Bayesian Spatial Analysis of Infectious diseases: models and metrics

Andrew Lawson
MUSC
USA

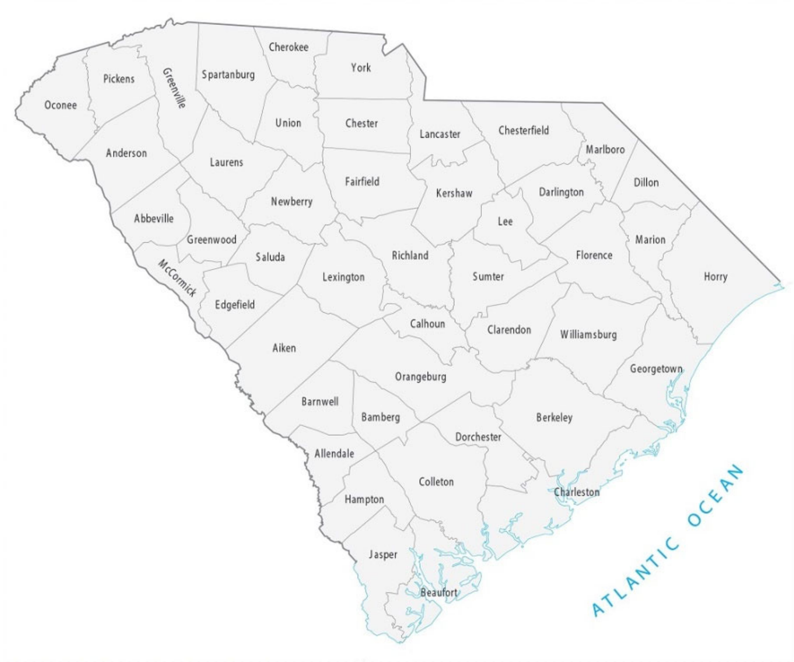
MUSC
MEDICAL UNIVERSITY
of SOUTH CAROLINA

Acknowledgement

- Including collaborative work with Joanne Kim MUSC

Motivation: location

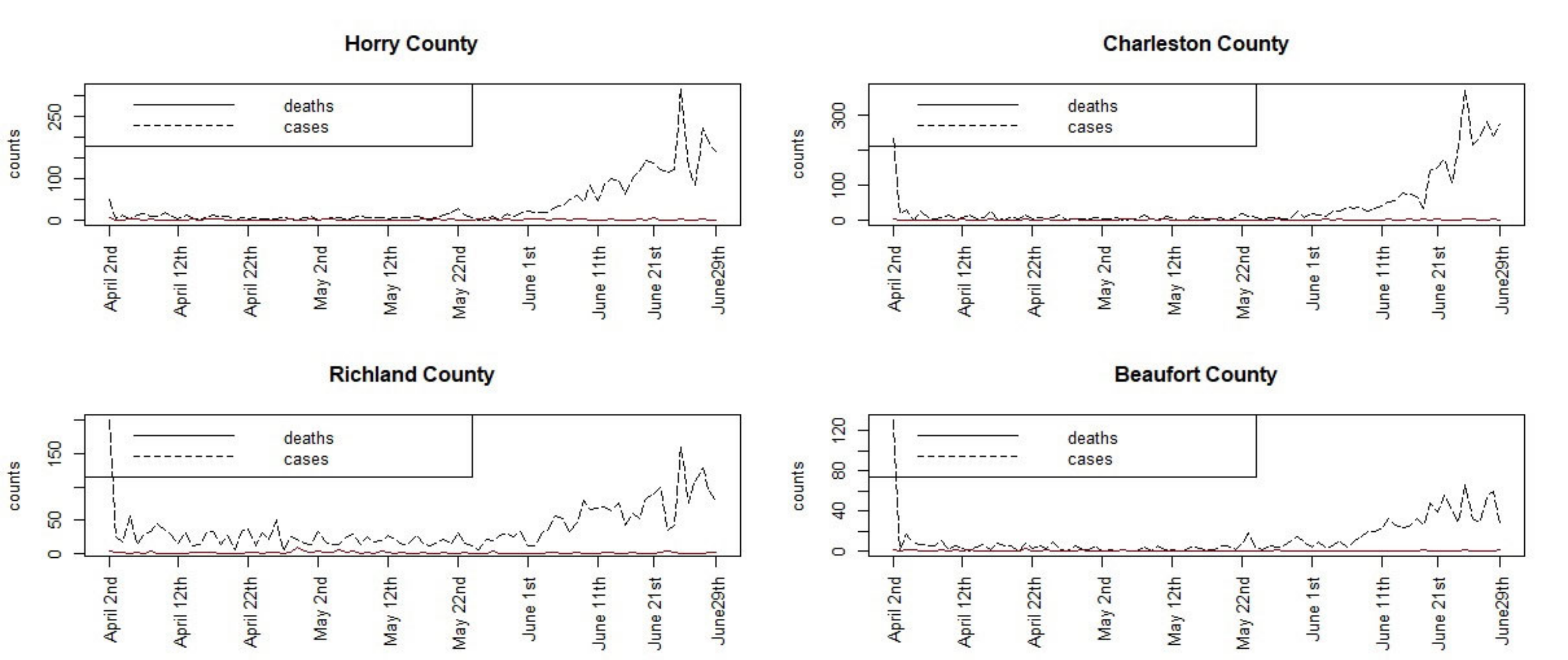
- SC state in USA
- 46 counties



SOUTH CAROLINA
© WWW.FREEWORLDMAPS.NET

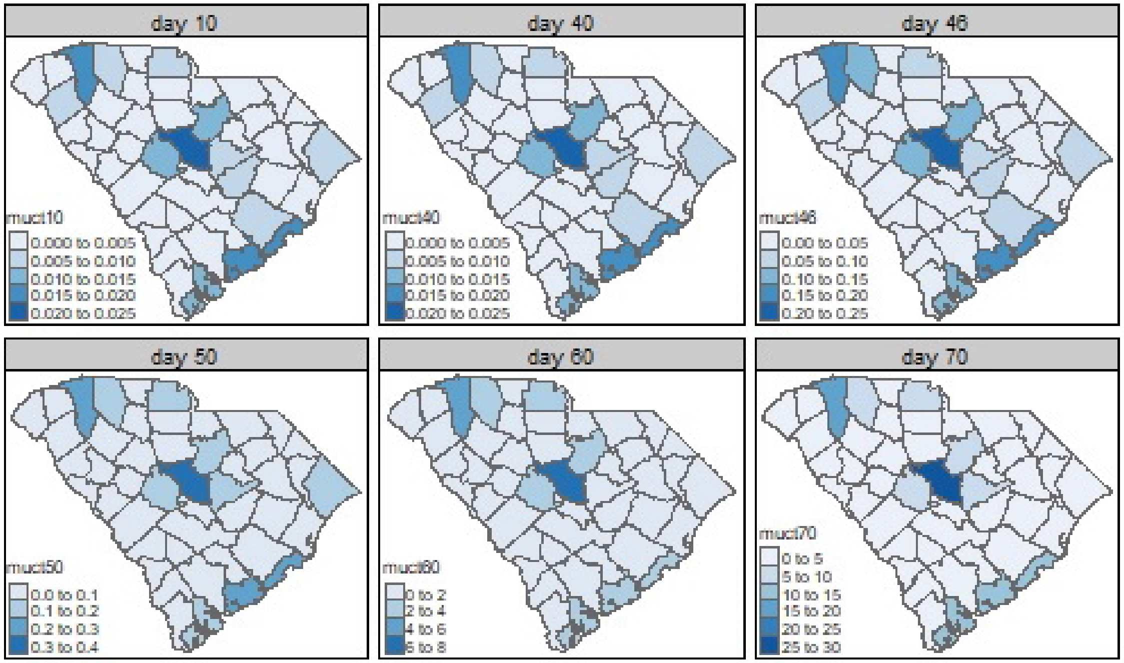
Motivation: Covid-19 in SC counties 2020: April 2nd – June 29th

- Dynamic Profiles
- 4 selected counties



Spatial background

- Modelled estimates of risk



Background

- Infectious disease models:
  - Descriptive
  - Mechanistic
- Descriptive models are usually random effect based
- Mechanistic models deal with transmission and are typically compartment models such as SIR or SEIR. They can be agent-based also.
- Individual or aggregate?

Individual models

- Classic susceptible-Infected-removed (SIR) models
- Differential equation models
- Difference versions can lead to lagged models
- R_0 is a function of coupled ODEs
- Often data is not available at the individual level in that observing all infected/removed could be problematic: under-ascertainment is common
- Very few epidemics where the complete realisation of the process is observed. (Hagelloch measles epidemic, **1861** is an exception)

Some assumptions

- Each person has independent risk of being infected

at *i* th location and time *j* with probability

$$p_{ij} = 1 - \exp\{-\lambda_{ij}\}$$
$$\lambda_{ij} = f(\text{susceptibility}, \text{closeness of infectives}, \text{contextual effects})$$

- Models for this probability can be constructed with likelihoods and hence the Bayesian paradigm is available.
- Usually models are based on contact distances in some way.
- Can also be contextual
- Examples are found in Deardon et al (2010) and elsewhere.

Aggregate count models

- Often counts of new infectives are more readily available.
    - Due to confidentiality concerns in human populations
    - Also stability concerns.
- Typical models assumed are:
    - Poisson in large populations with low probability of infection
    - binomial in finite populations
    - Poisson approximation is often adequate for SIR models
    - Negative binomial models are sometimes assumed to accommodated overdispersion. However these are not needed when using BHMs with a Poisson data model and random effects

SIR time series count models

- Morton and Finkenstädt (2005) Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society* C, 54, 575-594 (M&F)
- A good example of time series modeling of measles outbreaks in London, UK. Assumes a SIR model where infective counts depend on previous infectives and susceptible pool.

Modelling constructs

- Transmission model
  - Observed new infectives at *i* th location and time *j* : y_{ij}
  - True infective count at *i* th location and time *j*: I_{ij}
  - Susceptible pool at *i* th location and time *j* : S_{ij}
  - Removal at *i* th location and time *j* : R_{ij}
- Accounting equation
  - Idealised

$$S_{ij} = S_{i,j-1} - I_{i,j-1} - R_{i,j-1}$$

Under-ascertainment

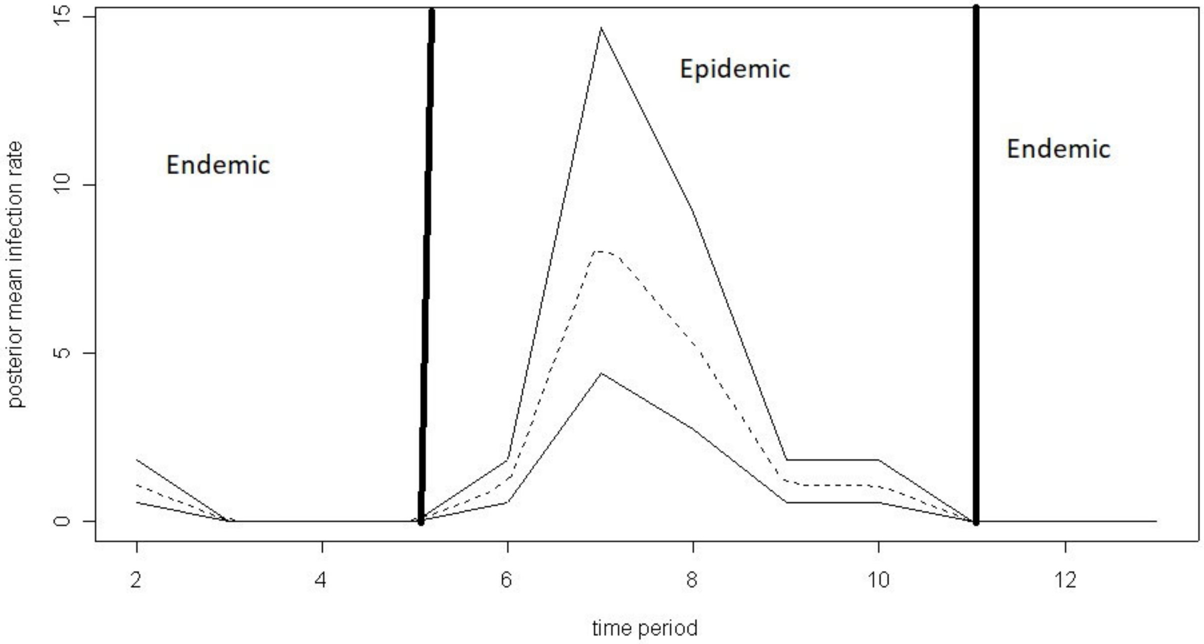
- Both true case count and true removal count could be under-ascertained
- The whole epidemic is seldom observed
- What to do?
  - Treat observed as the modelled data
    - Treat the true as a scaled version (simplest option)
  - Model the observed but estimate the true as a latent component (more complex and computationally difficult)
  - M&F assumed $y_{ij} \sim bin(\rho, I_{ij})$

Endemicity/Epidemicity

- For some disease there could be endemic behavior as well as epidemic.
- Endemicity occurs when a disease remains in the population but is not in an epidemic state
- Seasonal flu is slightly endemic (as some cases occur outside the main seasons)
- Outbreaks in the winter months represent an epidemic phase
- Models can be constructed where we have

$$E(y_{ij}) = \mu_{ij} = En_{ij} + Ep_{ij}$$

Endemic-Epidemic



Asymptomatic/Symptomatic

- With some viruses there can be asymptomatic and symptomatic cases (e.g. Covid19, typhoid)
- Confirmed tests of those with symptoms establish symptomatic case counts. Of course the degree of testing undertaken determines how many are found.
- Asymptomatic cases can infect others.
  - It is not clear how to assess these cases as they are not observed
  - Surveys of populations can yield proportions of asymptomatics, but these need to be repeated frequently. This is why wide spread testing in the Covid19 pandemic is important
  - In models the proportion of asymptomatics can be assumed known (but varied to assess effects).

Spatio-temporal modeling

- Mechanistic models are relevant
- SIR models are possible
- Here I describe an spatial extension of the M&F time series model, which allows for the spatial correlation and also neighborhood direct effects

Model structure I

$$y_{ij} \sim Poiss(\mu_{ij})$$

$$\mu_{ij} = S_{ij} f(y_{i,j-1};\boldsymbol{\theta})$$

S_{ij} are current susceptibles at start of time priod

$$\mathrm{S}_{ij} = S_{i,j-1} - I_{i,j-1} - R_{i,j-1}$$

Should $I_{i,j-1}$ be replaced by $y_{i,j-1}$?

- Can also use a binomial model if the sparseness is not too great

Model Structure II

- How is $f(y_{i,j-1};\boldsymbol{\theta})$ parameterised?
- Different models could be envisaged
  - Direct dependence on previous counts
  - Neighborhood dependence
  - Also dependence on predictors such as population density or % under the poverty line
  - Addition of spatially-structured noise? ICAR ?

SC Flu season county data

- Some previous work on the 2004 Flu season in SC led to an extension of the M&F model with spatial structure.
- Lawson and Song (2010) Bayesian hierarchical modeling of the dynamics of spatio-temporal influenza season outbreaks *Spatial and Spatio-temporal Epidemiology*, 1, 187-195 (L&S)
- Data: biweekly influenza C+ lab notifications
  - 13 periods
  - Iceberg effect

Flu models: Model I

$$y_{ij} \sim bin(\rho, I_{ij})$$

- Model assumes that the observed count is a proportion of the true infectives
- The true infectives depend on the previous true infective count

Flu models: Model II

- Accounting model

$$
\begin{aligned}
&I_{ij} \sim Pois(S_{ij} f(I_{ij-1})) \\
&S_{ij+1} \sim N(\mu_{ij+1}, \sigma_s{}^2) \\
&\mu_{ij+1} = S_{ij} - I_{ij} - R_{ij} \\
&R_{ij} \sim N(\beta I_{ij}, \sigma_R{}^2)
\end{aligned}
$$

Simpler version: Model 2

$$
\begin{aligned}
&I_{ij} \sim Pois(\pi_{ij}) \\
&\pi_{ij} = S_{ij} f(I_{ij-1})) \\
&S_{ij+1} = \mu_{ij+1} \\
&\mu_{ij+1} = S_{ij} - I_{ij} - R_{ij} \\
&R_{ij} = \beta I_{ij}
\end{aligned}
$$

Model 2

- How to parameterize the dependence on the previous infectives?

$$\log \pi_{ij} = \log S_{ij} + \log f(I_{ij-1})$$

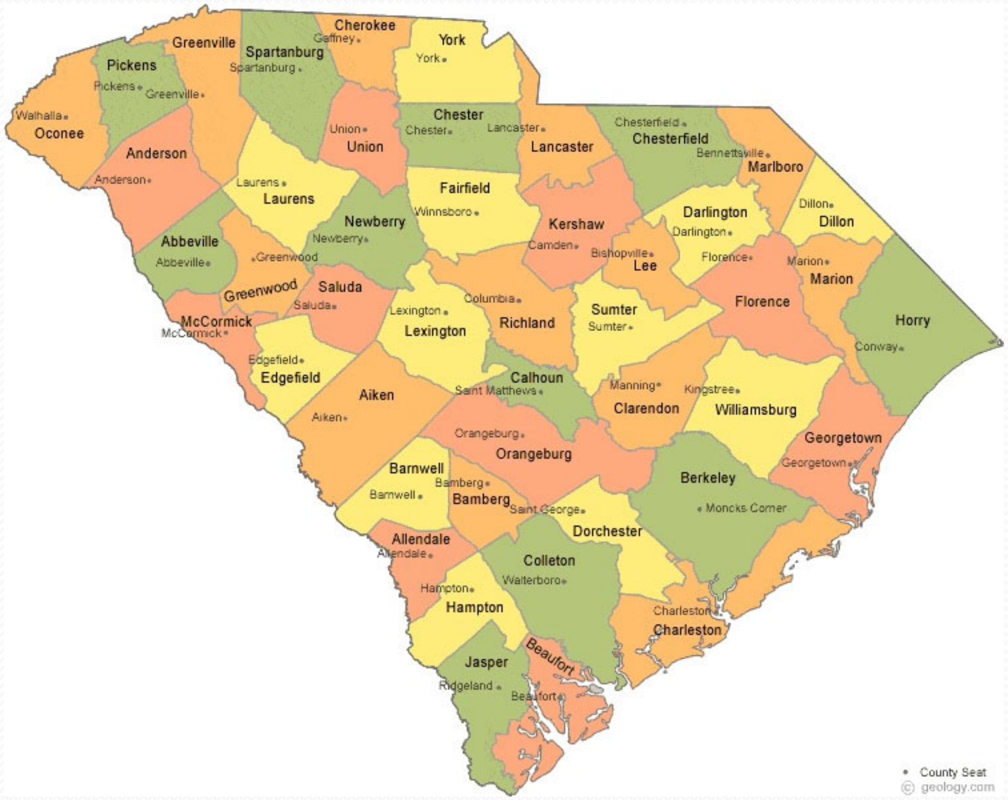
Dependencies

$$1)\log f(I_{ij-1}) = \log I_{ij-1} + b_0 + b_i$$

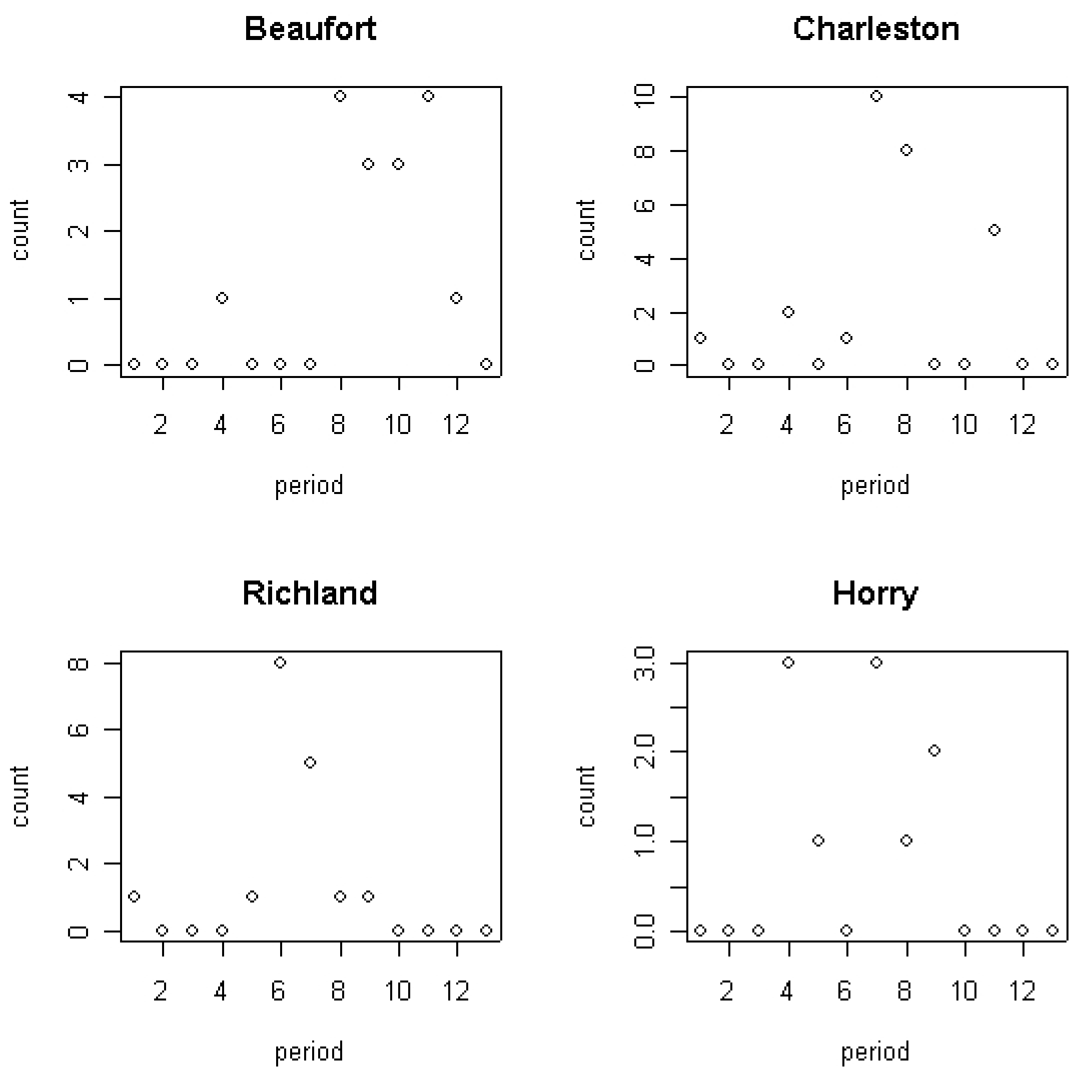
$$OR$$

$$2)\log f(I_{ij-1}) = b_0 + \log[I_{ij-1} + \sum_{l\in\delta_i} I_{lj-1}]$$

SC counties

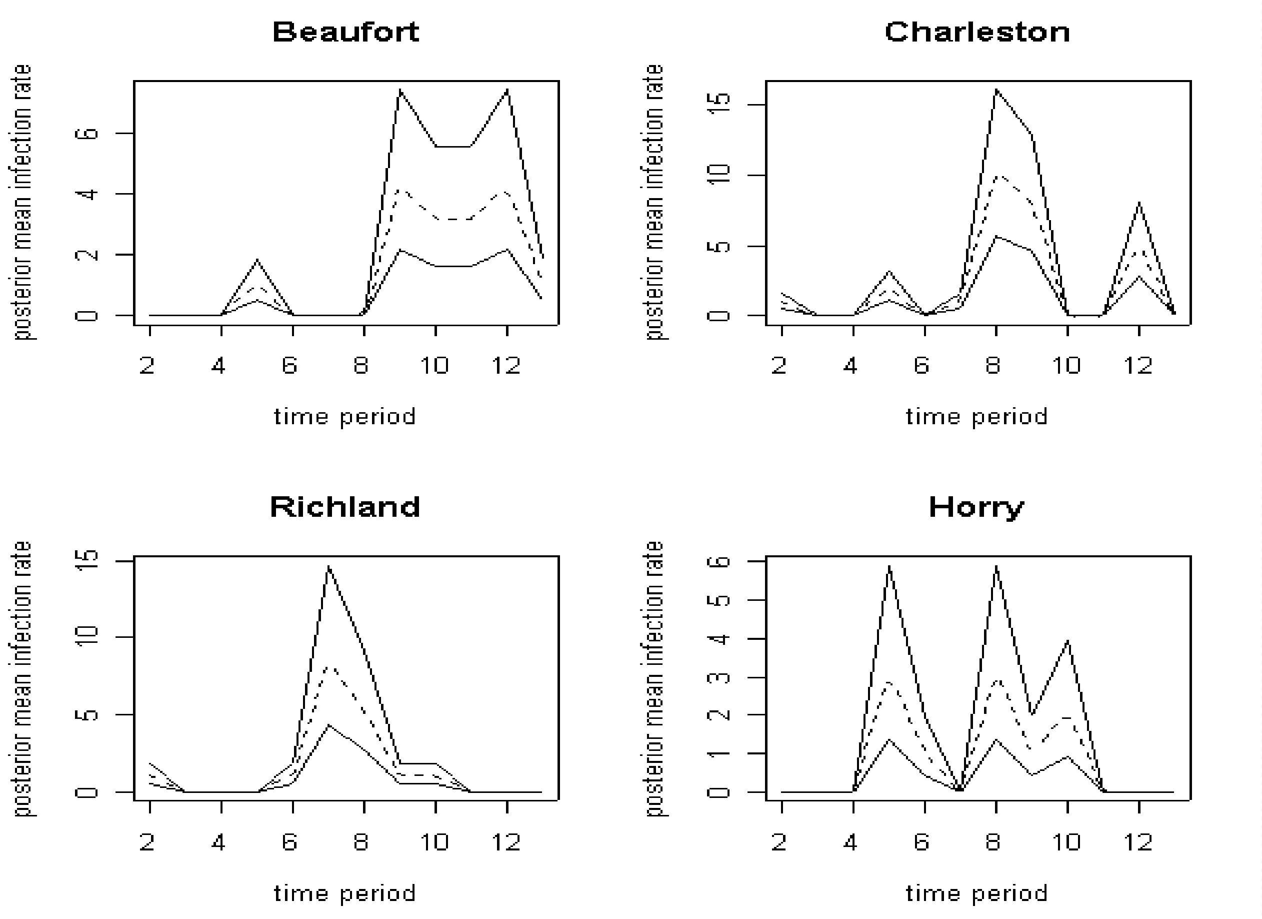

Cherokee
Gaffney
Greenville
Spartanburg
Spartanburg
York
York
Pickens
Pickens
Greenville
Walhalla
Oconee
Chester
Chester
Lancaster
Union
Union
Chesterfield
Chesterfield
Lancaster
Anderson
Bennettsville
Marlboro
Anderson
Laurens
Laurens
Fairfield
Winnsboro
Darlington
Darlington
Dillon
Dillon
Newberry
Newberry
Kershaw
Camden
Abbeville
Abbeville
Greenwood
Bishopville
Lee
Florence
Marion
Marion
Greenwood
Saluda
Saluda
Columbia
Lexington
Lexington
Richland
Sumter
Sumter
Florence
Horry
McCormick
McCormick
Conway
Edgefield
Edgefield
Calhoun
Saint Matthews
Manning
Clarendon
Kingstree
Williamsburg
Aiken
Aiken
Orangeburg
Orangeburg
Georgetown
Georgetown
Barnwell
Barnwell
Bamberg
Bamberg
Berkeley
Moncks Corner
Saint George
Dorchester
Allendale
Allendale
Colleton
Walterboro
Hampton
Hampton
Charleston
Charleston
Beaufort
Jasper
Ridgeland
Beaufort
County Seat
© geology.com

Data: 4 counties

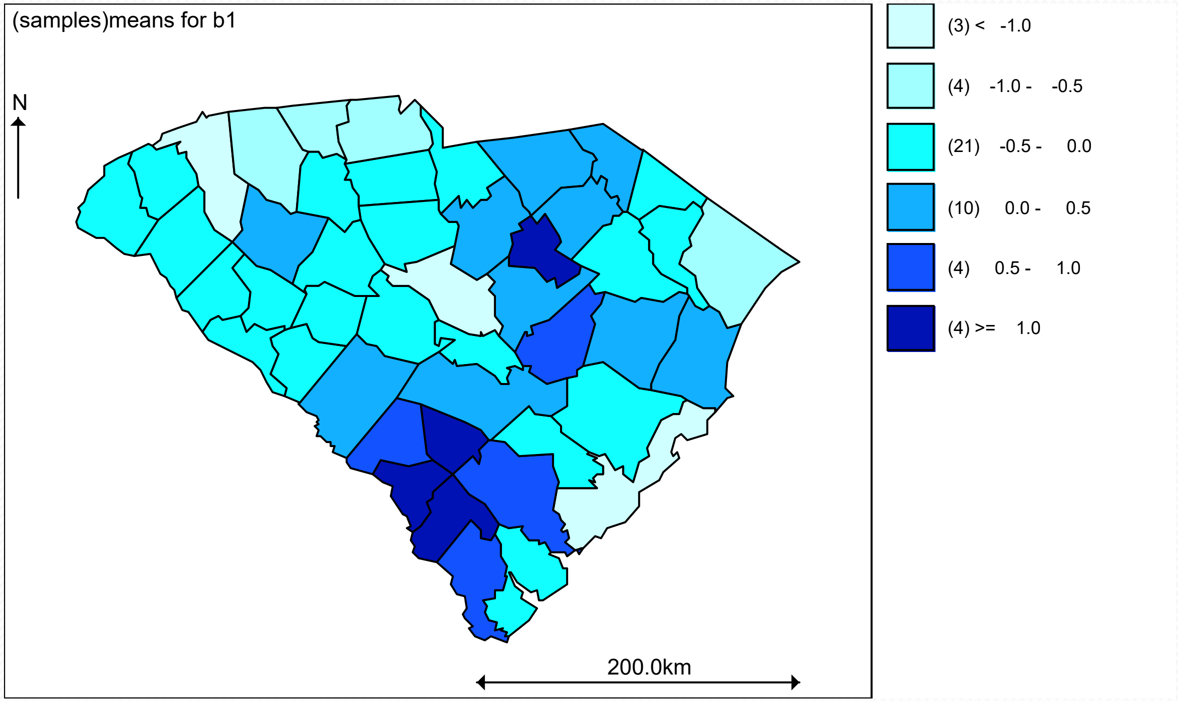


©Andrew B Lawson GEOMED webinar March 2021

Results: posterior mean risk profiles



Spatial heterogeneity: model 1



Covid19 in South Carolina

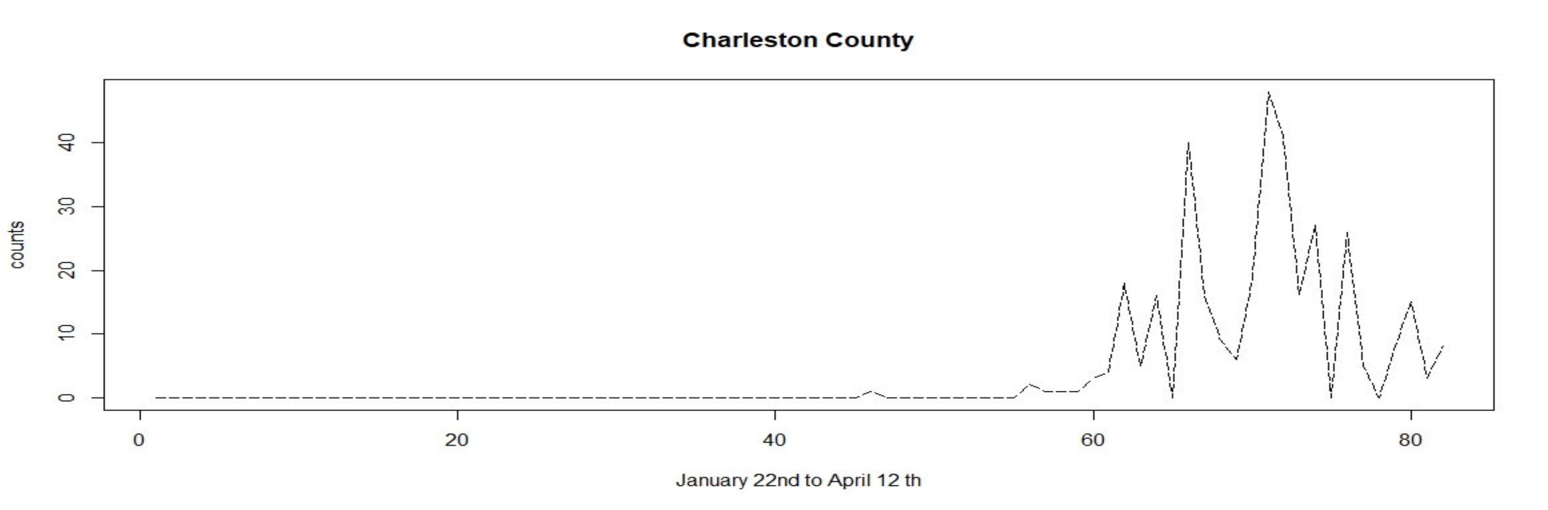
- Covid19 virus has gripped the world since late 2019.
- It is now a pandemic and few areas are unaffected
- Both cases and deaths are being recorded in various countries
- Scale is an issue: publicly available data is at province /state, county level but not below, although post codes are reported routinely on web sites.
- Individual level data is not publicly available right now

Aggregate data

- In the US both Johns Hopkins GitHub site and the New York Times site have county level daily cases and deaths.
- The JHU site is a amalgam of 10 different sources
- NYT site is based on National Center for Health Statistics (NCHS) US deaths data and case data from state health departments, and is well documented
- We have tried to use the JHU data but it is not well documented
- The NYT data seems to be more stable and well doucmented.

SC county level daily data

- Daily data on new cases
- Daily data on deaths
- Starting in January 2020
- Data from January 22nd to April 12th



Cases and Deaths

- Reporting issues
  - Cases: under-reporting as only severe symptoms in most cases are tested (symptomatic cases)
    - Asymptomatic not reported
    - Symptomatic under-ascertained
  - Cumulative count reporting errors:
    - Misallocation to wrong county
    - Testing
  - Deaths: only hospital Covid19 deaths recorded (initially) and also serious under-reporting especially in care homes.

Models

- Approach:
  - Assume that there is no endemic component (as yet!) and so we only have a propagator model coupled to accounting equation for symptomatic (*sym*) cases
  - Asymptomatics (*asym*) are largely unknown and so we either estimate the latent component or use scaling
  - As new infections must be a function of sym and asym

we must include both in any dependence model

Covid19 county level models

- Cases

$$sym_{ij} \sim Pois(\mu_{ij})$$

$$\mu_{ij} = S_{ij} f(sym_{i,j-1}, asym_{i,j-1}, \text{predictors, spatial confounding})$$

$$S_{ij} = S_{i,j-1} - I_{i,j-1} - R_{i,j-1}$$

where $I_{i,j-1}$ is the total case load

R_{ij} current removal (death, recovery)

- Asymptomatic assumed a proportion of symptomatic

Specific model components

- A variety of model variants have been run but the results cited here are for the following setup.
  - asym is 25% of sym
  - ICAR model assumed as spatial confounder
  - % under poverty line included as predictor
  - Removal consists recovered (0.1 of sym) and deaths
  - Best model found using DIC comparisons
  - Weakly informative gamma priors assumed for precisions

Models examined

$$1)\log f(I_{ij-1}) = \log I_{ij-1} + b_0 + b_i$$

$$2)\log f(I_{ij-1}) = b_0 + b_1 \log[I_{ij-1} + \sum_{l \in \delta_i} I_{lj-1}]$$

$$3)\log f(I_{ij-1}) = b_0 + b_1 \log I_{ij-1} + b_2[\%Pov] + b_i$$

$$4)\log f(I_{ij-1}) = b_{0j} + b_1 \log I_{ij-1} + b_2[\%Pov] + b_i$$

$$5)\log f(I_{ij-1}) = b_{0i} + b_1 \log I_{ij-1} + b_2[\%Pov]$$

DICs and a variant

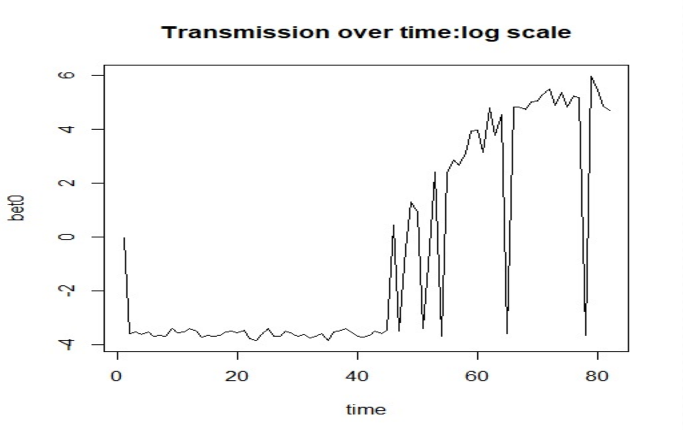
Model	DIC	pD (using SD(Dev)/2		
1	19449.2	72.16		
2	10,321,293.0	2499.6		
3	19431.0	55.99		
4	14327.2	122.74		*
4b	15057.6	36.97	minus ICAR	
5	19454.9	75.34		

Model 4 results

- Parameter estimates

Parameter	Posterior mean	SD
b1	-0.834	0.0088
b2	-0.699	0.0166

- b0



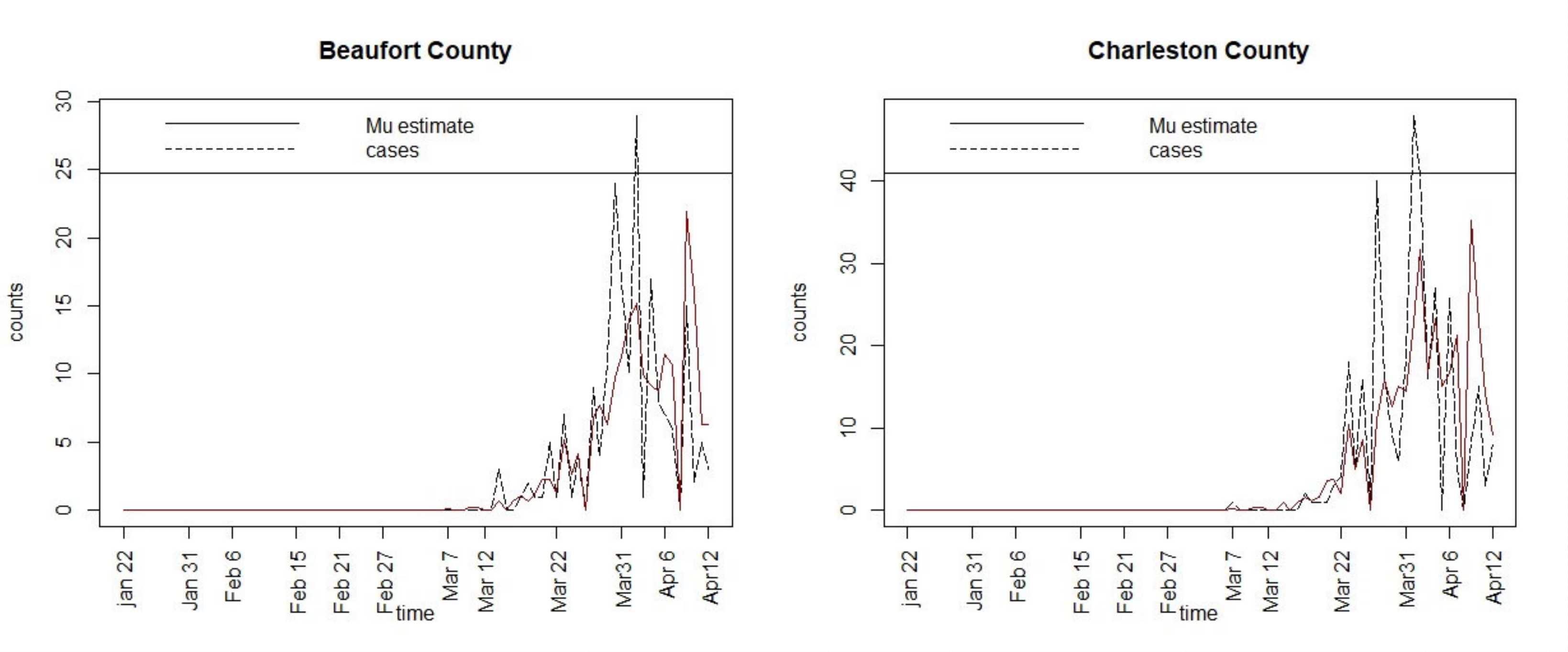
Key dates

- March 7th first case in Kershaw county SC
- March 12th WHO declares Pandemic
- March 13th SC state of emergency declared
- March 14th schools closed
- March 18th restaurants closed
- March 31st non essential businesses closed

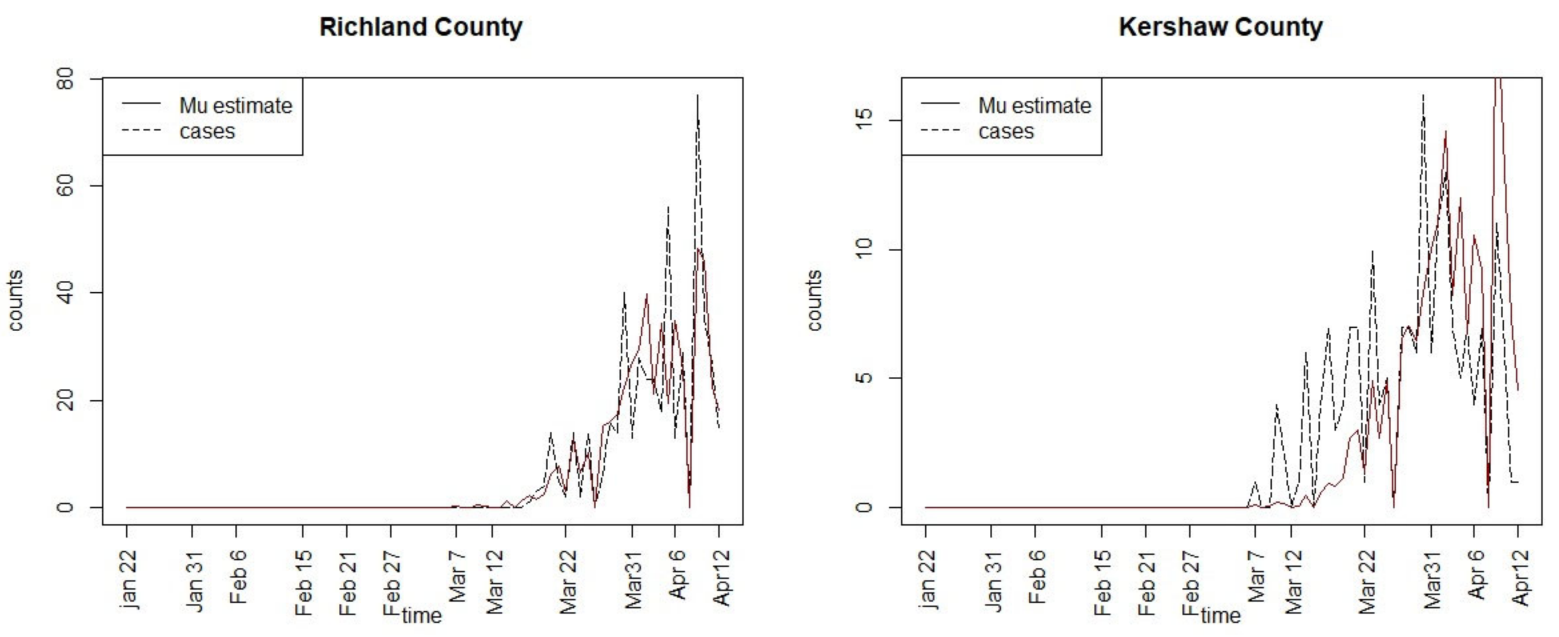
- April 6th stay at home order (lockdown)

- April 24th some lifting…some businesses opening
- May 1st lifted…open air restaurants open
- May 11th dining inside restaurants allowed
- May 12th stay at home order lifted

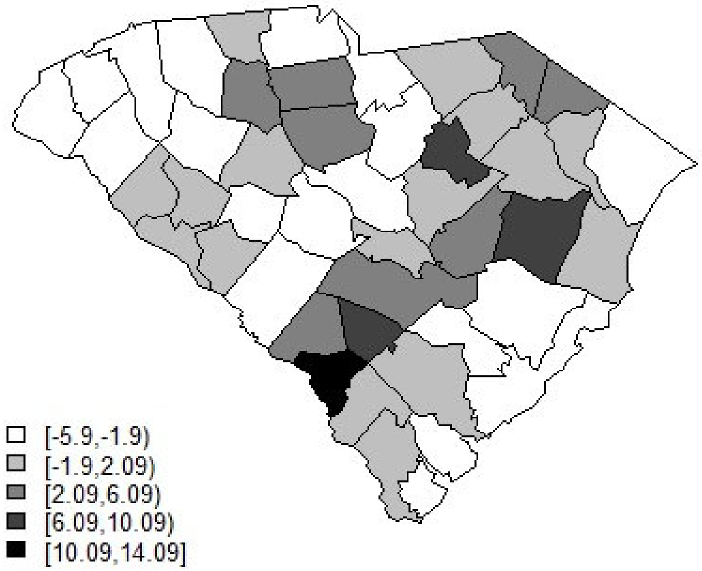
Posterior mean time profiles



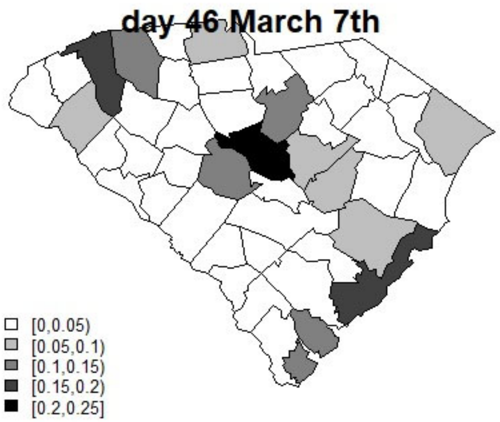
Posterior mean time profiles

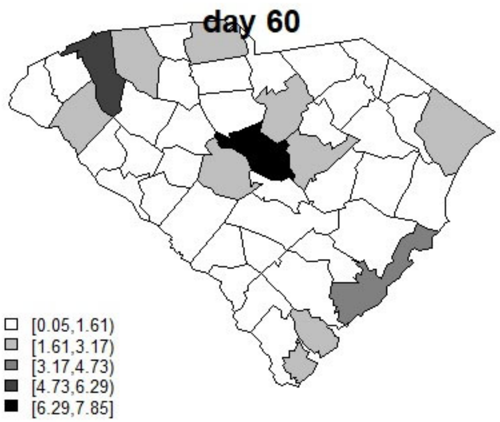


Spatial confounding: ICAR component

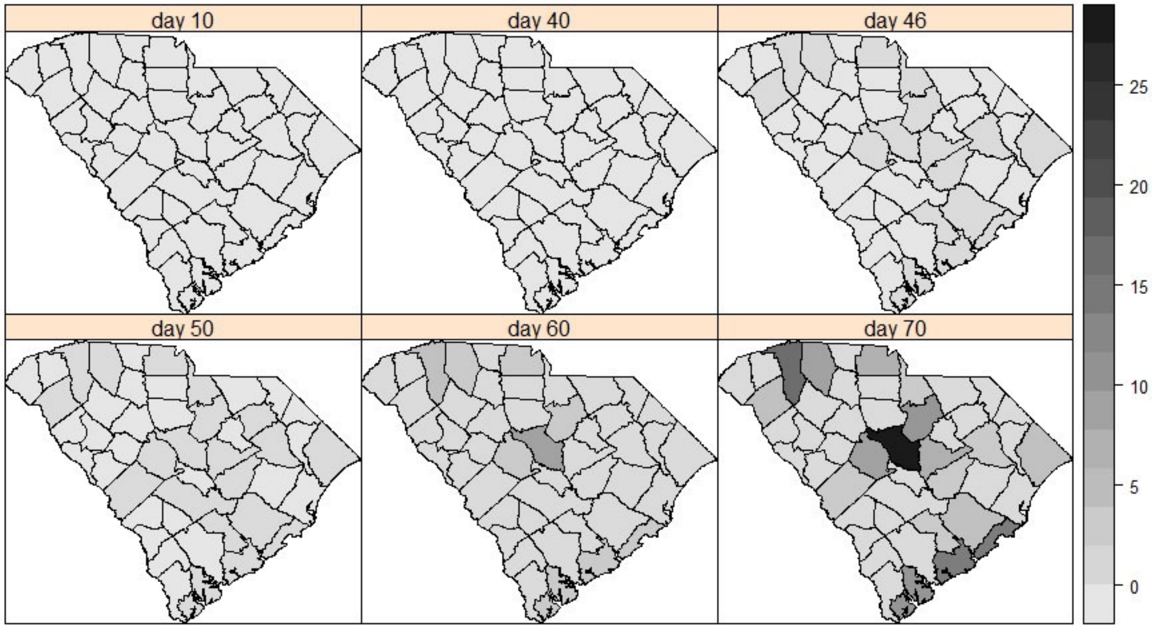


Posterior mean level maps: day 46 and 60





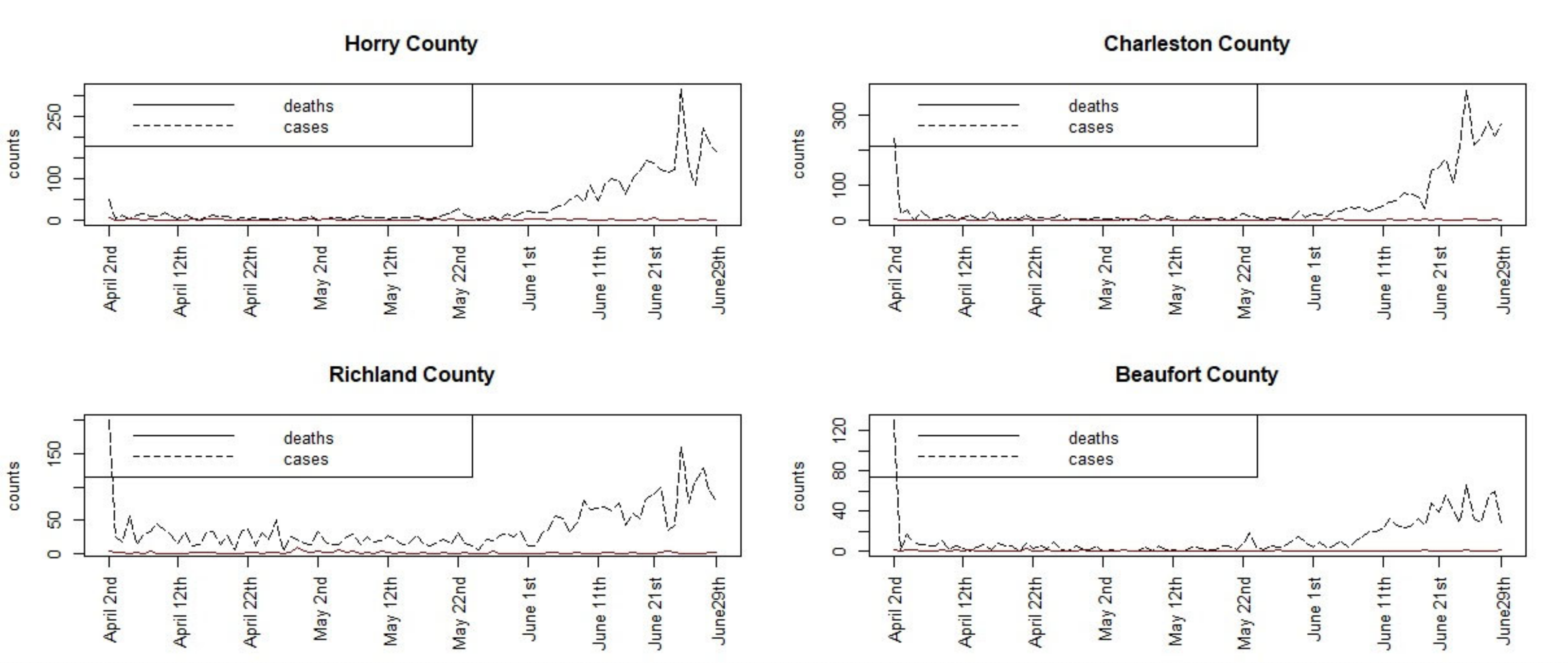
Sequence of posterior maps



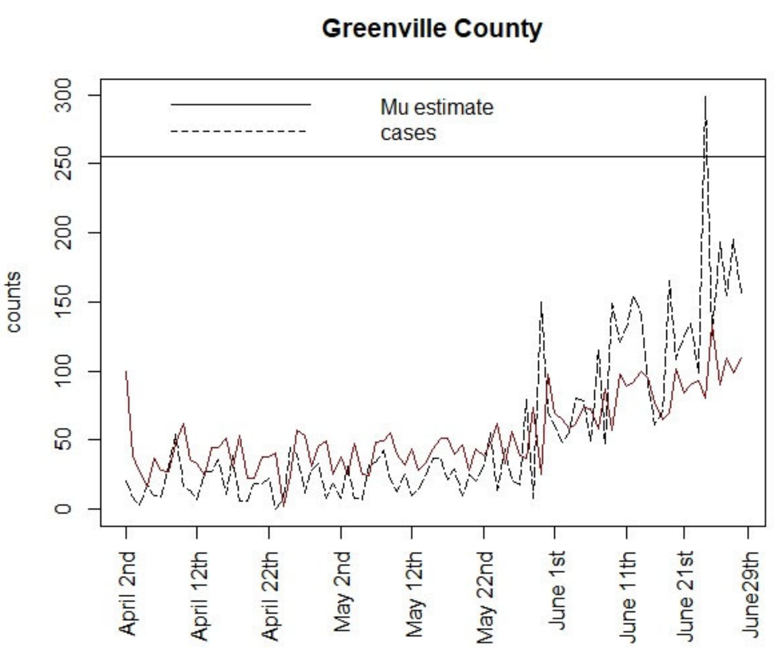
Later Periods and smoothed data

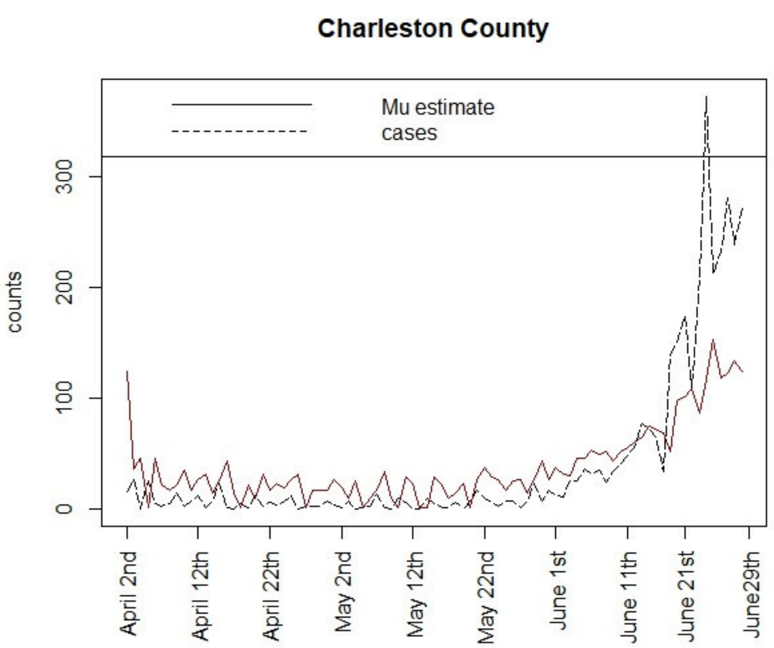
- A second period from April 2nd to June 29th was observed which included the second wave.
- Data for both periods were also subsequently smoothed using a 3 day smoothing.
- One step prediction was also enabled.
- Similar models were fitted to these additional data
- Model 3 with % poverty and spatial ICAR effect had best GOF

Second wave: April 2nd – June 29th

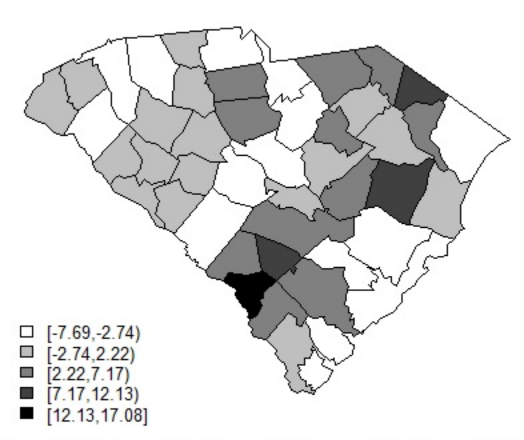


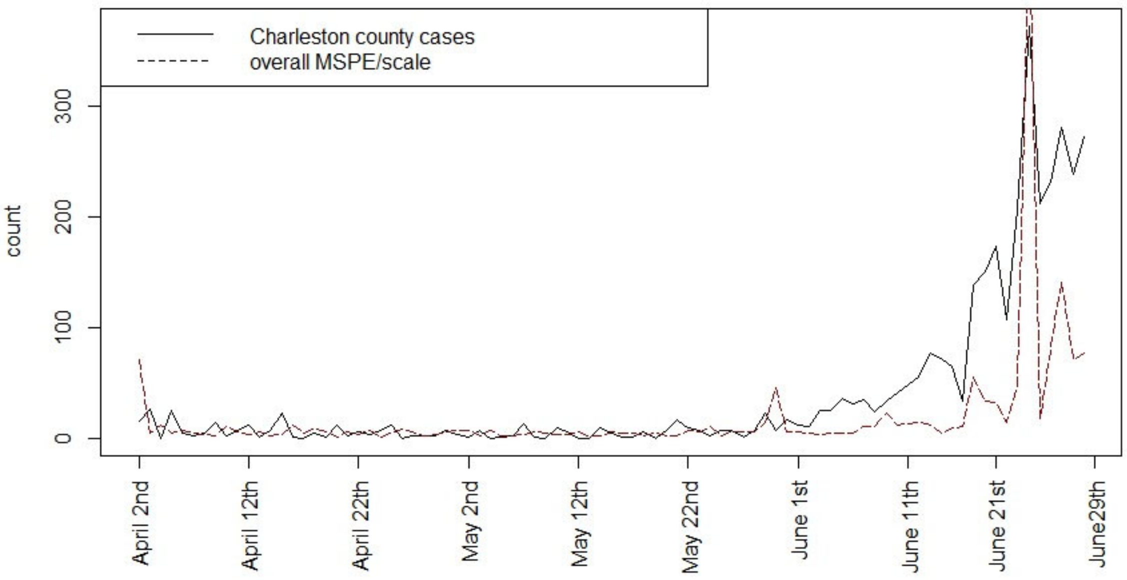
Model 3 estimates : second period



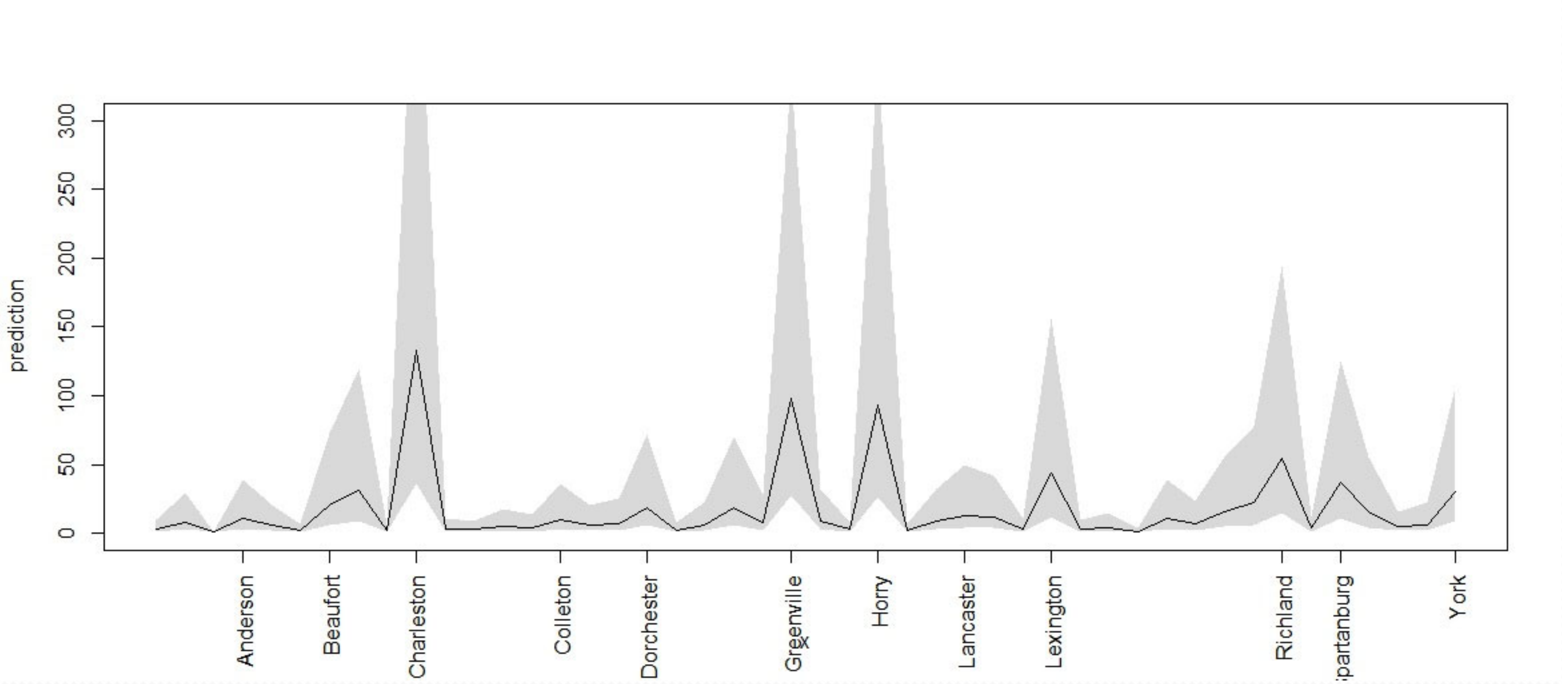


Spatial ICAR effect and MSPE for Charleston county

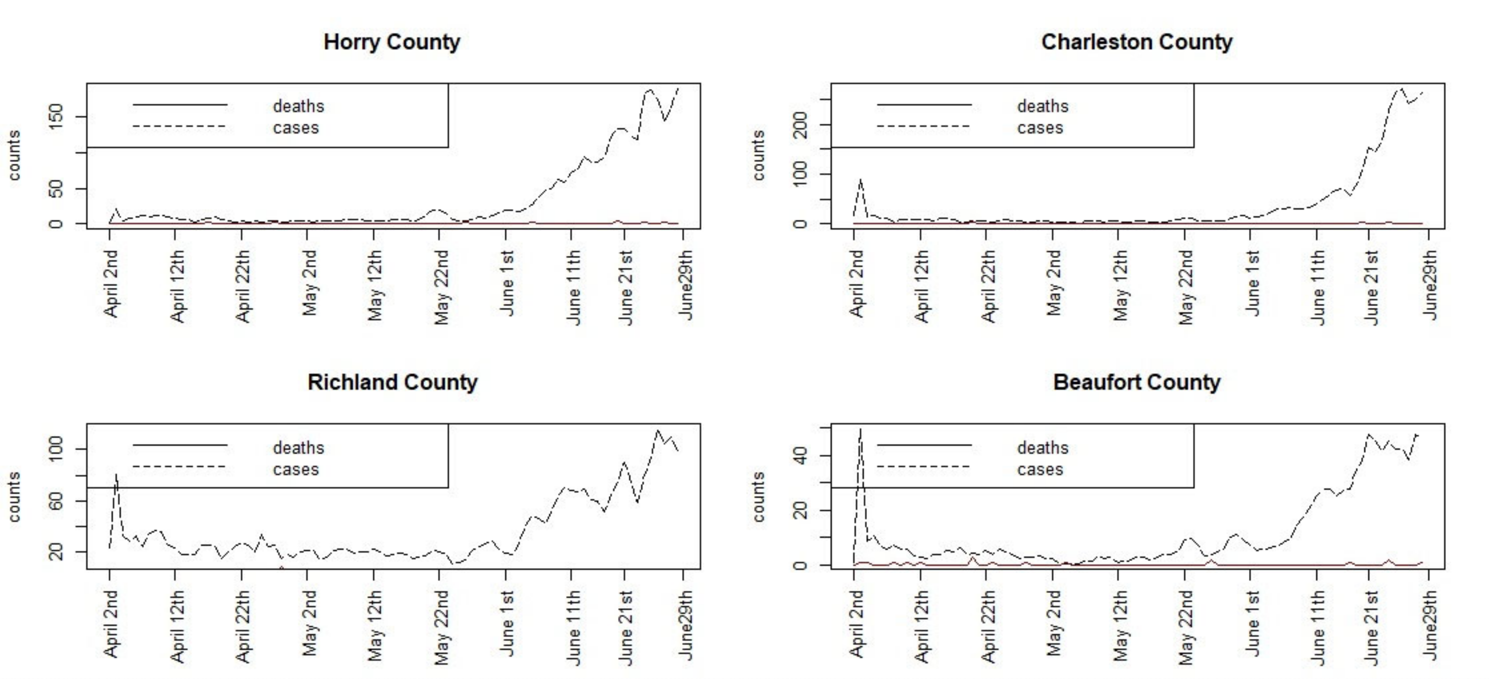




Prediction: one step June 30th

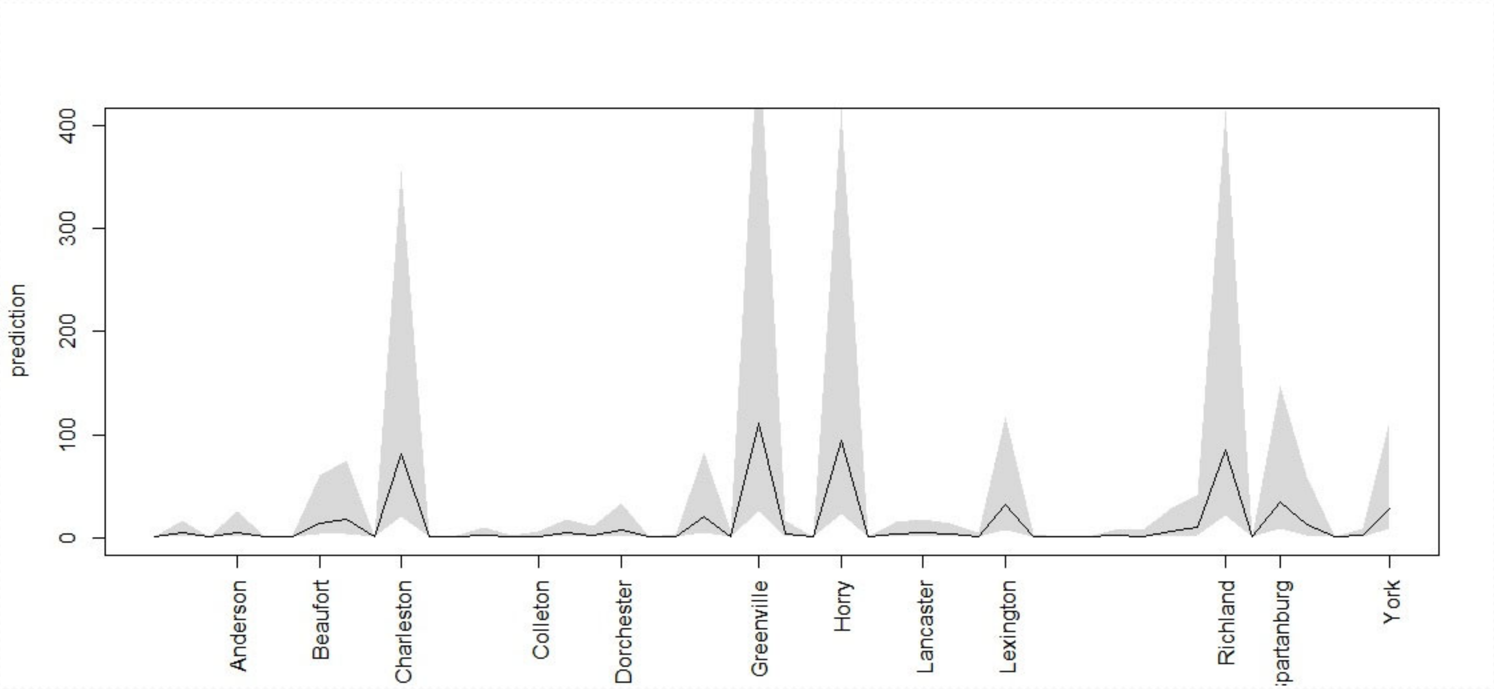


Smoothed Data (3-day average)



©Andrew B Lawson GEOMED webinar March 2021

Model results smoothed data



Comments

- SIR/SEIR Bayesian models can be estimated retrospectively for epidemic situations
- Some accounting for biases (weekends, county switches, underascertainment, asymptomatics) can be made via smoothing, scaling and asymptomatic modeling
- Death modeling could also be added as a joint model with lags
- Predictions can also be made
- Deprivation is a major factor in all best Covid-19 models

Paper

- Lawson, A. B. and Kim, J. (2021) Space-time Covid-19 Bayesian SIR Modeling in South Carolina
- https://www.medrxiv.org/content/10.1101/2020.11.03.20225227v1
- https://doi.org/10.1101/2020.11.03.20225227
- Accepted for PlosOne

NYT site

- https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html
- Github repository: https://github.com/nytimes/covid-19-data

Prediction and Surveillance

- A suitable retrospective model could be used for prediction but is not necessarily the best for surveillance
- Prediction is about getting close to true risk
- Surveillance is about detecting changes in risk

- Predictive distribution is useful for prediction within the observed data, and approximate prediction in the future
- However the confidence in prediction obviously decreases the further into the future moved.
- Better to have predictive evolving scenario so that data is incorporated as it arrives (prospectively)

Surveillance and metrics

- Surveillance models are often not optimal for retrospective studies
- Alternatively, the best fitting model for space-time retrospective data is not necessarily the best for detection of change.
- Imagine a situation where we have endemic and epidemic scenario, so that

$$E(y_{ij}) = En_{ij} + Ep_{ij}$$

where

$$Ep_{ij} = \alpha_0 + \alpha_1 p_{i,j-1}$$

and the propagator is $p_{i,j-1}$

Retrospective/Prospective

- A good retrospective model could have all the ingredients needed in En_{ij}
- For example we could have a descriptive RE ST model of the form

$$E(y_{ij}) = En_{ij} = \mu_{ij} = e_{ij}\lambda_{ij}$$

$$\log(\lambda_{ij}) = \alpha_0 + v_i + u_i + \gamma_j + \psi_{ij}$$

- However this could be too adaptive in time (!)
- We don’t want to model out the changes.

Detection of Change

- May want to detect $p_{i,j-1}$
- OR the existence of dependence via α_0, α_1
- One approach simply tests for positivity of these parameters
- Another approach uses metrics to detect changes
  - this is similar to the approach in quality control where processes are being monitored for compliance
  - In the disease case the monitoring has to be more sophisticated and account for at risk population.

Metrics

- A variety of measures are commonly employed to assess capabilities of models. E.g.
  - Predictive capability.....how well does the model predict the data and the future data (mspe, pmse,.......)
  - Residual deviation......using residuals to assess change
    - Bayesian predictive and surveillance residuals (see e.g. Vidal Rodeiro and Lawson (2006) Monitoring Changes in Spatio-temporal Maps of Disease. *Biometrical Journal*, 46,3,463-480)
- Metrics are posterior functionals that are intended to detect changes prospectively.

- Two different metrics have been proposed for Bayesian small area online monitoring
  - SCPO: Surveillance conditional Predictive Ordinate
  - SKL: Surveillance Kullback-Leibler measure

SCPO

- The SCPO is defined, for an MCMC sample of size G, as

$$SCPO_{ij} = \frac{1}{G}\sum_{g=1}^{G} Pois(y_{ij} \mid e_{ij}\lambda_{i,j-1}^{g})$$

- It is the average probability of the current data, given the previous time period's posterior sampled risk.
- If the SCPO is close to 1 then the data is closely predicted
- Often $1 - SCPO_{ij}$ is used for detection

SKL

- Kullback-Leibler divergence measures have two forms

and often the two are added together to give a total values: TSKL

- KL divergence is designed to assess differences in probabilities

$$p_1 = \sum_g Pois(y_{i,j}; e_{i,j}\lambda^{g}{}_{i,j-1}) / G$$

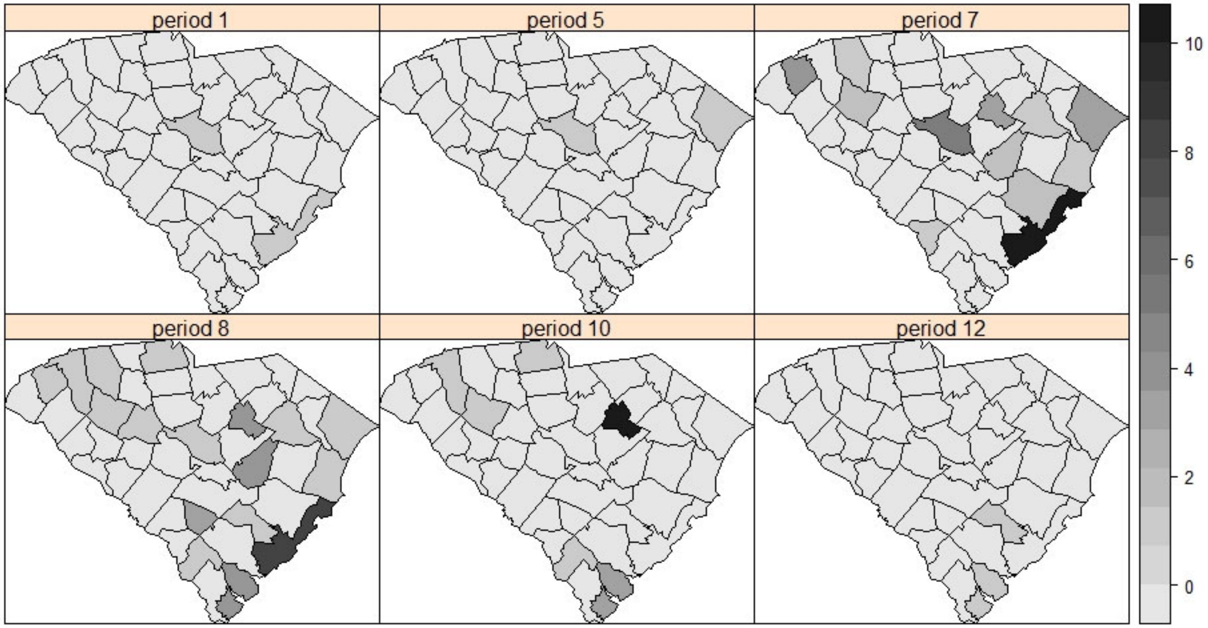
$$p_2 = \sum_g Pois(y_{i,j-1}; e_{i,j-1}\lambda^{g}{}_{i,j-1}) / G$$

$$skl_1 = p_1 \log\left(\frac{p_1}{p_2}\right);\ skl_2 = p_2 \log\left(\frac{p_2}{p_1}\right)$$

$$Tskl = skl_1 + skl_2$$

SC Flu season 2004-2005

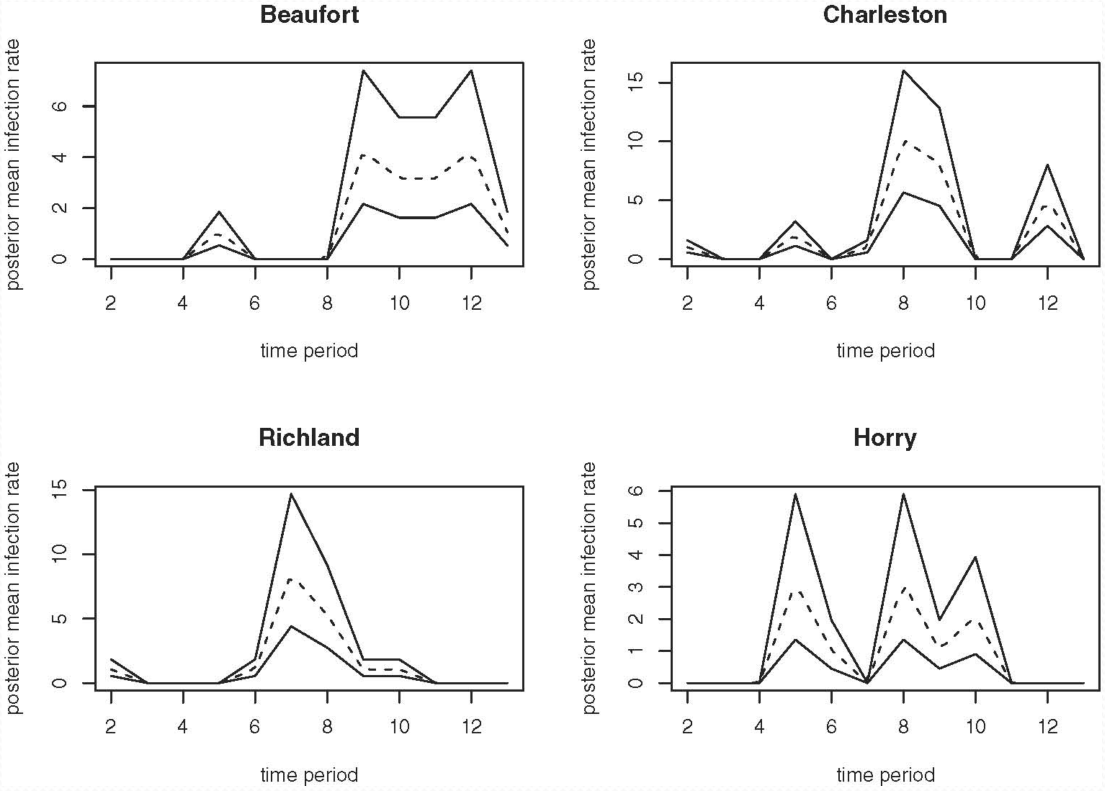
- County level C+ notifications (counts)
- 13 biweekly periods
- 46 counties



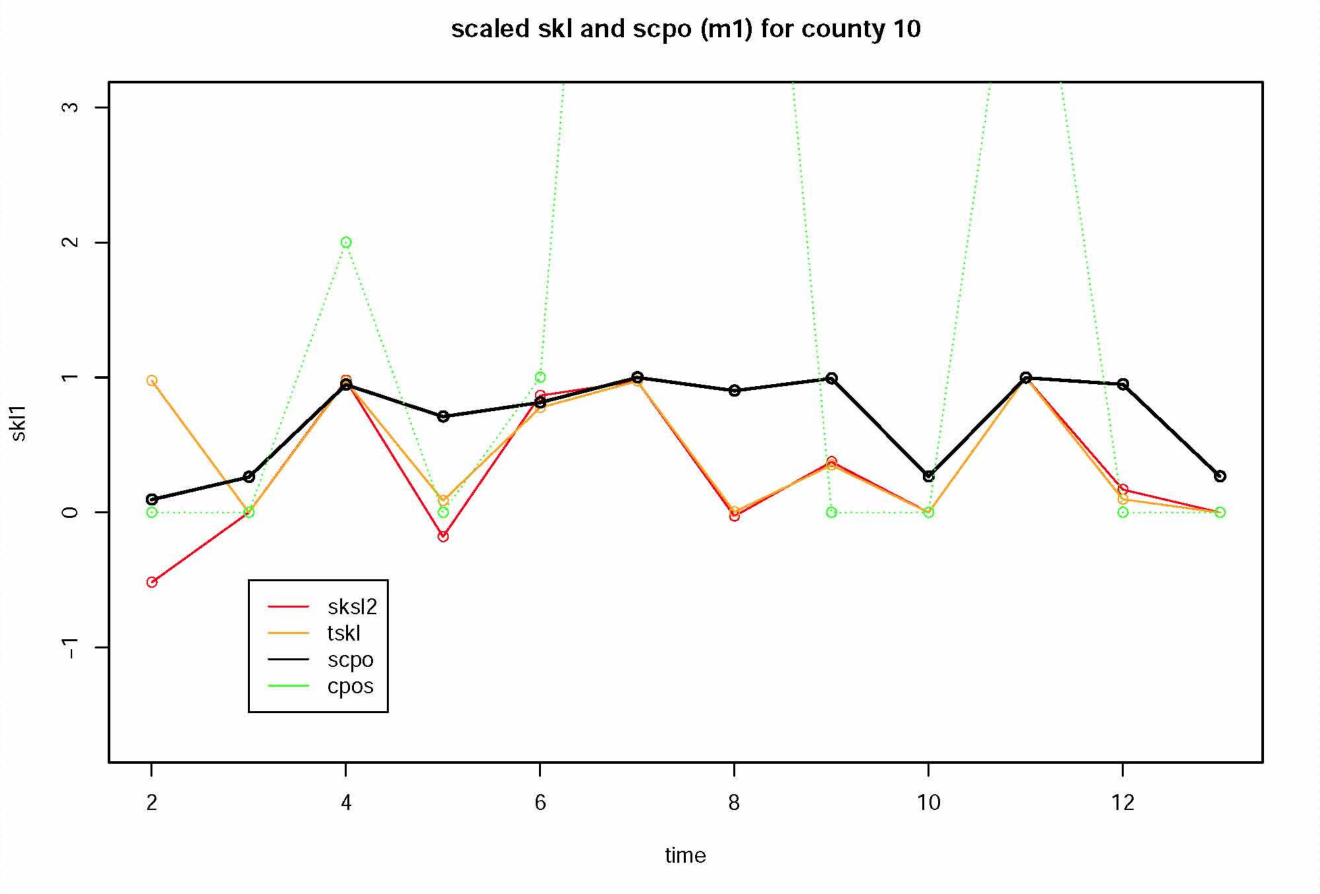
Some examples

- SC flu season 2004-2005
- County level
- Posterior mean risk

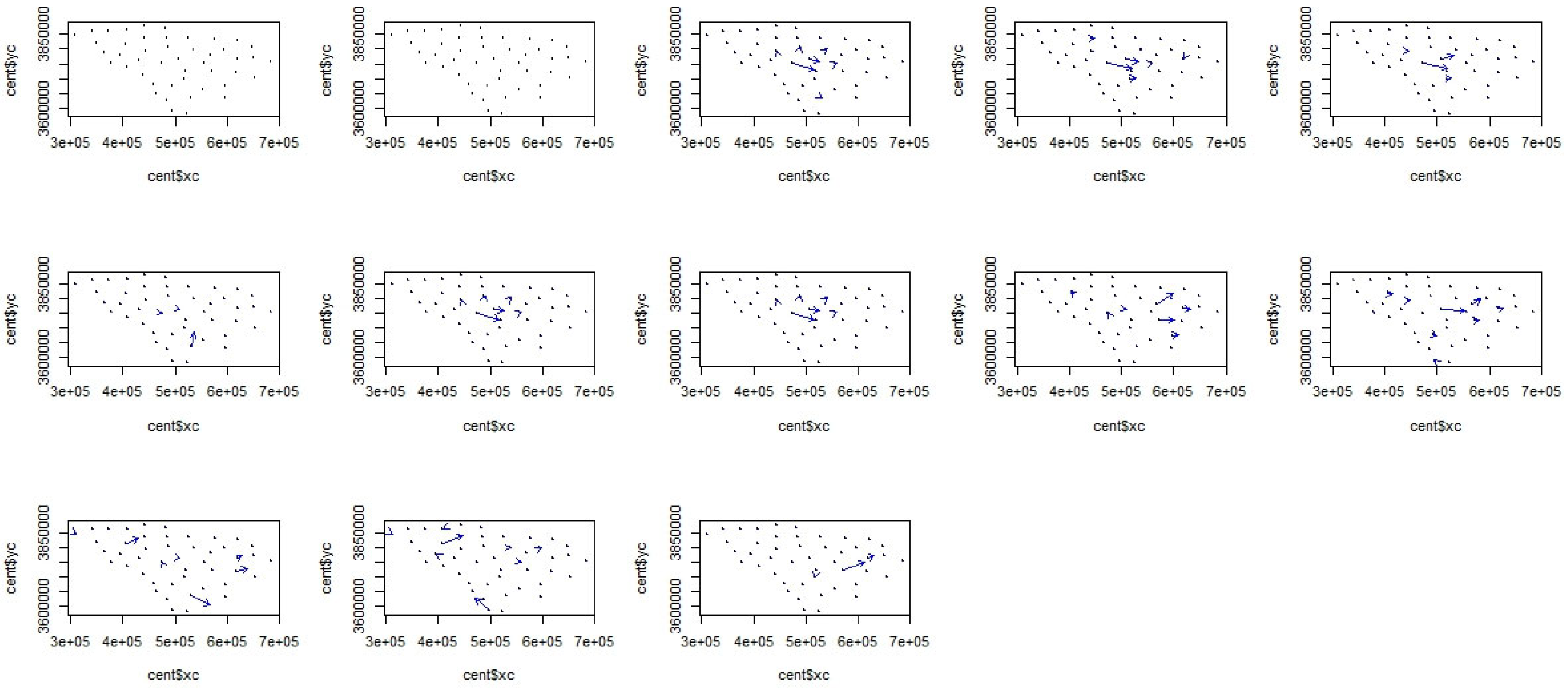
and 95 %credible intervals



Charleston county example



Directional potential modeling



Conclusions I

- Bayesian models can provide a range of tools for helping in the task of epidemic modeling and surveillance
- Descriptive models are limited in their predictive capabilities
- Endemic-epidemic models are a good avenue to pursue
- Lagged dependency models, which are approximations to difference equations, can fit reasonably well to daily respiratory infection data (e. g. Covid-19)

Conclusions II

- Infection rates can be estimated, neighborhood dependence and residual spatial effects can be included
- Predictors can be included also: % under poverty line was a significant correlate with infection risk.
- Metrics can be used to detect the start of epidemics and also their duration or termination
- Much work is needed in the calibration and tailoring of metrics for multivariate spatial data streams: a truly big BIG data application in public health.

References

- Keeling, M. and Rohani, P. (2008) Infectious Diseases in Humans and Animals. Princeton University Press.
- R. Deardon, S. P. Brooks, B. T. Grenfell, M. J. Keeling, M. J. Tildesley, N. J. Savill, D. J. Shaw& M. E. J. Woolhouse (2010), “Inference for individual-level models of infectious diseases in large populations” in Statistica Sinica, 20(1), 239-261.

Thank you