Modelling the spread of influenza in Western Australia*

Aldo F. Saavedra†
Centre for Data Science
The University of Sydney,
Sydney, NSW, Australia
Aldo.Saavedra@
sydney.edu.au

Sally Wood‡
Business School
The University of Sydney
Sydney, NSW, Australia
sally.wood@sydney.edu.au

Jemma L. Geoghegan
Marie Bashir Institute
for Infectious
Diseases and Biosecurity,
Charles Perkins Centre,
The University of Sydney,
Sydney, NSW, Australia
jemma.geoghegan@
sydney.edu.au

Edwards Holmes
Marie Bashir Institute
for Infectious
Diseases and Biosecurity,
Charles Perkins Centre,
The University of Sydney,
Sydney, NSW, Australia
edward.holmes@sydney.edu.au

Hugh Durrant-Whyte
Centre for Data Science
The University of Sydney,
Sydney, NSW, Australia
hugh.durrantwhyte
@sydney.edu.au

ABSTRACT

The daily number of confirmed cases of influenza throughout Western Australia, at a particular point in space and time is modelled nonparametrically by assuming that the observation is generated from a mixture of Poisson distributions where the log of the expected value of each component in the mixture has *a priori* a stationary Gaussian Process priors (GPP). The weights attached to the components in the mixture are parameterised to depend upon factors which are specific to a particular point in space and time, such as the population density, the time of year, the current weather conditions, the age distribution and human movement patterns as well as the attributes of the virus. This mixture model serves two purposes. First it seeks to identify the spread of influenza at a point in space-time as belonging to one of a finite but unknown number of possible influenza propagation *signatures*. Second it addresses the issue that the space-time covariance structure for influenza counts is likely to be nonstationary, by attaching different weights to the mixture components at different points in space and time. This allows the covariance to change across space-time. In this first study we focus on the data collected by a surveillance program run by Department of Health of Western Australia. Predictive distributions of influenza counts are obtained at a particular point in space and time, after accounting for the certainty surrounding the possible models as well as the uncertainty surrounding the parameters which prescribe these models. The multidimensional integration required to obtain these predictive distributions is performed using reversible jump Markov chain Monte Carlo (RJMCMC).

*This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10th of August 2015.

†Corresponding author

‡Corresponding author

1. INTRODUCTION

The spatial and temporal spread of an infectious diseases such as influenza are problems that readily lend itself to be modeled using non-parametric Bayesian methods[1].

Epidemics of influenza occur every year are a considerable cost to the community [2]. They tend to be seasonal, taking place during the winter in temperate zones [3] while in tropical regions the seasonality is less pronounced, and in some locations, non-existent. [4].

The relative isolation of the Australian continent, the diversity of its climate and the large distances between major urban centres provide a unique opportunity to study drivers behind the spread of influenza. Important insights have been gathered from simulation featuring air, surface transportation and human movement [5, 6, 7, 8, 3]. In developing a model for the spread of influenza wish it to have two properties (a) It is intrepretable and parsimonious and (b) Is flexible enough to give good estimates for a large range of functions and capture complex features. Choosing a parametric model for example, a linear model - satisfies the first requirement, but not the second, whereas a nonparametric model such as a GPP satisfies both requirements.

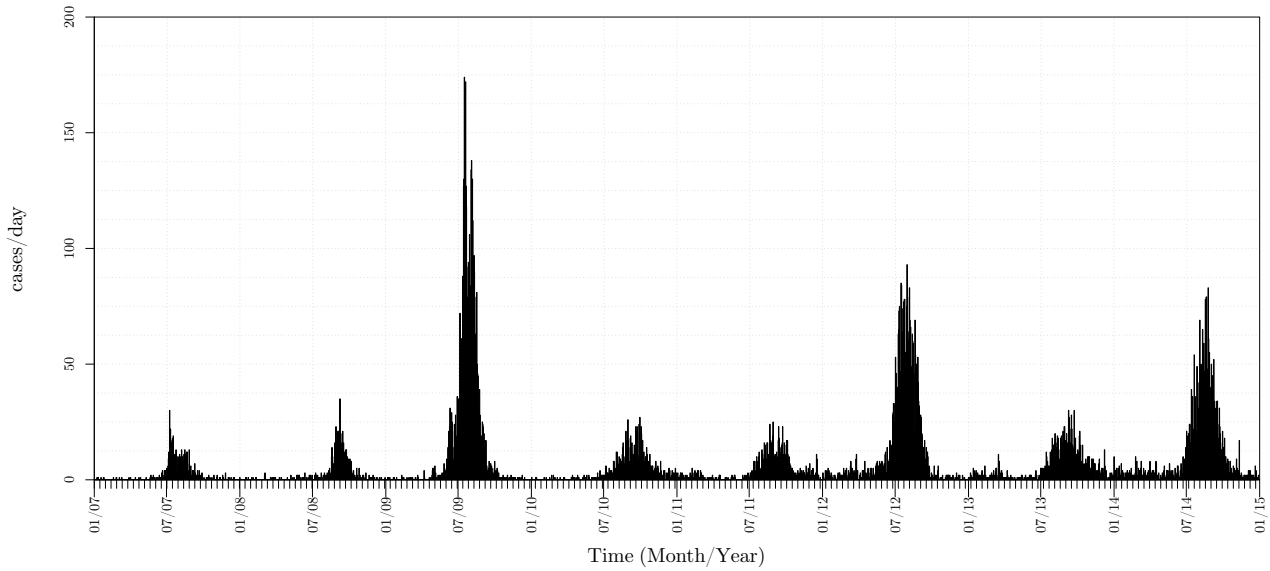

Figure 1: The number of confirmed cases determined by PathWest between 2007 and 2014.

2. DATA

The data was collected by a surveillance program run by Department of Health of Western Australia[9]. Medical centres distributed throughout Western Australia report on patients exhibiting symptoms of an 'Influenza like illness'. The date of the patient reporting the symptoms and their postcode of residence are recorded. A fraction of the reported cases undergo laboratory tests performed by PathWest[10] to confirm whether the illness is influenza. Only confirmed cases are included in this study and as such the method only samples a fraction of the total of the infected population at any one time.

Figure 1 shows the number of confirmed cases collected by the program between 2007 and 2014. The seasonal trend is the most prominent feature of the histogram with the peaks located around late August or early September each year. The severity of the influenza season cannot be determined from the height of peak alone. In figure 1 the highest peak corresponds to with the emergence of a novel swine-origin influenza A (H1N1) [11] which prompted more symptomatic people to seek a medical diagnosis.

Figure 2 shows the spatial distribution of influenza weekly counts a function of reported date for 2010. The left panel shows the distribution of the latitude of the postcode of residence of each case as a function of the reported date while the right panel shows the distribution corresponding to the longitude. The latitude and longitude associated with each postcode corresponds to an average of the positions of all the localities that share the postcode.

As expected, Figure 2, shows that the highest concentration of cases is in and around Perth (-31.9,115.9). The figure also shows that in additiona to spatial depency there is also a temporal dependence. For example, the number of reported cases in Albany (-35.0, 117.9), coincides with the peak in Perth although it is $\simeq$ 400 km south of the city while Karratha (-20.7,116.8), $\simeq$ 1500 km north of Perth, also reports a small number of cases in time with Perth.

An interesting feature found in Figures 1 and 2 is the number of cases reported between the seasonal peaks. The number of these 'out of season' cases is small but are persistently distributed across the state and not just limited to "humid-rainy" [12]. To understand the drivers behind such subtle features together with the main seasonal feature, the ultimate goal of this study is to build a nonparametric model that will include diverse data such as temperature, relative humidity, daylight hours, human movement and terrain. Such a model would provide insight into the drivers behind the temporal and spatial spread of influenza.

Our first step in our study is to outline the model and priors associated with this study.

3. MODEL AND PRIORS

3.1 Model for Mixture Components

Let $y_{tu_iv_i}$, be the observed disease count at time t, latitude u_i and longitude v_i, for $t = 1, \ldots, T$, $i = 1, \ldots, n$. We assume that this observation is generated from a mixture of a finite but unknown number of Poisson probability mass functions (PMF), with mean $\lambda_{tu_iv_i}$ which is a function of space and time. However for now, we limit our discussion to a fixed point in time, t, and develop a model for the spatial distribution of $\boldsymbol{y}_t = (y_{tu_1v_1}, \ldots, y_{tu_nv_n})$, and in what follows the subscript t will be dropped for ease of notation. The extension to a time varying distribution is the subject of future work.

For a fixed number of components in the mixture, r, each component has a different mean spatial surface $\boldsymbol{\lambda}_j = (\lambda_{u_1v_1j}, \ldots, \lambda_{u_nv_nj})$, for $j = 1, \ldots, r$ where subscript j denotes the component in the mixture. To place a prior on these mean surfaces we assume that, $\boldsymbol{\eta}_j = \log(\boldsymbol{\lambda}_j)$ is a

Gaussian process (GP), with covariance structure defined by the reproducing kernel for a thinplate smoothing spline [13, 14]. The overall log of the mean surface is modelled as a weighted average of these mixture components, where the weights attached to the components are parameterized to depend on space, as well as on other covariates, such as population density, current weather conditions, human movement patterns and virus type.

This parameterization has two purposes. First it allows the prior on η to be a nonstationary GP. This nonstationarity can arise because the covariance of η can change across space or it can arise in response to changes in other variables such as climate conditions. Second it allows for an overdispersed model, so that at a particular point in space the variance of the data need not equal the mean as must be the case if the data arose from a single Poisson component.

Suppose there are P possible variables which may give rise to nonstationarity, and that the values of these P covariates at a point in space are contained in the vector $\boldsymbol{z}_{uv} = (1, z_{1uv}, \ldots, z_{Puv})$. For a fixed number of r components, the PMF of y is

$$\Pr(y_{uv}{=}k|r, \tilde{\boldsymbol{\eta}}, \boldsymbol{z}_{uv}) = \sum_{j=1}^{r} \left\{ \begin{array}{c} \Pr(y_{uv}{=}k|\eta_{juv}, \gamma_{uv}{=}j, r) \\ \times \Pr(\gamma_{uv}{=}j|\boldsymbol{z}_{uv}) \end{array} \right\} \tag{1}$$

where

$$\Pr(y_{uv}{=}k|\eta_{juv}, \gamma_{uv}{=}j, r) = \frac{\exp(\eta_{juv}k)\exp(-\exp(\eta_{juv})}{k!},$$

$\tilde{\boldsymbol{\eta}}_r = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_r)$, and γ_{uv} is an indicator variable denoting the component in the mixture to which the observation belongs. Note that although all parameters should be indexed by r, to indicate that parameters will be of different dimension and take on different values for different values of r, we have omitted this subscript for clarity.

3.2 Model for Mixing Weights

The mixing weights are modelled using a multinomial logistic regression so that

$$\Pr(\gamma_{uv} = j) = \frac{\exp(z_{uv}\boldsymbol{\delta}_j)}{\sum_{l=1}^{r}\exp(z_{uv}\boldsymbol{\delta}_l)}$$

for $j = 1, \ldots, r$ and $\boldsymbol{\delta}_{jr} = (\delta_{0jr}, \delta_{1jr}, \ldots, \delta_{Pjr})$ are the regression coefficients with $\boldsymbol{\delta}_r$ set to $\mathbf{0}$ for identifiability.

3.3 The Likelihood

The likelihood function for the data for a fixed number of r components is

$$L_r = \prod_{i=1}^{n}\sum_{j=1}^{r}\left\{\frac{\exp(\eta_{ju_iv_i}y_{u_iv_i})\exp(-\exp(\eta_{ju_iv_i}))}{y_{u_iv_i}!}\Pr(\gamma_{u_iv_i}{=}j)\right\}$$

We note that observations are only recorded if there is at least one individual who has a confirmed case of influenza, so that the likelihood we work with is

$$\prod_{i=1}^{n}\sum_{j=1}^{r}\Pr(y_{u_iv_i}{=}k|y_{u_iv_i}{>}0, \eta_{ju_iv_i}, \gamma_{u_iv_i}{=}j, r)\Pr(\gamma_{u_iv_i} = j) \tag{2}$$

where

$$\Pr(y_{uv}{=}k|y_{uv}{>}0, \eta_{juv}, \gamma_{uv}{=}j) = \frac{\Pr(y_{uv}{=}k|\eta_{juv}, \gamma_{uv}{=}j)}{1-\exp(\exp(\eta_{juv}))}$$

and $\Pr(y_{uv}{=}k|\eta_{juv}, \gamma_{uv}{=}j)$ is given by (1).

3.4 Model for Number of Mixture Components

We consider the number of components in the mixture r to be a random variable. The PMF for y_{uv} unconditional on the number of components is,

$$\Pr(y_{uv}{=}k|\boldsymbol{\Theta}, \boldsymbol{z}_{uv}) = \sum_{r=1}^{R}\Pr(y_{uv}{=}k|r, \tilde{\boldsymbol{\eta}_{ruv}}, \Delta_r, \boldsymbol{z}_{uv})\Pr(r)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_R)$, is the set of all parameters which specify a mixture an unknown but finite number of components with $\boldsymbol{\theta}_r = (\tilde{\boldsymbol{\eta}_{ruv}}, \Delta_r)$ and $\Delta_r = (\boldsymbol{\delta}_{1r}, \ldots, \boldsymbol{\delta}_{rr})$ for $r = 1, \ldots, R$. The quantity $\Pr(r)$ is the prior probability that the mixture has r components.

The likelihood function now becomes

$$L = \sum_{r=1}^{R} l_r \Pr(r)$$

and L_r is given by (2).

3.5 Priors

For a given number of components, r, we write the log mean of the j^{th} component, η_j, as

$$\eta_j(u_i, v_i) = \alpha_{0j} + \alpha_{1j}u_i + \alpha_{2j}v_i + f_j(u_i, v_i)$$

and place a Gaussian Process prior on f_j, so that $f_j \sim N(0, \tau_j^2\Omega)$ where Ω is the covariance matrix corresponding to a thinplate spline prior, the elements of which are given in [13], pg 30. The parameter τ^2 controls the tradeoff between the goodness of fit and smoothness. If $\tau^2 = 0$, then η is linear in space, and as $\tau^2 \to \infty$, η interpolates the data. The prior for η is completed by specifying priors for $\boldsymbol{\alpha}_j = (\alpha_{0j}, \alpha_{1j}, \alpha_{2j})$ and τ^2. We assume that $\boldsymbol{\alpha}_j \sim N(0, c_\alpha I)$ and $\tau_j^2 \sim U[0, c_\tau]$ for some large c_α and c_τ, for $j = 1, \ldots, r$.

To place a prior on the parameters of the mixing functions, Δ_r, we followed [15] assume that $\boldsymbol{\delta}_j \sim N(0, c_\delta Z'Z^{-1})$.

We consider two priors for the number of components in the mixture. The first is $\Pr(r = k) = 1/R$ for $k = 1, \ldots, R$ and the second is a truncated Poisson with upper limit R and mean $(\mu_r) = 2$.

3.6 Predictive Distribution

The predictive distribution of the number of counts, y^* at location (u^*, v^*) given the data is

$$\Pr(y^*_{u,v} = k|\boldsymbol{y}) = \sum_{r=1}^{R}\int \Pr(y_{u,v} = k|\boldsymbol{y}, r, \boldsymbol{\theta}_r)p(\boldsymbol{\theta}_r|r, \boldsymbol{y})d\boldsymbol{\theta}_r \Pr(r|\boldsymbol{y})$$

and we use reversible jump Markov chain Monte Carlo (RJMCMC) to perform the required multidimensional integration.

The sampling scheme is broken into two steps; a between model move where the number of components potentially changes and a within model move where the parameters for a model of a fixed number of r components, $\boldsymbol{\theta}_r$ are updated

1. **Between Model Move**
This is the subject of future work, but the aim is to follow [16] as described below. Let the value of the current number of components in the chain be denoted by r^c, and value of the parameters which prescribe a model with r^c components be denoted by $\boldsymbol{\theta}^c_{r^c}$. We propose to move the chain from $(r^c, \boldsymbol{\theta}^c_{r^c})$ to $(r^p, \boldsymbol{\theta}^p_{r^p})$, by

by drawing $(r^p, \boldsymbol{\theta}^p_{r^p})$ from a proposal density $q(r^p, \boldsymbol{\theta}^p_{r^p}|r^c, \boldsymbol{\theta}^c_{r^c})$ and accepting this draw with probability

$$\alpha = \min\left\{1, \frac{p(r^p, \boldsymbol{\theta}^p_{r^p}|\boldsymbol{x}) \times q(r^c, \boldsymbol{\theta}^c_{r^c}|r^p, \boldsymbol{\theta}^p_{r^p})}{p(r^c, \boldsymbol{\theta}^c_{r^c}|\boldsymbol{x}) \times q(r^p, \boldsymbol{\theta}^p_{r^p}|r^c, \boldsymbol{\theta}^c_{r^c})}\right\},$$

where $p(\cdot)$ denotes a target density, which is the product of an approximate likelihood times prior densities. The proposal density $q(r^p, \boldsymbol{\theta}^p_{r^p}|r^c, \boldsymbol{\theta}^c_{r^c})$.

$$\begin{aligned} q(r^p, \boldsymbol{\theta}^p_{r^p}|r^c, \boldsymbol{\theta}^c_{r^c}) &= q(r^p|r^c) \times q(\boldsymbol{\theta}^p_{r^p}|r^p, r^c, \boldsymbol{\theta}^c_{r^c}) \\ &= q(r^p|r^c) \times q(\boldsymbol{\delta}^p_{r^p}, \boldsymbol{\tau}^{2p}_{r^p}, \boldsymbol{\eta}^p_{r^p}|r^p, r^c, \boldsymbol{\theta}^c_{r^c}) \\ &= q(r^p|r^c) \times q(\boldsymbol{\delta}^p_{r^p}|r^p, r^c, \boldsymbol{\theta}^c_{r^c}) \\ &\times\ q(\boldsymbol{\tau}^{2p}_{r^p}|\boldsymbol{\delta}^p_{r^p}, r^p, r^c, \boldsymbol{\theta}^c_{r^c}) \\ &\times\ q(\boldsymbol{\eta}^p_{r^p}|\boldsymbol{\tau}^{2p}_{r^p}, \boldsymbol{\delta}^p_{r^p}, r^p, r^c, \boldsymbol{\theta}^c_{r^c}). \end{aligned}$$

Thus, $(r^p, \boldsymbol{\theta}^p_{r^p})$ is drawn by first drawing r^p, followed by $\boldsymbol{\delta}^p_{r^p}$, $\boldsymbol{\tau}^{2p}_{r^p}$ and finally $\boldsymbol{\eta}^p_{r^p}$.

2. **Within Model Move**
For this type of move, r is fixed, and so the notation indicating the dependence on the number of components is dropped. Suppose the chain is at $\tilde{\boldsymbol{\eta}}^k = (\boldsymbol{\eta}^k_1, \ldots, \boldsymbol{\eta}^k_r)$, $\boldsymbol{\gamma}^k = (\gamma^k_{u_1,v_1}, \ldots, \gamma^k_{u_n,v_n})$, $\boldsymbol{\tau}^{2k} = (\tau^{2k}_1, \ldots, \tau^{2k}_r)$, and $\boldsymbol{\delta}^k$. We move the chain to $\tilde{\boldsymbol{\eta}}^{k+1}$, $\boldsymbol{\gamma}^{k+1}$, $\boldsymbol{\tau}^{2(k+1)}$, and $\boldsymbol{\Delta}^{k+1}$, via a kernel which consists of four parts;

   (a) **Drawing $\tilde{\boldsymbol{\eta}}^{k+1}$.**
   For $j = 1, \ldots, r$ a value of $\boldsymbol{\eta}^p_j$ is proposed from $q(\boldsymbol{\eta}_j|\boldsymbol{\gamma}^k, \tau^{2k}_j)$ where q is a Gaussian distribution with mean $\hat{\boldsymbol{\eta}}_j$ and covariance Σ_j where

   $$\hat{\boldsymbol{\eta}}_j = \arg\max_{\boldsymbol{\eta}_j} l(\boldsymbol{\eta}_j)$$

   with

   $$l(\boldsymbol{\eta}_j) = \sum_{i \in A_j} (y_i \eta_{ij} - \exp(\eta_{ij})) - 1/2 \boldsymbol{\eta}' D^{-1} \boldsymbol{\eta}$$

   and A_j is the set of integers, i, for which $\gamma_{u_i v_i} = j$, for $i = 1, \ldots, n$, and D the prior variance of $\boldsymbol{\eta}_j$. The covariance matrix, Σ_j, is equal to $\frac{\partial^2 l(\boldsymbol{\eta}_j)}{\partial \boldsymbol{\eta}_j{}^2}$ evaluated at $\hat{\boldsymbol{\eta}}_j$. Then with probability

   $$\alpha = \min\left\{1, \frac{p(\boldsymbol{\eta}^p_j|\boldsymbol{\gamma}^c, \boldsymbol{\tau}^{2c}) \times q(\boldsymbol{\eta}^c_j|\boldsymbol{\gamma}^c, \boldsymbol{\tau}^{2c})}{p(\boldsymbol{\eta}^c_j|\boldsymbol{\gamma}^c, \boldsymbol{\tau}^{2c}) \times q(\boldsymbol{\eta}^p_j|\boldsymbol{\gamma}^c, \boldsymbol{\tau}^{2c})}\right\},$$

   $\tilde{\boldsymbol{\eta}}^{k+1} = \tilde{\boldsymbol{\eta}}^p$ otherwise $\tilde{\boldsymbol{\eta}}^{k+1} = \tilde{\boldsymbol{\eta}}^k$

   (b) **Drawing $\boldsymbol{\tau}^{2(k+1)}$**
   Conditional on $\boldsymbol{\eta}^{k+1}_j$, $\tau^{(k+1)}_j$ is drawn from an inverse gamma distribution and accepted with probability 1.

   (c) **Drawing $\boldsymbol{\gamma}^{k+1}$**
   Conditional on $\tilde{\boldsymbol{\eta}}^{k+1}$, and $\boldsymbol{\Delta}^k$ we compute the probability that observation y_i is generated from component j, which is given by

   $$\begin{aligned} \Pr(\gamma_i{=}j|y_i, \boldsymbol{\Delta}, \tilde{\boldsymbol{\eta}}) &= \\ &= \frac{\Pr(y_i|\gamma_i{=}j, \boldsymbol{\eta}_j) \Pr(\gamma_i{=}j|\boldsymbol{\Delta})}{\sum_{k=1}^r \Pr(y_i|\gamma_i{=}k, \boldsymbol{\eta}_k) \Pr(\gamma_i{=}k|\boldsymbol{\Delta})} \\ &= \pi_{ij} \end{aligned}$$

   and γ^{k+1}_i is drawn from a multinomial distribution with probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ir})$ for $i = 1, \ldots, n$.

   (d) **Drawing $\boldsymbol{\Delta}^{k+1}$**
   Let $\boldsymbol{z}_i$ be an $r \times 1$ indicator vector with $z_{ji} = 1$ if $\gamma_i = j$ and $z_{ji} = 0$ otherwise, for $j = 1, \ldots, r$. A value of $\boldsymbol{\Delta}^p$ is proposed from $q(\boldsymbol{\Delta}|\boldsymbol{\gamma}^{k+1})$ where q is a Gaussian distribution with mean $\hat{\boldsymbol{\Delta}}$ and covariance Σ where

   $$\hat{\boldsymbol{\Delta}} = \arg\max_{\boldsymbol{\Delta}} l(\boldsymbol{\Delta})$$

   with

   $$l(\boldsymbol{\Delta}) = \sum_{i=1}^n \sum_{j=1}^r z_{ij} \log(\pi_{ij}) - \frac{1}{2c_\delta} \sum_{j=1}^r \boldsymbol{\delta}'_j \boldsymbol{\delta}_j.$$

   The covariance matrix Σ is $\frac{\partial^2 l(\boldsymbol{\Delta})}{\partial \boldsymbol{\Delta}^2}$ evaluated at $\hat{\boldsymbol{\Delta}}$. Then with probability

   $$\alpha = \min\left\{1, \frac{p(\boldsymbol{\Delta}^p|\boldsymbol{\gamma}) \times q(\boldsymbol{\Delta}^k|\boldsymbol{\gamma})}{p(\boldsymbol{\Delta}^k|\boldsymbol{\gamma}) \times q(\boldsymbol{\Delta}^p|\boldsymbol{\gamma})}\right\},$$

   $\boldsymbol{\Delta}^{k+1} = \boldsymbol{\Delta}^p$ otherwise $\boldsymbol{\Delta}^{k+1} = \boldsymbol{\Delta}^k$.

4. RESULTS AND DISCUSSION

We first illustrate the method and evaluate its performance on simulated data. To study the frequentist property of our technique we will generate 50 realizations from the model given by (1), with $n = 1600$, $r = 1, 2$, $\lambda_{i1} = 15 + 8\sin(6\pi u_i) \times \sin(6\pi v_i)$, $\lambda_{i2} = \mathbf{1}_n$,

$$\Pr(\gamma_i = 2) = \frac{\exp(17u_i + 17v_i - 24)}{1 + \exp(17u_i + 17v_i - 24)} = \pi_i$$

. For each realization we will compute the deviance measure

$$d^r = \sum_{i=1}^n (\log(\lambda_i ./ \hat{\lambda}^r_i) + (\lambda^r_i - \hat{\lambda}),$$

where λ_i, is the true mean function and equals $\lambda_{i1}(1-\pi) + \lambda_{i2}\pi$ and $\hat{\lambda}^r_i$ is the posterior mean function for a mixture of r components, for $r = 1, 2$. Initial results suggest that there is a considerable reduction in deviance in using a mixture of two GPP. The full results will be presented in a later version of this paper.

Figure 3 illustrates the method on a single simulated example. Panel (a) is a plot of the function with the data, panels (b) and c are plots of the estimated posterior mean surface for a one and two component respectively. Figure 3 shows that an estimate based on a mixture of two components clearly outperforms the single mixture estimate. The mixture of two component does a better job of capturing the peak in the mean function while remaining smooth.

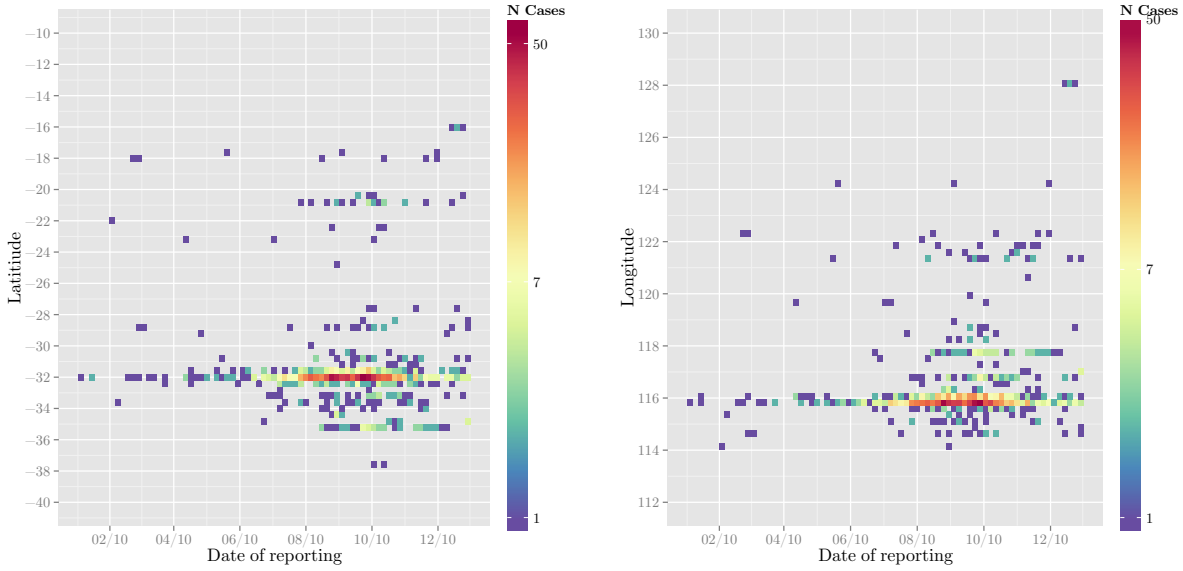

Figure 2: The distribution of flu cases as a function of latitude vs time (left) and longitude vs time (right).

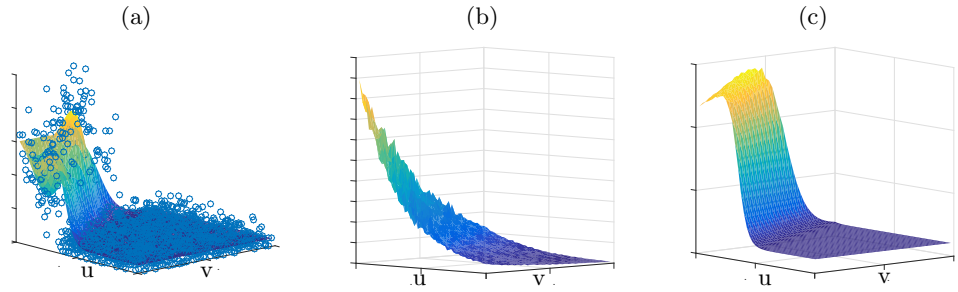

Figure 3: Illustration of the simulation example. Panel (a) shows the true 2d function as a surface and the data points obtained from the function. Panels (b) and (c) are plots of the estimated posterior mean surface obtained for a one and two component respectively.

To motivate the application of the method to disease mapping we provide preliminary analysis to demonstrate that influenza counts across WA exhibit very local structure. To do this we analyse influenza count data in four subsets of increasing size. The first subset are those counts corresponding to patients who reside in a postcode within 100km of the CBD in Perth; the second and third subsets correspond to those patients who live within 500km and 1000km from the CBD respectively. The fourth subsets is the data for the entire state of WA.

The model for the data is a single GPP with the reproducing kernel given by a thinplate smoothing spline prior. The data and the estimates of the posterior means for each of these four subsets are given in the figures

- *distance* $<$ 100 km. The data is shown in figure 4*a* and posterior mean in figure 4*c*.
- *distance* $<$ 500 km The data is shown in figure 4*b* and posterior mean in figure 4*d*.
- *distance* $<$ 1000 km The data is shown in figure 5*a* and posterior mean in figure 5*c*.
- No selection. The data is shown in figure 5*b* and posterior mean in figure 5*d*.

These figures show how sensitive modelling of influenza data require a method which can detect highly localized structure. In figure 4*c* one can clearly see the distribution of cases, north and south of the Perth and this is reflected in the contours obtained from the model in figure 4*c*. As the area associated with the data increases, these two peak merge into one dominant peak, with other main sources of cases included in the model.

For example in figure 5*c* small peaks correspond to Geraldton (-28.74,114.62) and Kalgoorlie (-30.75,121.47) and in figure 5*d*, which includes the data from the whole state, peaks corresponds with the area of Broome (-17.96,122.24) and Karratha (-20.74,116.85). The limitation of patients just reporting their postcode of residence becomes apparent in sparsely populated areas such as the north of Western Australia. For example, there are 8 localities in Broome that share the same postcode of 6725. This results in any spatial

structure that may have existed being artificially erased.

The results presented in this paper only include a small proportion of the possible data available for such study. It is expected that future results will include a greater amount of data that includes all the Australian states. The results show via simulation that the a mixture of GPP provides a better model for data which exhibits localized structure and that modelling disease counts needs a technique that is capable to do this. It is the subject of future work to apply the mixture model disease data across, where the mixing components are a function of space and other variables such as demographic data and climate data, to determine whether by employing a mixture of them will enable the model to be sensitive to both, small and large spatial structure. The last step in the study will be to develop the time component of the model.

References

[1] Larry P Ammann. Bayesian nonparametric inference for quantal response data. *The Annals of Statistics*, pages 636–645, 1984.

[2] Lone Simonsen. The global impact of influenza on morbidity and mortality. *Vaccine*, 17:S3–S10, 1999.

[3] Cécile Viboud, Ottar N Bjørnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *science*, 312(5772):447–451, 2006.

[4] Wladimir Jimenez Alonso, Julia Guillebaud, Cecile Viboud, Norosoa Harline Razanajatovo, Arnaud Orelle, Steven Zhixiang Zhou, Laurence Randrianasolo, and Jean-Michel Heraud. Influenza seasonality in madagascar: the mysterious african free-runner. *Influenza and Other Respiratory Viruses*, 9(3):101–109, 2015.

[5] I. M. Longini, P. E. Fine, and S. B. Thacker. Predicting the global spread of new infectious agents. *Am. J. Epidemiol.*, 123(3):383–391, Mar 1986.

[6] R. F. Grais, J. H. Ellis, and G. E. Glass. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *Eur. J. Epidemiol.*, 18(11):1065–1072, 2003.

[7] R. F. Grais, J. H. Ellis, A. Kress, and G. E. Glass. Modeling the spread of annual influenza epidemics in the U.S.: the potential role of air travel. *Health Care Manag Sci*, 7(2):127–134, May 2004.

[8] A. Flahault, S. Letrait, P. Blin, S. Hazout, J. MÃl'narÃl's, and A. J. Valleron. Modelling the 1985 influenza epidemic in france. *Statistics in Medicine*, 7(11):1147–1155, 1988.

[9] The Department of Health of Western Australia, 189 Royal Street, East Perth WA 6004, Australia.

[10] PathWest, J Block Hospital Avenue, QEII Medical Centre, Nedlands WA 6009, Australia.

[11] Emergence of a novel swine-origin influenza a (h1n1) virus in humans. *New England Journal of Medicine*, 360(25):2605–2615, 2009. PMID: 19423869.

[12] James D Tamerius, Jeffrey Shaman, Wladmir J Alonso, Kimberly Bloom-Feshbach, Christopher K Uejio, Andrew Comrie, and Cécile Viboud. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathog*, 9(3):e1003194, 2013.

[13] G. Wahba. *Spline models for observational data*. CBMS-NSF regional conference series in applied mathematics. Society for Industrial and Applied Mathematics, 1990.

[14] S A. Wood, Jiang Wenxin, and Martin Tanner. Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, 89(3):513 – 528, 2002.

[15] E J. Cripps, S A. Wood, R.E Wood, and J. Lau. Modelling the impact of personality on individual performance behavior with a time-varying mixture of monotonic random effects, 2014.

[16] S A. Wood, Ori. Rosen, and R.J. Kohn. Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics*, 20, 2011.

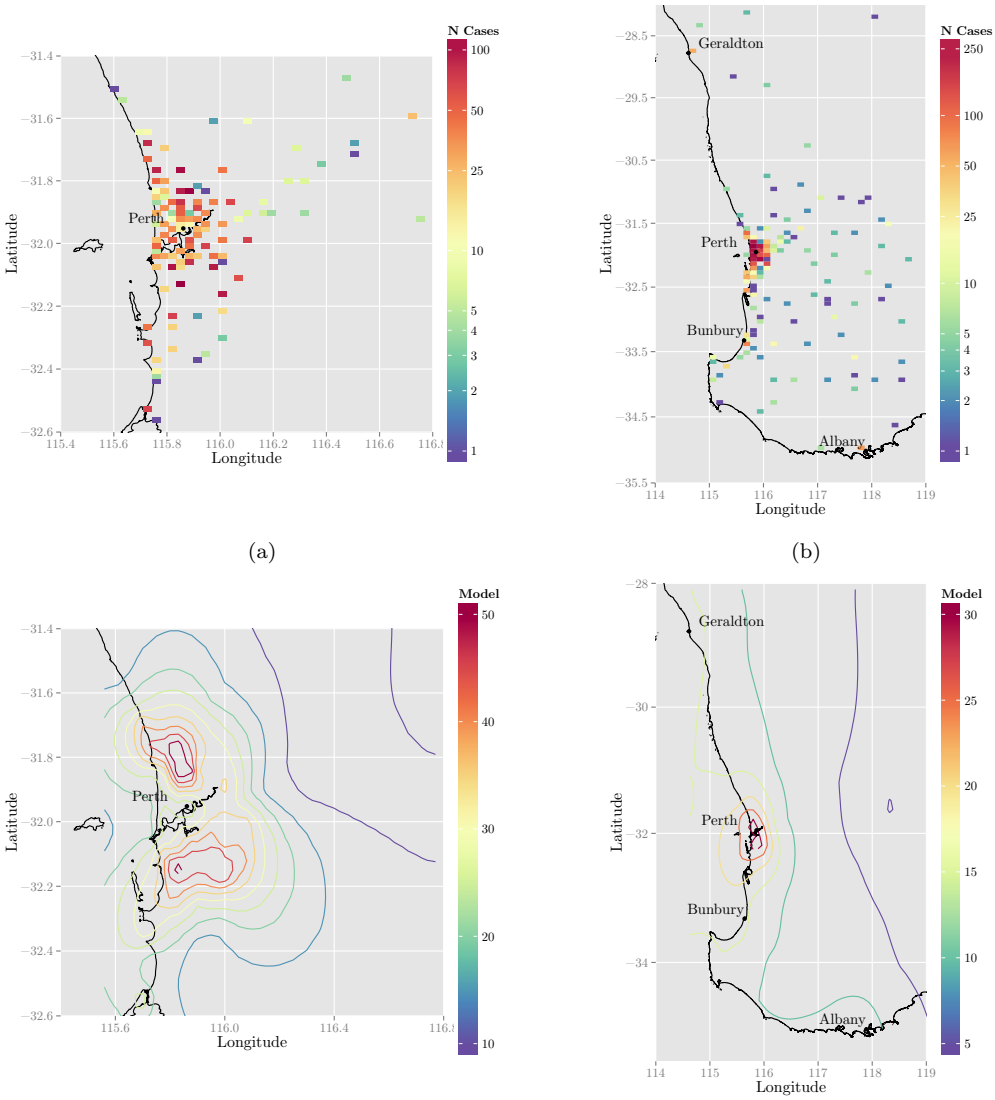

Figure 4: The spatial distribution of the number of cases between June and mid October for the year 2009 shown within a radius of 100 km (a) and 500 (b). The origin of the radius is the location of postcode 6000 (-31.92,115.91) which corresponds to the centre of Perth. Figures (c) and (d) show the contour of the model obtained using the data shown in figure (a) and (b) respectively.

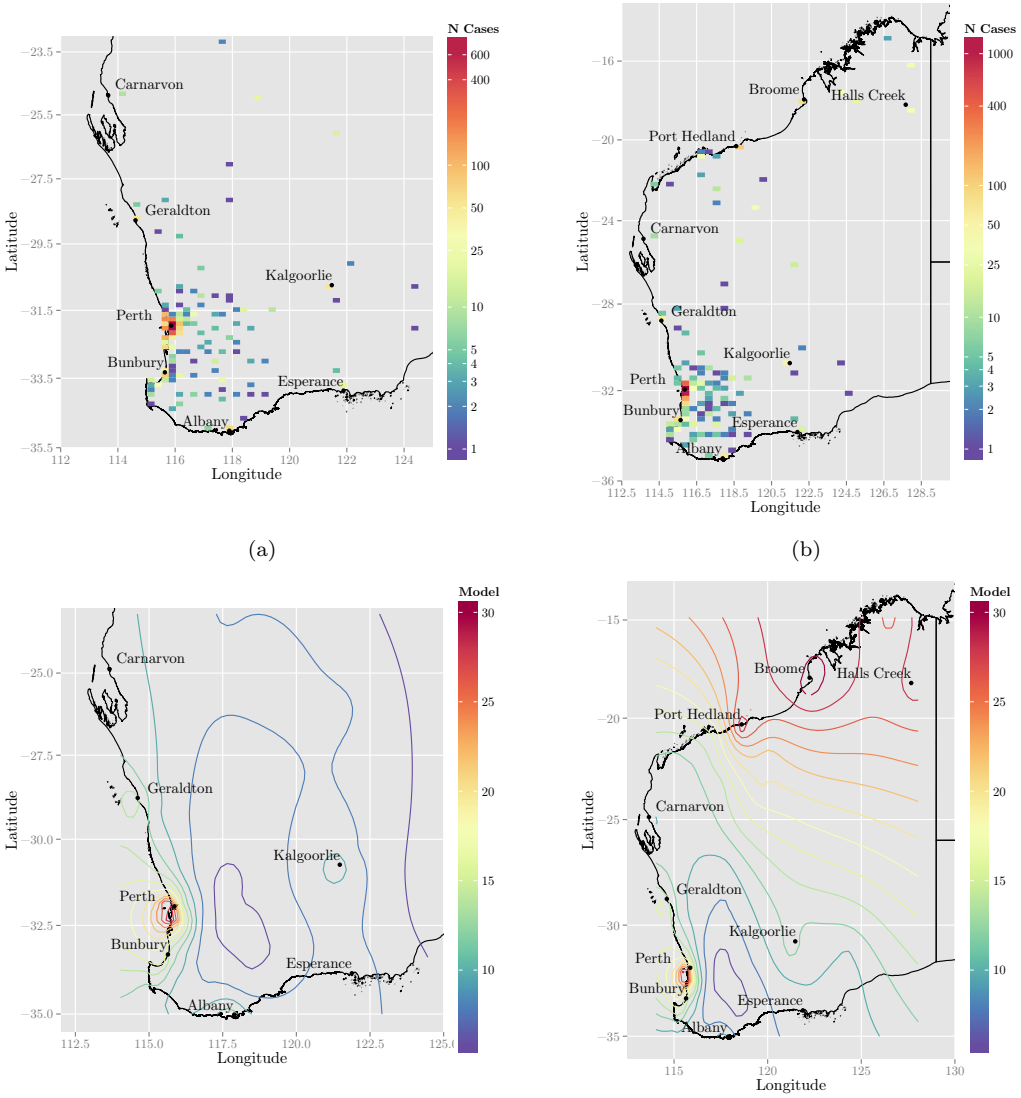

Figure 5: The spatial distribution of the number of cases between June and mid October for the year 2009 shown within a radius of 1000 km (a) and no restriction (b). The origin of the radius is the location of postcode 6000 (-31.92,115.91) which corresponds to the centre of Perth. Figures (c) and (d) show the contour of the model obtained using the data shown in figure (a) and (b) respectively.