significant horizontal transmission, with most infections believed to be caused by endogenous flora (Lafaix *et al.*, 1969; Flynn *et al.*, 1988; Baquero *et al.*, 2002).

Note that for such short time series of low-numbered counts, correlograms may be of limited use in model identification. Thus, while correlograms for R-GNR and MRSA are almost identical (Figure 1) there is a great difference in the fit of competing models for these two time series. We also found that

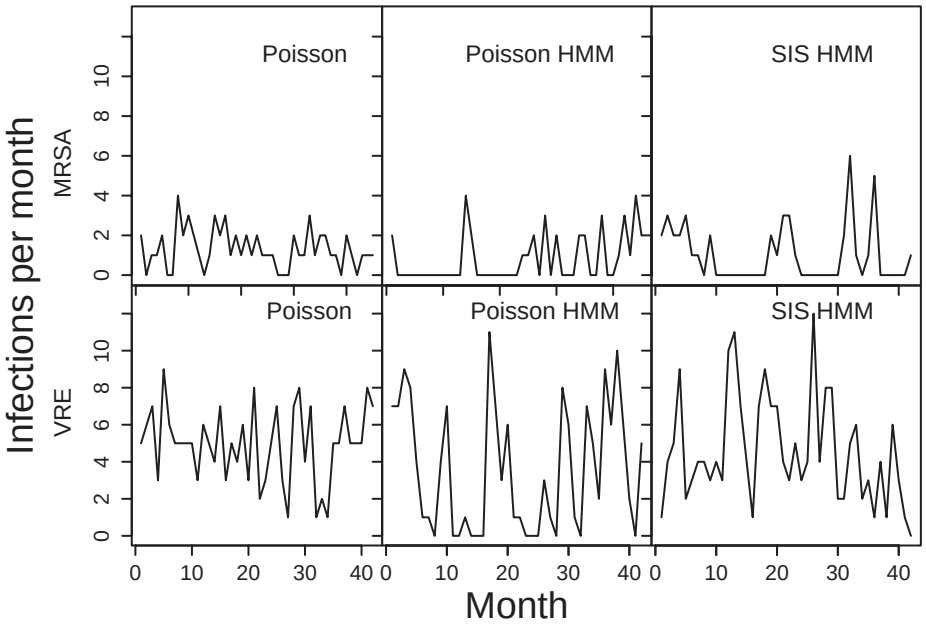

Fig. 2. Examples of simulated data using models fit to the monthly MRSA and VRE data.

simulated data from the structured hidden Markov model (using parameters obtained by fitting to the VRE and MRSA data) produced qualitatively similar correlograms to those in Figure 1, and rarely produced autocorrelation coefficients significant at the 5% level.

4.2 *Bootstrap results*

For both the positive/negative ratio and skewness (results not shown), actual values were found to be consistent with bootstrap results from all the fitted models (i.e. were within the 2.5–97.5 percentiles).

The bootstrap results for variance/mean and mean run length (shown in Table 1) suggest that the Poisson model is unlikely to be able to generate data consistent with the MRSA and the VRE time series. For the R-GNR time series all models were able to generate data consistent with the observed time series.

Figure 2 shows sample simulations from the models fitted to the MRSA and VRE data. The simulations of the MRSA data illustrate that the hidden Markov models are capable of modelling extended periods of fade-out, a feature that is apparent in the actual data. Most other models for time series of counts are not able to capture this behaviour, and in this sense the correspondence of the hidden states of unstructured Markov chains to epidemic and non-epidemic periods seems to have real value. However, when applied to the VRE data, simulations from the fitted Poisson hidden Markov models have an indication of a rather artificial alternation between periods of high and low incidence without an intermediate level. This is not seen in simulated data from the structured hidden Markov model.

4.3 *Estimation using structured hidden Markov models*

Table 2 shows estimated parameters for the structured hidden Markov model (SIS HMM) for the three time series. These suggest similar values for the transmission rate, β, in all three time series. However, the confidence intervals for β and ν obtained from variance estimates of log and logit transformed parameters

Table 2. *Parameter estimates [95% confidence intervals] for the structured hidden Markov model applied to the three time series. Units for rates are days*$^{-1}$

	$\hat{\beta}$ (transmission rate)		$\hat{\nu}$ (+ve on admission)		$\hat{\lambda}$ (infection rate)	
MRSA	0.329	[0.233, 0.465]	0.009	[0.002, 0.035]	0.349	[0.209, 0.490]
VRE	0.255	[0.177, 0.368]	0.028	[0.006, 0.128]	0.668	[0.448, 0.888]
R-GNR	0.323	[0.222, 0.470]	0.008	[0.001, 0.073]	0.150	[0.084, 0.216]

can be misleading, as the contour plots of likelihood surfaces in Figure 3 show. For the R-GNR data, where parameters may be on the bounds of the parameter space, the asymptotic approximation to the distribution of their estimators will not apply. A better approach might be to derive confidence intervals using a parametric bootstrap approach.

The contour plot for the R-GNR data also indicates an identifiability problem: the large plateau region shows that the data are not able to distinguish between a high transmission rate (high β) with little carriage on admission (low ν) and the converse. Such collinearity of parameter estimates is common in biologically plausible models (Brookhart *et al.*, 2002).

One natural question of interest is whether the observed data are consistent with the null hypothesis that there is no transmission ($\beta = 0$). We can test this against the alternative hypothesis, $\beta > 0$, using the likelihood ratio test for nested models. Setting $\beta = 0$ in the structured hidden Markov model gives a decrease in twice the log-likelihood of 22.90, 11.32, and 1.32 for the VRE, MRSA and R-GNR data respectively, with corresponding p-values (based on the chi-squared distribution, 1 d.f.) of 0.0000017, 0.00077 and 0.25. We therefore have no evidence to reject the assumption that transmission is not important for the R-GNR data, but strong evidence for rejecting the null in both the VRE and MRSA time series. This result is in accord with the earlier finding that the Poisson model provided an adequate description of the R-GNR data but could not account for the VRE or the MRSA data. Other causes of non-independence of outcomes for the MRSA and VRE cannot be ruled out, but given that these pathogens are known to be highly infectious in settings with high antibiotic use, patient-to-patient transmission is the most economical explanation.

The mass action mixing assumption used in the underlying Markov chain may be questioned. This assumption can, in fact, be relaxed and the model generalized to other mixing assumptions. For example, each susceptible patient could be allowed to become colonized at a rate given by $\frac{\beta}{N} C_t^{\kappa}$. For $\kappa = 1$ this reverts to the mass action assumption, while $\kappa = 0$ (for $C_t > 0$) gives a continuous time analogue of the Greenwood assumption (see, for example, Becker (1989)), where the chance that a susceptible patient becomes colonized in any small time interval does not change with the prevalence providing that there is at least one colonized patient. The Greenwood assumption would be appropriate if a single colonized patient caused saturated exposure amongst other patients. For $0 < \kappa < 1$ the model would allow for a decreasing contribution to the hazard rate for transmission per infectious source with each additional colonized patient. Values of κ greater than one would imply synergistic effects amongst the colonized patients. Though mechanisms for such effects can be conceived, this possibility seems biologically rather implausible. For the MRSA and VRE data the maximum likelihood estimates for κ were 2.67 and 2.11 respectively, but a likelihood ratio test gave no evidence for rejecting the null hypothesis that $\kappa = 1$ ($\chi_1^2 = 1.61$ for the VRE data and $\chi_1^2 = 0.43$ for the MRSA data), so we did not include this parameter in the final models. Furthermore, the Greenwood mixing assumption gave a lower log likelihood than the mass action model for both the VRE and MRSA time series (the differences in twice the log likelihood being 1.73 and 11.96 respectively)

Since the structured Markov models have a mechanistic interpretation, the underlying stationary

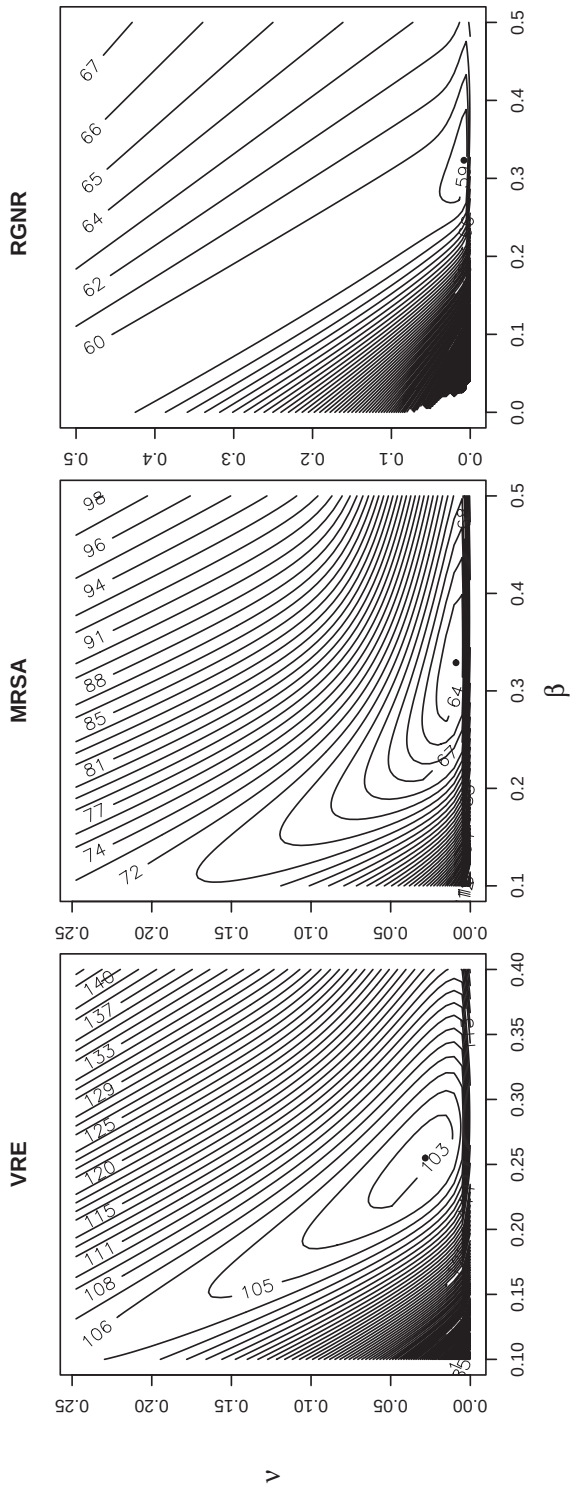

Fig. 3. Likelihood profiles for the structured hidden Markov (SIS HMM) model.

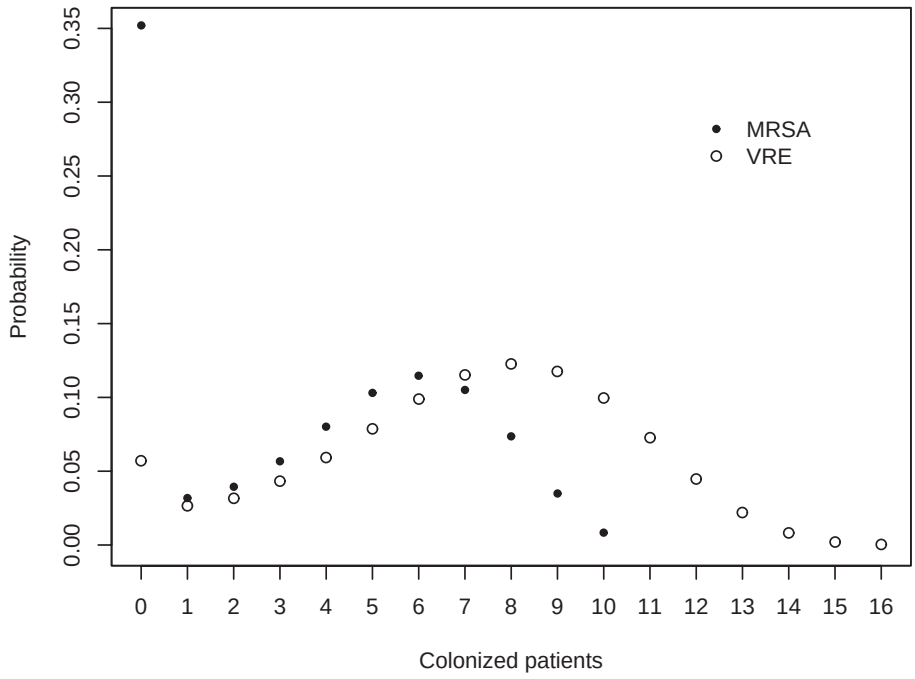

Fig. 4. Stationary distribution of colonized patients in the structured hidden Markov model fit to MRSA and VRE data.

distribution of the Markov chain is of interest in itself and can be interpreted as the equilibrium distribution of the ward-level prevalence. Stationary distributions for the chains estimated for the MRSA and VRE time series are shown in figure 4. The distribution for the MRSA data is more skewed to the right due to the higher transmissibility. The much higher probability of there being zero colonized patients in the MRSA distribution reflects both the lower introduction rate, ν, and the smaller ward size which makes stochastic fade-outs more likely and introductions rarer.

5. Discussion

Apart from potentially giving better fits to data, mechanistic models can provide insight into underlying processes and allow investigators to make inferences about key parameters. Such considerations led us to propose a structured hidden Markov model based on a continuous time epidemic model. The model showed a good fit to data and was able to capture observed patterns of fade-out. To our knowledge, this represents the first application of hidden Markov models to the study of epidemic data using a mechanistic transmission model for the underlying Markov chain.

There are a number of possible non-mechanistic approaches to modelling time series of counts not presented here. These include Markov regression models (Zeger and Qaqish, 1988), integer valued autoregressive models (McKenzie, 1985; Al-Osh and Alzaid, 1987), and time varying parameter (or 'state space') models (Harvey and Fernandes, 1989). We found that while some of these approaches clearly represented major improvements on the simple Poisson model, none showed improved fits to data from the MRSA and VRE time series compared to the structured hidden Markov model. Moreover, none had plausible biological interpretations. For the R-GNR data, none performed better than the Poisson model.

5.1 *Limitations*

Despite the advantages of the structured hidden Markov model approach, there are some limitations. As is typical of population dynamic models, collinearity between parameter estimates can lead to identifiability problems (Brookhart *et al.*, 2002). As a consequence, maximum likelihood estimates can sometimes correspond to highly implausible regions of the parameter space, and maximization algorithms may fail to converge. The problem becomes particularly severe when time series are short and data sparse. For time series with fewer than about 20 to 30 observations such problems may be the norm. The common practice of aggregating counts into weekly or monthly intervals further diminishes available information.

A further limitation is that while such a model may be appropriate for a single ward or unit, for larger hospital populations made up of several interacting units its value is not so clear. Moreover, when the state space becomes large (corresponding to a large number of beds) the algorithm becomes slow and numerical problems may occur. We found that our implementation worked adequately for a state space of dimension up to about fifty. Difficulties may also occur if there are large fluctuations in the total population size; while the observation model could readily cope with such fluctuations by varying denominators, changes in the dimension of the underlying Markov chain would be harder to accommodate.

More fundamentally, the assumptions of the transmission model may be questioned. In particular, the mass action assumption of random mixing is problematic. Social or spatial structuring can reduce the rate new cases occur below that achieved by random mixing as susceptibles close to infectives get 'used up' (i.e. become infected) faster than other susceptibles (Keeling and Grenfell, 2000). For many wards this may be unimportant: transmission rates and prevalence are generally low; patients are frequently moved within units, enhancing mixing; patient turnover is often high, again diminishing the importance of spatial effects; and most transmission is believed to be vector-borne (where health care workers act as the vector, transferring organisms between patients), also diminishing spatial clustering effects. Within ICUs, however, the nature of staff–patient contact patterns may be more likely to favour transmission between adjacent patients, transmission rates and ward-level prevalences are typically higher, and the possibility of such spatial clustering cannot be excluded. Nonetheless, we did not find evidence for rejecting the mass action assumption in our data.

A further limitation of the SIS HMM approach is the assumption that all patients are equivalent. In fact, one may expect variation in patients' vulnerability to infection. One way to deal with this variability would be to use a negative binomial observation model. In practice, we found that this did not give an improved fit to our data (results not shown).

It is also important to note that the structured hidden Markov model presented assumes all autocorrelation is generated by the epidemic model. In fact, other factors may cause autocorrelation. For example, staff–patient ratios, antibiotic use, and staff hand hygiene are all likely to affect transmission rates and may all vary with time. Such covariates, if recorded, can easily be included in the hidden Markov models, but if not adjusted for could lead to overestimates of the transmission rate.

5.2 *Future developments*

Although we have restricted our attention to time series of counts from a single hospital unit, many of the methods considered can be extended to repeated measures data. Other characteristics of some time series such as trends and seasonality could also be readily accommodated within the hidden Markov framework (see, for example, Le Strat and Carrat (1999)).

Extensive prior information is often available for important parameters which, nonetheless, may not be of primary interest. It seems likely that many of the problems with the proposed structured hidden Markov models might be overcome by adopting a Bayesian formulation. For example, community prevalence studies and admission screening of patients at high-risk of carrying organisms such as MRSA can provide

lower and upper bounds for the proportion of patients positive on admission. Similarly, studies where extensive screening is performed allow estimates of the rate of progression from colonization to infection. Using such prior information may overcome problems with collinearity and help ensure that only feasible regions of parameter space are explored. This approach is likely to be particularly valuable when data are sparse.

Recently, Markov chain Monte Carlo methods have emerged as valuable tools for fitting mechanistic models of infectious disease when the epidemic process is only partially observed (Auranen *et al.*, 2000; O'Neill and Roberts, 1999). This approach seems well-suited to hospital epidemiology and can be used to fit hidden Markov models within a Bayesian framework (Scott, 2002). Such an approach may also overcome the numerical difficulties when the state space is large, allowing similar models to be applied to community pathogens.

We conclude that structured hidden Markov models are a promising tool for analysing hospital infection count data for transmissible pathogens and can represent a marked improvement on current practice. However, they are not without problems. Many of these may be overcome by working within a Bayesian framework. Rather than necessarily introducing a subjective element into the analysis, this would allow relevant external information to be used, allowing better estimates of those parameters which are of primary interest.

Acknowledgements

We thank James Robins for helpful advice; Scott Fridkin, John McGowan and the ICARE project team for providing data, and Charles Huskins, Matthew Samore, Don Goldmann, Allan Donner and Dennis Wallace for stimulating discussions and helpful suggestions. The developers of the R software and contributors of R packages used in this paper are also thanked. This work was funded by NIH grant 1R21 AI55825-01.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csàki, F. (eds), *Second International Symposium on Inference Theory*, Budapest: Akadémiai Kiadó, pp. 267–281.

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

Almuneef, M., Baltimore, R., Farrel, P., Reagan-Cirincione, P. and Dembry, L. (2001). Molecular typing demonstrating transmission of gram-negative rods in a neonatal intensive care unit in the absence of a recognized epidemic. *Clinical Infectious Diseases* **32**, 220–227.

Auranen, K., Arjas, E., Leino, T. and Takala, A. K. (2000). Transmission of pneumococcal carriage in families: a latent Markov process model for binary longitudinal data. *Journal of the American Statistical Association* **95**, 1044–1053.

Austin, D. J., Bonten, M. J., Weinstein, R. A., Slaughter, S. and Anderson, R. M. (1999). Vancomycin-resistant enterococci in intensive-care hospital settings: transmission dynamics, persistence, and the impact of infection control programs. *Proceedings of the National Academy of Sciences USA* **96**, 6908–13.

Bailey, N. T. J. (1975). *The Mathematical Theory of Infectious Diseases*. London: Charles Griffin.

Baquero, F., Coque, T. and Canton, R. (2002). Allodemics. *Lancet Infectious Diseases* **2**, 591–2.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* **41**, 164–171.

BECKER, N. (1989). *Analysis of Infectious Disease Data*. London: Chapman and Hall.

BROOKHART, M. A., HUBBARD, A. E., VAN DER LAAN, M. J., COLFORD, JR., J. M. AND EISENBERG, J. N. (2002). Statistical estimation of parameters in a disease transmission model: analysis of a Cryptosporidium outbreak. *Statistics in Medicine* **21**, 3627–38.

BROWN, S., BENNEYAN, J., THEOBALD, D., SANDS, K., HAHN, M., POTTER-BYNOE, G., STELLING, J., O'BRIEN, T. AND GOLDMANN, D. (2002). Binary cumulative sums and moving averages in nosocomial infection cluster detection. *Emerging Infectious Diseases* **8**, 1426–32.

BUREAU, A., SHIBOSKI, S. AND HUGHES, J. P. (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine* **22**, 441–62.

CAMERON, A. C. AND TRIVEDI, P. K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

CARDINAL, M., ROY, R. AND LAMBERT, J. (1999). On the application of integer-valued time series models for the analysis of disease incidence. *Statistics in Medicine* **18**, 2025–39.

COOPER, B. S., MEDLEY, G. F. AND SCOTT, G. M. (1999). Preliminary analysis of the transmission dynamics of nosocomial infections: stochastic and management effects. *Journal of Hospital Infection* **43**, 131–47.

COOPER, B. S., STONE, S. P., KIBBLER, C. C., COOKSON, B. D., ROBERTS, J. A., MEDLEY, G. F., DUCKWORTH, G. J., LAI, R. AND EBRAHIM, S. (2003). Systematic review of isolation policies in the hospital management of methicillin-resistant *Staphylococcus aureus*. *Health Technology Assessment* **7**, 1–194.

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

FLYNN, D. M., WEINSTEIN, R. A. AND KABINS, S. (1988). Infections with gram-negative bacilli in a cardiac surgery intensive care unit: the relative role of enterobacter. *Journal of Hospital Infection* **11 Suppl A**, 367–73.

FRIDKIN, S. K., STEWARD, C. D., EDWARDS, J. R., PRYOR, E. R., MCGOWAN, JR, J. E., ARCHIBALD, L. K., GAYNES, R. P AND TENOVER, F. C. (1999). Surveillance of antimicrobial use and antimicrobial resistance in United States hospitals: project ICARE phase 2. Project Intensive Care Antimicrobial Resistance Epidemiology (ICARE) hospitals. *Clinical Infectious Diseases* **29**, 245–52.

GRIMMET, G. R. AND STIRZAKER, D. R. (1992). *Probability and Random Processes*. Oxford: Oxford University Press.

GRUNWALD, G. K., HYNDMAN, R. J., TEDESCO, L. AND TWEEDIE, R. L. (2000). Non-gaussian conditional linear AR(1) models. *Australian and New Zealand Journal of Statistics* **42**, 479–495.

HARVEY, A. C. AND FERNANDES, C. (1989). Time series models for count or qualitative observations. *Journal of Business and Economic Statistics* **7**, 407–417.

IHAKA, R. AND GENTLEMAN, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

KEELING, M. J. AND GRENFELL, B. T. (2000). Individual-based perspectives on R(0). *Journal of Theoretical Biology* **203**, 51–61.

LAFAIX, C., CAMERLYNCK, P., MAR, I., BALDE, I. AND REY, M. (1969). Notion de pseudo-épidémie hospitalière. [The concept of a hospital pseudo-epidemic]. *Bulletin de la Société Médicale d'Afrique Noire de Langue Française* **14**, 713–21.

LE STRAT, Y. AND CARRAT, F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine* **18**, 3463–78.

LINDSEY, J. K. (1999). *Models for Repeated Measurements*. Oxford: Oxford University Press.

LIPSITCH, M., BERGSTROM, C. T. AND LEVIN, B. R. (2000). The epidemiology of antibiotic resistance in

hospitals: paradoxes and prescriptions. *Proceedings of the National Academy of Sciences USA* **97**, 1938–43.

LOPEZ-LOZANO, J. M., MONNET, D. L., YAGUE, A., BURGOS, A., GONZALO, N., CAMPILLOS, P. AND SAEZ, M. (2000). Modelling and forecasting antimicrobial resistance and its dynamic relationship to antimicrobial use: a time series analysis. *International Journal of Antimicrobial Agents* **14**, 21–31.

MACDONALD, I. L. AND ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman and Hall.

MCKENZIE, E. (1985). Some simple models for discrete variate time-series. *Water Resources Bulletin* **21**, 645–650.

O'NEILL, P. D. AND ROBERTS, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A* **162**, 121–129.

PELUPESSY, I., BONTEN, M. J. AND DIEKMANN, O. (2002). How to assess the relative importance of different colonization routes of pathogens within hospital settings. *Proceedings of the National Academy of Sciences USA* **99**, 5601–5605.

PLOWMAN, R., GRAVES, N., GRIFFIN, M., ROBERTS, J., SWAN, A. V., COOKSON, B. AND TAYLOR, L. (1999). *The Socio-economic Burden of Hospital Acquired Infection*. London: PHLS.

SATTEN, G. A. AND LONGINI, I. M. (1996). Markov chains with measurement error: estimating the true course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics* **45**, 275–309.

SCHNABEL, R. B., KOONTZ, R. B. AND WEISS, B. (1985). A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software* **11**, 419–440.

SCOTT, S. L. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st century. *Journal of the American Statistical Association* **97**, 337–351.

SÉBILLE, V., CHEVRET, S. AND VALLERON, A. J. (1997). Modeling the spread of resistant nosocomial pathogens in an intensive-care unit. *Infection Control and Hospital Epidemiology* **18**, 84–92.

SMITH, T. AND VOUNATSOU, P. (2003). Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Statistics in Medicine* **22**, 1709–24.

TSAY, R. S. (1992). Model checking via parametric bootstraps in time series analysis. *Applied Statistics* **41**, 1–15.

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

ZEGER, S. L. AND QAQISH, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics* **44**, 1019–31.

[*Received June 3, 2003; revised August 15, 2003; accepted for publication September 25, 2003*]