Spatial and temporal modelling of Ross River virus in Queensland

[1]Wraith, Darren,[1]Mengersen, Kerrie, [1]Low Choy, Sama, and [2] Tong, Shilu

[1]School of Mathematical Sciences, [2] School of Public Health,
Queensland University of Technology, GPO Box 2434, Brisbane Qld 4001, Australia. E-Mail: d.wraith@qut.edu.au

Keywords: *Bayesian; Ross River virus; Disease modelling; Spatial; Temporal*

EXTENDED ABSTRACT

Ross River virus (RRv), also known as Epidemic Polyarthritis, is a debilitating disease and is the most prevalent vector-borne disease in Australia (Lin et al. 2002). The virus can survive and replicate in humans and other vertebrae hosts, and is transmitted by a variety of mosquito vectors (Russell and Dwyer 2000). The disease in humans is nonfatal and infections can be either asymptomatic or symptomatic, with symptoms including polyarthritis, rash, fever, myalgia, and lethargy (Harley et al. 2001).

There has been much recent research into the spatial and temporal nature of Ross River virus in Queensland (Gatton et al. 2004; Kelly-Hope et al. 2004; Tong and Hu 2002). A recent paper by Gatton et al. (2004) focussed on the spatial and temporal nature of outbreak periods, where outbreak periods are defined by comparison against long term incidence rates specific to that area. The spatial and temporal nature of outbreak periods is of public health importance as increased understanding will lead to more targeted public health interventions (Tong 2004).

In this paper, we use a Bayesian mixture model to analyse weekly cases of Ross River virus in Queensland from 1984 to 2001. RRv notification data was obtained from the Communicable Diseases Section of Queensland Health. An exploratory analysis revealed an association between climate variables and cases of RRv, so we aggregated the data to fifteen homogenous climate zones representing Queensland.

We explore a mixture model to separate the RRv data over time into a number of states or components, and use model choice criteria to choose which number of components is preferable. This is an extension of previous work on RRv which has focussed on two components or states, an outbreak state and non-outbreak state. The method also allows the data to indicate the component (state) in which it belongs, and thereby avoid possibly subjective decision rules. Extensions to more than two components is expected to offer flexibility in cases where, for example, hyperoutbreak periods can be identified.

The choice between competing models of a different number of components invariably involves a selection criteria that will take into account both measures of fit and complexity. In this paper we use methodology developed in Celeux et al. (2003) and choose between competing models based on Deviance Information Criterion (DIC) estimates. The parameters for the different models were estimated by Markov Chain Monte Carlo (MCMC) using the software package WinBUGS (Spiegelhalter et al. 2002).

We focussed the analysis on two different climate zones which appeared to display different temporal behaviour, and found much variability in the results, with a lower number of components preferred for data from the zone which appeared to show a more distinctive pattern.

1. INTRODUCTION

Ross River virus (RRv), also known as Epidemic Polyarthritis, is a debilitating disease and is the most prevalent vector-borne disease in Australia (Lin et al. 2002). It was first identified in 1958 from mosquitoes collected at Ross River, Townsville, by the Queensland Institute of Medical Research and since then has become common in Queensland. The virus can survive and replicate in humans and other vertebrae hosts, and is transmitted by a variety of mosquito vectors (Russell and Dwyer 2000). The disease in humans is nonfatal and infections can be either asymptomatic or symptomatic, with symptoms including polyarthritis, rash, fever, myalgia, and lethargy (Harley et al. 2001

In this paper we explore a mixture model to separate the RRv data over time into a number of states or components, and use model choice criteria to choose which number of components is preferable. This is an extension of previous work on RRv which has focussed on two components or states, an outbreak state and non-outbreak state. The method also allows the data to indicate the component (state) in which it belongs, and thereby avoid possibly subjective decision rules. Extensions to more than two components is expected to offer flexibility in cases where, for example, hyperoutbreak periods can be identified.

2. METHOD

2.1. Data

Ross river virus disease notification data from 1984 to 2001 was obtained from the Communicable Diseases section of Queensland Health. A notification was reported if serologic testing indicated a four-fold change in antibody titer between paired acute and convalescent sera, or if IgM and IgG antibody levels against RR virus were consistent with acute infection. Each complete notification included place of residence (location and street/road), date of onset, age and sex of the patient. Place of residence was further geocoded by the Queensland Department of Local Government and Planning into Statistical Local Areas (SLA) and later grouped to Local Government Areas (LGA).

An exploratory analysis of the data at the residence level and recent research indicates a strong relationship between the incidence of RRv virus and climate related variables such as rainfall, temperature, humidity, Southern Oscillation Index, and sea levels (McFallan 2001; Tong and Hu 2002; Kelly-Hope et al. 2004). On this basis, we decided to aggregate the data to 15 climate zones as identified by the Australian Bureau of Meteorology (See Figure 1).

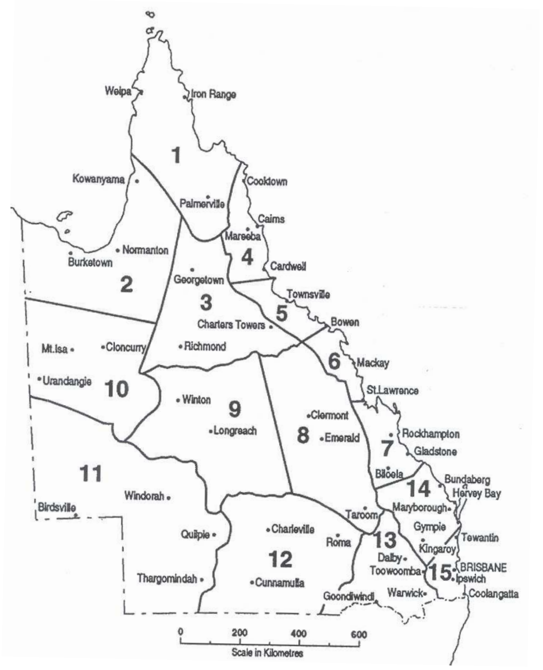

Figure 1. Queensland climate zones - Bureau of Meteorology

A summary of the data for the fifteen climate zones is provided in Table 1.

Table 1. Summary results - all zones

Zone	Min	Q1	Median	Mean	Q3	Max
1	0.00	0.00	0.00	0.08	0.00	2.00
2	0.00	0.00	0.00	0.14	0.00	5.00
3	0.00	0.00	0.00	0.33	0.00	7.00
4	0.00	0.00	2.00	4.17	5.00	42.00
5	**0.00**	**0.00**	**2.00**	**5.10**	**5.00**	**59.00**
6	0.00	0.00	1.00	2.07	2.00	37.00
7	0.00	0.00	1.00	3.21	3.00	48.00
8	0.00	0.00	0.00	0.85	1.00	14.00
9	0.00	0.00	0.00	0.29	0.00	9.00
10	0.00	0.00	0.00	0.42	0.00	20.00
11	0.00	0.00	0.00	0.07	0.00	3.00
12	0.00	0.00	0.00	0.73	1.00	15.00
13	0.00	0.00	1.00	2.07	2.00	59.00
14	0.00	0.00	1.00	2.33	2.00	44.00
15	**0.00**	**2.00**	**5.00**	**14.42**	**11.00**	**307.00**

We decided to focus our analysis on data from two climate zones, Zone 15 and 5, which appear to show quite different temporal behaviour. Figures 2 & 3 show the weekly number of RRv cases over time for Zones 15 and 5 respectively. The data from Zone 15 appears to show a more distinctive outbreak pattern than from Zone 5, and thus it is of interest to see how the results of applying a mixture model to these zones separately may differ.

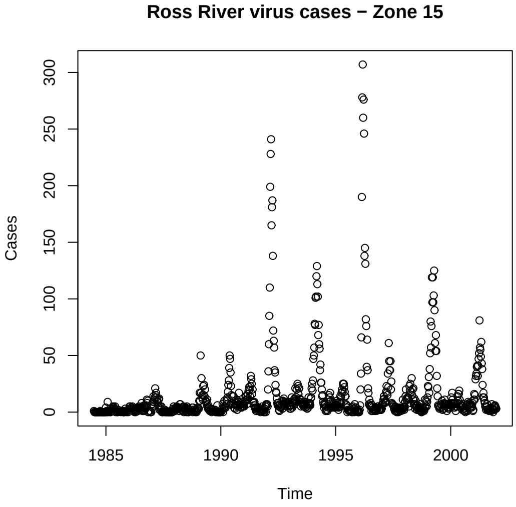

Figure 2. Time plot of weekly cases - Zone 15

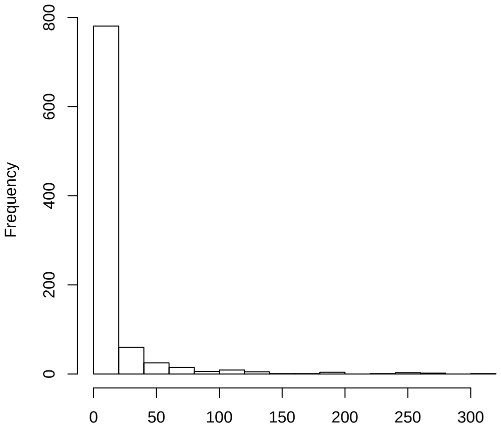

Figure 4. Histogram of weekly cases - Zone 15

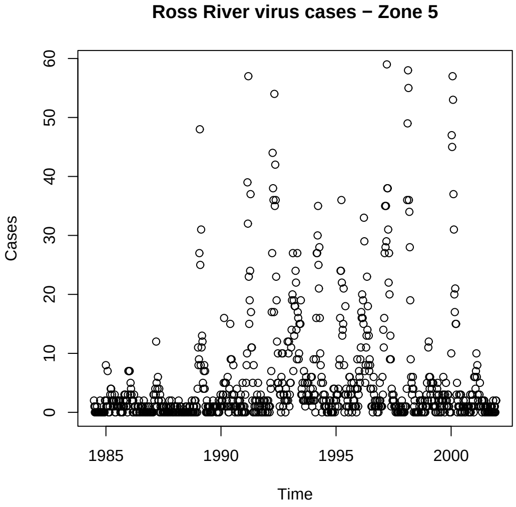

Figure 3. Time plot of weekly cases - Zone 5

2.2. Mixture models

The use of mixture distributions comprising a finite or infinite number of components, possibly of different distributional types, to describe different features of data has attracted a great deal of recent research interest (Marin et al. 2005; McLachlan and Peel 2000).

The mixture model can be formulated as,

$$p(y|\theta) = \sum_{j=1}^{k} w_j f(y|\theta_j)$$
$$\sum_{j=1}^{k} w_j = 1, \qquad k > 1 \tag{1}$$

where k is the number of components, w_j is the probability of being allocated to component j, and where the allocation of each observation y_i to one of the components is represented by a latent variable z_i ($z_i \in \mathbb{N}$ (discrete case))

$$p(z_i = j) = w_j$$
$$Z \sim \text{Multinomial}(1, p_1 \ldots p_k) \tag{2}$$

Choice between competing models of a different number of components invariably involves a selection criteria that will take into account both measures of fit and complexity. For example, a mixture model with a large number of components may fit the data well, but suffer from a lack of interpretability of the parameters. In this paper we use methodology developed in Celeux et al. (2003) and choose between competing models based on Deviance Information Criterion (DIC) estimates.

2.2.1. Application to RRv data

We first fitted a Poisson distribution to the RRv data for Zones 15 and 5, and found that due to the large number of zeros, this distribution does not offer a good fit (See Figures 4 & 5). There are a range of methods available to handle the case of a large number of zeros for count data (See for example Dalrymple et al. (2003)), here for simplicity we chose to let log(y_t+1) follow a normal distribution.

For RRv data relating to Zones 15 and 5 we specified,

$$\log(y_t + 1) \sim \text{N}(\mu_{Z_t}, \tau)$$
$$Z_t \sim \text{Categorical}(P_{1,\ldots,k})$$

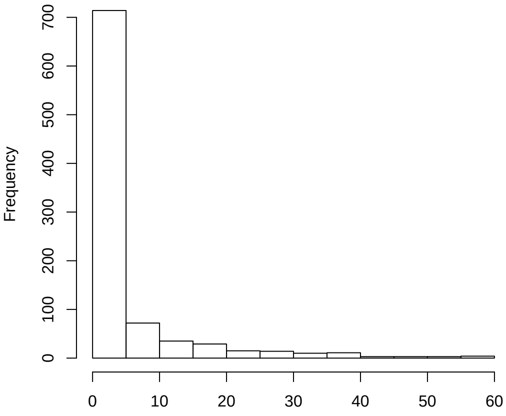

Figure 5. Histogram of weekly cases - Zone 5

and

$$\begin{aligned} \mu_{Z_t} &\sim \mathrm{N}(1, 0.01) \\ P_{1,\ldots,k} &\sim \text{Dirichlet}(\alpha) \\ \tau &= 1/\sigma^2 \\ \sigma &\sim \mathrm{U}(0.1, 20) \end{aligned}$$

where y_t are the observed cases, and μ is restricted to $\mu_1 < \mu_2 < \ldots \mu_k$, in order to prevent label switching.

The parameters for the different models were estimated by Markov Chain Monte Carlo (MCMC) using the software package WinBUGS (Spiegelhalter et al. 2002). Estimates are based on runs of 10,000 iterations, or until evidence of convergence. Convergence was assessed by examining Monte-carlo error estimates and Gelman-Rubin statistics (Brooks and Gelman (1998)).

3. RESULTS

The results for Zones 15 and 5 are provided in Tables 2 and 3 respectively. The estimates for μ have been exponentiated for ease of interpretability with the original data.

Our model choice criterion is to examine the DIC estimates (lowest being preferable), and indirectly the effective number of parameters (pd). For Zone 15 (Table 2), the estimates for four components and above show weak signs of convergence in the MCMC runs (Brooks and Gelman (1998)). This is also an indication that we could be overfitting. On the basis of this, the three component model appears to be preferable as there is a reduction in the DIC estimate from the two component model (2,880 ($k = 2$) to 2,869 ($k = 3$)), without a large increase in the number of effective parameters (3.43 ($k = 2$) to 6.04 ($k = 3$)).

Table 2. Results for Log Normal - Zone 15

		Value	Credible Interval		Value
k=1	μ_1	5.45	(4.97,5.99)		
	τ	1.99			
	DIC	2933.17		pd	2.12
k=2	μ_1	3.89	(3.45,4.38)	w_1	0.88
	μ_2	48.01	(33.54,69.11)	w_2	0.12
	τ	3.08			
	DIC	2880.76		pd	3.43
k=3	μ_1	1.23	(0.85,2.07)	w_1	0.35
	μ_2	7.18	(6.01,8.56)	w_2	0.54
	μ_3	64.17	(51.93,80.53)	w_3	0.11
	τ	8.83			
	DIC	2869.26		pd	6.04
k=4*	μ_1	0.81	(0.56,1.15)	w_1	0.28
	μ_2	5.73	(4.59,7.08)	w_2	0.52
	μ_3	23.51	(13.54,42.38)	w_3	0.13
	μ_4	96.32	(68.83,146.97)	w_4	0.06
	τ	22.69			
	DIC	2860.85		pd	7.79
k=5*	μ_1	0.72	(0.41,1.07)	w_1	0.26
	μ_2	4.31	(1.37,6.39)	w_2	0.34
	μ_3	9.72	(5.42,24.30)	w_3	0.25
	μ_4	31.62	(17.38,73.66)	w_4	0.10
	μ_5	108.73	(74.57,178.83)	w_5	0.05
	τ	33.28			
	DIC	2859.14		pd	8.21

Note: * the estimates given for these components are unstable from the MCMC runs

Table 3. Results for Log Normal - Zone 5

		Value	Credible Interval		Value
k=1	μ_1	2.17	(1.97,2.40)		
	τ	2.51			
	DIC	2668.44		pd	1.90
k=2	μ_1	1.09	(0.97,1.21)	w_1	0.80
	μ_2	15.09	(13.10,17.41)	w_2	0.21
	τ	11.55			
	DIC	2491.57		pd	3.66
k=3	μ_1	0.57	(0.45,0.73)	w_1	0.56
	μ_2	3.62	(2.96,4.55)	w_2	0.30
	μ_3	20.41	(17.93,23.19)	w_3	0.15
	τ	73.85			
	DIC	2461.81		pd	5.81
k=4	μ_1	0.07	(0.01,0.14)	w_1	0.29
	μ_2	1.64	(1.52,1.79)	w_2	0.42
	μ_3	6.36	(5.70,7.13)	w_3	0.17
	μ_4	24.23	(22.48,26.36)	w_4	0.12
	τ	43044.94			
	DIC	2413.33		pd	10.62
k=5	μ_1	0.00	(0.00,0.03)	w_1	0.27
	μ_2	1.36	(1.29,1.43)	w_2	0.36
	μ_3	4.13	(3.88,4.45)	w_3	0.19
	μ_4	12.14	(11.15,13.22)	w_4	0.11
	μ_5	31.88	(29.45,34.41)	w_5	0.07
	τ	3478962659.69			
	DIC	2276.10		pd	15.39

The results for Zone 5 are quite different than for Zone 15, suggesting that up to five components could be fitted to this data. For a five component model there is a sizeable reduction in the DIC estimate compared to a four component model (2,413 ($k = 4$) to 2,276 ($k = 5$)), with an increase in the number of effective parameters (10.62 ($k = 4$) to 15.39 ($k = 5$)).

In the context of disease modelling, we could interpret the results for the number of components in the following way: Evidence for a two component model suggests that over time, incidence of RRv can be in either a background (or normal) state or an outbreak state, with the latter state associated with a higher mean value of cases than the former. We could similarly interpret a three component model as indicating an additional state to these two, and call this a 'hyper-outbreak' state. By this, we mean a state indicating an outbreak of greater magnitude than normal. It is less clear at this point in time how to interpret data best fitted by a model with four or more components.

4. DISCUSSION

We explored a Bayesian mixture model to analyse cases of RRv occurring in 15 climate zones throughout Queensland, with a focus of the analysis on two climate zones which appear to show quite different temporal behaviour. We found much variability in the results of applying a mixture model to these two climate zones, with a lower number of components preferred for data from the zone which appeared to show a more distinctive pattern.

The analysis could be extended to assume other distributional forms to take account of the large number of zeros in the data. Further research will investigate the impact these distributional forms have on the results.

5. REFERENCES

Brooks, S. P. and A. Gelman (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*, 434–455.

Celeux, G., F. Forbes, C. P. Robert, and M. Titterington (2003, June). Deviance information criteria for missing data models. Technical report, Institut National De Recherche en Informatique et en Automatique.

Dalrymple, M. L., I. L. Hudson, and R. P. K. Ford (2003). Finite mixture, zero-infated poisson and hurdle models with application to sids. *Computational Statistics & Data Analysis 41*, 491–504.

Gatton, M. L., L. A. Kelly-Hope, B. H. Kay, and P. A. Ryan (2004). Spatial-temporal analysis of ross river virus disease patterns in queensland, australia. *American Journal of Tropical Hygiene and Medicine 71*(5), 629–635.

Harley, D., A. Sleigh, and S. Ritchie (2001). Ross river virus transmission, infection, and disease: a cross-disciplinary review. *Clin Microbiol Rev 14*, 909–932.

Kelly-Hope, L. A., D. M. Purdie, and B. H. Kay (2004). Ross river virus disease in australia, 1886-1998, with analysis of risk factors associated with outbreaks. *J Med Entomology 41*, 133–150.

Lin, M., P. Roche, J. Spencer, A. Milton, P. Wright, and D. Witteveen (2002). Australia's notifiable disease status, 2000. annual report of the national notifiable diseases surveillance system. *Commun Dis Intell 26*, 118–175.

Marin, J. M., K. Mengersen, and C. P. Robert (2005). Bayesian modelling and inference on mixtures of distributions.

McFallan, S. (2001). *Climatic change and its impact on Ross River virus*. Masters thesis, School of Mathematical Sciences, Queensland University of Technology.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley and Sons Ltd.

Russell, R. C. and D. E. Dwyer (2000). Arboviruses associated with human disease in australia. *Microbes Infect 2*, 1693–1704.

Spiegelhalter, D. J., A. Thomas, and N. G. Best (2002). Winbugs version 1.4 user manual. research report.

Tong, S. (2004). Ross river virus disease in australia: epidemiology, socioecology and public health response. *Internal Medicine Journal 34*, 58–60.

Tong, S. and W. Hu (2002). Different responses of ross river virus to climate variability between coastline and inland cities in queensland, australia. *Occup Environ Med 59*, 739–744.