BAYESIAN SURVEILLANCE FOR DETECTION OF SMALL AREA HEALTH ANOMALIES

Andrew B. Lawson
Ana Corberán-Vallet

Medical University of South Carolina
University of Valencia

BACKGROUND

- Bayesian modeling
- The surveillance task
- Bayesian modeling of spatio-temporal health data
  - Risk models
  - Model fitting: MCMC and INLA
  - Prospective fitting issues

BAYESIAN MODELING

Bayesian models consist of two components:

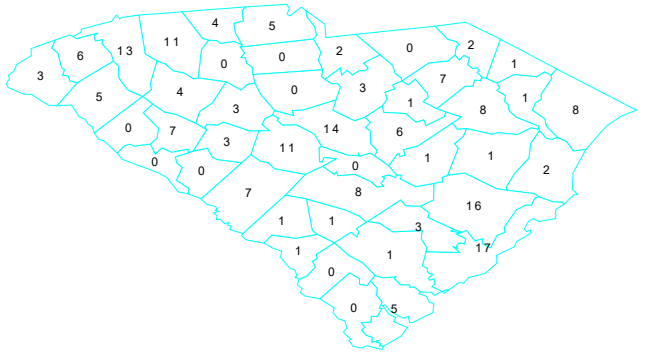
- *Likelihood* for the data
- *Prior distributions* for the parameters

These are combined to form a *posterior distribution* for the parameters

$$\underset{posterior}{\Pr(parameters \mid data)} \propto \underset{likelihood}{L(data \mid parameters)} . \underset{priors}{\Pr(parameters)}$$

THE SURVEILLANCE TASK

- Public health surveillance is the focus
- Health data usually consist of aggregated counts of disease within small areas (counties, districts, postal codes,…)
- Surveillance is essentially about change
- There are a number of things we focus on:
  - Development of clusters
  - Changes in trend
  - Geographical spread and jump diffusion
  - Detection of initiation of epidemics
- This has a huge impact on how we go about modeling



Example: SC congenital deaths 1990

BAYESIAN MODELING OF SPATIO-TEMPORAL HEALTH DATA

Count outcome in *m* small areas

$$\{y_i\}_{i=1,2,\ldots,m}$$

Poisson likelihood model

$$y_i \sim Po(\mu_i = e_i\,\theta_i)$$

e_i: expected count of disease representing the background population effect (fixed)

θ_i: unknown area-specific relative risk (focus of study)

BAYESIAN MODELING OF SPATIO-TEMPORAL HEALTH DATA

- Simple estimate of the relative risk: standardized incidence ratio (SIR), defined as the ratio of observed to expected counts

$$\hat{\theta}_i = y_i / e_i$$

This is a crude estimator and sometimes difficult to interpret and unstable

- We can assign a prior on θ_i or we can model its logarithm. The data likelihood forms a hierarchy with the parameter priors to give a hierarchical model (Bayesian hierarchical model)

RELATIVE RISK MODELS

$$\log(\mu_i) = \underset{offset}{\log(e_i)} + \log(\theta_i)$$

$$\log(\theta_i) = \ldots.\text{model terms}$$

A) Intercept (constant) model

B) Log–normal (random intercept) model

C) GLMM

D) Convolution model

RELATIVE RISK MODELS

D) Convolution model: Special case of GLMM that includes spatial correlation

$$\log(\theta_i) = \rho + \underbrace{u_i + v_i}_{\text{convolution}}$$

ρ: overall level of the relative risk

u_i: spatially structured effect

v_i: spatially unstructured extra variation

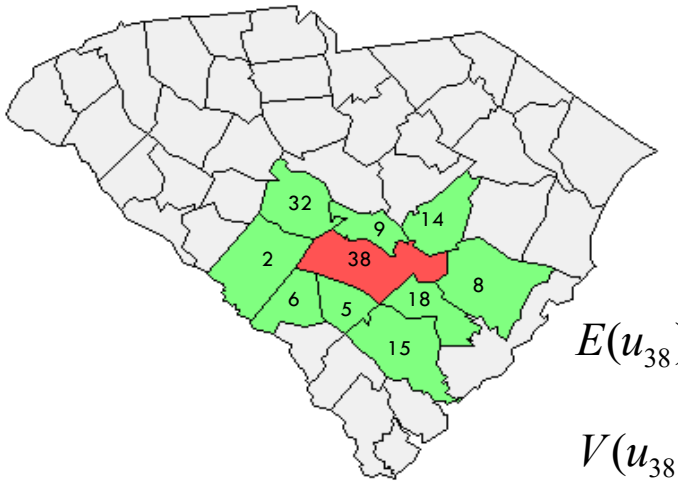
Adding covariates is straightforward:

$$\log(\theta_i) = \rho + \alpha_1 x_{1i} + u_i + v_i$$

RELATIVE RISK MODELS

The improper CAR model

$$u_i \mid u_{j \neq i} \sim N\left(\frac{1}{\mid n_i \mid} \sum_{j \in n_i} u_j, \frac{\sigma_u^2}{\mid n_i \mid}\right)$$



$$E(u_{38}) = \frac{u_5 + u_6 + u_2 + u_{32} + u_9 + u_{14} + u_8 + u_{18} + u_{15}}{9}$$

$$V(u_{38}) = \frac{\sigma_u^2}{9}$$

MODEL FITTING: MCMC AND INLA

- Conventionally Markov chain Monte Carlo is used to estimate posterior quantities for Bayesian models (such as the convolution or log-normal models)

- WinBUGS is designed to do this via two basic methods
  - Gibbs sampling
  - Metropolis –Hastings

- Approximation of posterior distributions has recently become available via Laplace approximation in the INLA package
  - Does not require iterative computation (unlike McMC)
  - Fast computation

PROSPECTIVE FITTING ISSUES

- Refitting at each new time point?
  - Could be computationally poor
  - Could use surveillance residuals
- Evolving model fitting
  - Endemic-epidemic approach
- Particle filtering
  - Resampling parameter values given new data

(Lawson and Kleinman, 2005, ch 4, ch 5)

OUR WORK

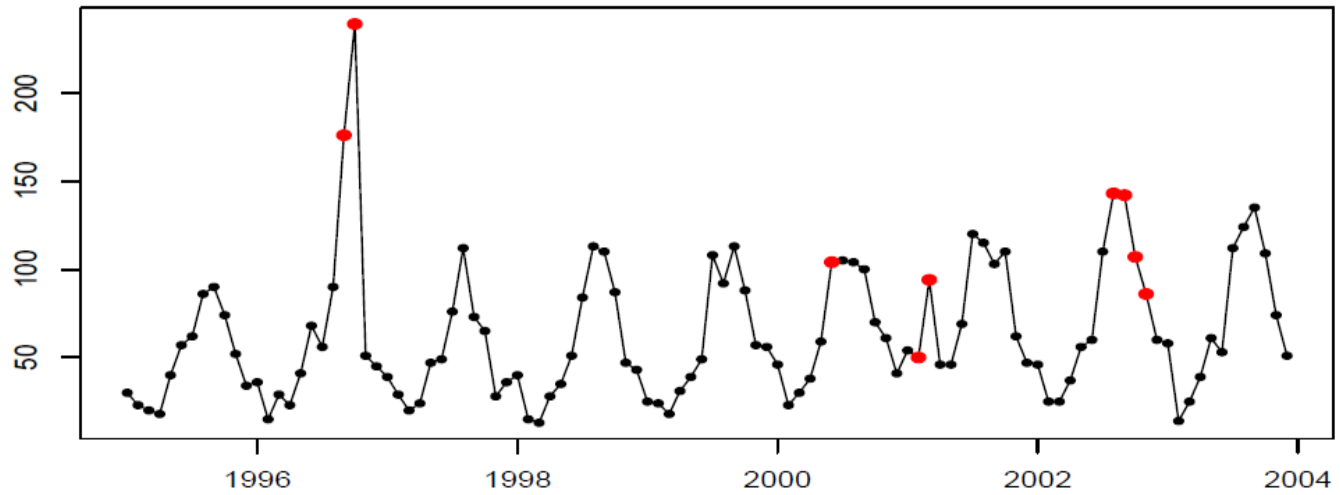
BAYESIAN DISEASE SURVEILLANCE

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}$$

Posterior probability | Likelihood | Prior probability

OBJECTIVE

- Detect disease outbreaks as soon as possible



Monthly counts of Salmonellosis cases in SC (1995-2003)

HOW?

- By using a model-based surveillance technique that incorporates both temporal and spatial information

- Idea: Use a statistical model to describe the overall behavior of disease in space and time under 'normal' conditions

  and

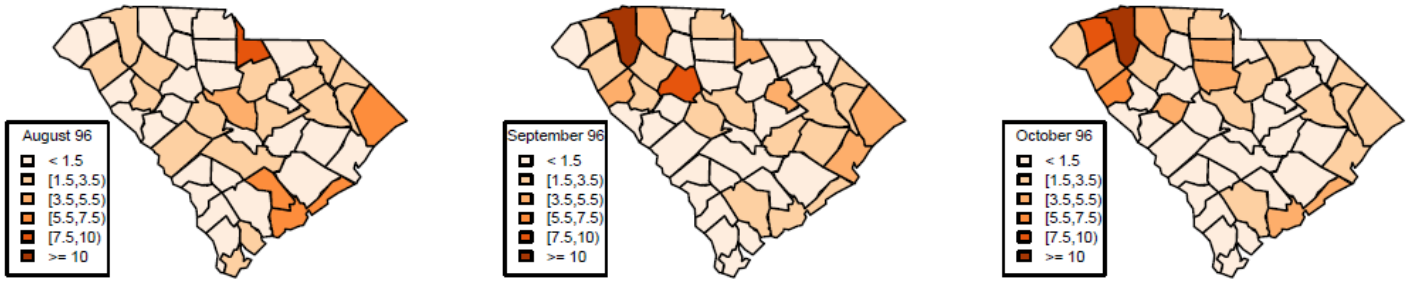
  detect unusual departures from predictable patterns based on the estimated model

ADVANTAGES

- Models: - allow covariate effects to be estimated

  - provide insight into etiology, spread, prediction and control of disease

- The use of spatial information increases the power to detect small localized outbreaks of disease



Spatial distribution of the SMR from August to October 1996

OUTLINE

- Univariate scenario
  - Model: The convolution model
  - Surveillance technique: SCPO
  - Case study: Salmonellosis

- Multivariate scenario
  - Model: The shared component model
  - Surveillance technique: MSCPO
  - Case study: ERD for respiratory diseases

- Can we go one step forward and anticipate disease outbreaks?
  - Syndromic information

UNIVARIATE SCENARIO: MODEL

- Monitor a map of *m* small areas over *T* time periods

$$\{y_{it}\} \quad i = 1,2,\ldots,m; t = 1,2,\ldots,T$$

- Bayesian hierarchical Poisson count model

$$y_{it} \sim Po(e_{it}\,\theta_{it})$$

e_{it}: expected counts of disease (background population effect)

θ_{it}: unknown area-specific relative risks

UNIVARIATE SCENARIO: MODEL

Convolution model (Besag et al., 1991, Lawson, 2013) vs Spatio-temporal model (knorr-Held, 2000)

$$\log(\theta_{it}) = \rho + u_i + v_i$$

$$\log(\theta_{it}) = \rho + u_i + v_i + \delta_{it}$$

ρ: overall level of the relative risk; $\rho \sim N(0, \sigma_\rho^2)$

u_i: spatially structured extra variation (improper CAR)

$$u_i \mid u_{j \neq i} \sim N\left(\frac{1}{|n_i|} \sum_{j \in n_i} u_j, \frac{\sigma_u^2}{|n_i|}\right)$$

v_i: spatially unstructured extra variation; $v_i \sim N(0, \sigma_v^2)$

δ_{it}: space-time interaction random effect; $\delta_{it} \sim N(0, \sigma_\delta^2)$

UNIVARIATE SCENARIO: MODEL

- We ran a simulation scenario (details in Corberán-Vallet and Lawson, 2011) to compare both models
- In terms of sensitivity, specificity and median time to detection, the convolution model outperformed the spatio-temporal model
- In a surveillance context:
  - The model must describe the behavior of disease under endemic conditions
  - It must be sensitive to temporal changes in the RR pattern of disease. A too complex model may absorb changes in risk in the model fit

UNIVARIATE SCENARIO: MODEL

- For seasonal data, in order to detect counts of disease higher than expected:

$$\log(\theta_{it}) = \rho + u_i + v_i + \sum_{s=1}^{12} \alpha_s I_s(t)$$

α_s: seasonal effects

$I_s(t)$: indicator function that takes the value 1 if time t corresponds to month s

- Different risks to account for seasonality, but the risks so defined are constant over time

UNIVARIATE SCENARIO: SURVEILLANCE TECHNIQUE

Surveillance Conditional Predictive Ordinate (Corberán-Vallet and Lawson, 2011)

$$SCPO_{it} = f(y_{it} \mid y_{1:t-1}) = \int f(y_{it} \mid \theta_i, y_{1:t-1})\, \pi(\theta_i \mid y_{1:t-1})\, d\theta_i$$

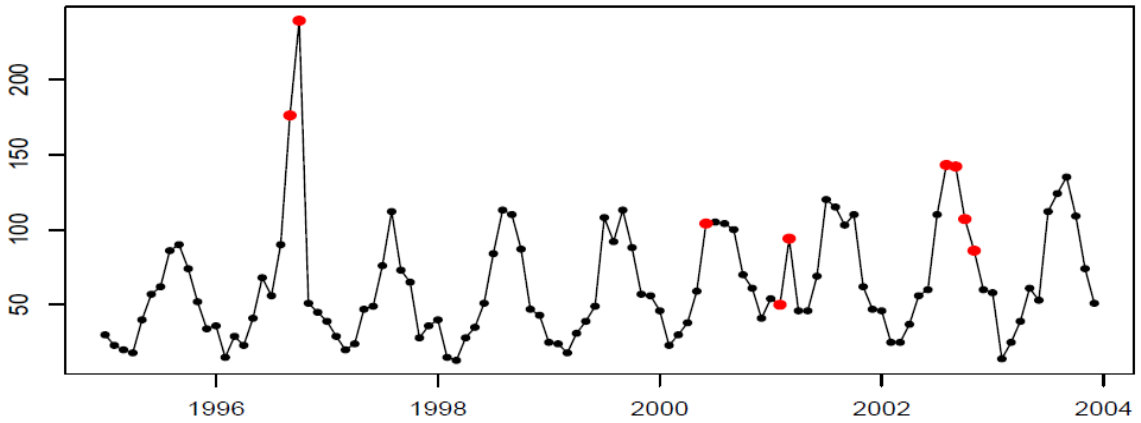
$$\approx \frac{1}{J} \sum_{j=1}^{J} Po(y_{it} \mid e_{it}\theta_i^{(j)})$$

$$\left\{\theta_i^{(j)}\right\}_{j=1}^{J} \sim \pi(\theta_i \mid y_{1:t-1})$$

If $SCPO_{it} < \alpha \Longrightarrow$ signal an alarm for area i

* See also: Surveillance Kullback Liebler divergence (SKL) (Rotejanaprasert and Lawson, 2016). Extension of the SCPO and behaves differently.

UNIVARIATE SCENARIO: CASE STUDY

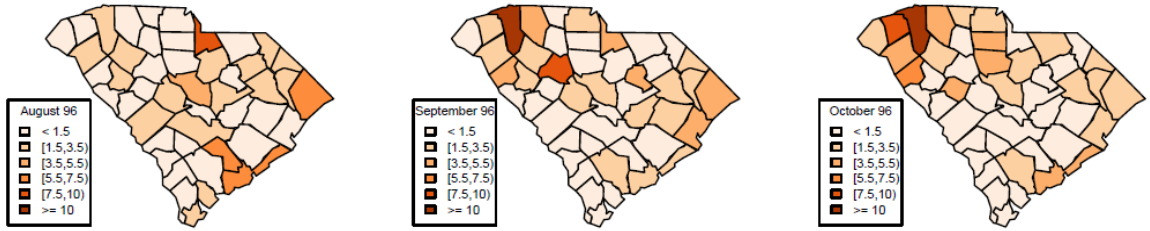


Monthly counts of Salmonellosis cases in SC (1995-2003)

Number of counties signaling an alarm at each time point during the surveillance period (1996-2003).
Data for year 1995 used to estimate the model.
Decision rule SCPO < 0.08

	J	F	M	A	M	J	J	A	S	O	N	D
1996	1	0	2	2	1	1	0	3	3	5	2	3
1997	0	3	2	2	4	2	5	2	0	1	0	1
1998	2	0	0	3	1	1	3	3	3	2	2	3
1999	0	1	2	2	1	1	3	2	0	0	2	0
2000	2	1	3	0	1	5	2	1	1	1	3	1
2001	2	5	3	2	1	1	1	2	1	1	0	1
2002	1	1	1	2	0	1	3	6	4	5	5	2
2003	1	0	0	0	2	2	1	2	4	3	0	2

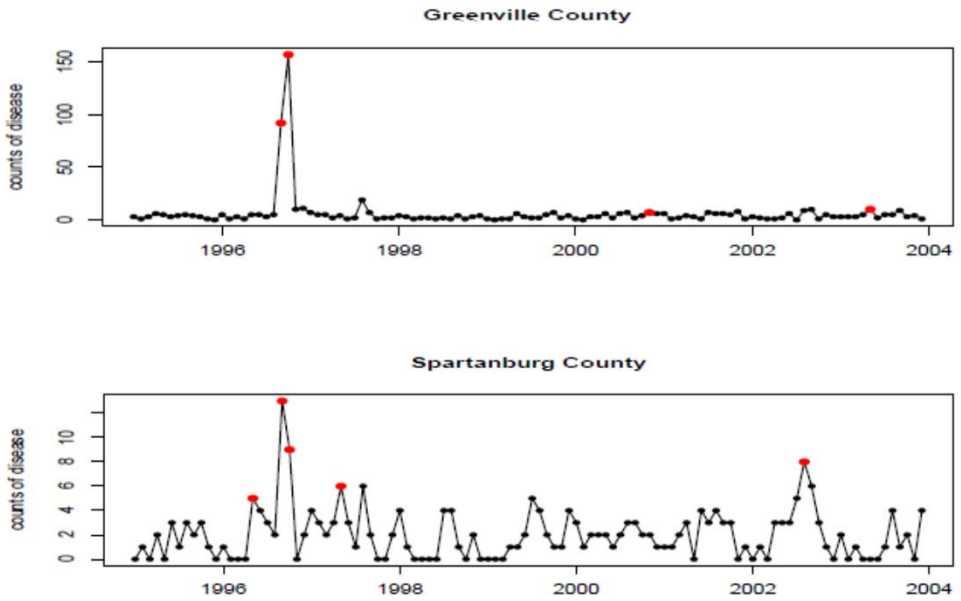
UNIVARIATE SCENARIO: CASE STUDY



Spatial distribution of the SMR from August to October 1996

Temporal plots for Greenville and Spartanburg counties

Red points represent alarms



MULTIVARIATE SCENARIO

- Surveillance systems are often focused on more than one disease within a predefined area
- A common approach is to monitor each disease separately: any correlation between diseases is ignored
- We present a multivariate extension of the proposed surveillance technique that
  - allows for correlation between diseases
  - can detect outbreaks happening in either one or a combination of diseases

MULTIVARIATE SCENARIO: MODEL

- A possibility to jointly model the endemic behavior of the multiple diseases is the shared component model (knorr-Held and Best, 2001)
- For the joint analysis of $k \geq 2$ diseases, Held el al. (2005) proposed a generalized SCM (only spatial information)

$$y_{ik} \sim Po(e_{ik}\,\theta_{ik})$$

$$\log(\theta_{ik}) = \rho_k + \sum_j \delta_{j,k}\, w_{j,i}$$

$$\sum_{l=1}^{n_{wj}} \log(\delta_{j,l}) = 0$$

$w_j = (w_{j,1}, w_{j,2}, \ldots, w_{j,m})$ is a spatial field (CAR component)

$\delta_{i,k}$: relative contribution of w_i to disease k

n_{wi} : number of relevant diseases for w_i

MULTIVARIATE SCENARIO: MODEL

Our shared component model formulation:

$$y_{itk} \sim Po(e_{itk}\,\theta_{ik})$$

$$\log(\theta_{ik}) = \rho_k + \sum_{l=1}^{L} \phi_{l,k}\,\delta_{l,k}\,w_{l,i} + \psi_{ik}$$

ρ_k : overall risk for disease k

L : number of spatial fields (CAR components) $w_l = (w_{l,1}, w_{l,2}, \ldots, w_{l,m})$

$\phi_{l,k} = 1$ if w_l has an influence on disease k, and $\phi_{l,k} = 0$ otherwise

$\delta_{l,k}$: weight

ψ_{ik} : spatial unstructured extra variation for disease k

Advantage: By using indicator variables, we do not have to specify the structure of the model in advance

MULTIVARIATE SCENARIO: SURVEILLANCE TECHNIQUE

For each small area *i* and time period *t*

$$\begin{pmatrix} y_{it1} \\ y_{it2} \\ \vdots \\ y_{itK} \end{pmatrix} \qquad \begin{pmatrix} e_{it1}\,\hat{\theta}_{i1} \\ e_{it2}\,\hat{\theta}_{i2} \\ \vdots \\ e_{itK}\,\hat{\theta}_{iK} \end{pmatrix}$$

$\hat{\theta}_{ik}$: posterior relative risk estimated at the previous time period (data up to time t-1)

counts higher than expected

$(y_{itk_1} \quad y_{itk_2} \quad \cdots \quad y_{itk_r})$

counts smaller than expected

MULTIVARIATE SCENARIO: SURVEILLANCE TECHNIQUE

A multivariate extension of the surveillance conditional predictive ordinate can be defined as (Corberán-Vallet, 2012)

$$MSCPO_{it} = f(y_{itk_1}, y_{itk_2}, \ldots, y_{itk_r} \mid y_{1:t-1})$$

$$= \int\int \ldots \int f(y_{itk_1}, y_{itk_2}, \ldots, y_{itk_r} \mid \theta_{ik_1}\ \theta_{ik_2} \ldots \theta_{ik_r})\ x$$

$$\pi(\theta_{ik_1}, \theta_{ik_2}, \ldots, \theta_{ik_r} \mid y_{1:t-1})\ d\theta_{ik_1}\ d\theta_{ik_2} \ldots d\theta_{ik_r}$$

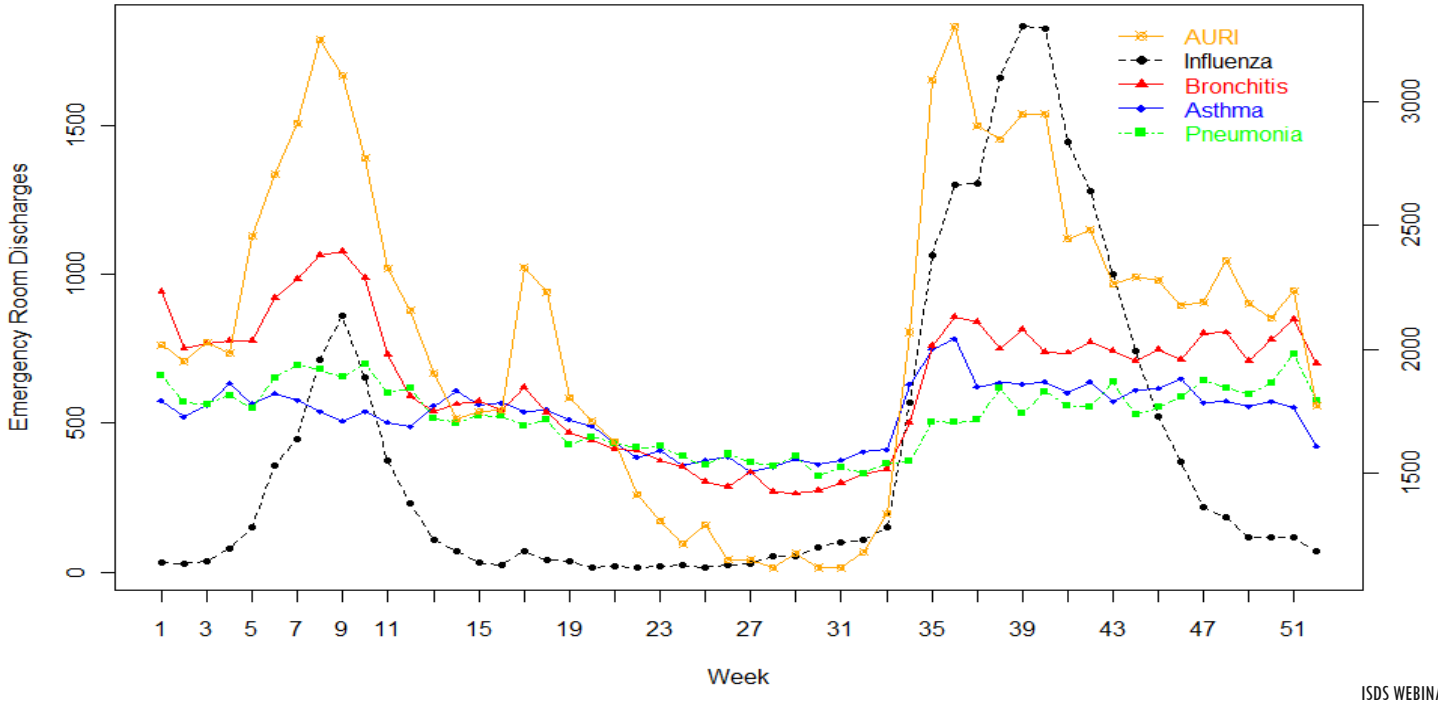
$$\approx \frac{1}{J}\sum_{j=1}^{J} Po(y_{itk_1} \mid e_{itk_1}\theta_{ik_1}^{(j)})\ Po(y_{itk_2} \mid e_{itk_2}\theta_{ik_2}^{(j)}) \ldots Po(y_{itk_r} \mid e_{itk_r}\theta_{ik_r}^{(j)})$$

$\{\theta_{ik}^{(j)}\}_{j=1}^{J}$; set of RR sampled from the posterior distribution at time t-1

If $MSCPO_{it} < \alpha$ $\Longrightarrow$ signal an alarm for area *i*

MULTIVARIATE SCENARIO: CASE STUDY

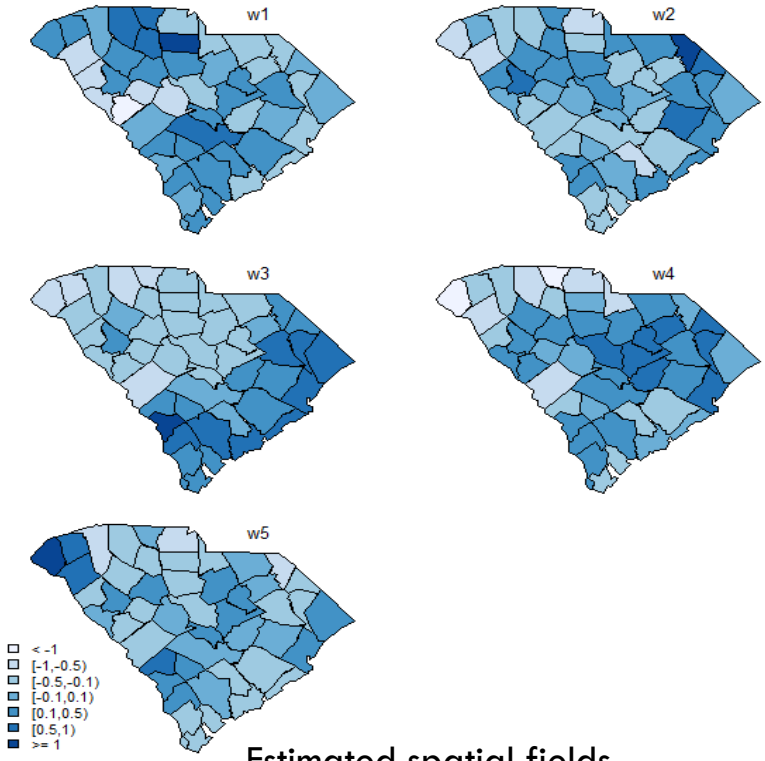
Weekly emergency room discharges for respiratory diseases in South Carolina in 2009



MULTIVARIATE SCENARIO: CASE STUDY

- We confine our analysis to data collected from week beginning June 28 to week beginning December 27 (weeks 26 – 52 in previous figure)
- 46 counties, 27 time periods, and 5 diseases
- Expected counts (constant) are calculated using the data from the first 3 weeks
- These data are also used to estimate the proposed SCM (we assume L = 10)
- The estimated model contains 5 spatial fields

MULTIVARIATE SCENARIO: CASE STUDY



Estimated spatial fields

Structure of the estimated model

$$\log(\theta_{i1}) = \rho_1 + \delta_{1,1}\, w_{1,i} + \delta_{2,1} w_{2,i} + \psi_{i1}$$

$$\log(\theta_{i2}) = \rho_2 + \delta_{3,2}\, w_{3,i} + \psi_{i2}$$

$$\log(\theta_{i3}) = \rho_3 + \delta_{1,3}\, w_{1,i} + \psi_{i3}$$

$$\log(\theta_{i4}) = \rho_4 + \delta_{1,4}\, w_{1,i} + \delta_{4,4} w_{4,i} + \psi_{i4}$$

$$\log(\theta_{i5}) = \rho_5 + \delta_{1,5}\, w_{1,i} + \delta_{5,5} w_{5,i} + \psi_{i5}$$

MULTIVARIATE SCENARIO: CASE STUDY

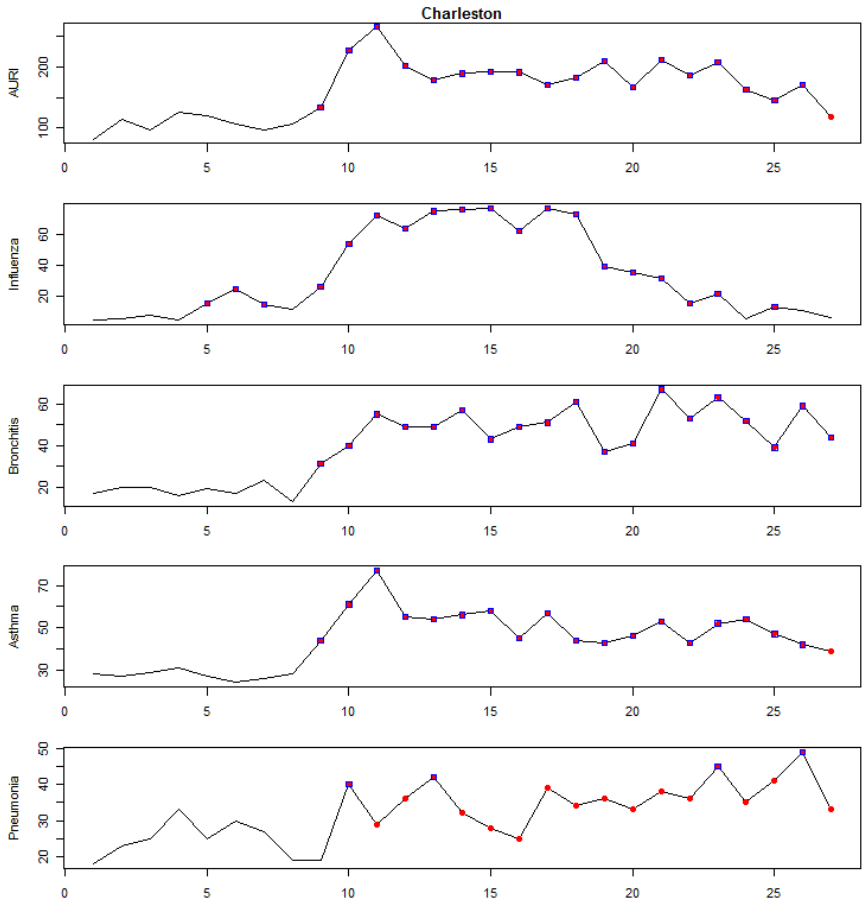
Goodness of fit: DIC (pD) for the proposed shared component model and five independent convolution models

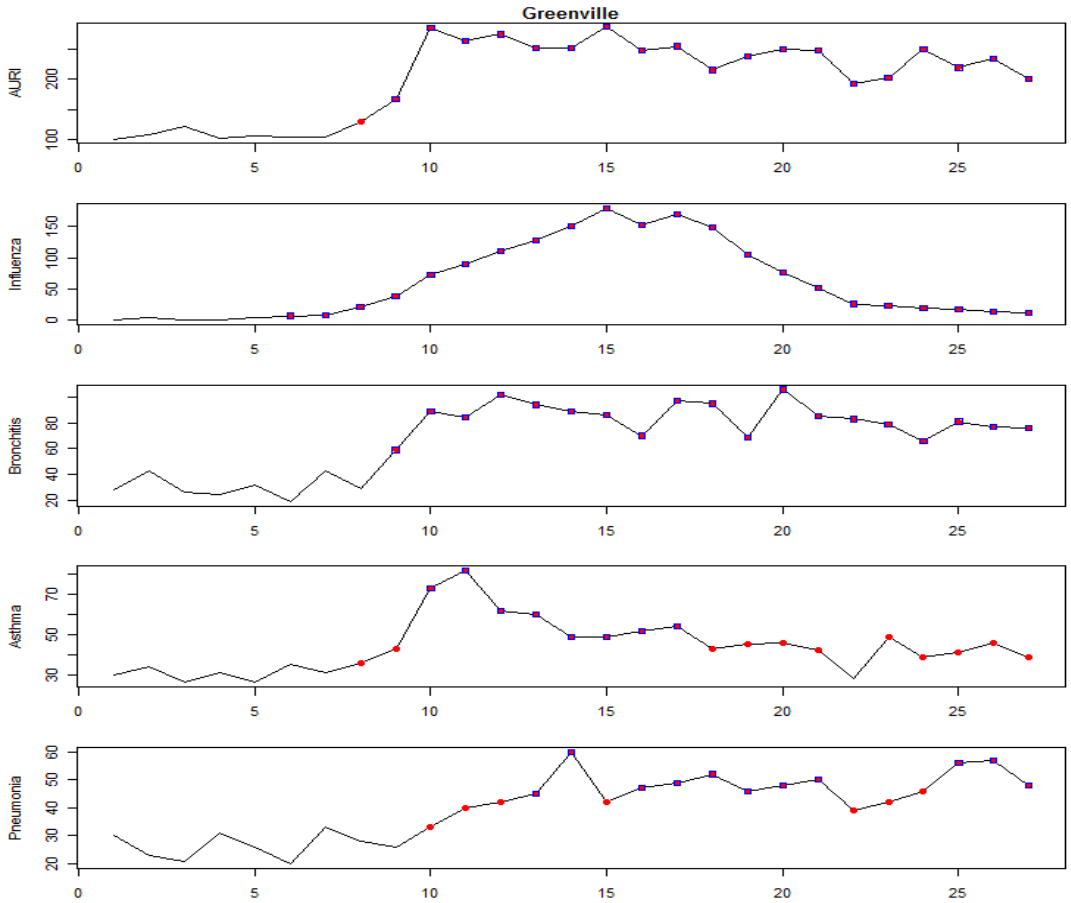
Model	Disease 1	Disease 2	Disease 3	Disease 4	Disease 5	Total
Proposed SCM	**808.18**	**268.04**	**578.84**	**657.64**	**652.43**	**2965.13**
	(39.85)	**(20.06)**	**(32.07)**	**(33.22)**	**(32.27)**	**(157.46)**
Convolution models	810.30	268.24	584.29	659.17	656.17	2978.16
	(40.80)	(19.92)	(33.88)	(33.86)	(32.62)	(161.08)

MULTIVARIATE SCENARIO: CASE STUDY

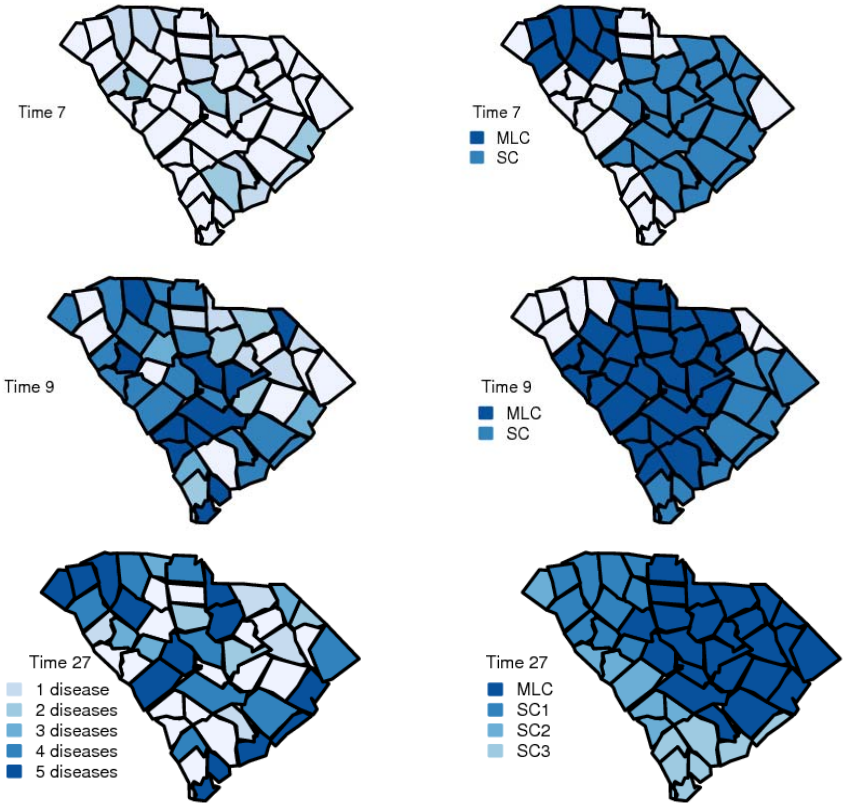
- For $t = 4, 5, \ldots, 27$, the SCM is estimated using the data observed up to $t-1$
- MSCPO values associated with the new data are analyzed to detect epidemic onsets
- An alarm for the i th county is sounded at time t if the $MSCPO_{it} < 0.05$
- Counts of disease detected as unusual are assumed to be missing when they become part of the history

MULTIVARIATE SCENARIO: CASE STUDY





MULTIVARIATE SCENARIO: CASE STUDY



A comparison with the multivariate scan statistic: Counties where an outbreak is declared

Left: Areas signaling an alarm based on the MSCPO

Right: Most likely cluster (MLC) and secondary clusters (SC) using the Poisson-based prospective space-time scan statistic

MULTIVARIATE SCENARIO: A COMPARISON

- The space-time scan statistic pinpoints the general time and location of the most likely cluster (and possible secondary clusters)
- Drawbacks:
  - Counties with no increased incidence can be included in the cluster
  - Some counties do not undergo an outbreak of disease for all the diseases reported in the cluster
  - Several large clusters covering practically all the study region are reported
- The MSCPO detects, at each time, counties with increased disease incidence and the diseases causing the alarm within each county
- It enables a timelier and more informed response

CAN WE ANTICIPATE DISEASE OUTBREAKS?

- We have developed a model-based surveillance technique to detect disease outbreaks as soon as possible
- But… can we predict disease outbreaks before they occur?
- The answer is based on the use of syndromic information
- However, we do not want to monitor syndromes or health-related data that precede diagnosis (these data can lead to false alarms)
- We want to develop a multivariate model that models both the disease of interest and syndromic information and helps to predict possible outbreaks

CAN WE ANTICIPATE DISEASE OUTBREAKS? A FIRST ATTEMPT

- The disease of interest is an infectious disease and we have information from a syndromic disease

$$y_{it} \sim Po(\mu_{it} + I_{it})$$

We want a model like this for the infection of interest

endemic component: describes the pattern of disease during non-epidemic periods

epidemic component: expected additive increase in disease counts due to an epidemic (depends on syndromic information)

CAN WE ANTICIPATE DISEASE OUTBREAKS? A FIRST ATTEMPT

y_{it}: number of cases of the disease of interest

$$y_{it} \sim Po(\mu_{it} + I_{it})$$

$$\mu_{it} = e_{it}\theta_{it}$$

$$\log(\theta_{it}) = \rho + u_i + v_i$$

y_{it}^{s}: number of cases of the syndromic disease

$$y_{it}^{s} \sim Po(\mu_{it}^{s} + I_{it}^{s})$$

$$\mu_{it}^{s} = e_{it}^{s}\theta_{it}^{s}$$

$$\log(\theta_{it}^{s}) = \rho^{s} + \psi u_i + v_i^{s}$$

during non-epidemic conditions, the two diseases may be influenced by common confounding factors (Wang and Wall, 2003) . Here $\psi \sim N(0, \sigma_{\psi}^{2})$

CAN WE ANTICIPATE DISEASE OUTBREAKS? A FIRST ATTEMPT

y_{it}: number of cases of the disease of interest

y_{it}^{s}: number of cases of the syndromic disease

$$y_{it} \sim Po(\mu_{it} + I_{it})$$

$$I_{it} = \beta_{it}\left(y_{i,t-1} + \gamma_i \sum_{j \in n_i} y_{j,t-1} \right) + \phi_i I_{i,t-1}^{s}$$

$$y_{it}^{s} \sim Po(\mu_{it}^{s} + I_{it}^{s})$$

$$I_{it}^{s} = \beta_{it}^{s}\left(y_{i,t-1}^{s} + \gamma_i^{s} \sum_{j \in n_i} y_{j,t-1}^{s} \right)$$

Component based on data up to time t-1. At time t we can make predictions for time t+1

CONCLUDING REMARKS

- We have presented a Bayesian model-based surveillance technique for on-line spatio-temporal public health surveillance
- As a local measure, different alarms are sounded for those areas of increased disease incidence
- It can be applied in any surveillance context where a model is used to describe the endemic behavior of diseases
- Simple spatial models are the key to allowing detection of change over time

CONCLUDING REMARKS

- The technique can be easily extended for the monitoring of multiple diseases
- The proposed SCM allows us to identify the number of latent spatial fields required to describe the correlation across both areas and diseases
- The multivariate surveillance technique improves outbreak detection when changes in disease incidence happen simultaneously
- Finally, we have presented a model that incorporates syndromic information to predict the start of epidemics
- Some preliminary results obtained in a prospective analysis of infectious disease data showed its good performance

REFERENCES

Besag J, York J and Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; 43: 1-59

Rotejanaprasert, C. and Lawson, A. B. (2016) Bayesian prospective detection of small area health anomalies using Kullback–Leibler divergence. *Statistical Methods in Medical Research* (to appear)

Corberán-Vallet A. Prospective surveillance of multivariate spatial disease data. *Statistical Methods in Medical Research* 2012; 21: 457-477

Corberán-Vallet A and Lawson AB. Conditional predictive inference for online surveillance of spatial disease incidence. *Statistics in Medicine* 2011; 30: 3095-3116

Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* 2000; 19: 2555-2567

Knorr-Held L and Best NG. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, series A* 2001; 164: 73-85

Held L, Natário I, Fenton SE, Rue H and Becker N. Towards joint disease mapping. *Statistical Methods in Medical Research* 2005; 14: 61-82

Lawson, A. B. (2013) Bayesian Disease Mapping: hierarchical modeling in Spatial Epidemiology 2nd Ed CRC Press, New York.

Lawson, A. B. And Kleinman, K. (eds) (2005) Spatial & Syndromic Surveillance for Pubic Health. Wiley, New York

Wang F and Wall MM. Generalized common spatial factor model. Biostatistics 2003; 4: 569-582