2021 SISMID Module 5
Lecture 7: Clustering and Cluster Detection for Count Data

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Overview of Clustering

Background reading: Chapter 8 of Elliott *et al.* (2000) and Chapters 6 and 7 of Waller and Gotway (2004).

We begin with an obvious statement: the distribution of the population across space is not uniform, and so even if cases occur completely at random amongst the population, the pattern of cases will not be uniform.

Informally clustering occurs when the spatial pattern of the cases is more "clumped" than the non-cases.

Mechanisms for clustering:

- Infectious diseases.
- Genetics.
- Risk factors, measured or unmeasured.
- Data anomalies (which may have spatial pattern).

A Definition of Clustering

(My) Definition of clustering: **A disease exhibits spatial clustering if there is epidemiologically-significant local spatial variation in residual risk.**

- Residual here acknowledges that known risk factors (e.g. age, gender) have been accounted for.
- Local recognizes that clustering is not simply large-scale trends. This is a subjective description.
- The epidemiologically-significant part is clearly also subjective but acknowledges that there will *always* be some level of residual variability.
- This definition is relative to the data we collect, and is not necessarily an intrinsic characteristic of the disease. For example, a particular set of data may have missing confounders, which induce clustering.

A Definition of a Cluster

(My) Definition of a cluster: **If a disease has increased residual risk in an area then this will lead in expectation to an 'excess' of cases – such a collection of cases is what we define as a** *cluster.*

- With this definition a cluster may be over a very large geographical area – some previous epidemiological definitions of a cluster are in terms of a realization of cases that are close in space.
- For example, Knox (1989) gives the definition, "a cluster is a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance".
- If a disease exhibits clustering then this may result in multiple clusters.
- Surveillance systems are built around cluster detection.

Running Example

We analyze data previously examined by Cressie (1991).

The data set also contains a neighbor list and data are available on the numbers of cases and on the number of births, both dichotomized by a binary indicator of race.

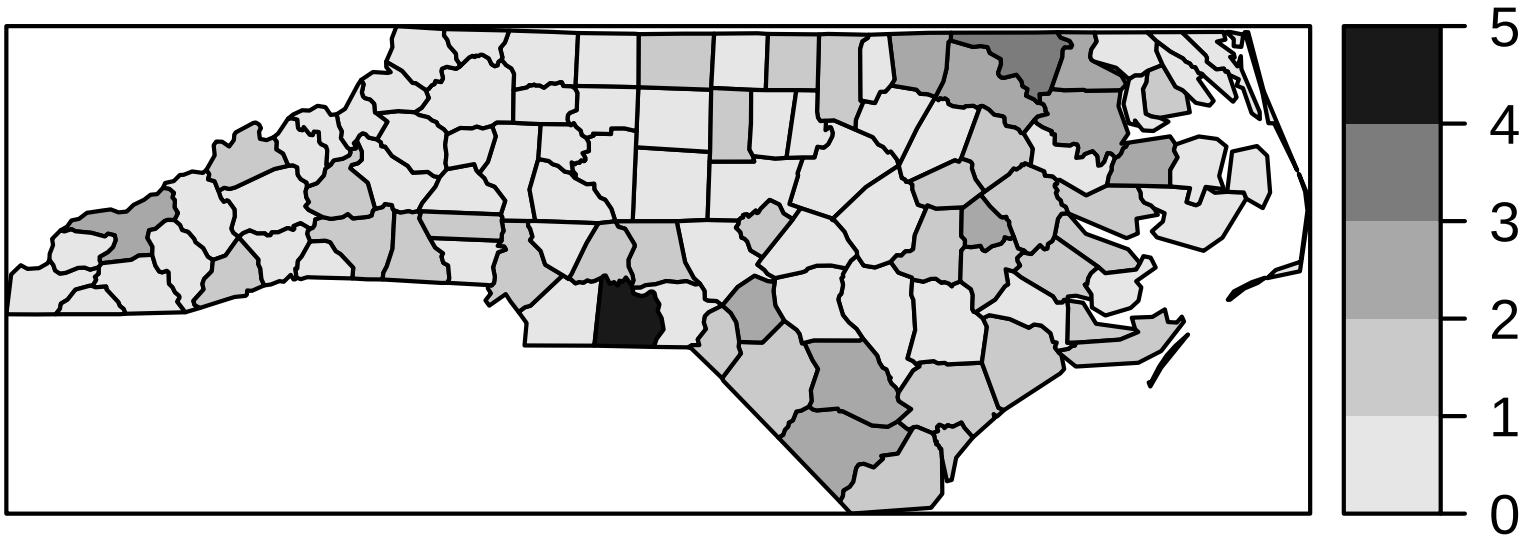

Figure 1: SIDS SMRs in North Carolina.

Clustering for Count Data

Overdispersion and Spatial Dependence

We first look at measures of overdispersion and spatial dependence for count data.

Due to unmeasured risk factors, data anomalies and within-area variability in confounders/exposures, it is usual for count data to exhibit overdispersion.

Overdispersion with rare events in the form of counts is often known as *excess-Poisson variability*, that is, independent counts with $\text{var}(Y_i) > \text{E}[Y_i]$ for $i = 1, \ldots, n$.

Spatial dependence is a different concept, namely, dependence between Y_i and Y_j that depends on the geographical positions of areas indexed by i and j, $i, j = 1, \ldots, n$, $i \neq j$.

Overdispersion

If we find evidence in the data that overdispersion is present then this is telling us that the data are not following the (Poisson) model that is often assumed.

The discrepancies may occur due to:

- unmeasured risk factors,
- the latter include infectious agents (which will often lead to spatial dependence also),
- data anomalies include under/count of disease cases and populations at risk,
- inaccurately measured exposures,
- model inadequacies.

We describe a number of statistics that may be used in exploratory analyses.

Methods for Detecting Overdispersion: Pearson's χ^2

- Pearson's chi-squared statistic is one measure of overdispersion.
- Suppose we fit the quasi-likelihood model:

$$\begin{aligned} \mathrm{E}[Y_i] &= E_i\theta_i \\ \mathrm{var}(Y_i) &= \kappa \times \mathrm{E}[Y_i], \end{aligned}$$

  where $\theta_i = \exp(\beta_0 + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}_1)$ with $\dim(\boldsymbol{\beta}_1) = p - 1$.
- Then a common approach (for example, as described in McCullagh and Nelder, 1989) is to estimate the overdispersion via Pearson's chi-squared statistic

$$\widehat{\kappa} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - E_i\widehat{\theta}_i)^2}{E_i\widehat{\theta}_i}. \tag{1}$$

  where p is the number of parameters in $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_1]$.
- A large value of κ means subnational variation in risk, which may be subsequently analyzed to gauge spatial dependence.

Autocorrelation Statistics for Assessment of Clustering of Count Data

A number of approaches have been suggested for measuring spatial autocorrelation – these are global measures and so address "clustering" and not "cluster detection".

A large number of statistics have been suggested to assess global clustering, and are typically of the form:

$$T = c\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \times \text{Similar}_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{2}$$

where

- c does not depend on the area i,
- n is the number of areas,
- w_{ij} is a weight reflecting the proximity between areas i and j, and
- Similar_{ij} is a measure of the similarity between data values Z_i and Z_j in areas i and j.

Assessing Significance

For many common choices the mean and variance of the statistics under the null of no clustering are available, and asymptotic normality may be appealed to under certain assumptions – not reliable and to be avoided.

In a permutation test approach (also known as a randomization or exact test) a test statistic is evaluated under all possible permutations of the data.

Unless the data set is small this is usually too computationally expensive, and so under a Monte Carlo test the distribtion of the test statistic is evaluated under a large number of randomizations.

In a Monte Carlo approach the data Z_i, $i = 1, \ldots, n$, may be repeatedly randomly assigned to different areas, and the statistic calculated under each assignment, yielding a comparison distribution.

Under a bootstrap approach the data are sampled, with replacement from the observed data.

Measures of Proximity

As with disease mapping there are various ways of measuring the 'closeness' of two areas, for example:

- Take $w_{ij} = 1$ if areas i and j are adjacent (i.e. have a boundary in common) and 0 otherwise.
- In the previous version the weights may be standardized so that they sum to 1 for each area.
- Take $w_{ij} = 1$ if the centroids of areas i and j are within the q nearest of each other.
- Take $w_{ij} = d_{ij}^{-1}$ where d_{ij}^{-1} is the inverse distance between the area centroids of areas i and j.
- More generally, take $w_{ij} = d_{ij}^{-\alpha}$ for some power $\alpha > 0$.
- Take $w_{ij} = 1$ if the centroids of areas i and j are within a certain distance of each other.

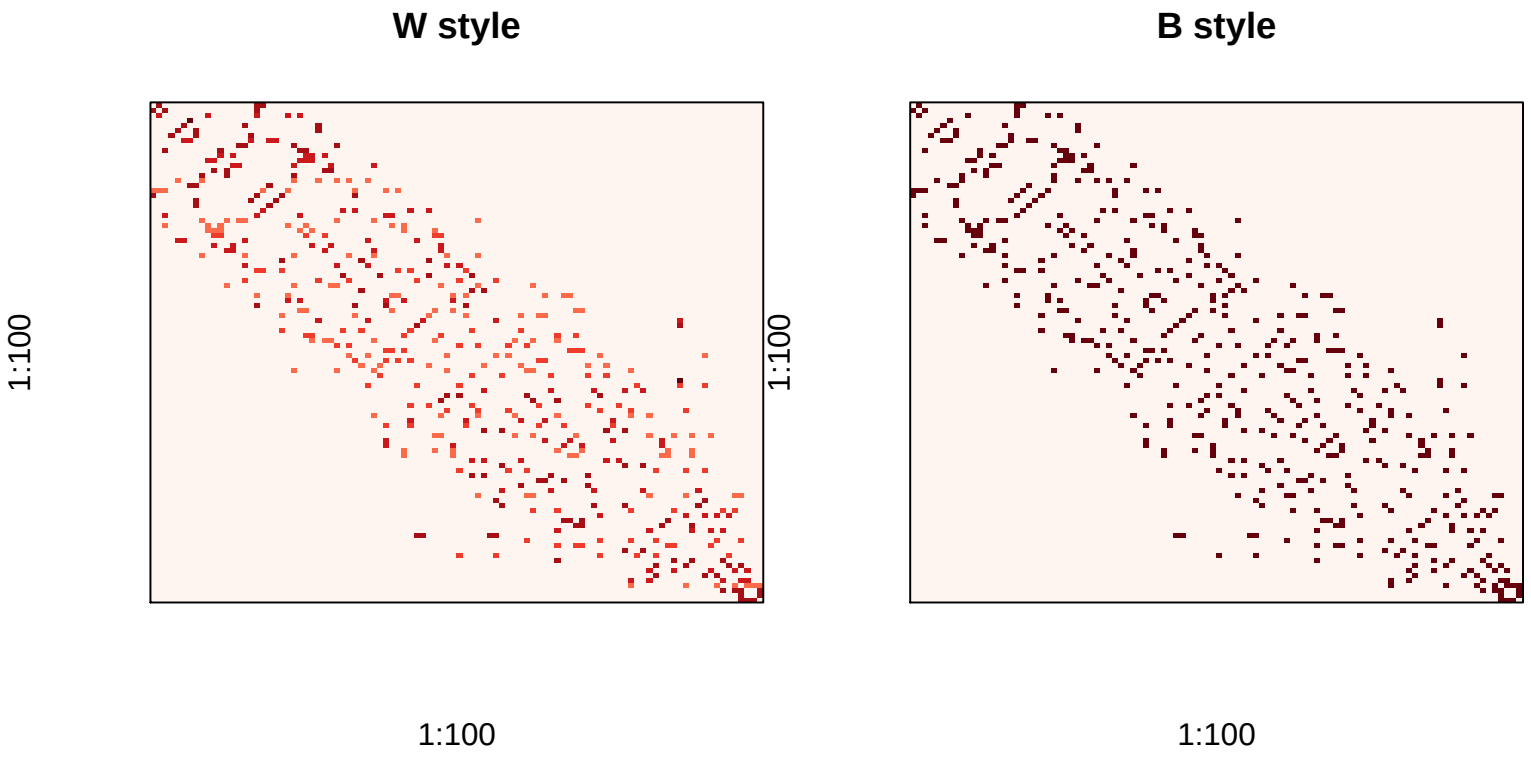

Figure 2: Neighborhood schemes. B has 0/1 corresponding to non-neighbor/neighbor – this means areas with many neighbors are more influential. W has rows standardized by the number of neighbors so that the sum for each row (area) is unity.

Measures of Proximity

The choice of weights depends on the type of spatial dependence that one is trying to detect.

For example, a distance-based measure may be appropriate if a smoothly-varying environmental pollutant is thought to be responsible for the clustering.

See Bivand et al. (2013, Sections 9.2, 9.3).

What to use as the “data”?

Considerations:

1. Standardization: We will almost always want to standardize the observations in some way (and not use the raw counts, since these are based on different population sizes).

   As an example we could take $Z_i = Y_i/N_i$ if we have counts within an age-gender stratum (e.g. men over 65).

   Alternatively, to control for confounders we might take $Z_i = Y_i/E_i$, the SMRs, of area i.

   Unfortunately the above choices do not yield data, Z_i, $i = 1, ..., n$, with the same variance which can induce anomalous behavior.

2. Detrending Spatial large-scale trends should be removed before the statistic is calculated, e.g., look at residuals after putting latitude and longitude in the model.

A Time Series Tangent

In a time series context with equally-spaced data the correlation between observations Z_i at lag $k = 1, 2, \ldots$ is

$$\rho(k) = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(Z_i - \overline{Z})(Z_{i+k} - \overline{Z})}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2}.$$

This can be rewritten as

$$\rho(k) = \frac{1}{S^2}\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - \overline{Z})(Z_j - \overline{Z})}{n} \tag{3}$$

where

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})^2$$

and w_{ij} are weights such that $w_{ij} = 1$ if $i + k = j$ and $= 0$ otherwise.

The $\rho(k)$ are plotted versus k to give a correlogram.

Space is more complex because it is 2D and the areas are irregular, but the form (3) suggests a way forward.

Moran's *I* statistic

Moran's *I* statistic (Moran, 1950) is,

$$I = \frac{1}{S^2} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}, \tag{4}$$

where

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (Z_i - \bar{Z})^2.$$

- If there is no spatial dependence *I* will be close to zero.
- If there is clustering then areas close together (as defined by w_{ij}) will tend to have responses that are similar and so the term $(Z_i - \bar{Z})(Z_j - \bar{Z})$ will be positive and the statistic *I* will be positive.
- The statistic is similar to the regular correlation coefficient though it need not lie in $[-1, +1]$. Under the null, $\mathrm{E}[I] = -1/(n-1)$.

Geary's c statistic

Geary's c statistic (Geary, 1954) is closely related to Moran's statistic and is,

$$c = \frac{1}{s^2} \frac{\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - Z_j)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}. \tag{5}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2.$$

- If there is spatial dependence, terms in the numerator will be small (similar responses in "close" regions) and the value of the statistic will be close to zero.
- The absence of spatial dependence leads to c close to 1, with $c = 0/2$ corresponding to perfect positive/negative correlation.

Issues with Assessment of Clustering

- These statistics were designed for situations in which the variance of Z is constant across space.
- If we use the SMR as our outcome, then the variance is clearly not constant, because the expected numbers are not constant.
- High or low values of Z_i will tend to occur in areas with small populations, i.e. in rural areas, and these are likely to be close together, inducing positive dependence.
- The problem is that under permutations under the null the spatial distribution of the expected numbers is not retained (we permute (Y_i, E_i) pairs) which would compensate for the increased variability.
- For a thorough discussion of spatial autocorrelation statistics, see (Schabenberger and Gotway, 2005, Section 1.3).
- Here we recommend using the residuals from a Poisson regression model – these have a variance that is closer to constant, across areas.

Moran's Test for North Carolina SIDS Data

Using the residuals from a Poisson regression with no covariates, and with the W weighting scheme we obtain p-value of 0.0074.

Including Eastings and Northings in the Poisson model (to remove large scale trend) we get a p-value of 0.016 – still suggests clustering, but reduced significance.

Under the B neighborhood scheme we obtain a p-value of 0.013.

Geary's statistic gives results that are consistent with Moran's.

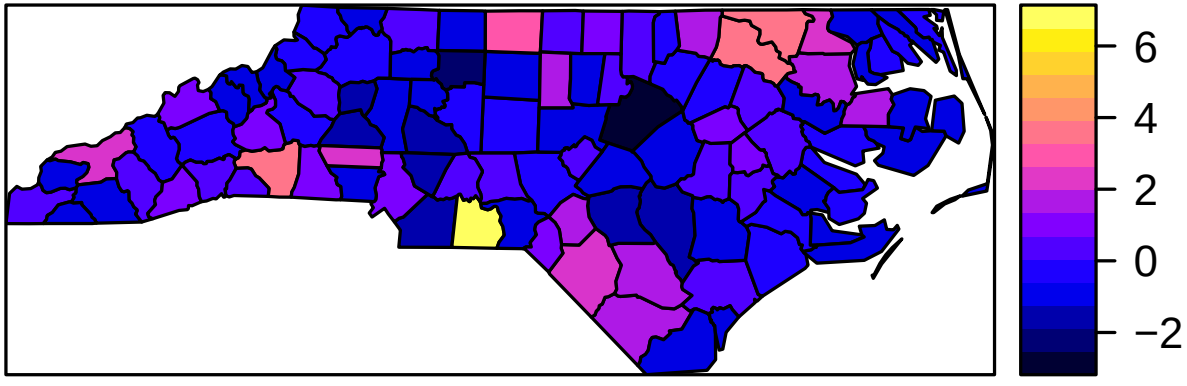

Figure 3: Residuals from Poisson model with linear trends.

Disease Mapping Models

We could fit the model,

$$\begin{aligned} Y_i|\theta_i &\sim \text{Poisson}(E_i\theta_i), \\ \log\theta_i &= \beta_0 + e_i \end{aligned}$$

with $e_i \sim \text{N}(0, \sigma_e^2)$ and priors on β_0, σ_e^2.

We then add ICAR random effects S_i, to give a BYM2 model (Riebler *et al.*, 2016).

IID model

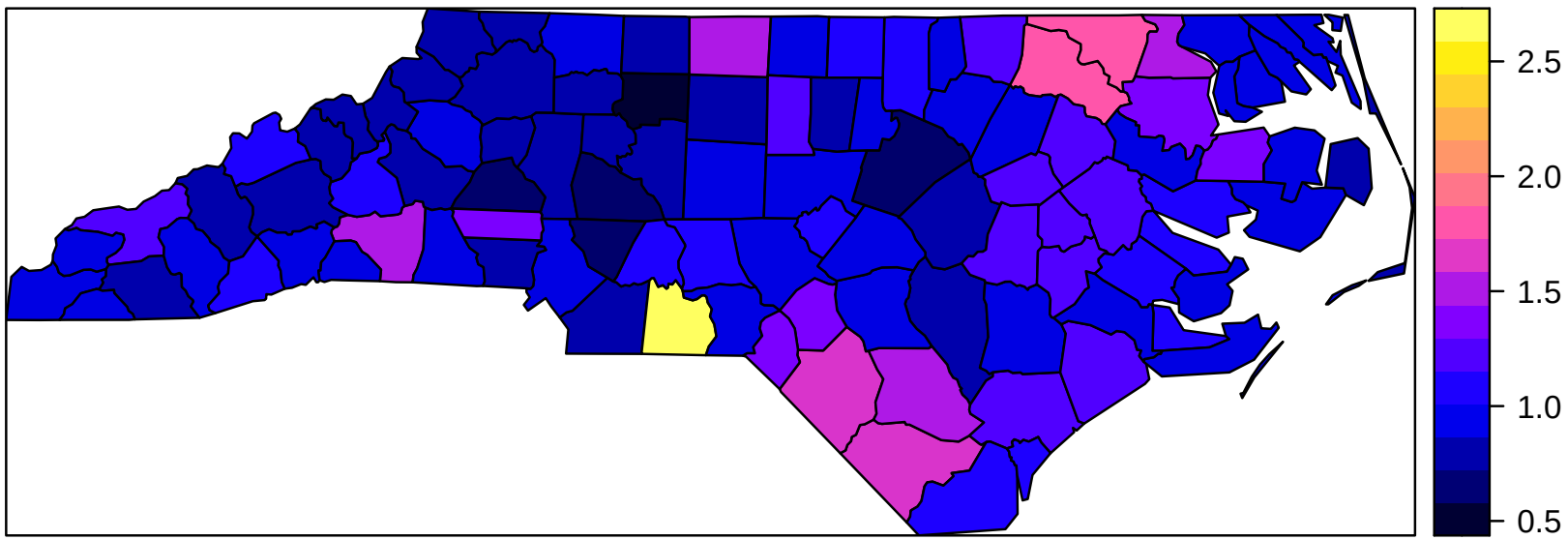

Figure 4: Posterior median of relative risks under IID hierarchical model.

BYM2 model

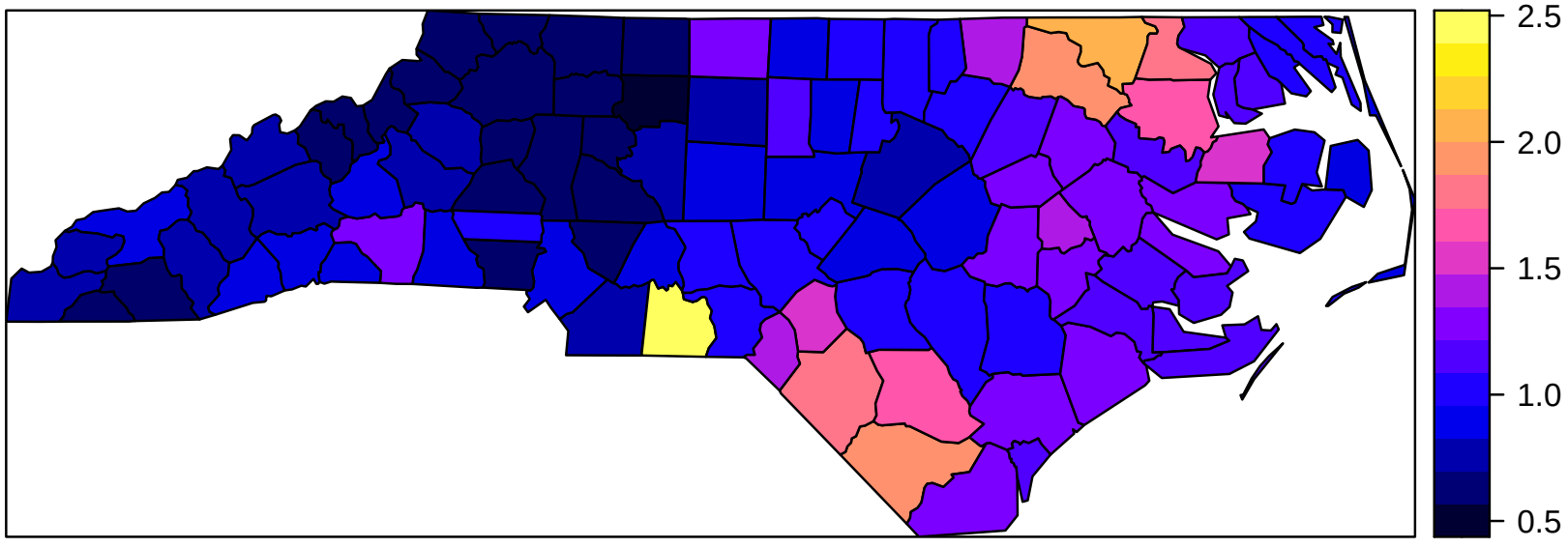

Figure 5: Posterior median of relative risks under BYM2 hierarchical model. The posterior median for the proportion spatial is 0.67 with 95% CI (0.18,0.97).

Comparison of hierarchical model

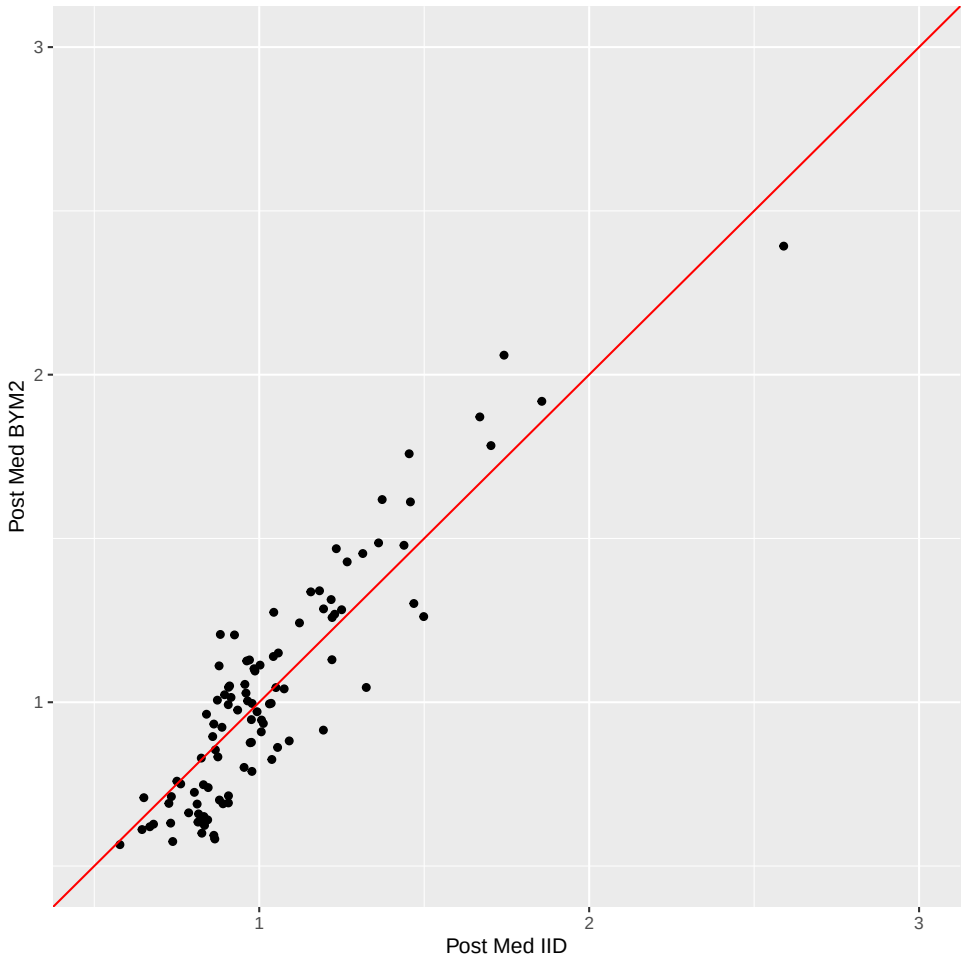

Figure 6: Comparison of posterior medians. The spatial model is producing differences in estimates from the IID model.

Comparison of hierarchical model

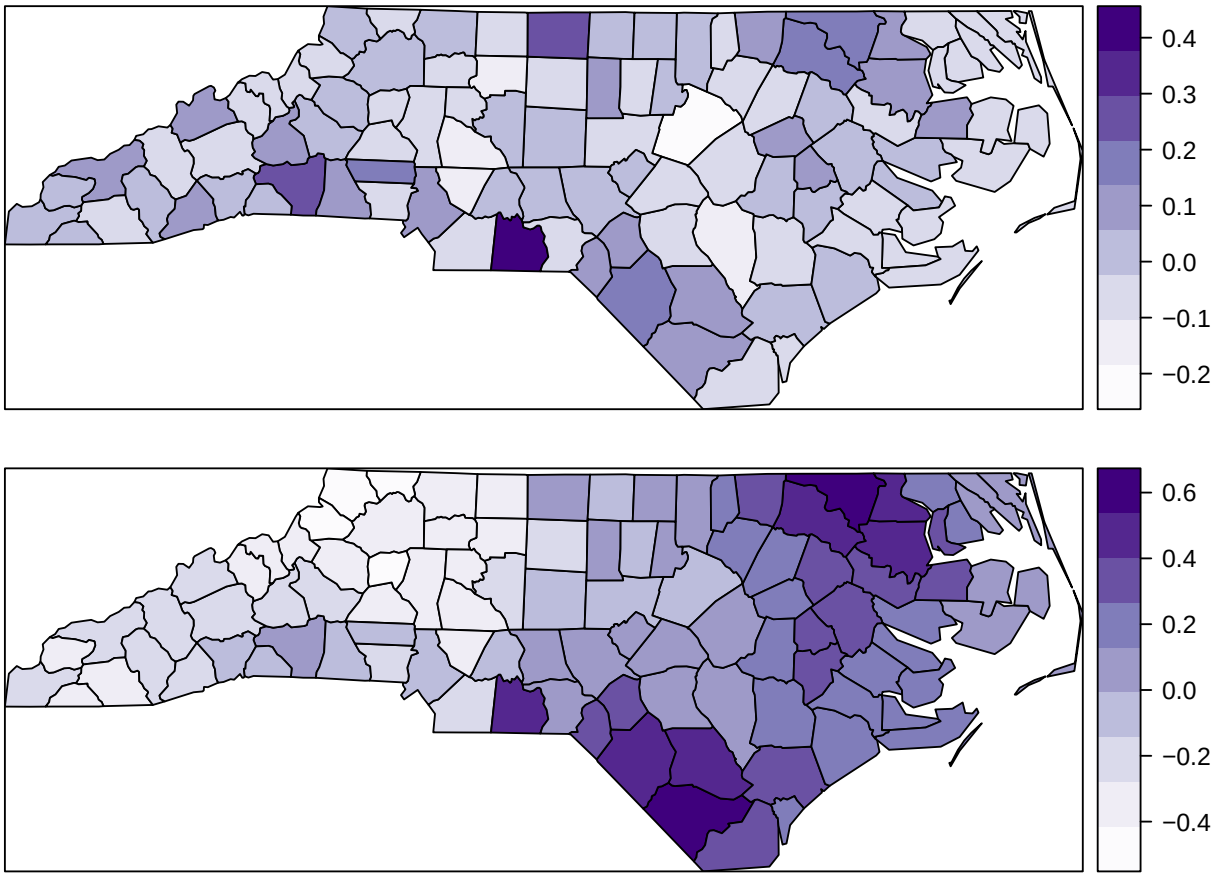

Figure 7: Non-spatial (top) and spatial (bottom) random effects. The spatial standard deviation is estimated as 0.43 with 95% interval (0.30,0.60). The proportion of the variation that is spatial, ϕ, is estimated as 0.67 with 95% interval (0.18,0.97). Hence, there is a large amount of excess Poisson variation, and the majority is spatial.

Clustering for Count Data Conclusions

General Approach

- We have defined a pair of statistics (Moran, Geary) to determine the level of clustering in a set of data.
- In the context of count data in spatial epidemiology these methods have some drawbacks, due to the non-constant variance of the response.
- We use the residuals to overcome this difficulty; the use of residuals from a model also allows the modeling of the mean function, so that the variable used (the residuals) are closer to stationary.
- I view these methods as useful in an exploratory first step in an analysis.
- The hierarchical model provides greater information but is based on many assumptions.

Cluster Methods for Count Data

Overview of Moving Window Methods

In this section we describe methods that superimpose a number of circular regions onto the study region and then determine the significance of the number of cases that fall within each circle – these methods assess cluster detection and may be used for surveillance.

Different methods define the circles in terms of:

- distance (Openshaw).
- the number of cases (Besag and Newell) and,
- the population size (scan statistics).

These methods may be used as screening devices by which particular regions may be highlighted and subsequently investigated.

We focus on scan statistics.

Scan statistics

- Scan statistics were originally developed to ‘scan’ across a time region of interest with the test statistic being the maximum number of events to occur within windows of constant size
- The fixed window and maximum number of the original formulation makes it clear that the statistic is being compared to an underlying intensity that is uniform.
- In a spatial context this is clearly unreasonable.

Scan statistics

- Turnbull *et al.* (1990) suggested an approach by which the 'windows' are defined to contain a constant population, N^*, and are centered on each area centroid.
- The maximum number of cases across the windows may then be used as a test statistic, i.e.

$$M = \max_j Y_j(N^*), \tag{6}$$

  where j indexes the areas as defined via the population N^*.
- As an alternative, Kulldorff and Nargarwalla (1995) suggested the use of the likelihood ratio test statistic.

Scan Statistics

- A Monte Carlo test may be performed under random distribution of cases across the study region.
- The approach therefore differs from those of Openshaw and Besag and Newell since the most significant circle over the whole study region is searched for instead of all circles significant at a certain level.
- Since only a single test is carried out it is straightforward to determine the correct statistical properties of the procedure.
- We describe for count data, for which numbers of cases and size of population are required along with the centroids of each area.
- If adjustment for covariates is required, then expected numbers should replace the population numbers.

Scan Statistics

- Potential clusters are defined as circles centered on the centroids of the areas (though grid lines can be given).
- The user is required to specify the maximum circle size – the default is 50% of the population.
- Responses are examined within circles that are centered on each centroid and ranging between zero and whatever the specified maximum is.
- Various probability models may be assumed, including Poisson and Bernoulli.

Scan Statistics

We concentrate on the Poisson model with adjustment for confounders within the expected numbers, for which for a given circle

$$\begin{aligned} Y_1 &\sim \text{Poisson}(E_1\theta_1) \\ Y_0 &\sim \text{Poisson}(E_0\theta_0) \end{aligned}$$

where

- Y_1 and Y_0 are the numbers of cases inside and outside the circle,
- E_1 and E_0 the respective expected numbers, and
- θ_1 and θ_0 the relative risks.

Scan statistics

- The approach is to evaluate a likelihood ratio statistic comparing the hypotheses

$$H_0 : \theta_1 = \theta_0, \quad H_A : \theta_1 > \theta_0$$

  for each circle c.
- The overall test statistic of the significance of the "most likely" statistic is then the maximum of these statistics, over $c = 1, \cdots, C$.
- For the Poisson model, the total number of cases $Y_+ = Y_0 + Y_1$ is conditioned upon, in which case

$$Y_1 | Y_+ \sim \text{Binomial}(Y_+, \pi)$$

  where

$$\pi = \frac{E_1\theta_1}{E_1\theta_1 + E_0\theta_0}.$$

Scan statistics

- Under the null, $\widehat{\pi}_0 = E_1/(E_1 + E_0)$ and under the alternative $\widehat{\pi}_A = Y_1/Y_+$.
- This gives the likelihood ratio statistic:

$$T = \frac{\Pr(Y_1|H_A)}{\Pr(Y_1|H_0)} = \left(\frac{Y_1}{E_1}\right)^{Y_1} \left(\frac{Y_0}{E_0}\right)^{Y_0} I(Y_1 > E_1)$$

- The significance level is assessed by carrying out a Monte Carlo procedure in which the pairs

$$(Y_i, E_i), \qquad i = 1, ..., n,$$

are randomly relabeled.

Scan statistics

- If the Poisson model is wrong then the procedure is not invalidated (since all the Poisson assumption is being used for is to define the test statistic) – but power will be reduced when compared to a statistic derived from the true distribution.
- Once the window with the greatest exceedence is identified, the sampling distribution of T is evaluated using a Monte Carlo test.
- The `SatScan` software, written by Martin Kulldorff, to implement the scan test statistic is available from

`http://srab.cancer.gov/satscan/`

Difficulties with Scan Statistics

- The choice of population size is somewhat arbitrary and there are no clear guidelines for a choice, Hjalmars *et al.* (1996) use 10% of the total population to define the windows while Kulldorff *et al.* (1997) use 50%.
- In practice the method is not just used to indicate a single cluster but a number of potential clusters are highlighted.
  Once this is done the properties of the procedure become unknown.
- The circles are also not completely comparable since it is populations and not expected numbers that are defining the choice of radii (although it is straightforward to use expected numbers).
- For this and all methods the choice of a p-value threshold is difficult.
- More subtly p-value thresholds should be a function of sample size, and so there should be different thresholds for different window sizes.

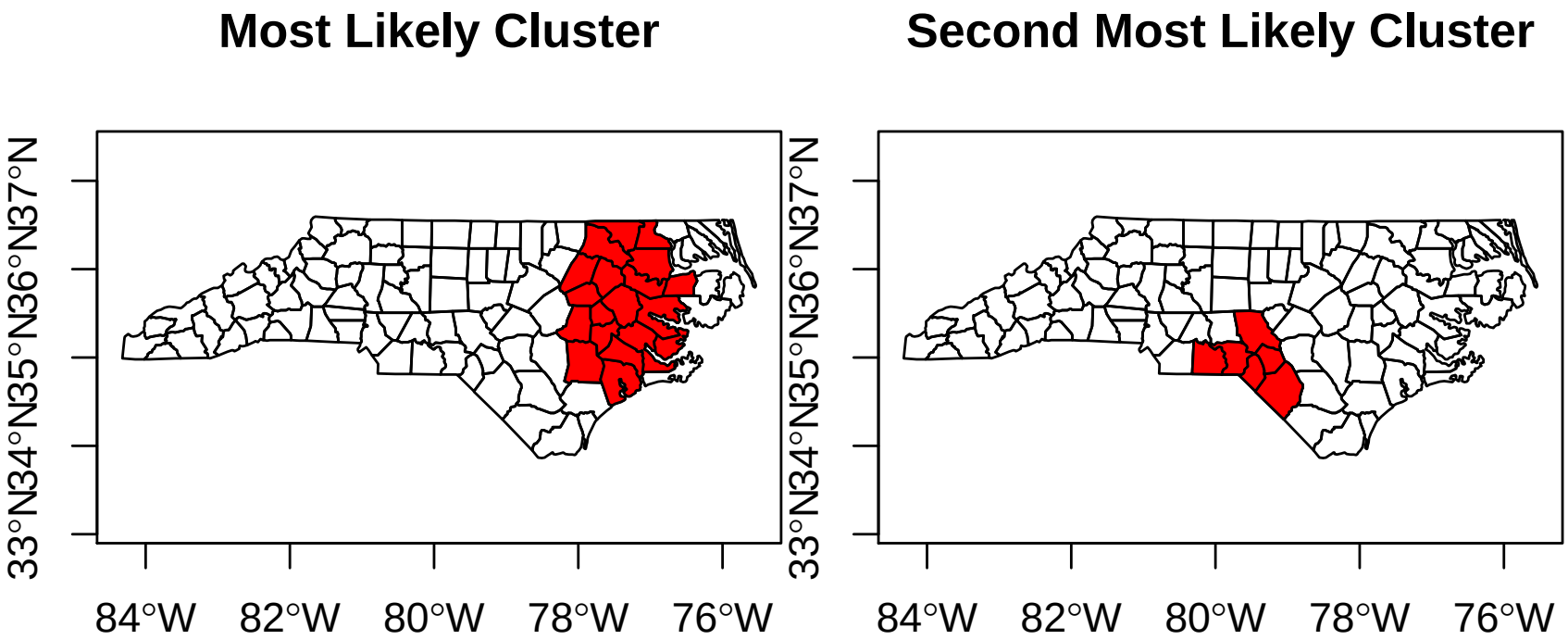

Figure 8: Significant clusters under SatScan for the NC SIDS data. Both had p-value < 0.01.

Hierarchical modeling

- Disease mapping, hierarchical models, are not designed for cluster detection.
- If there are isolated areas of high risk (e.g., a single area only), then unless the expected number is high (in which case the SMR will be a good summary anyway) then shrinkage will occur, and the excess will be missed.
- This makes sense intuitively, and has been borne out in extensive simulation studies (Richardson *et al.*, 2004).

IID model: relative risk threshold of 1.5

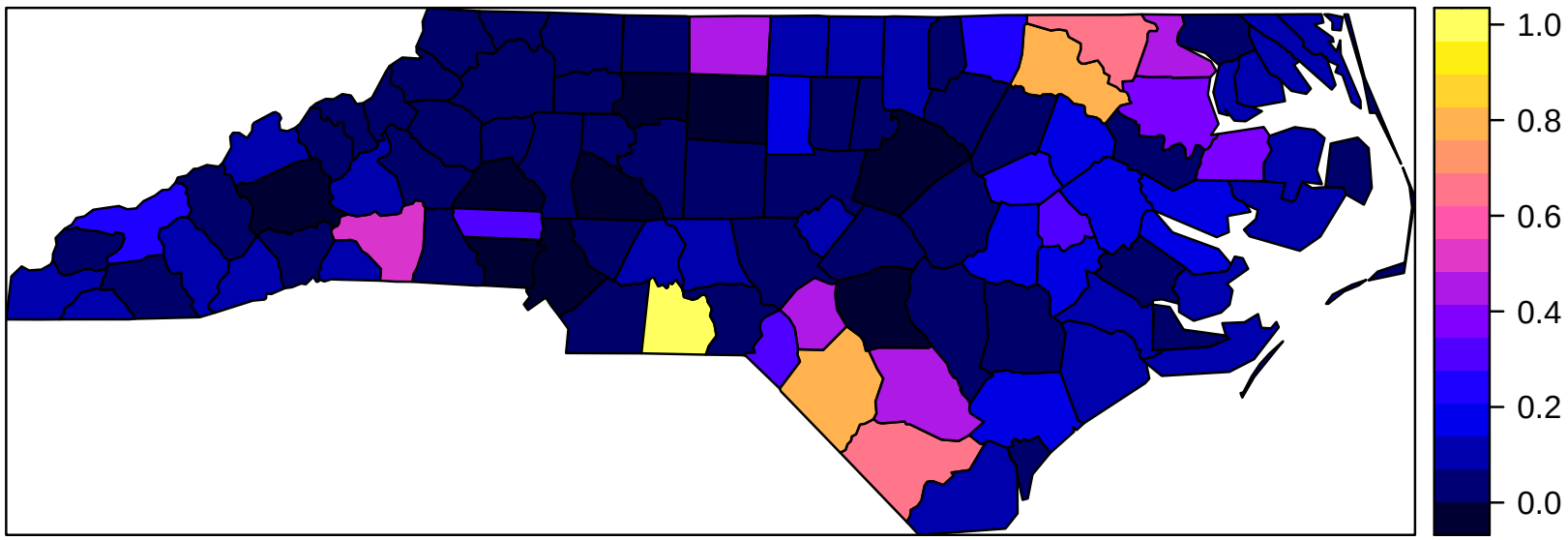

Figure 9: Posterior probability $\Pr(\text{RR} > 1.5|\boldsymbol{y})$.

IID model: relative risk threshold of 2.5

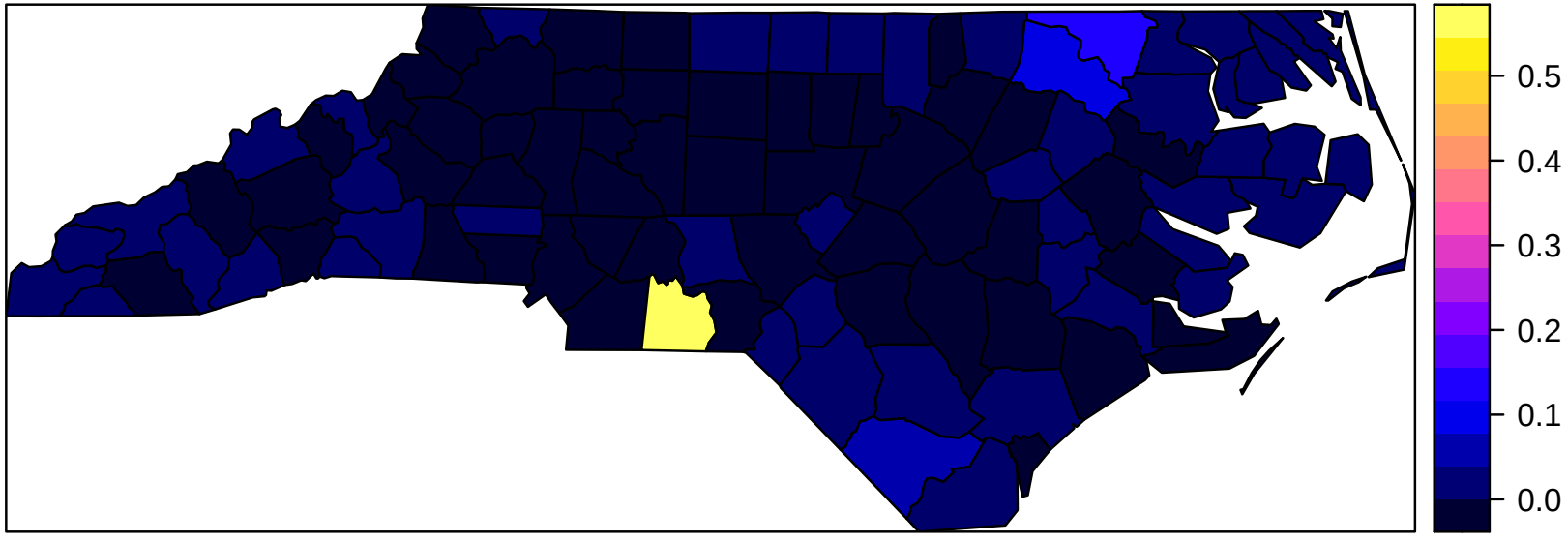

Figure 10: Posterior probability $\Pr(\text{RR} > 2.5|\boldsymbol{y})$ (note the change of scale from Figure 9).

BYM2 model: relative risk threshold of 1.5

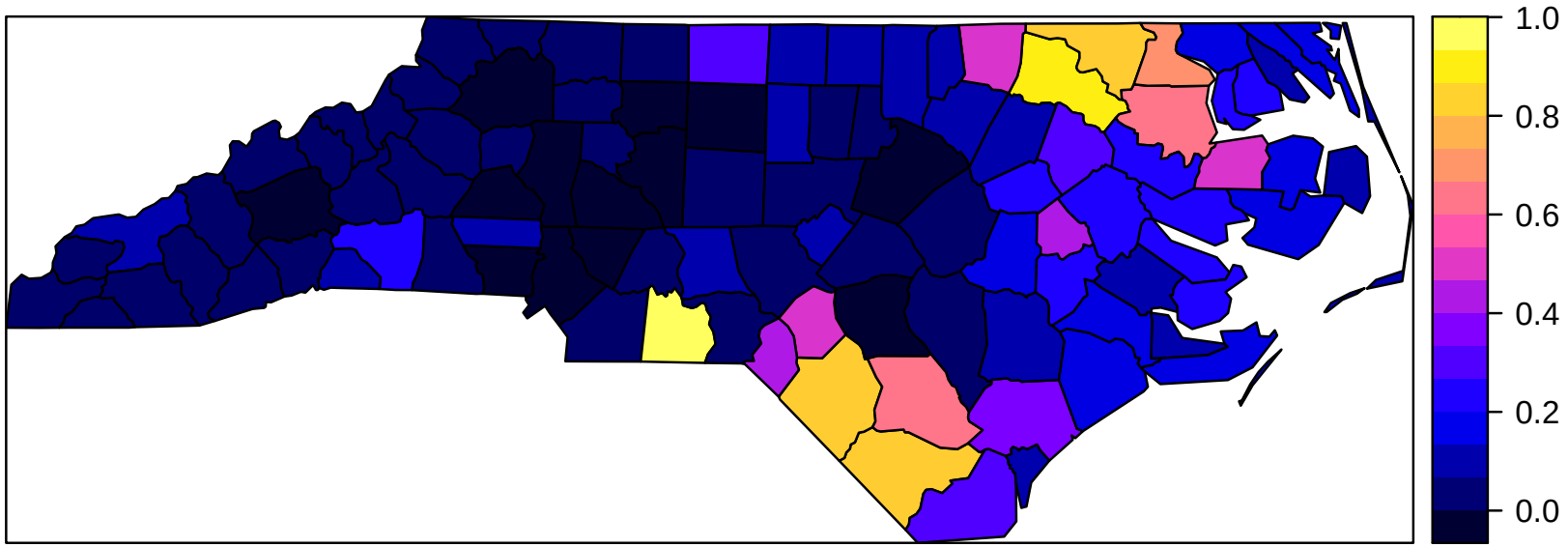

Figure 11: Posterior probability $\Pr(\text{RR} > 1.5|\boldsymbol{y})$.

BYM2 model: relative risk threshold of 2.5

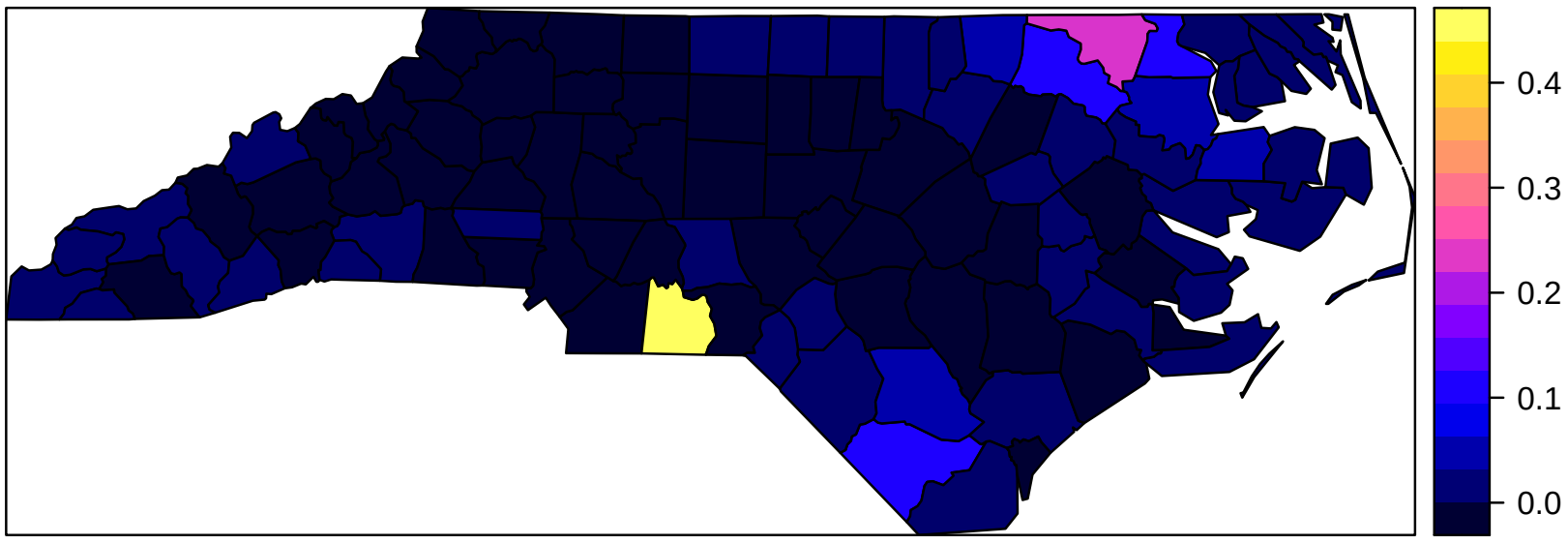

Figure 12: Posterior probability $\Pr(\text{RR} > 2.5|\boldsymbol{y})$.

Hierarchical Modeling Analysis

- Both the IID and BYM2 analyses suggest that there are clusters in the centre south (the strongest signal), and also in the north-east and south-east.
- These conclusions are consistent with those from the SatScan approach.
- The hierarchical modeling approach is basically estimation, versus the hypothesis testing approach of SatScan — my personal preference is always the former.

Cluster Detection for Count Data Conclusions

Of the frequentist moving window methods the Kulldorff procedure has the best statistical foundation but it has drawbacks.

How to deal with the possibility of multiple clusters?

- Original version simply compared the *p*-values of the second, third,... most significant zone (discarding those with overlap).
- Recent version Zhang *et al.* (2010) removes a significant zone, and then repeats...until no more significant zones found.

How to choose a significance level?

The power may be very different in different studies: no balancing of Type I and Type II errors if significance level α fixed in all studies.

Cluster Detection for Count Data Conclusions

Can also use hierarchical models for detecting clusters but be wary of shrinkage which could remove true clusters.

In terms of cluster detection:

- we can threshold the fitted surface and examine those areas that are highlighted (and the cases in these areas).
- For example, we could only plot those areas in which the odds of disease is greater than some critical value with a certain posterior probability.

Bayesian cluster method (Wakefield and Kim, 2013; Kim and Wakefield, 2016) has a probabilistic foundation, but many prior inputs required.

References

Bivand, R., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, 2nd Edition*. Springer, New York.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley, New York.

Elliott, P., Wakefield, J. C., Best, N. G., and Briggs, D. J. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford University Press, Oxford.

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5**, 115–145.

Hjalmars, U., Kulldorff, M., Gustafsson, G., and Nagawalla, N. (1996). Childhood leukaemia in sweden: using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine*, **15**, 707–715.

Kim, A. and Wakefield, J. (2016). A Bayesian method for cluster detection with application to five cancer sites in Puget Sound. *Epidemiology*, **27**, 347.

Knox, G. (1989). Detection of clusters. In P. Elliott, editor, *Methodology of Enquiries into Disease Clusters*, pages 17–22, London. Small Area Health Statistics Unit.

Kulldorff, M. and Nargarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, **14**, 799–810.

Kulldorff, M., Feuer, E., Miller, B., and Freedman, L. (1997). Breast cancer clusters in the northeast united states: a geographic analysis. *American Journal of Epidemiology*, **146**, 161–170.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall, London.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Richardson, S., Thomson, A., Best, N., and Elliott, P. (2004). Mini-monograph: Interpreting posterior relative risk estimates in disease mapping studies. *Environmental Health Perspectives*, **112**, 1016–1025.

Riebler, A., Sørbye, S., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, **25**, 1145–1165.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysi*. CRC press.

Turnbull, B., Iwano, E., Burnett, W., Howe, H., and Clark, L. (1990). Monitoring for clusters of disease: application to leukaemia incidence in upstate New York. *American Journal of Epidemiology*, **132**, S136–S143.

Wakefield, J. and Kim, A. (2013). A Bayesian model for cluster detection. *Biostatistics*, **14**, 752–765.

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons.

Zhang, Z., Assuncovcão, R., and Kulldorff, M. (2010). Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, page Article ID 642379.