2016 SISMID Module 16
Lecture 6: Clustering and Cluster Detection for Count Data

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Outline

Clustering Methods for Aggregated Count Data

Cluster Detection for Count Data: Moving Window Methods

Cluster Detection for Count Data: A Bayesian Moving Window Model

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 descriptor.
- 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.

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 $i = 1, \dots, n$.

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, $i, j = 1, \dots, n$, $i \neq j$.

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:

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

where $\theta_i = \exp(\beta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_1)$ 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 $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_1]$.

It is not straightforward to obtain the standard error of $\widehat{\kappa}$ — its difference from 1 can be assessed via simulation (though usually it's obvious!).

An alternative measure of lack of fit is the residual deviance.

Methods for Detecting Heterogeneity: Pearson's χ^2

High values of $\widehat{\kappa}$ will result if there is overdispersion.

If the Poisson model is adequate then both the deviance (likelihood ratio statistic) and Pearson's chi-square statistic have an asymptotic chi-square statistic on $n - p$ degrees of freedom, under certain assumptions.

Specifically, we need the number of "x-values" (predictors) in the model to remain fixed as the data grow (hypothetically) larger; this occurs when the predictors are factors with a fixed number of levels.

Alternatively (and preferably) significance may be assessed via calculation of a Monte Carlo p-value in which observations are randomly simulated under the null hypothesis and the test statistic is calculated under each simulation.

These are measures of unmodeled heterogeneity and say nothing about spatial dependence.

The residuals may be examined for clues to excesses at particular locations.

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:

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

where

- c is a constant,
- 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.

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.

As usual, 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, 1948)

Moran's I statistic (Moran, 1948) is given by

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

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, 1954)

Geary's *c* statistic is closely related to Moran's statistic and is given by

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

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.

Note the similarity of the numerator to the semi-variogram:

$$\frac{1}{2}\text{var}(Z_i - Z_j).$$

Non-parametric Statistic

Letting Z_i^* denote the ranks of the Z_i we may calculate a non-parametric measure of spatial dependence

$$D = \frac{\sum_i \sum_j w_{ij} |Z_i^* - Z_j^*|}{\sum_i \sum_j w_{ij}}, \tag{6}$$

with small values of D implying positive dependence.

Suggested by Cliff and Ord (1981, p. 46).

Issues with Assessment of Clustering

There are complications when using these measures for SMRs unless they are based on equal expected numbers.

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.

Disease Mapping Methods

Recall the Poisson log-linear model we used for disease mapping and spatial regression:

$$\begin{aligned} Y_i \mid \theta_i &\sim_{iid} \text{Poisson}(E_i \theta_i) \\ \log \theta_i &= \beta_0 + \beta_1 x_i + \epsilon_i + S_i \end{aligned}$$

where $\epsilon_i \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$ and S_i are spatial random effects with ICAR structure.

We can examine (for example) the fitted surface or probabilities above a threshold to find areas of high risk, i.e. clusters (but recall the smoothing aspect which may effectively remove clusters).

To determine clustering we may examine the magnitude of the variance of the spatial random effects.

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, for example.
- 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) has constant mean.
- 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.

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.

Openshaw's method

Openshaw *et al.* (1987) proposed a 'Geographical Analysis Machine' method in which a regular grid is superimposed on the study region and circles of constant radius are drawn on the intersections of the grid lines.

Typically a range of radii are constructed based on the scale at which clustering is expected and the grid-lines are such that adjacent circles overlap by 80%.

A common geography for populations and cases is established (census Enumeration Districts are recommended for the UK) and then 'jagged' circles are formed containing the area centroids.

Circles are 'flagged' if they attain a certain level of significance, under the assumption that cases within circles follow a Poisson distribution.

The p-value is small to address the multiple testing problem but there is no theoretical development of a particular size; $p = 0.005$ has often been used in applications.

Difficulties with Openshaw Method

Openshaw's method has been heavily criticized in the literature since there is clearly a huge multiple testing problem:

- There are a large number of tests, and
- The tests are dependent.

Obviously the size of p is crucial to the sensitivity (probability of flagging given a true cluster) and specificity (probability of non-flagging given a true non-cluster) of the method and the lack of guidelines for this choice remains a major drawback.

Since the different circles contain different numbers of cases and different populations at risk the power to detect clusters will vary across circles which makes interpretation difficult.

If overdispersion is present then the Poisson distribution is not appropriate (extension to negative binomial straightforward).

Openshaw's method is not recommended but of historical interest since it started the ball rolling...

Besag and Newell's Method

The method of Besag and Newell (1991) was developed to rectify some of the problems of Openshaw's method.

The first step in applying the method is to select a cluster size k. For each case in turn a circle is drawn, centered on that case, with radius such that the k-th nearest neighboring case is included.

As with Openshaw's method the expected number of cases is calculated for each circle.

Unlike Openshaw's method the circles are now more comparable since they are all based on k cases.

By defining the cluster in terms of the number of cases the method has a greater chance of detecting small rural clusters than the distance-based method of Openshaw.

Besag and Newell's Method

The expected number of clusters under the null may be calculated, and compared with the actual number found, as an aid to deciding whether any of the highlighted regions should be investigated further.

Specifically, consider:

- a generic circle containing k cases in addition to the case upon which the circle is centered,
- let Y represent the number of cases in this circle, and
- E the expected number of cases in the areas within which the k cases are found.

Besag and Newell's Method

Then

$$\Pr(Y \geq k | H_0) = 1 - \Pr(Y < k) = 1 - \sum_{s=0}^{k-1} \frac{\mathrm{e}^{-E} E^s}{s!}$$

where H_0 is the null that the cases are randomly distributed amongst the population at risk.

The choice of k is clearly vital and several values are typically selected. The more values that are chosen, the more difficulty in interpretation.

Still a multiple testing problem.

If overdispersion is present then the Poisson distribution is not appropriate (but relatively straightforward to extend to a negative binomial).

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.

$$M = \max_j Y_j(N^*), \tag{7}$$

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 is then 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.

However, in practice the statistic is repeated using various values of the population size upon which circle construction is based, thus producing a set of non-independent tests.

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.

Then circles are examined centered on each centroid and ranging between zero, to 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 $c = 1, ..., C$.

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 (in common with the methods of Openshaw and Besag and Newell).

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.

A Bayesian Model

We describe the method of Wakefield and Kim (2013); Kim and Wakefield (2015), which is available in the `SpatialEpi` packages.

Built on previous work of Gangnon and Clayton (2000, 2003); Gangnon (2006); Gangnon and Clayton (2007).

Partition the study region so each area is either within a cluster/anti-cluster or is null.

Single clusters defined as in `SatScan`: we call these zones.

Multiple clusters are formed as combinations of single clusters.

A Bayesian Model

The number of clusters/anti-clusters can be $j = 0, ..., K$, with K fixed: with each consisting of a single zone.

No overlap allowed, and there has to be a buffer between any two zones.

We place priors on numbers of clusters (usually strongly encouraging zero or few clusters).

Relative risks associated with null areas arise from a "narrow" gamma, centered at 1.

Relative risks associated with cluster areas arise from a "wide" gamma, centered at 1.

Details of Bayesian Model

We define a configuration as a legal collection of single zones, and for $j = 0, ..., K$ suppose there are N_j configurations of such zones. For $j = 0$ (no clusters/anti-clusters) we set $N_0 = 1$ for notational consistency.

We label the null configuration as c_{01} and c_{jl} as the l-th configuration of j single zones, for $j = 1, ..., K$, and $l = 1, ..., N_j$.

c_{jl} denotes a collection of indices of single zones:

- $c_{01} = \phi$,
- $c_{1l} = l$, $l = 1, ..., N_1$,
- $c_{2l} = \{\ l(1), l(2)\ \}$ for the pair of single zones that correspond to configuration l. These labels range over all pairs that are "legal", i.e. non-overlapping with a buffer between. There are N_2 such pairs.
- $c_{3l} = \{\ l(1), l(2), l(3)\ \}$, etc...
- The indices $l(k) \in \{1, 2, ..., N_1\}$, i.e. are from the collection of single zone labels.

A Bayesian Model

Basic model is again:

$$Y_i|\theta_i \sim \text{Poisson}(E_i\theta_i)$$

where θ_i is the relative risk associated with area i.

The prior assigned the θ_i depends on whether **area** i is null, or lies within a cluster/anti-cluster.

If area i is null, assume a narrow gamma prior $\theta_i \sim \text{Ga}(a_{\text{N}}, b_{\text{N}})$ with $a_{\text{N}}, b_{\text{N}}$ fixed. This allows some "wobble" about 1.

Consequently, $\Pr_{\text{N}}(y_i)$ is $\text{NegBin}(a_{\text{N}}, b_{\text{N}})$.

Under the null configuration:

$$\Pr(\mathbf{y}|c_{01}) \quad = \quad \prod_{i=1}^{n} \Pr_{\text{N}}(y_i).$$

A Bayesian Model

For those areas in zone z, $\theta_i = \theta^\star \sim \mathrm{Ga}(a_{\mathrm{W}}, b_{\mathrm{W}})$, a wide prior with $a_{\mathrm{W}}, b_{\mathrm{W}}$ fixed.

Marginally, y_i is $\mathrm{NegBin}(a_{\mathrm{W}}, b_{\mathrm{W}})$.

Non-null configuration c_{jl}, contains j single zones each with vectors of counts and expected numbers $\mathbf{y}_z^{\mathrm{Z}}$, $\mathbf{E}_z^{\mathrm{Z}}$, and summed counts and expected numbers, $y_z^{\mathrm{Z}}, E_z^{\mathrm{Z}}$, $z = 1, ..., j$.

The associated likelihood for $\mathbf{y}$ is composed of two parts corresponding to null and non-null areas:

$$\Pr(\mathbf{y}|c_{jl}) \quad = \quad \prod_{\text{null areas}} \Pr_{\mathrm{N}}(y_i) \times \prod_{z \in c_{jl}} \left\{ \Pr_{\mathrm{W}}(y_z^{\mathrm{Z}}) \times \Pr(\mathbf{y}_z^{\mathrm{Z}} | y_z^{\mathrm{Z}}) \right\}.$$

$\Pr(\mathbf{y}_z^{\mathrm{Z}} | y_z^{\mathrm{Z}})$ is a multinomial distribution with dimension the number of areas in zone z, $|\mathbf{y}_z^{\mathrm{Z}}|$, total y_z^{Z} and vector of probabilities $\mathbf{E}_z^{\mathrm{Z}} / E_z^{\mathrm{Z}}$.

A Bayesian Model

It is useful to recognize that

$$\frac{\Pr(\mathbf{y}|c_{jl})}{\Pr(\mathbf{y}|c_{01})} = \prod_{z \in c_{jl}} \text{BF}(z) \tag{8}$$

where

$$\text{BF}(z) = \frac{\Pr_{\text{W}}(y_z^{\text{Z}}) \times \Pr(\mathbf{y}_z^{\text{Z}}|y_z^{\text{Z}})}{\prod_{\text{null areas}} \Pr_{\text{N}}(y_i)}$$

This is the Bayes factor comparing the distribution of the data under configuration c_{jl} to that under the null model.

Hence, for a configuration with j zones (8) is the product of j Bayes factors.

One consequence of this expression is that computation is vastly simplified since we only need to consider calculations for single zones.

Bayes Factors and Likelihood Ratios

In the case of a single zone, the Bayes factor

$$\text{BF}(z) = \frac{\prod_{z \in c_{jl}} \Pr_{\text{W}}(y_z^{\text{Z}}) \times \Pr(\mathbf{y}_z^{\text{Z}} | y_z^{\text{Z}})}{\prod_{\text{areas in zone } z} \Pr_{\text{N}}(y_i)}$$

may be compared with the LR statistic of `SatScan`:

$$\begin{aligned} \text{LR}(z) &= \frac{\Pr(\mathbf{y}| \text{ alternative })}{\Pr(\mathbf{y}| \text{ null })} = \frac{\prod_{\text{areas not in zone } z} \Pr(y_i|\theta_i = 1) \times \Pr(y_z^{\text{Z}}|\widehat{\theta}_z)}{\prod_{i=1}^n \Pr(y_i|\theta_i = 1)} \\ &= \frac{\Pr(y_z^{\text{Z}}|\widehat{\theta}_z)}{\prod_{\text{areas in zone } z} \Pr(y_i|\theta_i = 1)} \end{aligned}$$

- In the denominator: conditions on $\theta = 1$, Bayes integrates over the narrow prior.
- In the numerator, maximizes over θ_z (subject to $\theta_z > 1$), Bayes integrates over the wide prior.
- The multiple zone version is based on sequential LR statistics, Bayes based on products of Bayes factors.

Prior Distribution on Single Zones

Priors for all configurations are constructed from single zones.

For single zone z there are various possible priors, but a simple on is uniform on the N_1 possibilities.

We typically place a mass close to 1 on zero clusters.

Priors on multiple (legal) zones are proportional to the product of the individual single zone prior probabilities.

What to Report?

The obvious quantity is Pr(configuration $|\mathbf{y}) = \Pr(c_{jl}|\mathbf{y})$, but these will typically be small: if a true cluster, lots of overlapping zones.

A useful summary is

$$\Pr(\text{ number of clusters } = j|\mathbf{y}), \qquad j = 0, \dots, K.$$

Maps of Pr(area is "high" $|\mathbf{y}$).

Influence of prior may be removed via looking at Bayes factors comparing posterior to prior odds of an area being high.

Operating characteristics may be examined via simulation.

Need to trade-off sensitivity and specificity.

Prior Choices: Narrow range of RRs: $(0.848, 1.169)$. Wide range of RRs: $(0.037, 5.323)$.

Computation via MCMC, sampling over the possible configurations.

Computation

Single parameter is c_{jl}: sample using MCMC. Given c_{jl} we propose a new configuration $c^{\star}_{jl}$ via one of five moves:

1. **Growth**: The index set c_{jl}, is increased by aggregating nearest free neighboring areas to the single zone's centering area.
2. **Trim**: The index set c_{jl}, is reduced by dropping areas that are furthest from the centering area. Trim moves are reciprocal to growth moves.
3. **Replacement**: Replace an element c_{jl} with another single zone with a different centering area.
4. **Death**: drop one of the j single zones to form a configuration of $j-1$ single zones.
5. **Birth**: add a new single zone to c_{jl} to form a a new configuration of $j+1$ single zones. Birth moves are reciprocal to death moves.

   Configurations $c^{\star}_{jl}$ are proposed randomly via one of two mechanisms:

   - Uniformly from the N_1 single zones, i,e. N_1^{-1}, or
   - Proportional to the posterior probabilities $\Pr(c^{\star}_{jl}|\mathbf{y})$, i.e. $\propto \prod_{k=1}^{j} \mathrm{BF}(k)$.

Computation

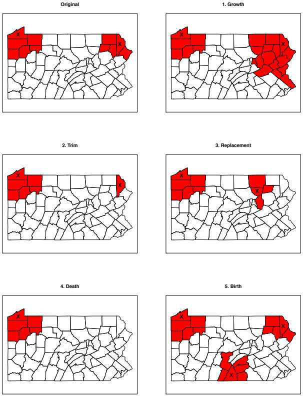

Figure 1 : MCMC Move types.

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 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 (Richardson *et al.*, 2004).

In terms of clustering:

- We can examine the random effects S_i and examine maps of these (to compare with maps of ϵ_i).
- We can also examine the empirical variance of the S_i's and compare to σ_ϵ^2.

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 has a probabilistic foundation, but many prior inputs required.

Kulldorff (2001) proposes a space-time version of `SatScan`, while Li *et al.* (2012) describe a Bayesian model for space-time data.

References

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, **154**, 143–55.

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

Cliff, A. and Ord, J. (1981). *Spatial processes: Models and applications*. Pion, London.

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

Gangnon, R. (2006). Impact of prior choice on local Bayes factors for cluster detection. *Statistics in Medicine*, **25**, 883–895.

Gangnon, R. and Clayton, M. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics*, **56**, 922–935.

Gangnon, R. and Clayton, M. (2003). A hierarchical model for spatially clustered disease rates. *Statistics in Medicine*, **22**, 3213–3228.

Gangnon, R. and Clayton, M. (2007). Cluster detection using bayes factors from overparameterized cluster models. *Environmental and Ecological Statistics*, **14**, 69–82.

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. (2015). A Bayesian method for cluster detection with application to five cancer sites in Puget Sound. *Epidemiology*. To Appear.

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. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A*, **164**, 61–72.

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.

Li, G., Best, N., Hansell, A., Ahmed, I., and Richardson, S. (2012). Baystdetect: detecting unusual temporal patterns in small area data via bayesian model choice. *Biostatistics*, **13**, 695–710.

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

Moran, P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, **10**, 243–251.

Openshaw, S., Charlton, M. Wymer, C., and Craft, A. (1987). A mark i geographical analysis machine for the automated analysis of point data sets. *International Journal of GIS*, **1**, 335–358.

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.

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.