Final report: SCIG-MAS-001

Development and application of new tools for the analysis of *Campylobacter* surveillance data: identifying the spatial and temporal determinants of raised notifications in New Zealand

October 2009

prepared for Dr Donald Campbell
New Zealand Food Safety Authority

by

Dr Jonathan Marshall
Dr Simon Spencer
Professor Nigel French

Hopkirk Institute
Institute of Veterinary, Animal and Biomedical Sciences
College of Sciences
Massey University
New Zealand
Email - N.P.French@massey.ac.nz
Phone - +64 (06) 356 9099 extn 81188
Fax - +64 (06) 350 5714

Contents

1 Executive Summary

In this study, *Campylobacter* notifications from 2001 to 2008 across three regions of New Zealand were analysed, and spatial and temporal trends were identified. Risk factors associated with these trends were investigated, and several relationships were observed. In urban areas, the Social Deprivation Index was a risk factor for notifications, with areas of high deprivation having low notification rates. In contrast, the SDI had no clear association with notifications in rural areas, where areas of high ruminant (sheep and dairy) density were more closely aligned with notification rates. Differences were shown in notification rates across age groups, with children under 5 years of age having significantly higher notification rates than other age groups, with the majority of these notifications coming from rural populations. The MultiLocus Sequence Typing of isolates from the Manawatu indicate a clear difference in the spatial distribution of sequence types associated with poultry compared to those associated with ruminants, with poultry associated isolates more prevalent in urban areas. Several meterological variables were also investigated, and were shown to be associated with the temporal variation in notification rates, though peaks in the weather variables lagged behind corresponding peaks in date of notification by several weeks.

A mathematical model was developed to account for the average spatial and temporal trends of notification risk, allowing any unusual deviations at particular spatial locations in particular time periods to be detected in the residuals. Epidemic indicator variables were introduced to model these anomalous events, and two separate priors were considered for these variables. The independent indicators method was found to be more discriminatory at identifying outbreaks of campylobacteriosis.

Finally, the model was adapted to be useful as a surveillance tool, and a simple user interface was developed to facilitate the running of the model. The tool is designed to be run weekly as notification data are accumulated, allowing automated early detection of possible outbreaks, which may then

be followed up by health professionals. A retrospective application of the surveillance tool to four of the larger outbreaks from 2005 and 2006 in the Manawatu region showed that it could reliably detect such outbreaks in the first week of data availability.

2 Introduction

New Zealand has one of the highest per-capita incidence rates of campylobacteriosis in the world (Baker et al., 2006). In order to reduce the incidence, we must have a thorough understanding of the epidemiology of infection, and thus need to understand the spatial and temporal determinants of raised notifications.

There are several risk factors known to be associated with *Campylobacter* notifications, and many of these are spatially and temporally structured. Social deprivation is a risk factor for many infectious diseases, however, a UK study has indicated that campylobacteriosis notifications are higher among those with higher socioeconomic status (Gillespie et al., 2008), suggesting that social deprivation may be a protective effect. Gender is also considered a risk factor for campylobacteriosis, with males in general notifying higher than females (Sneyd and Baker, 2003; Baker et al., 2007). Based on data from 2001-2004, Baker et al. also showed that significantly higher rates of notification and hospitalisation were seen among urban populations in comparison to rural populations. However, when children and adults were treated separately, children under the age of 15 years in rural areas notified significantly higher than their urban counterparts. This difference between children and adults is also seen in a comparison of notification rates of different age groups. Children aged 1-4 years old had the highest notification rates, followed by 20-29 year olds and infants, with other age groups being at least 25% lower.

Campylobacteriosis case rates are highly seasonal, although there is large variation between years in the timing, duration and extent of peak inci-

dence. Baker et al. (2007) examined notification and hospitalisation rates from 2001-2003, and found the highest rates occurred during the summer months. However, there have also been increased notifications during early winter (Williman et al., 2008), an example of which occurred in winter 2006, where a nationwide increase was observed (McTavish et al., 2008). The majority of cases in New Zealand are considered sporadic (Wilson, 2005), however there are also brief localised outbreaks in which cases share a common exposure. Identifying these short periods of increased incidence through close proximity of the cases in space and time is of particular interest to pubic health professionals, who can use this information to follow cases up and determine the exposure, and perhaps prevent future cases from occurring in the same way.

The high degree of unpredictability in case rates presents unique statistical challenges when analysing background temporal and spatial trends, which must be well understood before anomalous outbreaks can be identified. Many existing methods, such as SaTScan™[1], rely on using data from a period containing no known outbreaks to identify the baseline case rates. For a disease such as campylobacteriosis, however, such periods do not exist. A method that can determine the baseline case rates in the presence of sporadic cases, therefore, is required.

The goal of this project is to analyse campylobacteriosis notification data spanning the years 2001 to 2008 for three regions of New Zealand: Auckland, Canterbury and the Manawatu. The aims are four fold:

1. Develop a model for the background spatial and temporal trends in the data.

2. Identify potential risk factors that might be associated with these trends.

3. Develop a method of identifying anomalous outbreaks over and above

[1]SaTScan™is a trademark of Martin Kulldorff. The SaTScan™software was developed under the joint auspices of Martin Kulldorff, the National Cancer Institute and Farzad Mostashari at the New York City Department of Health and Mental Hygiene.

these trends.

4. Produce a tool suitable for use in the surveillance of campylobacteriosis in New Zealand.

An overview of the data is given in the following section. The mathematical model is then described in Section 4 and investigation of risk factors to the underlying spatial and temporal trends is given in Section 5. The model is then extended to include outbreak indicators in the spatio-temporal term in Section 6 and we describe the use of the model as a potential surveillance tool in Section 7. Implementation details for the model are given in Appendix A.

3 Data overview

The data consist principally of a list of cases with a notification date and an approximate spatial location. The data span the years 2001-2006 in the Auckland and Canterbury regions, with data available from 2001-2008 in the Manawatu region. The spatial information associated with each notification is the census meshblock, which are small areas of New Zealand that normally contain between 0 and 200 people in their usually resident population. Meshblocks therefore vary in size, with those in urban areas giving a more precise spatial location than those in rural areas. In addition, we have epidemiological data from the EpiSURV database for each case in the Manawatu region, including the sequence type (ST) of the *Campylobacter* strain obtained by MultiLocus Sequence Typing (MLST) as well as the age and gender of the case.

The data for potential risk factors include the Social Deprivation Index (SDI), age and gender, weather information, poultry sales information, and proximity to large poultry farms and livestock densities. The SDI was calculated at the meshblock level from the 2006 census data, and has proved a key risk factor in urban areas. In rural areas, proximity to high livestock densities such as dairy cattle is also a risk factor. Data on farm locations and size was

Year	Auckland	Canterbury	Manawatu
2001	2.232	2.269	2.049
2002	3.312	3.105	2.109
2003	3.114	3.834	2.651
2004	3.032	2.685	1.845
2005	2.976	4.111	2.453
2006	4.093	3.879	2.598
2007	3.363	3.943	2.003
2008	1.778	1.486	1.183

Table 1: Number of notifications per 1,000 people per year in each region.

taken from the Agribase^{TM}[2] database from January 2006. Weather data from weather stations throughout each region was obtained from NIWA, such as sunshine hours, maximum and minimum temperature, rainfall and storm water run-off. There were 16 weather stations used throughout the Manawatu, 46 in Auckland, and 33 in Canterbury, though not all of these had readings available for all variables considered. In addition, combined weather information, such as patterns of weather that were suitable for barbecues were also investigated. Finally, retail sales information of fresh poultry from all four major poultry companies was provided by the poultry industry. All spatial and temporal data was interpolated or aggregated to the meshblock and week level to facilitate comparison with the notification data.

Table 1 summarises the notification data across each region. Of interest is the marked drop off in notification rates across all regions in 2008.

4 Mathematical model of spatial and temporal trends

A Bayesian statistical model is developed based on a collection of models discussed by Diggle et al. (2002).

[2] Agribase^{TM} is a product of AsureQuality.

Assume that the number of notifications $Y_{i,t}$ in meshblock i in week t has a Poisson distribution with mean $n_i\lambda_{i,t}$, where n_i is the number of people resident in meshblock i, and $\lambda_{i,t}$ is the probability an individual residing in meshblock i presents with campylobacteriosis in week t. Next, assume that the log of the risk may be split into separate components, so that

$$\log(\lambda_{i,t}) = R_t + U_i + W_{i,t}, \tag{1}$$

where R_t is a purely temporal component, U_i is a purely spatial component, and $W_{i,t}$ is a spatio-temporal interaction.

We provide structure on these components by specifying prior distributions on R_t and U_i so that they represent the underlying temporal and spatial trends in the data. For the temporal trend parameters, we first specify R_1 and R_2 using improper flat priors, and then describe the trend using a second order Gaussian random walk. Thus, given R_{t-1} and R_t, we assume R_{t+1} has a normal distribution,

$$R_{t+1}|R_t, R_{t-1}, \kappa_R \sim R_t + (R_t - R_{t-1}) + N(0, \kappa_R^{-1}).$$

The assumption here is that the risk in week $t+1$ is equal to the risk in week t plus the most recent change in risk, with some uncertainty governed by κ_R. This is equivalent to interpolating R_{t+1} using a linear trend through R_t and R_{t-1}, and then adding uncertainty from a normal distribution with variance κ_R^{-1}. This allows a linear trend to fit the data where the data suggest it, whilst allowing changes in that trend to emerge at a later point in the time series. The hyperparameter κ_R is assumed to have a non-informative conjugate Gamma prior. Notice that, given data up to week T, this model allows the number of cases in week $T+1$ to be predicted – it is just the usual linear trend through the previous two time points.

For the underlying spatial trend parameters, an intrinsic Gaussian Markov random field prior is used (Besag et al., 1991). This means that U_i is depen-

dent only on the risk in neighbouring meshblocks $n(i)$, and thus

$$U_i|\{U_j : j \in n(i)\}, \kappa_U \sim \frac{1}{|n(i)|}\sum_{j\in n(i)} U_j + N\left(0, \frac{1}{\kappa_U|n(i)|}\right).$$

The prior here ensures that meshblock i takes the average value of the surrounding meshblocks, with some additional uncertainty. The $|n(i)|$ divisor in the variance ensures that those meshblocks with fewer neighbours (such as those on the edge of regions) have the same relative variance as those meshblocks in the interior. Once again, a conjugate Gamma prior is used for κ_U.

The purely temporal and spatial terms act to absorb the underlying temporal and spatial trends in the data, leaving the last term $W_{i,t}$ to include changes over and above this trend, which allow the fitting of epidemic indicators as in Section 6.

In addition, to assess hypotheses regarding spatial risk factors for notifications in rural areas, a multiple Poisson regression model with log link function is used in Section 5.2.

5 Identifying potential risk factors for increased notifications

The purpose of this section is to examine the underlying temporal and spatial trends in the data, in order to generate hypotheses about likely risk factors. To do this, we begin by setting $W_{i,t} \equiv 0$ in equation 1, leaving only the R_t and U_i terms. Figure 1 shows the weekly case rates per 1,000 people estimated by the model. The summer peaks can be seen to be irregular in height, duration and timing. Canterbury has the most seasonal variation with several severe but short epidemics in the summer. Auckland has longer summer seasons and shorter winters, with the notifications appearing to come in short bursts during the summer. The Manawatu has a significantly lower

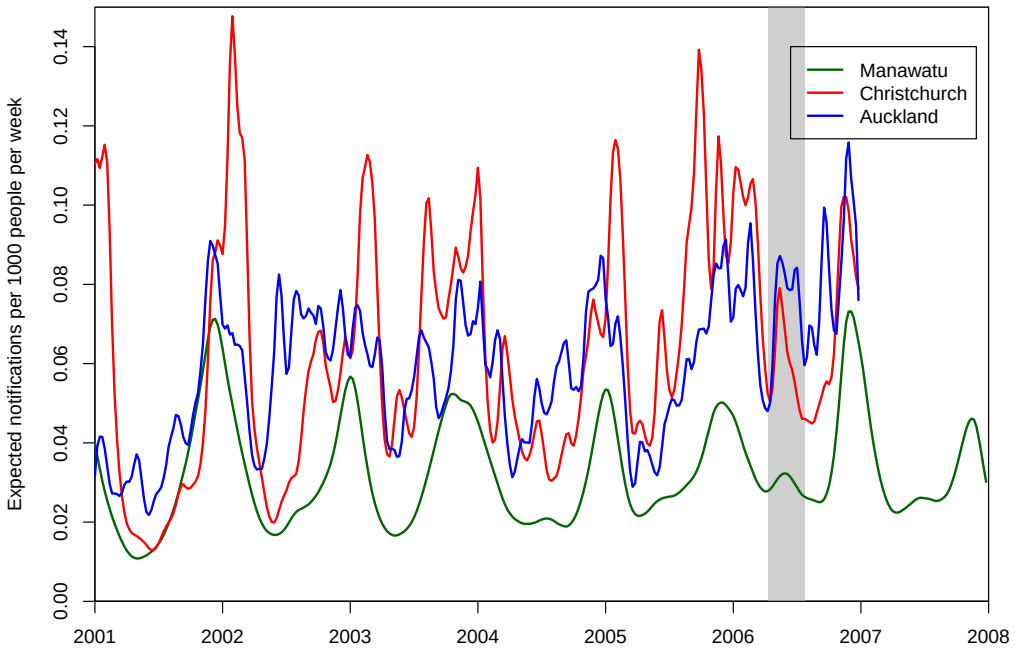

Figure 1: The expected number of cases per 1000 people per week people for the Manawatu, Canterbury and Auckland regions. The grey background highlights the 2006 winter epidemic.

number of cases per person than the other two regions, and the notification rates appear smoother through time.

Figure 1 also shows an increase in the number of winter cases across all regions, culminating in the winter epidemic of 2006. The differences across the regions may be due to differing climates. The larger variation between summer and winter notification rates for Canterbury, for instance, may be attributable to the larger difference in temperatures between summer and winter compared to the Manawatu and Auckland.

The spatial trends can be seen in Figures 2-4. The highlighted spot around Foxton in the Manawatu (Figure 2) may be due to occupationally acquired infections from workers at the poultry processing plant located there, although

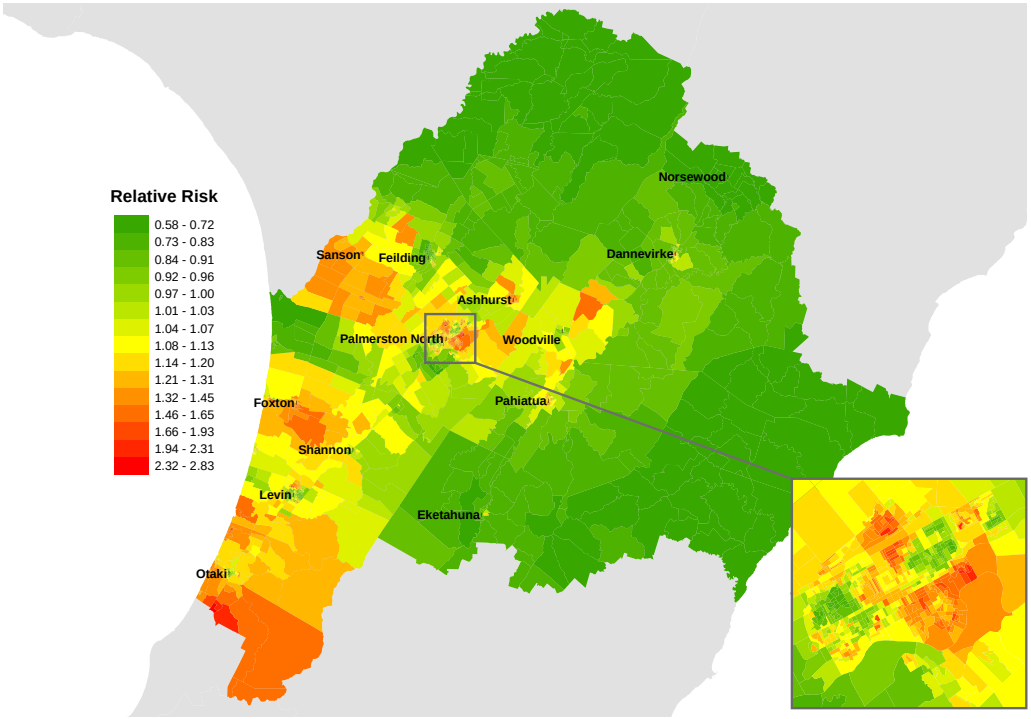

Figure 2: The estimated relative risk surface for the Manawatu region.

most of these cases occur in 2001-2003. The Canterbury region shows the most variation in risk, with some meshblocks registering more than 40 times the per-person risk than others. The large green region in South Auckland is a lower socio-economic area.

Based on the observed trends above, and other trends highlighted in the literature, we have identified the following potential risk factors for *Campylobacter* notifications in the above regions.

1. Social Deprivation Index (SDI).
2. Distance from areas of high livestock (ruminant) densities or large poultry farms.
3. Sequence type.
4. Age and gender.
5. Variation in weather (rainfall, sunshine hours, temperature and storm water run-off).

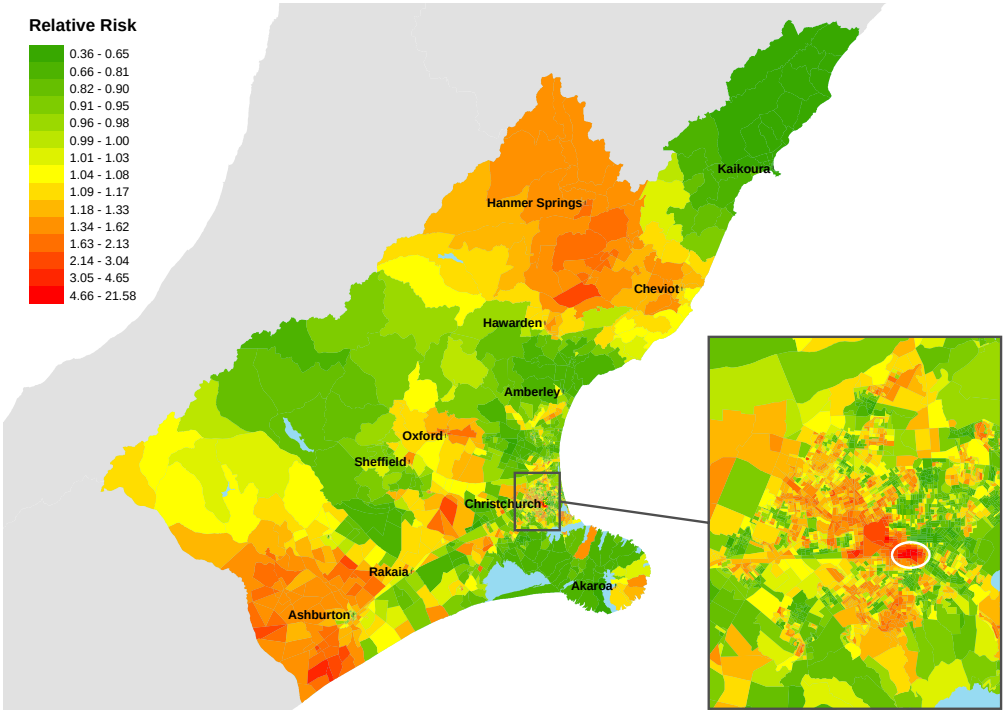

Figure 3: The estimated relative risk surface for the Canterbury region.

6. Poultry retail sales.

We address each of these in turn.

5.1 Social Deprivation Index

Figures 2-4 indicate that notifications are in general higher in the more affluent areas of the three regions, particularly within the urban centres. In order to investigate this further, the SDI and the urban/rural classification of each meshblock was computed from the 2006 census. Figure 5 shows the deciled relative risk from the model, side by side with the SDI for Auckland. There is a clear similarity, with the SDI appearing highly negatively associated with risk – the more affluent areas having higher notification rates. This pattern, however, is not repeated in rural areas.

We confirmed this relationship by replacing the spatial term U_i in equation 1 with a set of fixed effects – one for each of the 10 levels of deprivation. The linear predictor thus becomes $\log(\lambda_{i,t}) \sim R_t + V_{D(i)}$, where $D(i)$ is the SDI of

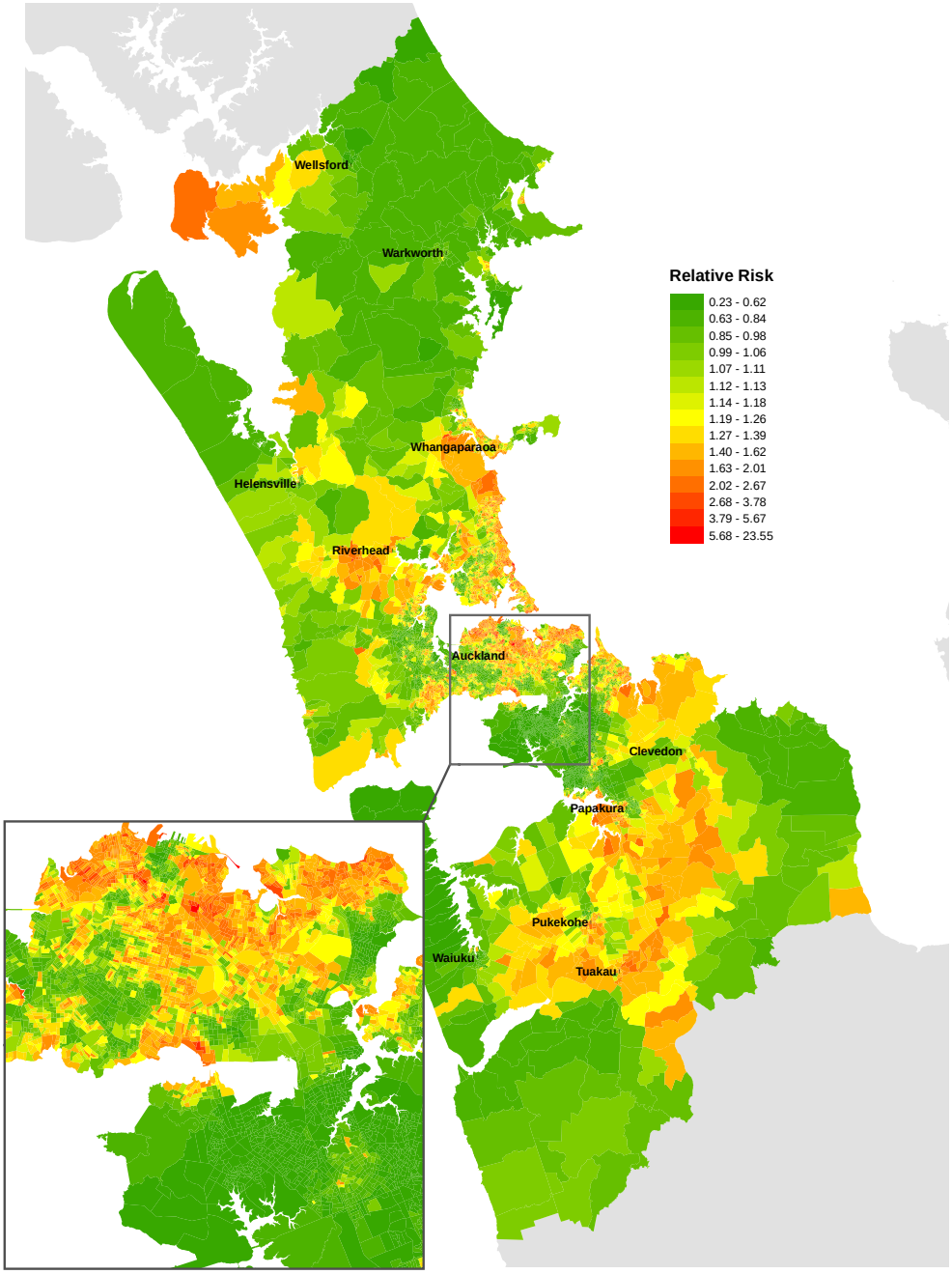

Figure 4: The estimated relative risk surface for the Auckland region.

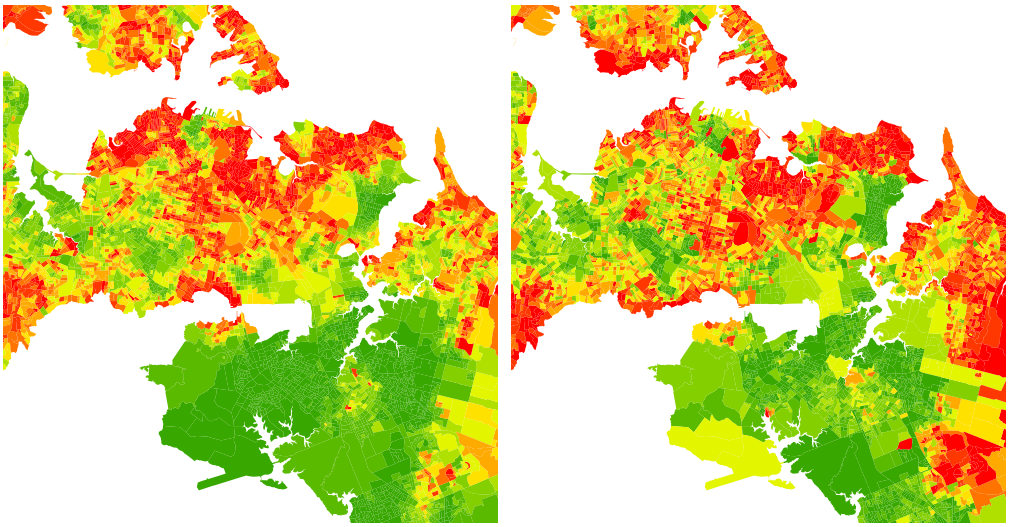

Figure 5: Deciles of the estimated relative risk for central Auckland (left) compared with the Social Deprivation Index (right).

meshblock i. The fixed effects $V_1, V_2, \ldots, V_{10}$ are given a first order Gaussian random walk prior, i.e. given $V_1, \ldots, V_j$,

$$V_{j+1} \sim N\left(V_j, \frac{1}{\kappa_V}\right).$$

This is equivalent to assuming that each level of deprivation will have a similar level of risk as adjacent levels. The hyperparameter κ_V is presumed to have a conjugate Gamma distribution.

The results are given in Figure 6, where we see a clear drop in risk in deprived areas, particularly for indices greater than 7. The largest trend is observed in the Auckland region, possibly due to having a larger urban area, and thus comparatively more urban meshblocks free from rural influences.

There is one exception to this strong SDI relationship in central Christchurch, where, in the area marked by the white circle on Figure 3, a large number of cases occur in an apparently small population. This part of Christchurch

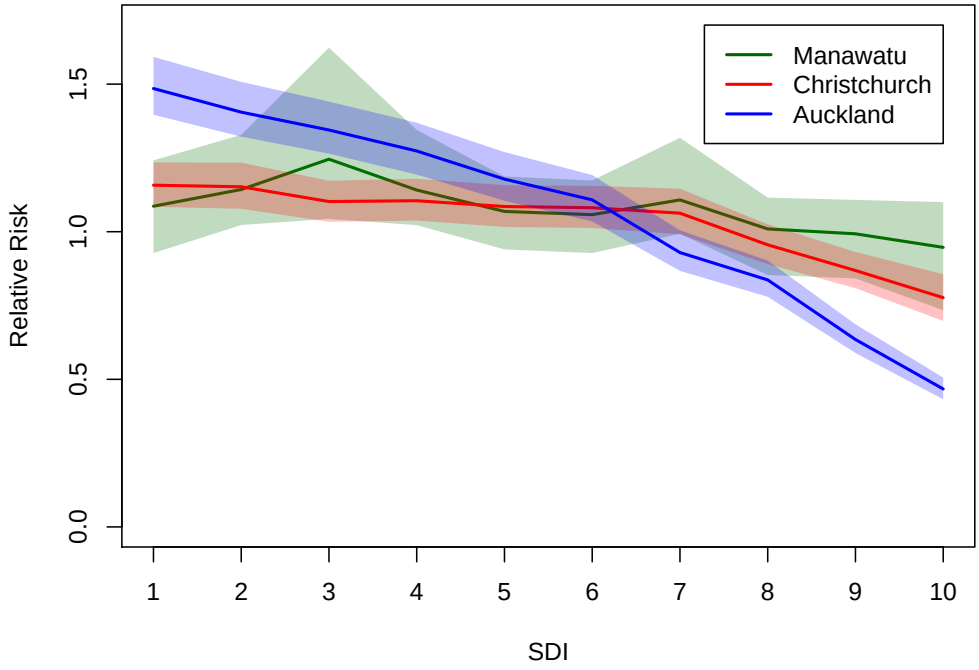

Figure 6: Social Deprivation Index (SDI) against notification risk with 95% credibility envelopes for urban meshblocks.

contains low-cost, long-stay accommodation, and thus the usually resident population recorded in the census may not accurately reflect the number of individuals who might nominate this area as an address should they get campylobacteriosis, hence artifically increasing the relative risk per person in this area.

5.2 Livestock density

In rural areas, the relationship between notifications and the Social Deprivation Index appears less strong, suggesting that the risk factors for rural areas may differ from urban areas. Furthermore, a χ^2 test in the Canterbury and Auckland regions shows urban areas notify at a significantly higher

rate than rural areas ($p = 0.006$ for Canterbury, and $p < 0.0001$ for Auckland.) Given these differences, the association between animal density, SDI and campylobacteriosis notifications in rural meshlocks was investigated as an alternative explanation for the spatial pattern.

Using Agribase^{TM}[3], the number of dairy cattle, beef cattle and sheep in each meshblock was estimated by overlaying farm information on meshblocks, based on data from January 2006. Since the majority of poultry were confined to large poultry farms (greater than 1000 birds), crude estimates for the proximity of a meshblock to a large poultry farm were found by counting the number of meshblocks that must be passed through to get from the meshblock to the poultry farm. For instance, meshblocks containing large poultry farms were given distance zero, and their neighbours were given distance one and so on. We preferred this measure of distance over Euclidean distance to the meshblock centroid because some meshblocks are very large in rural areas, and thus a part of the meshblock could be quite close to a poultry farm, yet the distance to its centroid would be large. Figures 7 and 8 show the spatial distribution of dairy, beef and sheep densities, as well as proximity to a large poultry farm for the Canterbury and Manawatu regions respectively.

A multiple Poisson regression model with log link function was fitted using the generalised linear model function in R (glm). The response variable was the number of cases in each rural meshblock, with the population of the meshblock used as an offset, and the variables SDI, poultry farm distance and each of the animal density variables used as predictors. A backward stepwise regression procedure was used to eliminate variables that had p-value greater than 0.15. Note that rural meshblocks only were used for this analysis.

The results are given in Table 2 where the regression coefficients and p-values for each variable that was not eliminated from the regression model in each of the three regions are given. The beef cattle density was eliminated from all three models, and in Auckland the remainder of the livestock

[3]Agribase^{TM}is a product of AsureQuality.

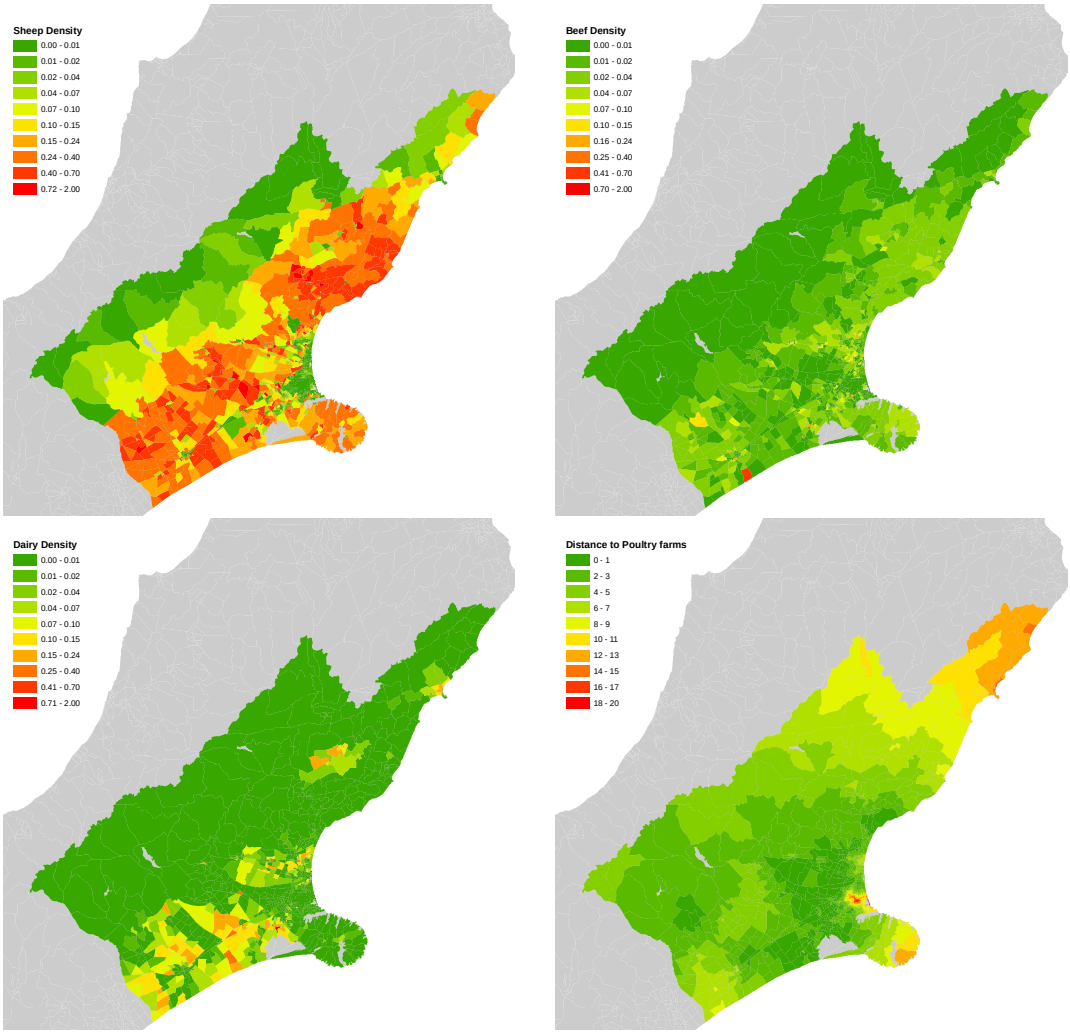

Figure 7: Densities of sheep, beef, and cattle, as well as proximity to a large poultry farm, for the Canterbury region in 2006.

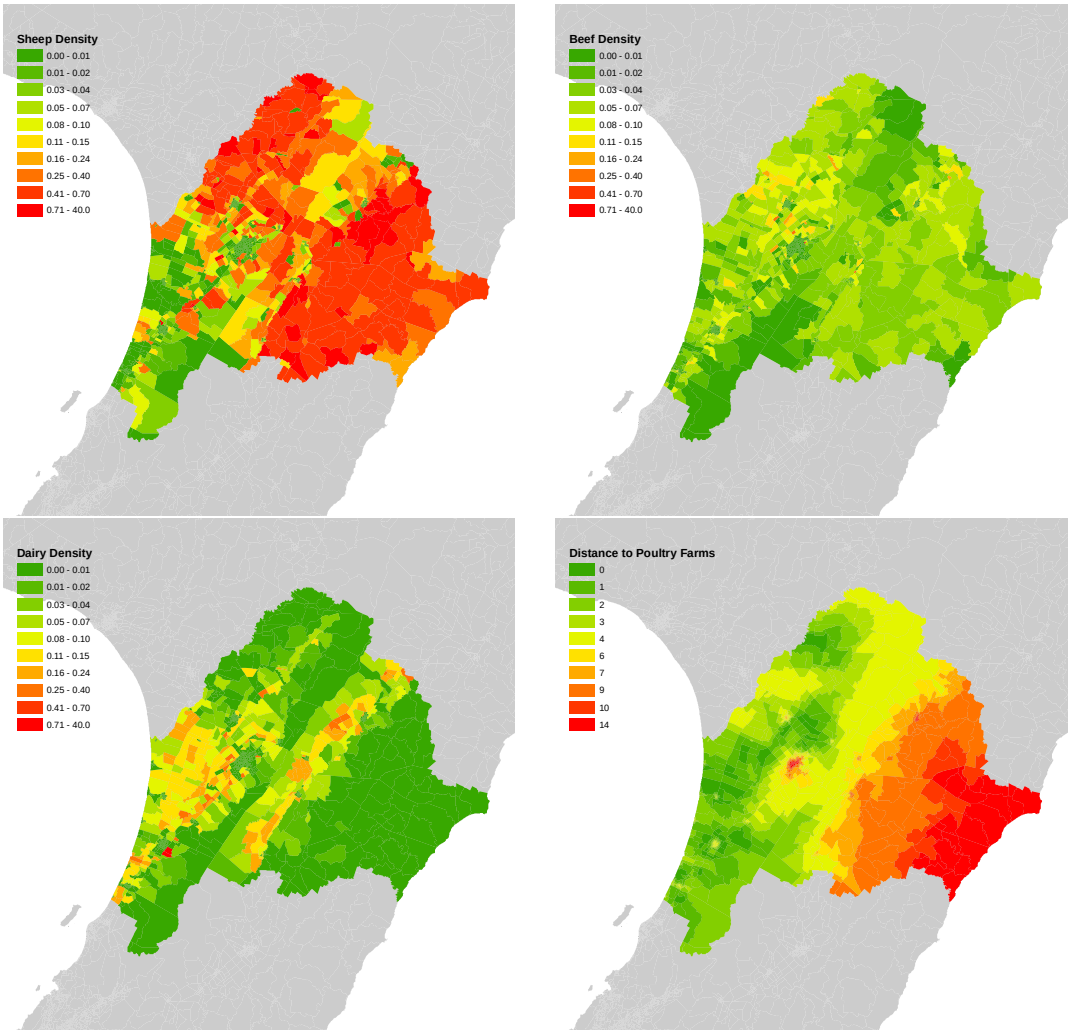

Figure 8: Densities of sheep, beef, and cattle, as well as proximity to a large poultry farm, for the Manawatu region in 2006.

Variable	Manawatu	Canterbury	Auckland
SDI	-	-	-0.046 (0.0039)
Poultry farm distance	-0.049 (0.0236)	0.0242 (0.0203)	-
Sheep density	-	0.376 (0.006)	-
Dairy cattle density	1.51 (0.0508)	0.652 (0.0198)	-

Table 2: Regression coefficients and p-values (in brackets) for each of the predictor variables from the final multiple Poisson regression model for rural notifications after the stepwise elimination of non-informative variables. Beef cattle density was eliminated from all models.

density variables were also eliminated, leaving only the SDI as predictor. In the Manawatu and Canterbury regions, however, SDI was eliminated as a predictor in favour of the poultry farm distance and dairy cattle density. Poultry farm distance was negatively correlated with notification rates in the Manawatu, while being positively correlated in Canterbury. Cattle density was positively correlated with notification rates in both regions, with sheep density also being positively correlated in the Canterbury region.

5.3 Age and gender

For the Manawatu, we have additional data for each case from the EpiSurv database, including gender and age group. Denominator information for these variables was obtained from the 2006 census. The cases were divided into subgroups by each of these variables, allowing analysis of the spatial distribution of each group separately using the model in section 4. Where there were very few notifications in a particular age group, the data were sparse, and hence there were some problems with the convergence of the MCMC algorithm. In these cases, a more informative prior for the hyperparameter κ_U was used ($\kappa_U \sim \text{Gamma}(1.5, 0.15)$). In addition, some of the age groups were also pooled, and chains were run longer in order to obtain convergence.

Table 3 shows the yearly case rates per 1,000 people in the Manawatu. The rates for 0-4 year olds is twice as high as almost all other age groups, and

Group	Notification rate
0-4 years	4.75
5-9 years	1.61
10-14 years	1.58
15-19 years	1.95
20-29 years	2.80
30-39 years	2.39
40-49 years	2.06
50-59 years	2.24
60 and older	2.01
Male	2.73
Female	2.11

Table 3: Notification rates per 1,000 people subdivided by age and gender for the Manawatu.

there appears a slight increase in the 20-29 age group. There is a significant difference in notification rates between males and females ($P < 0.0001$ from a χ^2 test), with males more likely to notify than females.

The estimated spatial distributions of cases in the two youngest age groups (0-4 and 5-9 year olds) is given in Figure 9. We see that cases in 0-4 year olds are much more likely to occur in rural areas, whereas this trend is reversed in the 5-9 year olds, with most cases occurring in urban Palmerston North.

The spatial distributions for gender (not shown) are similar, but the male risk appears slightly more rural with a cluster of increased risk around Foxton.

5.4 Sequence type

With the recent sequence typing of many of the isolates from cases in the Manawatu, we have been able to explore the epidemiology of *Campylobactor* strains with different host associations (Marshall and French, 2009). Figures 10-12 show the estimated relative risk surfaces for ST 474 (a strain associated with one particular poultry company), all other poultry related STs, and ruminant associated STs. The ruminant associated strains have

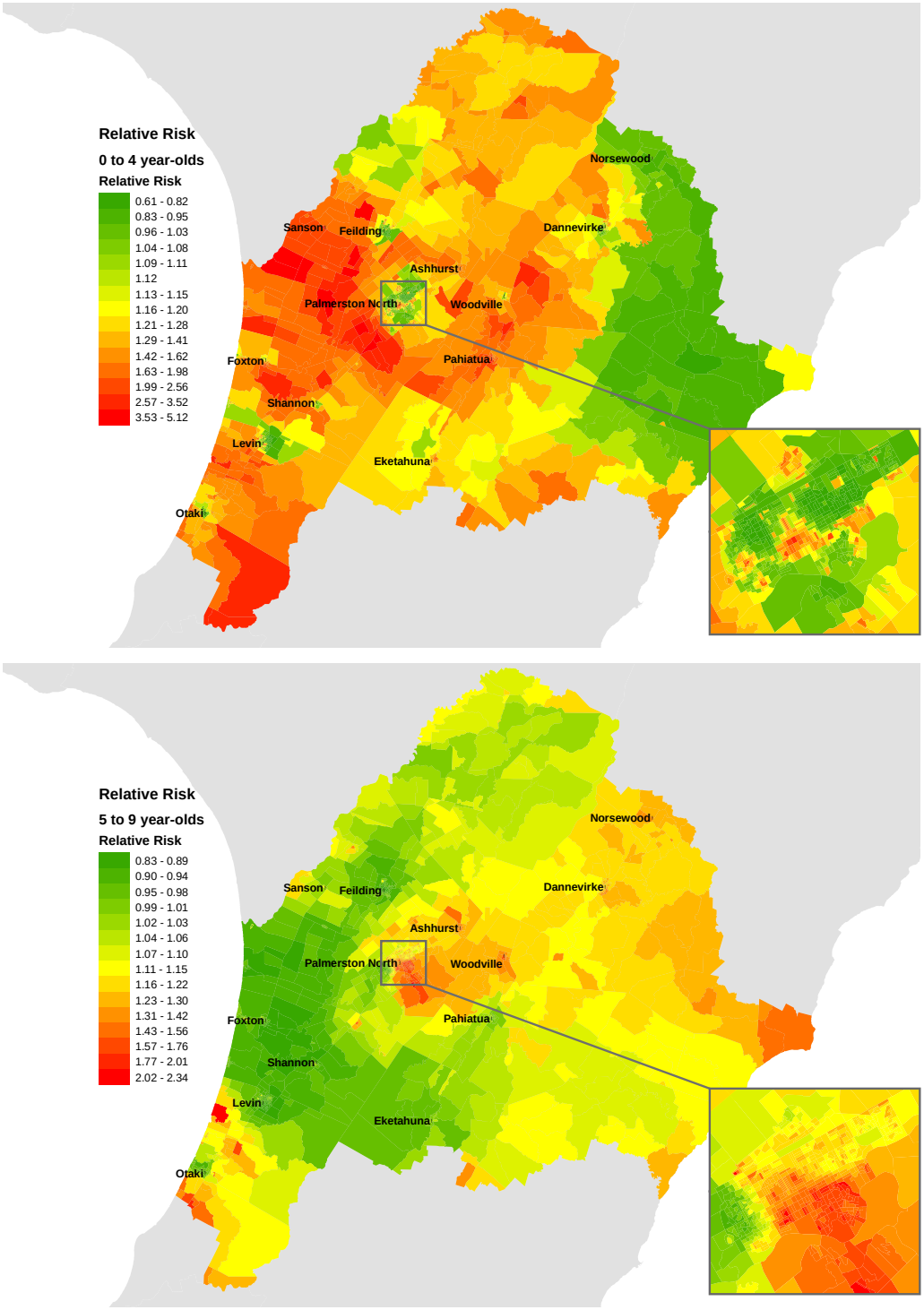

Figure 9: Relative risk for 0 to 4 year-olds (top) and 5 to 14 year-olds (bottom).

increased prevalance in rural areas, and poultry associated strains appear more frequently in urban areas.

5.5 Weather

The relationships between meteorological variables and campylobacteriosis notifications have been assessed in various studies across various locations, particularly with reference to the underlying causes of seasonality (Hearnden et al., 2003; Bi et al., 2008). It is clear from the literature that location affects the way in which weather relates to notifications, possibly due to different epidemiological pathways being more prominent in different locations, with each pathway being affected by the weather differently. Temperature in particular is often suggested to affect notifications. A less sophisticated but more direct approach is taken here, by jointly plotting the case series and weather data to explore potential connections.

The weather data analysed consists of daily maximum temperature, minimum temperature, rainfall, storm water run-off, and hours of sunshine. These data were taken from as many weather stations as possible spread throughout, and just outside, each of the three regions. Daily observations were then averaged to provide a weekly mean. The datasets were centred and normalised by subtracting the mean and dividing through by the standard deviation to allow comparisons to be made. We also applied some simple smoothing by replacing the observation at time t with the mean of any observations between $t-3$ and $t+3$.

Figures 13 to 15 show the time series of campylobacteriosis notifications, mean maximum temperature and mean hours of sunshine. The peaks in notifications appear to occur earlier than the peaks in sunshine hours and temperature. Note that the notification data is likely to be delayed by several days from the actual infection date.

To explore this further, maximum temperature, minimum temperature, rain-

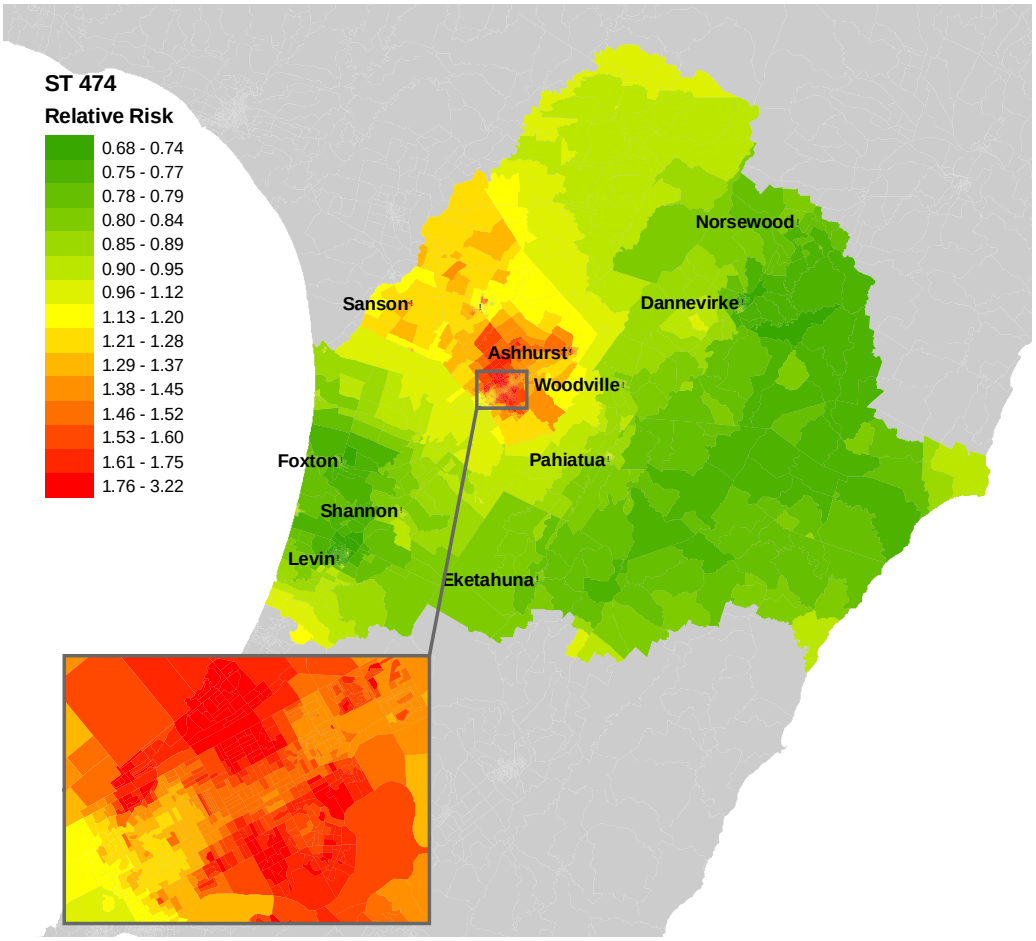

Figure 10: Relative risk maps of the Manawatu region for ST 474 (associated with one particular poultry company).

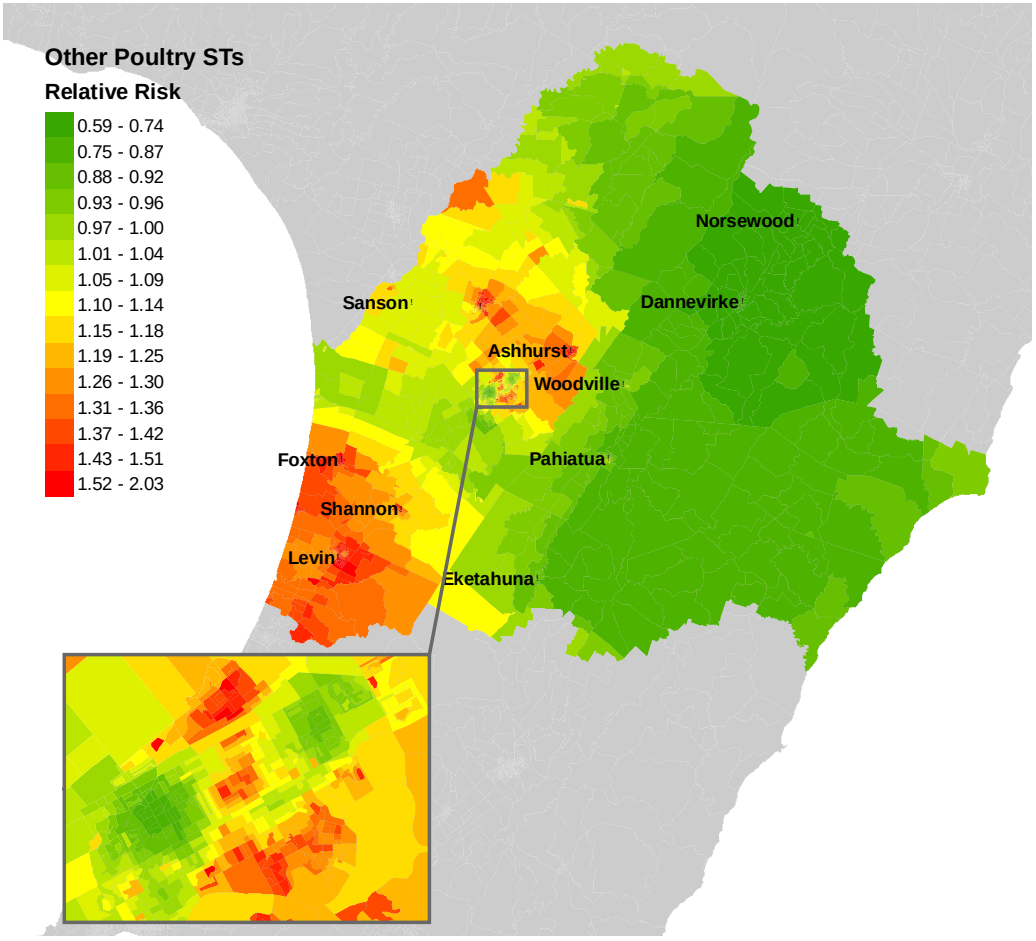

Figure 11: Relative risk maps of the Manawatu region for poultry associated sequence types other than ST 474.

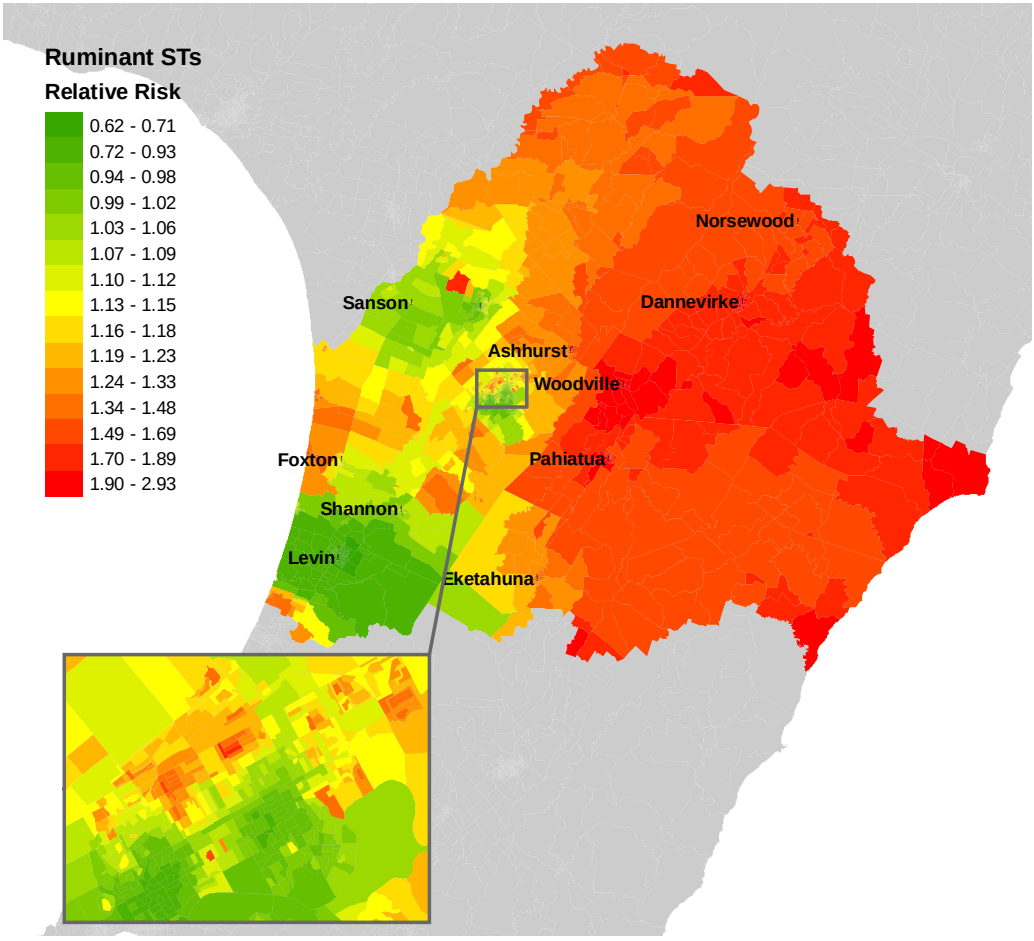

Figure 12: Relative risk maps of the Manawatu region for ruminant associated sequence types.

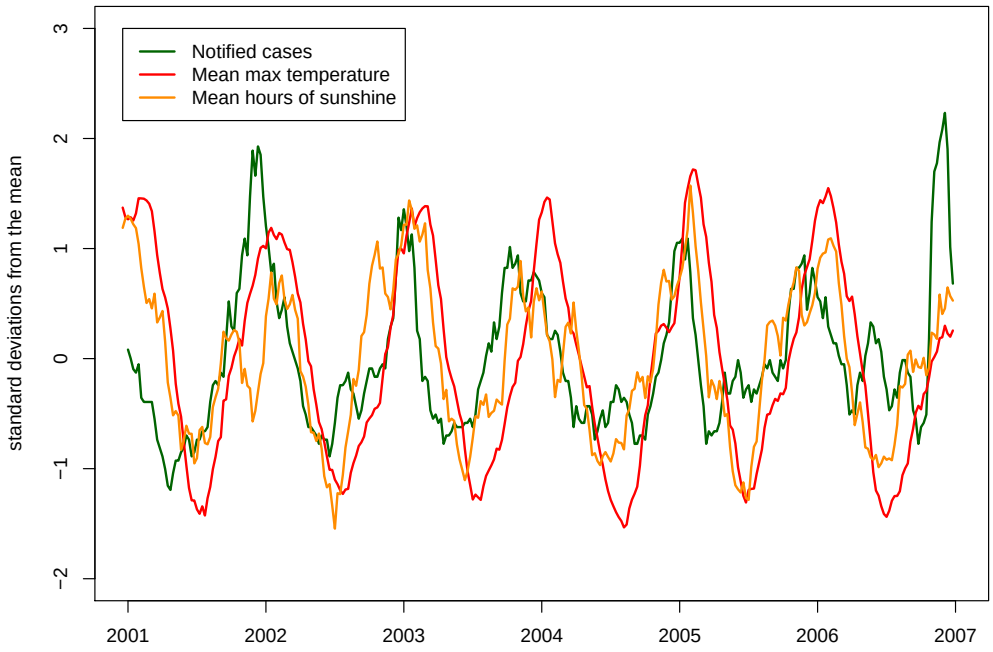

Figure 13: Centred, normalised and smoothed time series of notifications, mean maximum temperature and mean hours of sunshine for the Manawatu region.

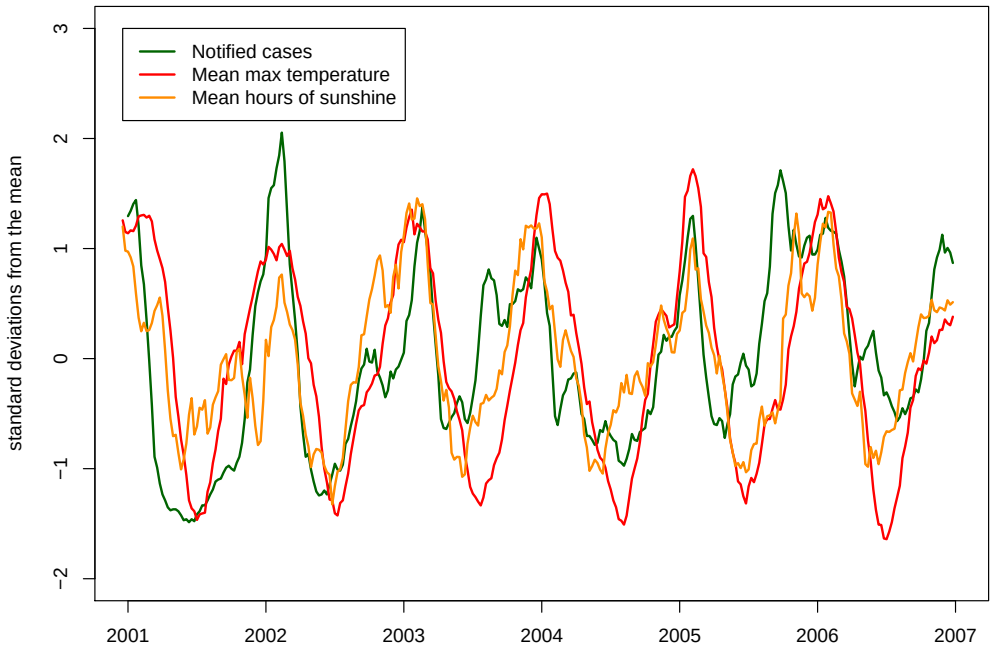

Figure 14: Centred, normalised and smoothed time series of notifications, mean maximum temperature and mean hours of sunshine for the Canterbury region.

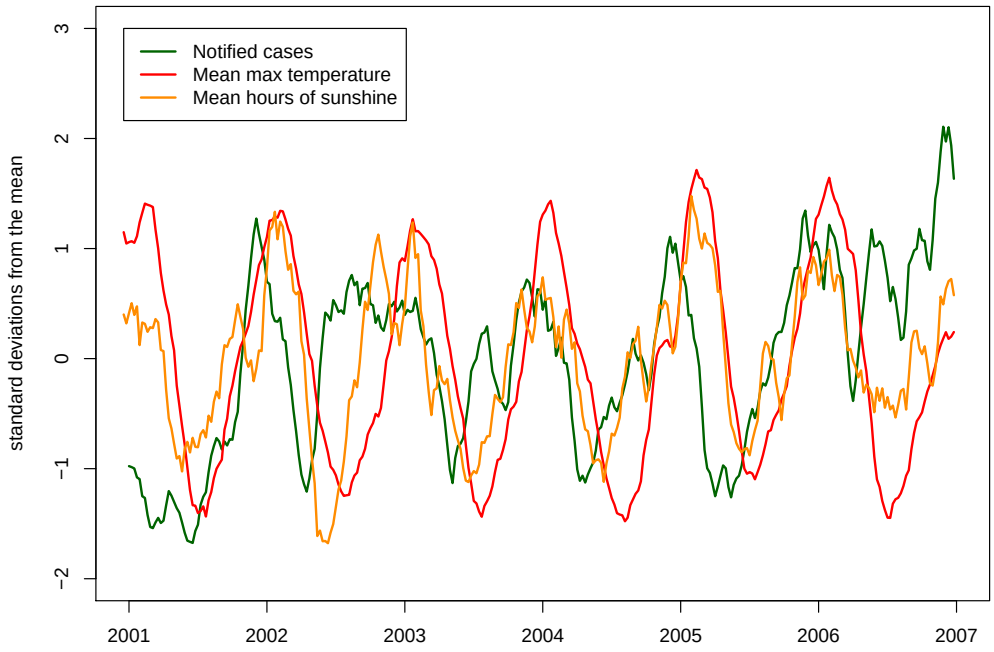

Figure 15: Centred, normalised and smoothed time series of notifications, mean maximum temperature and mean hours of sunshine for the Auckland region.

fall, storm water run-off, and hours of sunshine were interpolated spatially across meshblocks and added as covariates to the model with time lags of zero, one or two weeks. This was accomplished by setting the $W_{i,t}$ term in Equation 1 to

$$W_{i,t} = Z_{L(i,t)} + Z_{L(i,t-1)} + Z_{L(i,t-2)},$$

where $L(i,t)$ is the level of weather in meshblock i at time t and the Z_k are fixed effects. The priors on the Z_k are first order Guassian walks as was done for the fixed effects for SDI,

$$Z_{k+1}|Z_1, \ldots, Z_k, \kappa_Z \sim N\left(Z_k, \frac{1}{\kappa_Z}\right),$$

where κ_Z is again given a Gaussian conjugate prior. However, no consistent relationship with campylobacteriosis notification risk was found for any of these four variables.

Another suggested explanation for the increase in campylobacteriosis notifications during the summer months is barbecue use, where unfamiliar catering practises may increase the risk of cross contamination or under-cooking. Unfortunately, barbecue-use data is difficult to obtain or estimate. As a proxy, the number of days of warm, dry weather each week was calculated, with weekends and public holidays having a higher weight. This variable could also approximate other things, such as exposure to environmental *Campylobacter* through swimming in rivers. Although the relationship is far from straightforward, the risk of getting campylobacteriosis does appear to increase one week after the presence of warm, dry weather in Auckland and the Manawatu, as seen in Figure 16. This trend is not observed in Canterbury, however, suggesting different sources of infection in this region.

The main problem when attempting to predict notifications using weather data is that the data are usually available via various summary statistics, such as weekly average rainfall or daily maximum and minimum temperatures. These statistics may fail to capture the diverse and rapidly changing range of weather conditions that effect both bacteria and, perhaps more importantly,

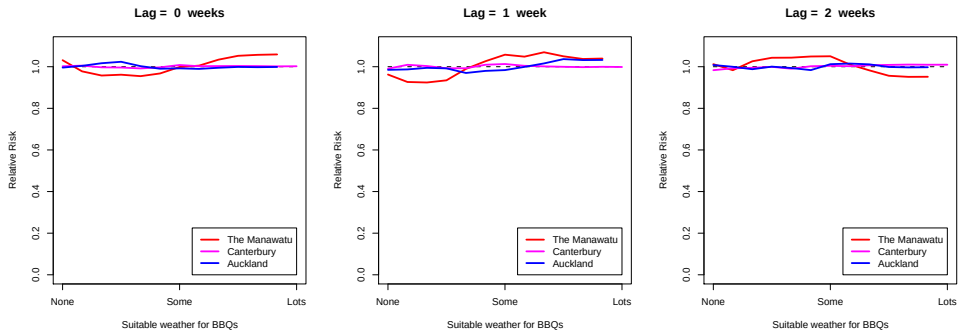

Figure 16: Relative risks for different amounts of weather suitable for barbecues.

human behaviour. This difficulty is highlighted in the literature, in which evidence can be found for the opposing hypotheses that warmer temperatures can both increase or decrease notifications (Bi et al., 2008).

A further complication arises from the inherent delay and variable smoothing present in the notification data itself. The length of time between infection and notification is not well understood, but is likely to vary between one and three weeks for the majority of reported cases. This may flatten any sharp peaks in notifications, and thus obscure any causal relationships. Any temporal trends with a longer time frame, such as seasonal patterns, however, would not be hidden in this way, and thus we should expect these to be easier to identify. While there is a relationship to be seen, as in Figures 13-15, this relationship is not conducive to predictive modelling.

The EpiSurv database allows for the date of onset to be recorded. Although still after the date of infection, the onset of symptoms may give a more accurate picture of when infection has occured, and thus allow associations to be identified more precisely. Unfortunately, for the dataset used in this project, the onset date has not been recorded in the majority of cases, and the missing data appears non-uniformly within the dataset. The completion dates in Auckland and Canterbury in particular are very low, so there is a spatial pattern to onset date completion.

Such biases make it difficult to identify any associations between onset date and weather events across all regions, as the onset date completion trends predominate. However, within the Manawatu data, in particular in the last 3 years, we have reasonable completion of onset date available. This subset of the data was analysed to determine the correlation between weather variables and notification rates at different time offsets.

With onset date information there is no need to group at the weekly level, as there is less likely to be a day of week effect than with the notification date, which occur only Monday to Friday. Thus, for each possible lag (from 50 days to -50 days) between weather and notifications, we offset the daily weather data, and then smoothed both weather and notification data with a 6 week moving average. We centered and normalised each time series, and then computed the correlation. The smallest lag with the highest correlation occurred when the sunshine hours data lagged behind the notification data by 17 days ($\rho = 0.52$). This suggests that, in addition to these meterological variables, there may be other drivers of the seasonal trend in human notifications.

5.6 Poultry retail sales

Recently, it has been shown that the majority of campylobacteriosis cases in New Zealand (Mullner et al., 2009), Scotland (Sheppard et al., 2009) and the United Kingdom (D.J et al., 2008) are poultry associated strains.

Poultry retail sales data was made available by the Poultry Industry Association of New Zealand from March 2005 through to October 2008. Retail sales information from all four major poultry companies (Companies A through D) was available at differing temporal and spatial resolutions. The data from company D was excluded due to having summary data only at the year and region level. The data from company C was also excluded due to retail sale information being on an as yet undetermined scale. Company A sells the majority (70%) of their fresh poultry to two large abattoirs (one in Auck-

land and one Christchurch), which are likely then to distribute to the major supermarkets throughout the country. The absense of spatial resolution for these data meant that it has also been excluded. Thus, the remaining data represents approximately 30% of the retail data from company A, and full retail data from company B, and therefore may not represent the true spatial or temporal distribution of retail poultry sales throughout the three regions.

Given that the data may not be representative, modelling notifications using this data is unwise, and we instead elected to do some straightforward comparative analysis against notifications. The data were interpolated spatially and temporally to the meshblock/week level, allowing direct comparison to the notification data. Figures 17 and 18 show the spatial distribution of poultry sales in the Manawatu and Canterbury regions respectively for the two poultry companies. There does not appear to be any significant correlation with the relative risk of Figures 2 and 3, and we see that the large urban areas of Palmerston North and Christchurch have very little sales data recorded.

The temporal distribution of the Manawatu sales data is given in Figure 19. Of interest here is the apparent peak in retail sales from company A which corresponds to the winter epidemic of 2006 (McTavish et al., 2008).

6 Identifying anomalous events

In order to model temporary periods of increased risk, Knorr-Held and Richardson (2003) propose using indicator variables $X_{i,t}$ that switch from zero to one during epidemic periods. The advantage of using such indicators in the Bayesian setting is that the posterior distribution consists of the probability that meshblock i is experiencing an epidemic in week t for each i and t. This allows simple interpretation of these spatio-temporal indicators, allowing the indicators with high posterior probability to trigger further investigation into the cases at that time and location.

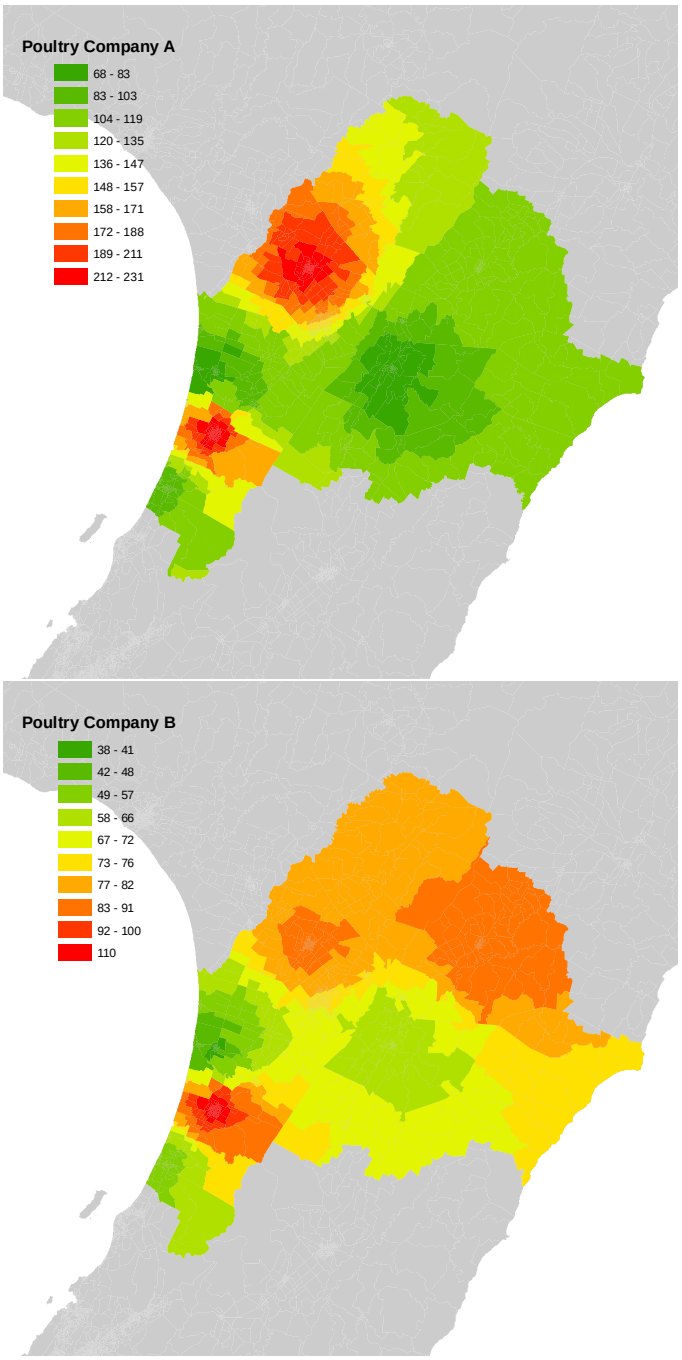

Figure 17: Average monthly poultry retail sales (kg) for two companies in the Manawatu.

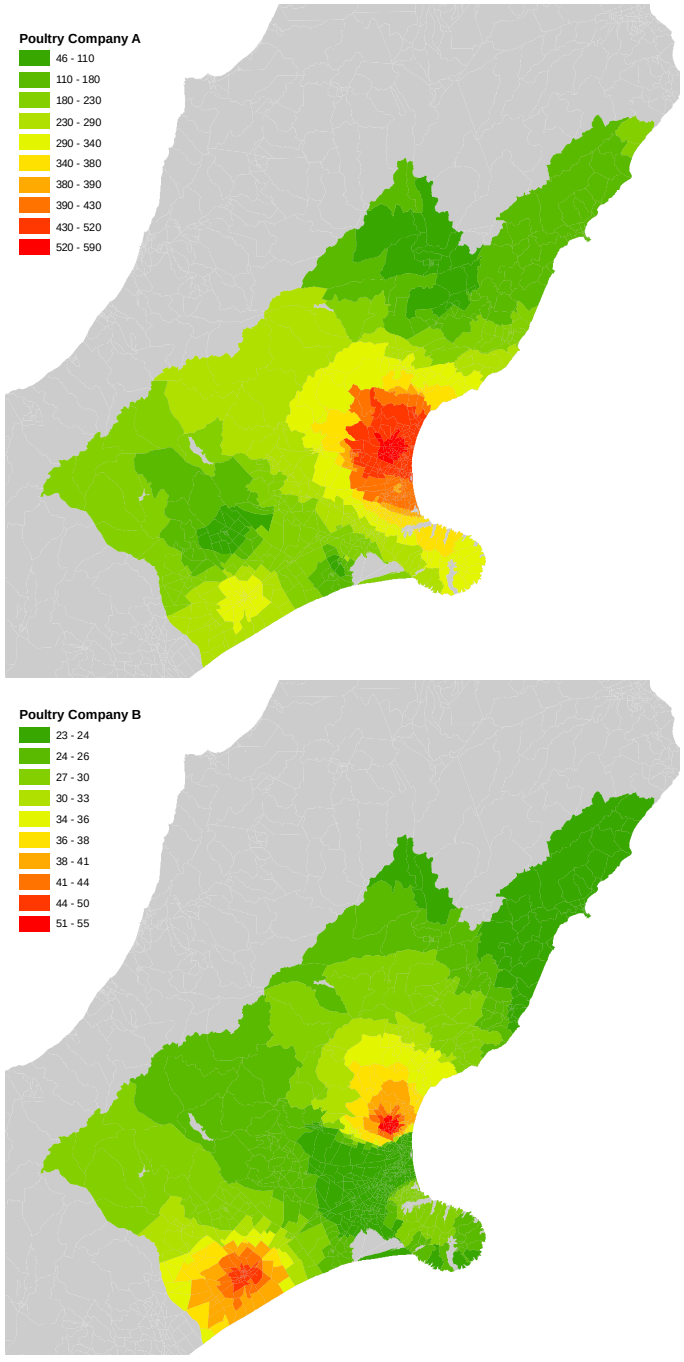

Figure 18: Average monthly poultry retail sales (kg) for two companies in the Canterbury.

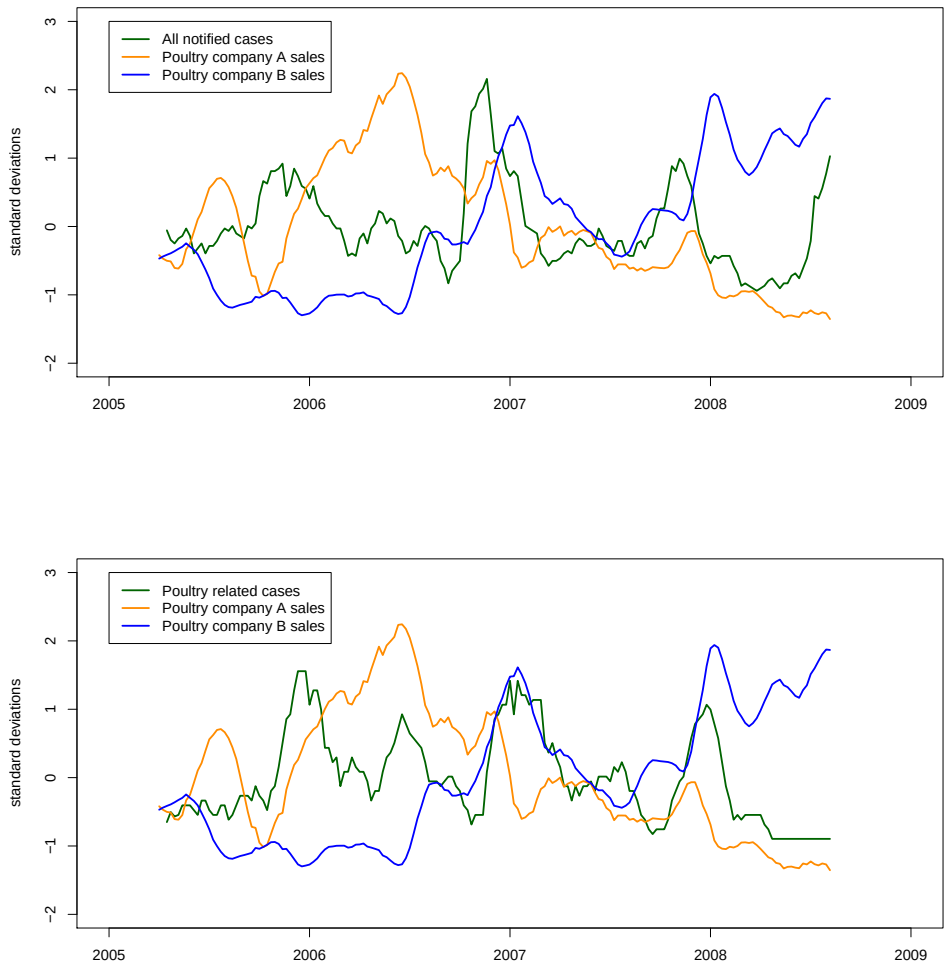

Figure 19: Centred, normalised and smoothed time series of all notifications (top) and poultry related notifications (bottom) compared with fresh retail poultry sales from the Manawatu region.

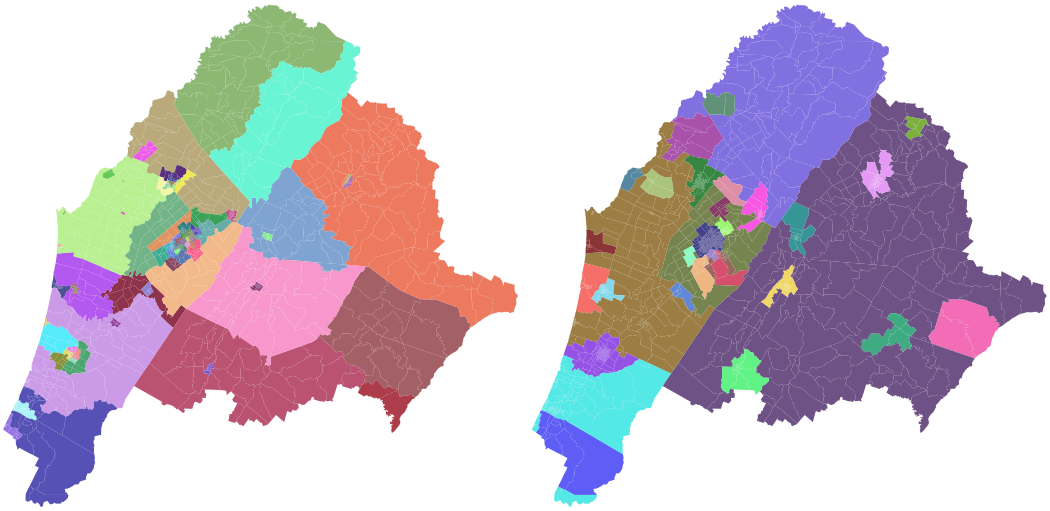

Figure 20: Maps showing the area unit regions (left) and the water supply regions (right) used for detecting localised outbreaks in the Manawatu.

Due to the large number of meshblocks and time periods (over 9000 meshblocks in Auckland, over 300 time periods), it is necessary to combine several meshblocks into 'regions' in which outbreaks are to be detected, thus reducing the number of outbreak indicators, in order to improve the performance of the estimation. We have used two separate schemes for this. The first scheme uses the census area units to group meshblocks into regions, and the second is based on water supply, so that meshblocks that share the same water supply are in the same region. As an example, the two regions for the Manawatu are shown in Figure 20.

There are two models proposed: the Knorr-Held and Richardson model, which uses temporally correlated priors, and the independent indicators model. Both models use the same basic structure to estimate the underlying spatial and temporal components, differing only in the priors. Given equation 1 we set

$$W_{i,t} = \beta_{r(i)} X_{r(i),t},$$

where $r(i)$ is the index of the region containing meshblock i, $X_{r,t}$ is the

indicator for the presence or absence of an outbreak, and β_r reflects the size of the increase in risk for outbreaks in region r. For the outbreak size parameters, we assume that $\beta_r \sim \text{Exp}(1)$ to assist with convergence.

The Knorr-Held and Richardson model uses priors on the $X_{r,t}$ that take the form of a Markov chain with transition probabilities $p[a, b]$, where a and b are 0 or 1. $p[a, b]$ then, represents the probability that the indicator variable goes from $X_{r,t-1} = a$ to $X_{r,t} = b$. Hence,

$$X_{r,t} \sim \text{Bernoulli}(p[X_{r,t-1}, 1]).$$

The priors for the transistion probabilities are $p[0, 0], p[1, 1] \sim U(0, 1)$, which then define the other two parameters as $p[0, 1] = 1 - p[0, 0]$ and $p[1, 0] = 1 - p[1, 1]$.

The independent indicators model, on the other hand, has a simpler prior structure that removes the temporal dependence. We set

$$X_{r,t} \sim \text{Bernoulli}(p),$$

where p is given the prior $p \sim \text{Beta}(1, 51)$. This prior suggests that we expect on average one outbreak a year in each region.

The lack of temporal correlation in the independent indicators model should allow new outbreaks to be detected in their first observable week. In contrast, the temporal correlation in the Knorr-Held and Richardson model could prevent outbreaks being detected in this first week. The Knorr-Held and Richardson model was initially applied to meningococcal disease incidence data, which involves person to person transmission. As there is no transmission between people with campylobacteriosis, the temporal correlation is not as applicable. The independent indicators method, therefore, is likely to be more suited to applications in which real-time monitoring of outbreaks is required.

Figures 21 and 22 show the output from the Knorr-Held and Richardson and

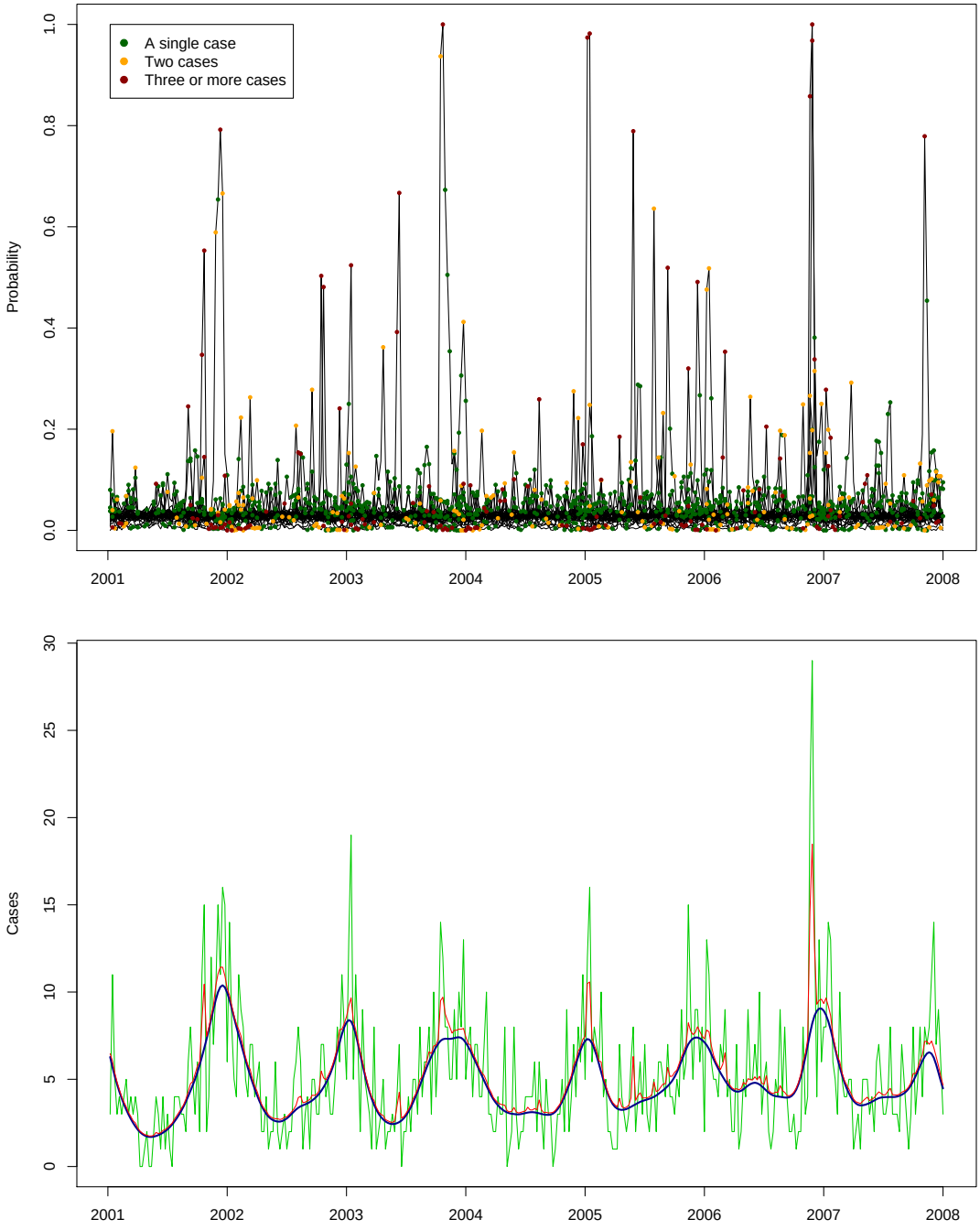

Figure 21: The posterior probability of an outbreak with all regions overlayed (top) and the estimated number of cases each week (bottom) for the Knorr-Held and Richardson method. Both graphs use the water supply regions for the Manawatu data.

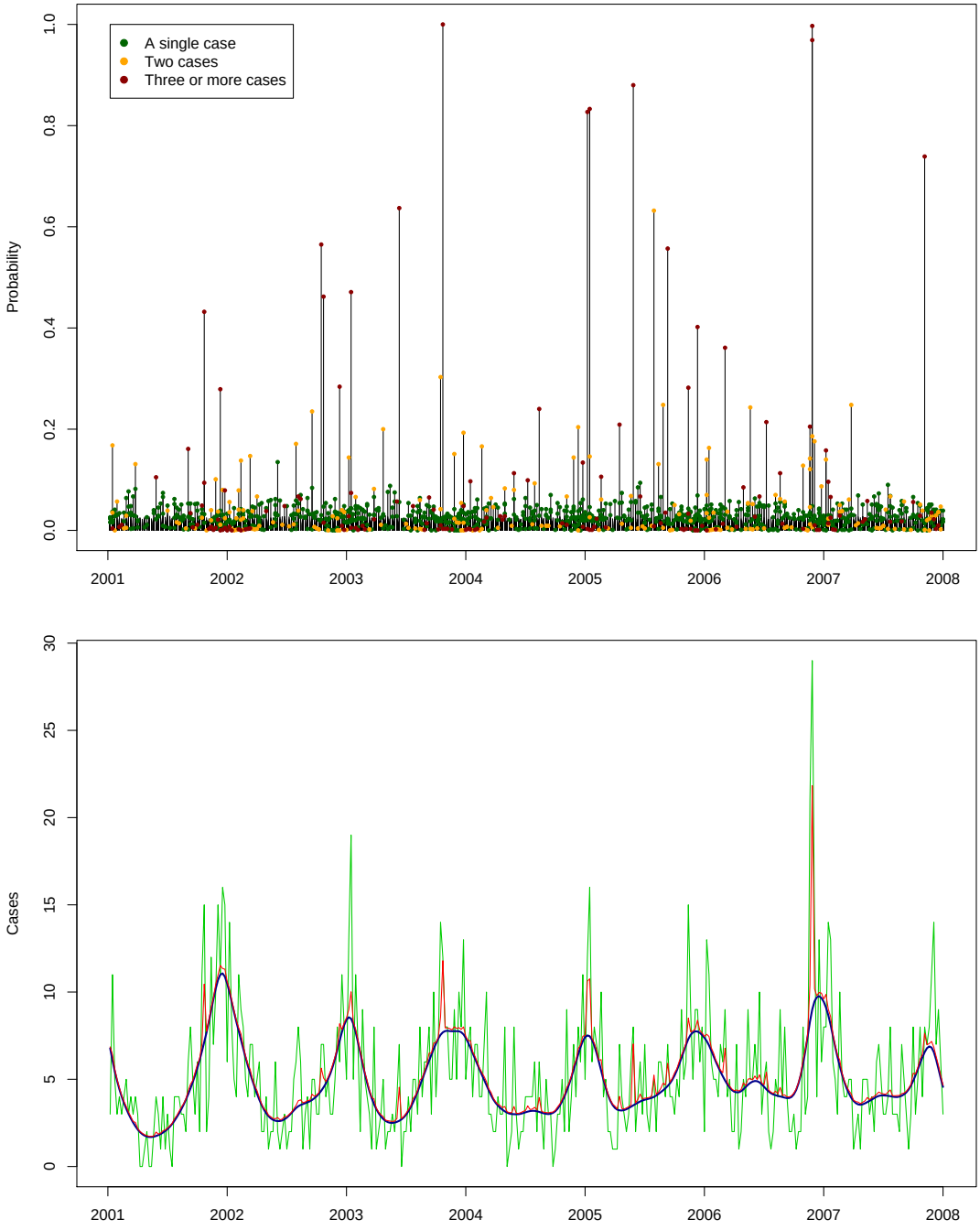

Figure 22: The posterior probability of an outbreak with all regions overlayed (top) and the estimated number of cases each week (bottom) for the independent indicators method. Both graphs use the water supply regions for the Manawatu data.

independent indicators models respectively for the Manawatu region. The coarser grid of the water supply regions has been used, and these appear to be more useful than the area units at detecting outbreaks. This may be due to their larger size, which may increase the chance at detecting outbreaks where the geographical location of cases are not as important, such as an outbreak caused by contaminated food at a restaurant. The two models agree for the most part, with the independent indicators model appearing to be slightly more discriminatory. Note that the number of outbreaks detected is not overwhelmingly large, so that any further investigation of epidemiological factors associated outbreaks is minimized. Also, note that, in the bottom figures, some of the spikes in the case series (in green) are explained as outbreaks (estimated number of cases in red). Not all the spikes are explained, however – some of them may just be random variation, or possibly outbreaks spread over too large a region to be detected by the model.

The independent indicators model output for the Canterbury and Auckland regions are shown in Figures 23 and 24.

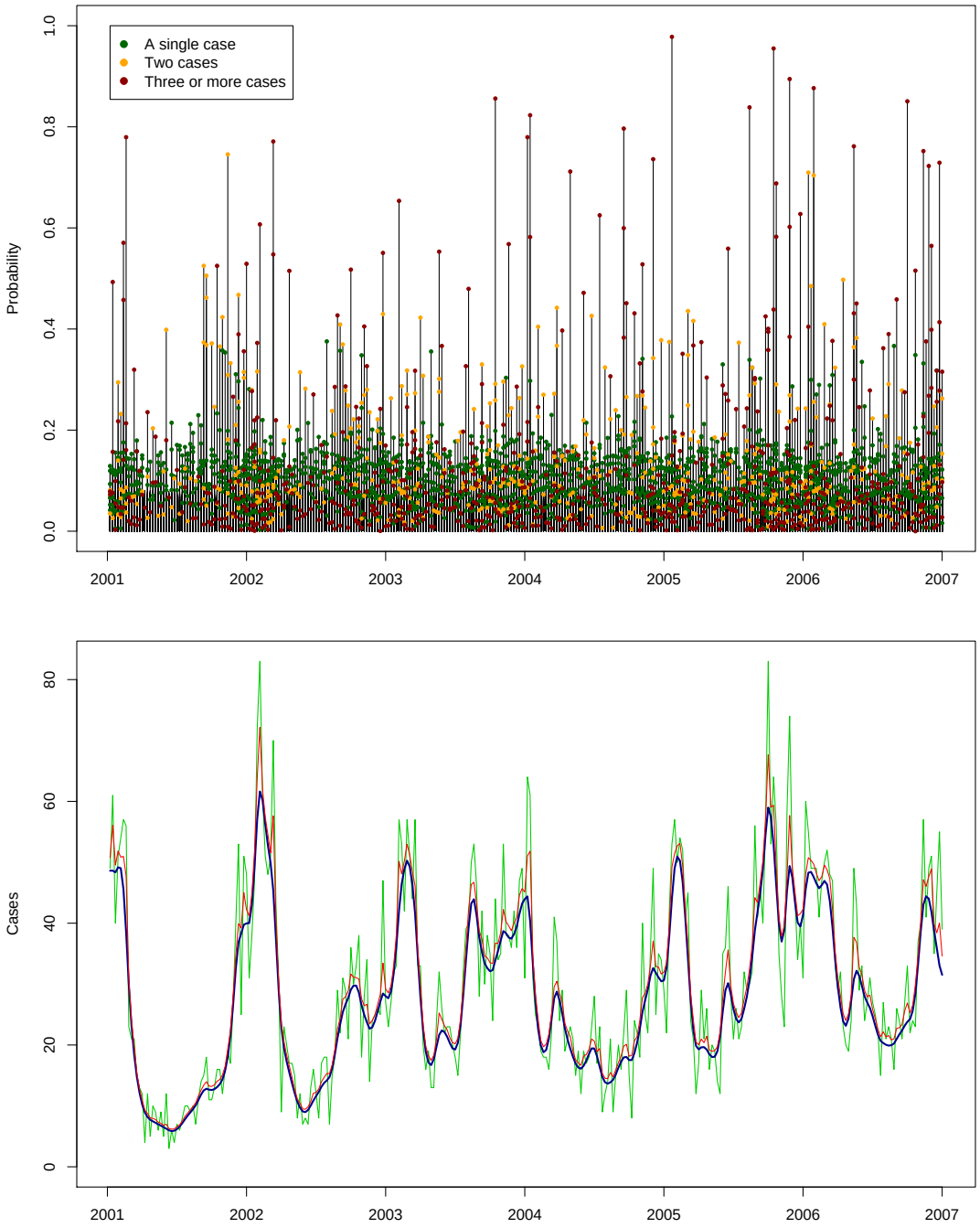

Figure 23: The posterior probability of an outbreak with all regions overlayed (top) and the estimated number of cases each week (bottom) for the independent indicators method. Both graphs use the water supply regions for the Canterbury data.

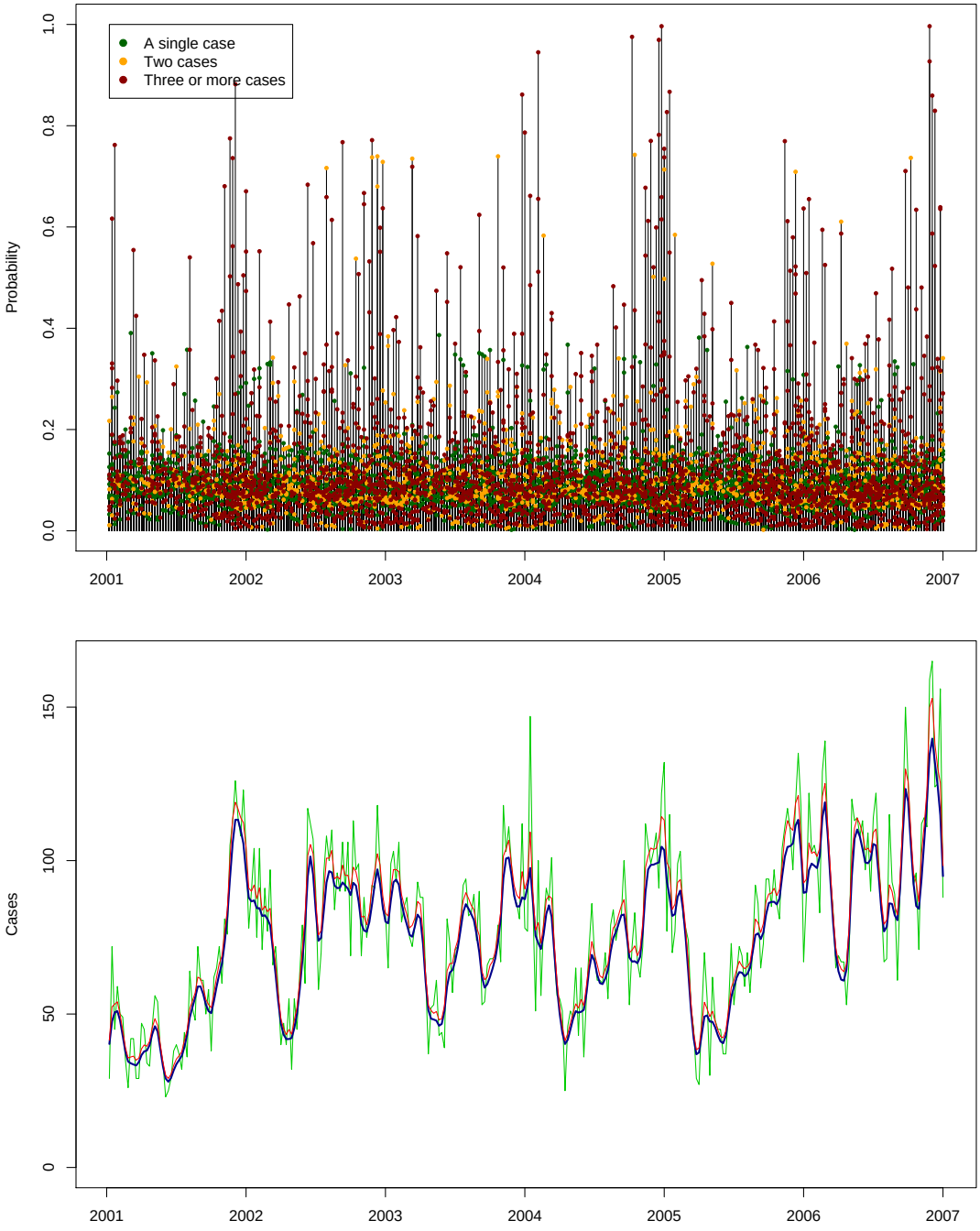

Figure 24: The posterior probability of an outbreak with all regions overlayed (top) and the estimated number of cases each week (bottom) for the independent indicators method. Both graphs use the water supply regions for the Auckland data.

7 A potential surveillance tool

A tool has been developed to provide an easy to use interface for running the model. The tool provides a traditional Windows dialog box interface that allows the input datasets to be described, and the basic parameters for the model to be specified. The tool then allows the model to be run in the background and provides feedback to the user during processing. Once finished, the output data files are provided along with graphs similar to those in Figures 22-24, and analysis can then be done in order to investigate possible outbreaks. Details on the convergence of the model are also output.

A screenshot of the user interface is shown in Figure 25. The required input data files consist of

1. A meshblock file, specifying the meshblock numbers and the population within each meshblock.
2. A neighbours file, linking each meshblock with its neighbours $n(i)$.
3. A regions file, specifying the map $r(i)$ from meshblock to region.
4. A cases file, specifying the date of the case, the meshblock, and optionally the accuracy of the meshblock assignment.

The first three files may be setup once for each region and then reused with each model run. These are plain text files that are reasonably easy to generate using census and GIS data. The last file, the cases file, is the only file that is expected to change between runs of the model, and thus the requirements for this file are less stringent. It is a comma separated file with headers included. When the user selects the file, each of the drop down boxes for selecting which field should be interpreted as the meshblock or date fields are populated, allowing simple selection by the user. A third, optional field giving the accuracy of the assignment of the case to a meshblock may be specified, and a set of allowed values chosen. This allows those cases that have inaccurate spatial information (e.g. they have been assigned to a meshblock

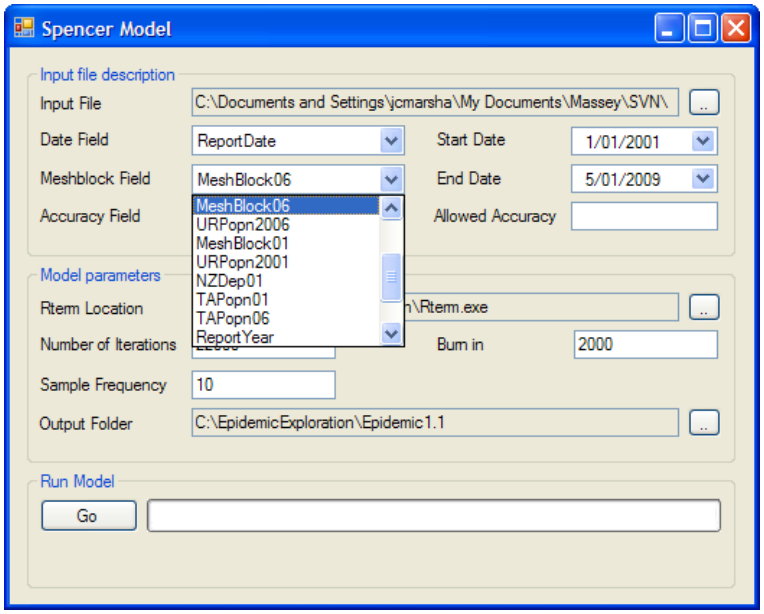

Figure 25: Selecting the field for meshblocks in the surveillance tool.

based on territorial authority information, rather than true address details) to be ignored by the model.

Once the input files are selected, and the appropriate fields are chosen, the model parameters can be defined, which includes the number of iterations to run, how many of these should be treated as the burn-in period, and the thinning rate used for sampling from the posterior distributions. Once specified, the user selects the "Go" button, and the model starts running. Feedback is provided via the number of iterations complete and a progress bar as shown in Figure 26. Once complete, an analysis phase runs, generating a set of plain text files, specifying potential outbreaks and relative risk values for each meshblock, and pdf graphs of the temporal model. A potential outbreak is noted if the posterior probability $X_{r,t}$ exceeds a predefined threshold ($\hat{p} = 0.1$) and there are more than one cases in region r at time t. These files and others allowing further analysis are placed in a folder selected by the user.

In order to assess the use of the tool in a surveillance scenario, we selected

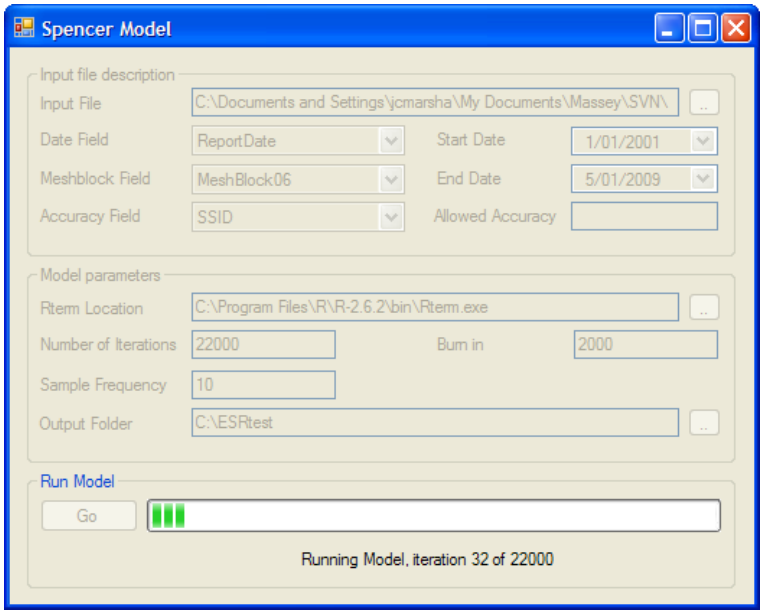

Figure 26: The surveillance tool during a model run.

Outbreak	Region(s) Affected	Number of Cases
11 January 2005	1	8
25 May 2005	6	7
15 November 2006	2, 6	8
21 November 2006	4, 6, 11, 22	23

Table 4: Outbreaks used for epidemic detection with the surveillance tool.

four outbreaks previously identified by the MidCentral District Health Board and ran the model on the case data surrounding each outbreak. We ran the independent priors model on the water supply regions using case data spanning two years up to and including the week in which the outbreak was identified. We also ran the model including two weeks of data after the outbreak was identified. Details of the outbreaks used are given in Table 4.

The model was run using 22000 iterations with 2000 iterations burnin, sampling from the posterior distributions every 10th iteration. Such runs take approximately 12 hours on a standard 2GHz desktop computer.

Outbreak	Region	Probability	Number of Cases
11 January 2005	1	1.0000	10
25 May 2005	6	0.4785	8
	6	0.948	9
15 November 2006	2	0.620	6
	1*	0.104	2
	6	1.000	26
21 November 2006	11	0.999	6
	22	0.303	2
	1*	0.121	4

Table 5: Likely outbreaks from the surveillance tool for four epidemics in the Manawatu region. Starred regions are those which are unlikely to be part of an outbreak, due to their probabilities being low.

Output from the model runs is shown in Table 5. As can be seen, the tool correctly detects all four outbreaks in the week in which they occurred. The probabilities given are from the model runs with data up to and including the week in which the outbreak occurred. With the addition of an extra two weeks data after the event, some of the posterior probabilites are increased (e.g. the likelihood of the 25 May 2005 outbreak increases to 0.7825). A comparison with Table 4 shows that the tool reliably picks up the regions present in each of the outbreaks, other than those regions in which only one case was observed (such as region 4 in the 21 November 2006 outbreak). Although these results are encouraging, care must be taken in ensuring that the threshold for detection $\hat{p}$ is of a sufficient level so as to eliminate false positives, while still allowing detection of actual outbreaks. The value used in this modelling of $\hat{p} = 0.1$, for instance, would not be applicable to the Canterbury or Auckland datasets, where a level of $\hat{p} = 0.25$ may be more appropriate (see Figures 23 and 24.)

8 Discussion

The purpose of this study was four fold: Firstly to develop a model for the background spatial and temporal trends in the data, secondly to identify potential risk factors that may be associated with those trends, thirdly to develop a method of identifying anomalous outbreaks over and above these trends, and finally to produce a tool suitable for use in the surveillance of campylobacteriosis in New Zealand.

8.1 Spatial and temporal trends and risk factors

The temporal and spatial components of the model in Section 4 effectively yield the background spatial and temporal trends in the data. By specifying prior information for each time period and spatial location in terms of the neighbouring points, we smooth the data, bringing the long term trends to the forefront.

There is a clear difference in spatial risk factors between urban and rural areas. In urban areas, the Social Deprivation Index is highly associated with notifications, with deprived areas reporting fewer notifications. This relationship may be caused by underreporting due to differing access to health care, as the main route of notification in New Zealand is via General Practitioners, and there is a cost associated with visiting a GP. A study of utilisation of GP services (Scott et al., 2003) suggested that, while government subsidies targeted at low-income families reduce the barrier posed by doctors fees, they does not fully compensate these costs. Furthermore, there may be other barriers preventing low-income people from consulting GPs, such as a lack of transport or communication difficulties (Lake et al., 2009). There may also be differences in food consumption patterns between high and low income groups. For example, people in lower socioeconomic groups may consume less fresh poultry, instead favouring less expensive frozen poultry that has is associated with a lower risk of *Campylobacter* infection.

In rural areas, on the other hand, animal densities and proximity to poultry farms play a larger role, with increased notifications occurring in areas with a high dairy cattle density in the Manawatu and in areas with high dairy or sheep densities in Canterbury. The MLST results support this, with a clear increase in ruminant strains in rural areas. This suggests that rural infections may originate through increased exposure to faecal material in the environment, rather than through food sources. Occupational exposures in rural areas, such as the handling and treatment of animals on farm, as well as the exposure of meat workers to faecal matter during slaughter (Gilpin et al., 2008) support this hypothesis, as does the noticeably higher risk among 0-4 year olds in rural areas.

The temporal trend in *Campylobacter* notifications is associated with variation in weather. There is more variation in notification rates in areas in which the climate allows larger variation in temperature through the year, such as in Canterbury. The number of sunshine hours and temperature appear associated with notification rates, with the correlation strongest when the weather data lags onset date data by 17 days. Though these associations are present, the weather variables prove less useful as predictors in the model. However there appears to be a slight increase in notifications in the week following a period of warm, dry weather in the Manawatu and Auckland. Such periods may be associated with increases in activities in which the exposure to *Campylobacter* may be higher, such as swimming in rivers or barbecuing.

8.2 Identifying anomalous events and surveillance

With the spatial and temporal components of the model accounting for the background trends in the notification data, the residual information then contains any peaks over and above the average. Modelling these peaks in terms of epidemic indicator variables allows the probability of an epidemic occurring at a particular time in a particular region to be estimated.

Two prior proposals for the epidemic indicators were compared, both per-

forming similarly, with the independent indicator model appearing to be slightly more discriminatory. Two separate region layouts were also compared, with the larger water regions giving more useful output, particularly in the Manawatu and Canterbury regions. It is important to note that, while the method detects a number of possible outbreaks, and thus accounts for some of the spikes in cases through time, it does not explain all such spikes – some may be due to outbreaks spread out across too large a region.

Given the potential use of the model as a surveillance tool, a graphical front-end was developed to assist and automate the repeated running of the model on a weekly basis. The tool produces summary output intended to be useful in triggering additional investigation of possible outbreaks immediately following the week in which they occur. Retrospective model runs on the Manawatu data have shown that it can reliably detect larger outbreaks, and with the output including the probability of an epidemic, a threshold can be set to reduce the chance of false positives. Allowing outbreaks to be flagged at an early stage may assist health professionals in enacting followup investigations in a timely manner.

9 Recommendations

The surveillance tool is a useful addition to New Zealand's current surveillance system (EARS[4]). Once the spatial information is defined for a particular region, the model is straightforward to setup and use and, on completion of a model run, provides a summary output file detailing cases that may be linked. If run regularly, such output may allow outbreaks to be identified as little as a week after they occur, allowing for timely follow-up investigation by health professionals.

We have found the water regions appear more useful than the area unit regions at identifying outbreaks, particularly where rural regions are included.

[4]EARS is a product of the Centers for Disease Control and Prevention, USA.

Other groupings of meshblocks may be preferred in predominantly urban areas such as Auckland.

Further work investigating the efficacy of the model's outbreak identification would be useful across additional regions, in order to identify appropriate groupings of meshblocks, and appropriate thresholds for detection, in each region. In addition, investigation into running the model on a national scale would be useful. While the model could be run in parallel across individual District Health Boards (DHBs), there may be interest in applying a single run of the model at a national level, under a suitable transformation of the spatial units. One could replace the current meshblock level information with area units, for example, and replace the water regions with DHB boundaries. This would allow the flagging of anomalous events across a particular region, and thus suggest further follow-up investigation within that region, in a similar vein to what EARS currently provides. Additionally, the probability associated with the anomalous event may allow a graduated response to be adopted.

10 Acknowledgements

This work was funded by the New Zealand Food Safety Authority and was done in collaboration with ESR Ltd, Kenepuru and MidCentral Public Health Services. We acknowledge the following individuals who contributed to the writing of this report: Dr Daniel Wilson (University of Chicago, formerly Lancaster University, UK), Dr Anne Midwinter, Dr Julie Collins-Emerson, Petra Mullner and Tui Shadbolt. The work was carried out by the above and Rebecca Pattison, Rukhshana Akhter, Errol Kwan, Lynn Rogers, Isabel Li, Jim Learmonth, Anthony Pita, Sarah Moore, (Massey Molecular Epidemiology group, Hopkirk Institute), Dr Phil Carter, Ruth Pirie and Dr Tecklok Wong (ESR), Dr Grant Hotter (AgResearch). We also acknowledge the following individuals and organisations for their contribution to the study: Palmerston North Hospital, MidCentral PHS, MedLab Central, Gra-

ham McBride (NIWA), Poultry Industry, Professor Mike Hedley, Dr Geoff Jones, Dr Alasdair Nobel, Professor Martin Hazelton, Dr Michael Baker, Diane Richardson, Horizons Regional Council, and the farmers in the Manwatu who allowed us to sample cattle and sheep.

References

M.G. Baker, N.A. Wilson, R. Ikram, S. Chambers, P. Shoemack, and G. Cook. Regulation of chicken contamination is urgently needed to control New Zealand's serious campylobacteriosis epidemic. *N Z Med J.*, 119 (1243), 2006.

M.G. Baker, E. Sneyd, and N.A. Wilson. Is the major increase in notified campylobacteriosis in New Zealand real? *Epidemiology and Infection*, 135 (1):163–70, 2007.

J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43:1–21, 1991.

Peng Bi, A.S. Cameron, Y. Zhang, and K.A. Parton. Weather and notified *Campylobacter* infections in temperate and sub-tropical regions of Australia: An ecological study. *Journal of Infection*, 57(4):317–323, 2008.

P. Diggle, L. Knorr-Held, B. Rowlingson, T-L. Su, P. Hawtin, and T.N. Bryant. On-line monitoring of public health surveillance data. In R. Brookmeyer and D.F. Stroup, editors, *Monitoring the health of populations*. Oxford University Press, 2002.

Wilson D.J, Gabriel P, Leatherbarrow A.J, Cheesbrough J, Gee S, Bolton E, Fox A.J, Fearnhead P, Hart CA, and P. Diggle. Tracing the source of campylobacteriosis. *PLoS Genetics*, in press, 2008.

I.A. Gillespie, S.J. O'Brien, C. Penman, D. Tompkins, J. Cowden, and T.J. Humphrey. Demographic determinants for Campylobacter infection in

England and Wales: implications for future epidemiological studies. *Epidemiology and Infection*, 136(12):1717–25, 2008.

B.J. Gilpin, P. Scholes, B. Robson, and M.G. Savill. The transmission of thermotolerant Campylobacter spp. to people living or working on dairy farms in New Zealand. *Zoonoses Public Hlth.*, 55:352–360, 2008.

M. Hearnden, C. Skelly, R. Eyles, and P. Weinstein. The regionality of campylobacteriosis seasonality in New Zealand. *International Journal of Environmental Health Research*, 13(4):337–348, 2003.

L. Knorr-Held. Conditional prior proposals in dynamic models. *Scand. J. Statist.*, 26:129–144, 1999.

L. Knorr-Held and S. Richardson. A hierarchical model for space-time surveillance data on meningococcal disease incidence. *Appl. Statist.*, 52(2):169–183, 2003.

R. Lake, B. Adlam, and S. Perera. Acute Gastrointestinal Illness (AGI) Study: Final study report. Technical report, Institute of Environmental Science and Research Limited, 2009.

J.C. Marshall and N.P. French. Dynamic modelling of *Campylobacter* sources in the Manawatu. Technical Report RM11178, Massey University, August 2009.

S.M. McTavish, C.E. Pope, C. Nicol, K. Sexton, N.P. French, and P.E. Carter. Wide geographical distribution of internationally rare Campylobacter clones within New Zealand. *Epidemiology and Infection*, 136: 1244–1252, 2008.

P. Mullner, G. Jones, A. Noble, S. Spencer, S. Hathaway, and N.P. French. Source attribution of food borne zoonosis in new zealand: A modified hald model. *Risk Anal*, page in press, 2009.

K.M. Scott, J.C. Marwick, and P.R. Crampton. Utilization of general practitioner services in New Zealand and its relationship with income, ethnicity

and government subsidy. *Health Services Management Research*, 16:45–55, 2003.

S.K. Sheppard, J.F. Dallas, N.J. Strachan, M. Macrae, N.D. McCarthy, D.J. Wilson, F.J. Gormley, D. Falush, I.D. Ogden, M.C. Maiden, , and K.J. Forbes. Campylobacter genotyping to determine the source of human infection. *Clin. Infect. Dis.*, 48:1072–78, 2009.

E. Sneyd and M.G. Baker. Infectious diseases in new zealand: 2002 annual surveillance summary. Technical report, Institute of Environmental Science and Research Limited, 2003.

J. Williman, P. Cressey, and R. Pirie. Annual report concerning foodborne disease in New Zealand 2007. Technical report, Institute of Environmental Science and Research Limited, 2008.

N.A. Wilson. A systematic review of the aetiology of human campylobacteriosis in New Zealand. Technical report, Food Safety Authority of New Zealand, 2005.

A Implementation details

The model is implemented in R using a custom Monte Carlo Markov Chain (MCMC) procedure. The spatial components U_i are updated using a mixture of Metropolis-Hastings proposals and single site conditional prior proposals (Knorr-Held, 1999). The temporal components R_t, however, are highly correlated, and thus single site updating with Metropolis-Hastings proposals alone performs poorly. Instead, block updates with conditional prior proposals with blocks of length 4, 5, 9 and 11 are used, which improves mixing. The hyperparameters κ_U and κ_R are updated using Gibb's sampling. The results in Section 6 were obtained with multiple chains run from randomly generated starting values for 40,000 iterations, sampling every 20 iterations, after a burn-in period of 2000 iterations.

The surveillance tool interface in Section 7 was implemented using a combination of C++ for setting up the model, and Managed C++ for the Windows user interface. The tool takes user supplied data, creates the text files needed by the R based model, copies this to a writable folder and executes R with a predefined script that starts the model running. The model in R then writes its progress to a file which is monitored periodically by the Windows user interface, thus communicating the progress to the user. Once the model has completed, an analysis stage is run which generates output, including graphs of the temporal variation, text files describing the spatial relative risk for mapping, and descriptive text files identifying any potential sporadic outbreaks.

The source code for the model is available from the authors.