J Geograph Syst (2005) 7: 189–206
DOI: 10.1007/s10109-005-0156-5

ORIGINAL PAPER

V. Gómez-Rubio · J. Ferrándiz-Ferragud · A. López-Quílez

Detecting clusters of disease with R

Received: 23 July 2003 / Accepted: 28 June 2004

Abstract One of the main concerns of Public Health surveillance is the detection of clusters of disease, i. e., the presence of high incidence rates around a particular location, which usually means a higher risk of suffering from the disease under study (Aylin et al. 1999). Many methods have been proposed for cluster detection, ranging from visual inspection of disease maps to full Bayesian models analysed using MCMC. In this paper we describe the use and implementation, as a package for the R programming language, of several methods which have been widely used in the literature, such as Openshaw's GAM, Stone's test and others. Although some of the statistics involved in these methods have an asymptotical distribution, bootstrap will be used to estimate their actual sampling distributions.

Keywords Spatial statistics · Epidemiology · Disease cluster detection · R programming language

JEL Classification C60 · C88

We would like to thank co-editor Dr. Manfred M. Fischer and four anonymous referees for their suggestions and comments to improve this paper. The help of Dr. Roger Bivand has also been of great value. Furthermore, this work has been partly funded by Consellería de Sanitat and EUROHEIS Project (code SI2.329122, 2001CVG2-604). The authors wish to express their regard and gratitude to Prof. Juan Ferrándiz-Ferragud who died during the revision of this paper. Juan was the main researcher of the Spanish EUROHEIS group, and was really a master for all the people involved in the project.

V. Gómez-Rubio (✉) · J. Ferrándiz-Ferragud · A. López-Quílez
Departament d'Estadística i Investigació Operativa,
Facultat de Matemàtiques, C/ Dr. Moliner 50,
46100 Burjassot, València, Spain
E-mail: virgil@uv.es

1 Introduction

Clusters of disease can be defined in several ways, but probably the simplest one is to say that a cluster is a set of neighbouring areas where far more cases than expected appear during a concrete period of time. For this reason, Public Health Authorities have always been concerned with the investigation of these kind of clusters.

The beginning of Spatial Epidemiology is without doubt Snow's study (Snow 1854) of an outbreak of cholera in London, the focus of which he found by plotting the location of those people affected on a city map, noticing that most cases were concentrated around a water pump. After this study, many methods have been developed to detect spatial clusters of disease.

Most methods described in this paper have repeatedly appeared in the bibliography and have been widely used in real studies. Furthermore, we describe the use of a new package for the R statistical programming language (Ihaka and Gentleman 1996) called DCluster which implements routines to use all these methods. Although there are several packages for clustering available in R, they provide general methods and none of them is devoted to spatial clusters of disease.

DCluster can be downloaded from CRAN, at http://cran.r-project.org, or from the first author's web site at http://matheron.uv.es/~virgil/Rpackages/, where extra documentation is also available.

The paper is structured as follows. Section 2 introduces the general structure of the data available for the problem of cluster detection followed by Sect. 3, the most popular and used statistical models. Next, we will briefly describe methods implemented in DCluster (Sect. 4), bootstrap procedures (Sect. 5) and a brief outline of package DCluster (Sect. 6). We explore the use of these methods using real data in Sect. 7. Finally, the conclusions of this paper are available in Sect. 8.

2 Data structure

Let us suppose that our study area is divided in to n non-overlapping regions (which may be, for e.g., counties, provinces or municipalities). Data are usually available as counts, i.e., number of deaths or affected people in each region.

O_i will represent the observed number of cases in region i, E_i its expected number, which can be calculated in several ways, and P_i the population at risk in region i. By O_+, E_+ and P_+ we will represent the sums of observed cases, expected cases and population over all regions in the study area.

Usually, population is stratified according to age and sex and, sometimes, a measure of deprivation or poverty. This stratification is useful in order to control for the effect of these variables which are known to be important to the analysis. Other covariates can be incorporated into standardisation in a similar way (Ferrándiz et al. 2004).

P_{ij} will represent population at risk at stratum j in region i. It is clear that $P_i = \sum_j P_{ij}$. O_{ij} and E_{ij} can be defined in a similar way.

E_{ij} is often calculated using indirect standardisation (Jenicek and Cléroux 1982). Briefly, if we have a reference population from which we know their incidence rates($r_{ij} = O'_{ij}/P'_{ij}$) at each stratum, then $E_{ij} = P_{ij}\, r_j$.

When the reference population is the same as the population under study, standardisation is called *internal* and it holds that $O_+ = E_+$ (Morris and Wakefield 2000).

Finally, regions will be located by their centroids, which mark the centre of the total area. These centroids are usually not taken as the geometrical centre, but are weighted according to the actual distribution of the population within the region.

3 Statistical models for diseases

As a first approximation, we will consider the O_i's to be independent and drawn from a Poisson distribution whose mean is $\theta_i E_i$ (Wakefield et al. 2000a), where θ_i is the relative risk, which measures the local deviation of the disease. If the relative risk is much higher than 1 it is likely that a risk excess exists in the region.

The maximum likelihood estimator for θ_i, called the *standardised mortality ratio* (SMR), is $\widehat{\theta}_i = O_i/E_i$. This estimation can be used to create thematic maps to show the spatial risk of the disease.

Unfortunately, the variance of this estimator is proportional to $1/E_i$, so values arising from rare diseases or areas with small populations (where the number of expected cases is really low) may lead to poor estimates.

Conditioning on O_+ leads to a Multinomial model (Whittermore et al. 1987), in which the size is O_+ and probabilities are given by expression (E_1/E_+,..., E_n/E_+). This model is often used when performing Monte Carlo simulations to estimate distributions of different statistics (Best et al. 2001).

Notice that this model is equivalent to randomly distributing the total of observed cases among all the regions proportionally to E_i.

Sometimes the Poisson model is too strict in the sense that it imposes mean and variance to be equal. When data exhibits some kind of overdispersion, the Poisson distribution is unlikely to be the best choice (Dean 1992).

Clayton and Kaldor (1987) propose the use of a hierarchical Bayesian model in which relative risks θ_i are drawn from a Gamma distribution with two fixed hyperparameters. Conditioned on θ_i, observed counts O_i are independent realizations of a Poisson distribution whose mean is $\theta_i E_i$:

$$\begin{aligned} O_i|\theta_i &\sim Po(\theta_i E_i) \\ \theta_i &\sim Ga(v, \alpha) \end{aligned} \tag{1}$$

As a consequence, O_i is distributed following a Negative Binomial with size v and probability $\alpha/(\alpha + E_i)$. v and α are usually estimated via Empirical Bayes using equations proposed by Clayton and Kaldor (1987).

The M.L.E. for θ_i is now $(O_i + v)/(E_i + \alpha)$, which provides a smoothed estimator of the relatives risks. These estimators are frequently used when performing disease mapping.

4 Implemented procedures

Methods implemented in package DCluster can be classified as general and focused, as discussed by several authors, such as Besag and Newell (1991) and Tango (1995). This distinction is made depending on whether the method is used to locate clusters in the study area or to assess the presence of a cluster around a given region.

Furthermore, we have considered other groups of statistics that provide a global measurement of clustering, homogeneity among relative risks or spatial autocorrelation.

Section 5 describes how tests described below can be carried out by means of bootstrap following a unified, general and straightforward procedure.

4.1 Tests for homogeneity

These methods can be used as a first approach to the problem and to investigate if relative risks are homogeneous (i.e., equal) in the study area. Different relative risks may lead to zones where they tend to be higher (or lower) than expected and, hence, a cluster may appear.

4.1.1 Pearson's chi-square statistic

This statistic is used for testing goodness of fit to a given distribution. Basically, it compares observed and expected data in the following way:

$$T = \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{E_i}. \tag{2}$$

Test hypotheses are as follows:

$$\begin{aligned} H_0 &: \quad \theta_1 = \cdots = \theta_n = \lambda \\ H_1 &: \quad \text{Not } H_0. \end{aligned}$$

In the case where λ is unknown, E_i must be substituted by $E_i\, O_+ / E_+$ in expression (2) and statistic T is asymptotically distributed as a Chi-square with $n-1$ degrees of freedom (Potthoff and Whittinghill 1966a, b).

Usually, λ is supposed to be 1. In this case, no modification to E_i is needed and the degrees of freedom are n.

When internal standardisation is used the case is slightly different. Since $O_+ = E_+$, λ must be 1 and the degrees of freedom are $n-1$.

Note that this statistic is also sensitive to low observed cases and that non-homogeneity may not only be related to high relative risks but also to low ones.

4.1.2 Potthoff-Whittinghill's test

Potthoff and Whittinghill (1966b) assume that data come from a Multinomial distribution and consider the locally most powerful test for the following hypotheses:

$$H_0: \quad \theta_1 = \cdots = \theta_n = \lambda$$
$$H_1: \quad \theta_i \sim Ga(\lambda^2/\sigma^2, \lambda/\sigma^2).$$

It should be noted that the alternative hypotheses means that relative risks are drawn from a Gamma distribution with mean λ and variance σ^2.

The statistic involved in the test is:

$$PW = E_+ \sum \frac{O_i(O_i - 1)}{E_i}, \tag{3}$$

which asymptotically is normally distributed, with mean $O_+(O_+ - 1)$ and variance $2(n-1)O_+(O_+ - 1)$.

This is a general test for homogeneity and λ is supposed to be unknown. Note that if internal standardisation was carried out, then the hypotheses of homogeneity imply λ to be equal to 1.

4.2 Spatial autocorrelation

Statistics presented in this section measure spatial autocorrelation (Cliff and Ord 1981) of the data. Usually the quantities involved are SMRs or residuals from a previously fitted model, such as those described in Section 3 or, for example, any other suitable generalised linear model (McCullagh and Nelder 1989). By working with residuals, we search for correlation among what remains unexplained by our primary model. When using SMRs we expect to find regions where they tend to be higher (or lower).

As stated in Cliff and Ord (1981), when using the two statistics described below, the null hypothesis for the theoretical tests is different for residuals and SMRs. However, bootstrap overcomes this problem since it provides a unified way of assessing significance by means of a Monte Carlo test.

Spatial Autocorrelation may be a source of overdispersion in the data (Cressie 1993). For this reason, testing for autocorrelation also is important to decide which model better represents our data, as discussed in Sect. 5.

4.2.1 Moran's I statistic

Moran (1948) proposes a statistic, called the *I statistic*, which is very close to the correlation coefficient between two variables:

$$I = \frac{n \sum_i \sum_j W_{ij}(Z_i - \overline{Z})(Z_j - \overline{Z})}{2(\sum_i \sum_j W_{ij}) \sum_k (Z_k - \overline{Z})^2}. \tag{4}$$

As mentioned before, Z_i may be either residuals ($O_i - E_i$) or SMRs (O_i/E_i). W is a matrix which measures proximity between regions and it can be defined in different ways. For example, W_{ij} can be 1 if regions i and j have a common boundary (and 0 otherwise) or the inverse of the distance between their centroids.

It has been shown that this statistic is quite robust to changes in sampling distribution (Zoellner and Schmidtmann 1999). Several authors point out that Moran's I performs badly when heterogeneous populations are involved (Walter 1993) and have proposed modifications to this statistic (Oden 1995; Waldhör 1996; Assunçao and Reis 1999).

4.2.2 Geary's c statistic

Geary's c statistic (Geary 1954) is defined in a similar way to Moran's I:

$$c = \frac{(n-1)\sum_i \sum_j W_{ij}(Z_i - Z_j)^2}{2(\sum_i \sum_j W_{ij}) \sum_k (Z_k - \overline{Z})^2}. \tag{5}$$

Note that now the differences between two values are computed instead of their deviation from the mean. W is, again, a matrix that measures proximity between regions.

4.3 General clustering

These methods provide a general measurement of clustering in the whole area. For this reason, they are not suitable for detecting localised clusters. In addition, these methods may fail to detect global clustering when actual clusters are small or scattered in the study area.

4.3.1 Whittermore's statistic

The statistic proposed by Whittermore et al. (1987) is based on the distance between all pairs of cases and is defined as:

$$W = \frac{n-1}{n} r^{\mathrm{T}} D r \begin{cases} r^{\mathrm{T}} = [O_1/O_+, \ldots, O_n/O_+] \\ D = (d_{ij}) \text{ distance between centroids.} \end{cases} \tag{6}$$

This statistic has been very criticised by Tango (1999) because only the observed number of cases are taken into account and not the discrepancies between observed and expected cases.

4.3.2 Tango's statistic for general clustering

Was proposed by Tango (1999) as a modification to Whittermore's statistic by incorporating expected cases as follows:

$$T = (r-p)^{\mathrm{T}} A (r-p) \begin{cases} r^{\mathrm{T}} = [O_1/O_+, \ldots, O_n/O_+] \\ p^{\mathrm{T}} = [E_1/E_+, \ldots, E_n/E_+] \\ A = (a_{ij}) \text{ closeness matrix} \end{cases} \tag{7}$$

Tango suggests $a_{ij} = exp\{-d_{ij}/\phi\}$, where d_{ij} is the Euclidean distance between regions i and j (i.e., between their centroids) and ϕ is a positive constant used to measure the strength of dependence between zones.

4.4 Scan statistics

These methods define a window which is moved across the whole study area to test whether regions lying inside this window are a cluster. Some of these methods, specially GAM, have been highly criticised because they perform many non-independent tests. Openshaw et al. (1987) argue that, on the other hand, the level of the local tests can be corrected and there is no bias in the investigation of cluster locations because data have not been explored a priori.

Unfortunately, the overall level of the simultaneous tests carried out is not controlled in the case of GAM and Besag and Newell's statistic. This overall control is the main contribution by Kulldorff and Nagarwalla, who propose the selection of the most likely cluster.

4.4.1 Openshaw's GAM

This was probably the first scan method proposed (Openshaw et al. 1987). It is based on creating a grid over the study region and centring circles of a given radius at these points.

For each circle, a local test is performed to decide whether it is a cluster or not. Those circles which are found to be a cluster are drawn on the map. In this way, we can get an idea of where clusters may be by inspecting those areas where more circles were drawn.

By default, the test implemented in this package compares the local observed number of cases to the quantile of level α of a Poisson distribution whose mean is the local expected number of cases. Local observed and expected number of cases are just the sum over these quantities along regions whose centroids fall within the circle.

4.4.2 Besag & Newell

This method was developed by Besag and Newell (1991) to detect clusters of size $k+1$, that is, regions that, when grouped together, reach $k+1$ observed cases.

Taking each case as the centre of a possible cluster, the other regions are sorted according to the distance from this, and the number of regions needed until k cases are found are computed (L_i). The observed number of regions to obtain k cases will be called l_i.

Then, it is tested whether l_i is low enough to be a cluster or, what is equivalent, the probability of finding k or more cases in these l_i regions. When data come from a Poisson distribution this probability is:

$$p\text{-value} = P(L_i \leq l_i) = P(\text{Cases} \geq k | \lambda = E_i^*) = 1 - \sum_{s=0}^{k-1} \frac{\exp(E_i^*)(E_i^*)^s}{s!} \quad (8)$$

λ represents the mean of the underlying Poisson distribution while E_i^* is the sum of the expected number of cases over these l_i regions.

4.4.3 *Kulldorff & Nagarwalla*

Kulldorff and Nagarwalla (1995) also create a grid and they consider, for a given point, the set (Z) of all possible circles centred there containing up to a fraction of the total population. For each one of these circles, let us call it z (with $z \in Z$), they are concerned with the probability of there being a case given the population at risk is inside this circle (p_z) and the probability of there being a case given the population at risk is outside z (q_z). If p_z is much higher than q_z then circle z can be thought of as a cluster.

For this reason, they propose the following test at each point:

$$\begin{aligned} H_0 &: \quad p_z = q_z \\ H_1 &: \quad \text{There is a cluster } z \text{ such that } p_z > q_z. \end{aligned}$$

They compute the maximum likelihood ratio, under the assumption of a Poisson model and conditioning on the total number of observed cases. It is equivalent to considering this statistic:

$$KN = \max_{z \in Z} \frac{L(z)}{L_0}, \quad (9)$$

where L_0 and $L(z)$ are defined as:

$$L_0 = \frac{O_+^{O_+}(P_+ - O_+)^{P_+ - O_+}}{P_+^{P_+}} \quad (10)$$

$$L(z) = \begin{cases} \phi(O_z, P_z)\phi(O_+ - O_z, P_+ - P_z) & \text{if } \frac{O_z}{P_z} > \frac{O_+ - O_z}{P_+ - P_z} \\ L_0 & \text{if } \frac{O_z}{P_z} \leq \frac{O_+ - O_z}{P_+ - P_z} \end{cases}, \quad (11)$$

where $\phi(O, P) = O^O (P-O)^{P\text{-}O}/P^P$.

O_z (P_z) represents the sum of the observed number of cases (population at risk) over all regions whose centroids lie within circle z.

Probability values can be calculated by means of bootstrap or Monte Carlo simulations, as explained in Sect. 5.

4.5 Focused tests

Unlike scan methods, the method presented here considers a single pre-established or previously known region around which the hypothesis of

clustering is tested. This region usually contains a putative pollution source thought to affect public health. Examples of such sources are nuclear plants (Stone 1988), waste incinerators (Diggle et al. 1997) and petrochemical complexes (Pekkanen et al. 1995).

A bias will be introduced in the study if these methods are employed in regions that have been suggested after looking through data. This is due to the fact that we try to assess whether the observed number of cases is significantly high after knowing that it is actually high. Then, the probability of rejecting null hypotheses will be increased.

4.5.1 Stone's Test

Supposing that all regions are sorted according to distance to the central region, Stone (1988) proposes to check for a descending trend from the source with this test:

$$\begin{aligned} H_0: &\quad \theta_1 = \cdots = \theta_n = \lambda \\ H_1: &\quad \theta_1 \geq \cdots \geq \theta_n, \end{aligned}$$

which is performed with this statistic:

$$T = \max_{1 \leq j \leq n} \frac{\sum_{i=1}^{j} O_i}{\sum_{i=1}^{j} E_i}. \tag{12}$$

Again, if λ is supposed to be unknown, the expected number of cases must be multiplied by O_+/E_+.

5 Bootstrap

Since the sampling distributions of statistics described in Sect. 4 can be difficult to derive, we propose the use of bootstrap to estimate them. The idea is to choose a suitable model or distribution for the data under the null hypotheses and to simulate the observed number of cases in every region. For each simulation the value of the statistic being used is calculated.

After a number of simulations have been computed, we have an approximation to the sampling distribution under the null hypotheses of this statistic and probability values can be easily calculated.

In other words, bootstrap provides a set of values that are an approximation to the sampling distribution of the statistic whatever the method or sampling model. Thus, the (one-tailed) probability value is obtained as the number of samples whose value is higher than the observed value of the statistic, divided by the total number of samples plus one.

Four possible procedures (which are explained below) seem to be adequate: permutation (non-parametric) bootstrap, Multinomial bootstrap (Wakefield et al. 2000b), Poisson bootstrap (Morris and Wakefield 2000) and Negative Binomial bootstrap (Clayton and Kaldor 1987) from the Poisson-Gamma model.

The last three proposals are based on models explained in Section 3. Note that the use of these methods is conditioned by whether our data exhibit extra-variation (Dean 1992). If this is not the case, Poisson or Multinomial bootstrap may be used. If overdispersion (also called extra-variation) may be related to analysed data, Negative Binomial sampling is a better choice.

If O_+ is high compared to the number of regions n, then Multinomial and Poisson bootstrap will produce similar results, because the Multinomial distribution can be obtained from n Poisson distributions by conditioning on O_+. When O_+ is high, small variations will not strongly affect the Multinomial distribution.

Permutation bootstrap is based on redistributing relative risks or residuals among all regions without replacement. It has been used when assessing spatial dependence between neighbouring regions by means of spatial autocorrelation (Cliff and Ord 1981).

6 DCluster overview

First of all, data must be stored in a data frame with at least the following columns: Observed (number of cases), Expected (number of expected cases), Population (total population at risk), x (centroid easting coordinate) and y (centroid northing coordinate). Those functions related to Spatial Autocorrelation and General Clustering also need to be passed a list of boundaries and weights, stored in an object of class listw, implemented in package spdep.

Package spdep contains a number of structures and routines focused on the assessment of spatial dependence. Moran's I and Geary's c are calculated using functions moran and geary, available in this package, and object listw is used to describe neighbours and weights.

Whittermore's statistic and Tango's statistic also use this kind of object to store neighbours and weights. Although weights used are globally standardised, significance is not affected because this standardisation is performed by dividing all weights by the same constant. Since significance is based on bootstrap, it is not affected by this transformation because all simulated values are also rescaled.

Package boot has been used to compute bootstrap by means of function boot. This function needs the data frame mentioned before as input, the statistic to compute, the basic model for sampling data, and the number of replicates to be done.

For every statistic presented in Sect. 4 several functions have been implemented. Basically, one to compute its value given a data set and two more to be used in bootstrap, be it non-parametric (permutation) or parametric (Multinomial, Poisson or Negative Binomial).

Once bootstrap is performed, an object of type *boot* is returned by function boot. This object can be plotted to obtain a histogram of the simulated values (where the observed value is also marked) and a normal qq-plot. This graphic gives a quick and easy answer to whether observed data are significant or not.

For scan statistics there is a main function called opgam, which implements a standard Openshaw GAM. This function basically needs a data set, a way to build the grid (which can be done in several ways) and a function, which we call iscluster, to assess whether the local area being inspected at each point of the grid is a cluster or not. This provides a general framework that has been used in the implementation of other scan methods.

For every scan statistic described above, a version of iscluster has been implemented following general guidelines (which are explained in package documentation). The object returned by these functions is a data frame containing coordinates (x and y) of points marked as clusters, values of the statistic (statistic), associated probability values (p value) and size of the clusters (expressed as number of regions forming the centre). No boot object is returned this time, since they are only used in local calculations.

Function opgam returns all this information for the points found to be significant according to the significance level chosen by the user. Nothing is returned for the points that were not clusters.

It is worth saying that for Besag and Newell's statistic, exact p-values are calculated when sampling from Multinomial or Poisson distributions. In future, exact calculation of the p-value for Stone's Test will also be added (Stone 1988).

7 Example

In order to illustrate the use of package DCluster, a brief example using real data is provided below. The data employed are the number of cases of sudden infant death syndrome (SIDS) in North Carolina between years 1974 and 1978. They are described, for example, by Cressie and Chan (1989) and Cressie (1993).

These data are available in package spdep, and they have been reformatted to accomplish DCluster requirements. Population at risk is the number of births, while the expected number of cases have been calculated by $P_i O_+ / P_+$. Neighbours used in the analysis are those provided in package spdep, while map boundaries have been downloaded from the U.S. Census Bureau web site (at http://www.census.gov) and maps have been created with package RArcInfo (Gómez-Rubio and López-Quílez 2005).

Figure 1 shows a histogram and a boxplot, which provide a brief summary of SIDS data.

Figure 2 shows relative risks estimators (SMRs) and smoothed relative risks estimators. There it is shown how areas with extremely high or low relative risks are smoothed. Two clusters are clearly found on these maps to the south and northeast.

In order to choose a suitable sampling model, a likelihood ratio test was performed between a fitted Negative Binomial model and a fitted Poisson model, showing that the Negative Binomial fitted the data better (p value of 0). Tests based on statistics P_B and P_B' proposed by Dean (1992) were also carried out and their resulting p-values were both 0. These results led us to use a Negative Binomial distribution when bootstrapping.

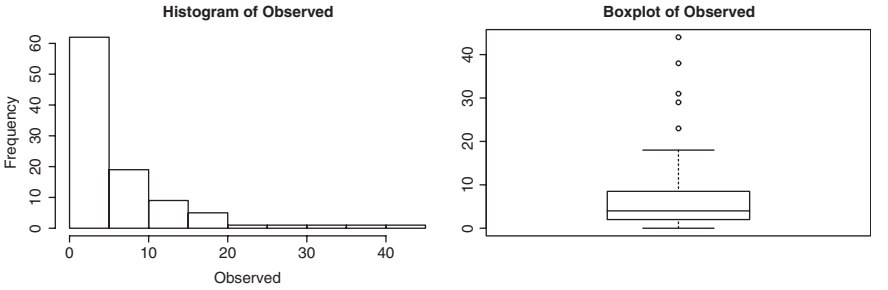

Fig. 1 Histogram and boxplot of SIDS data

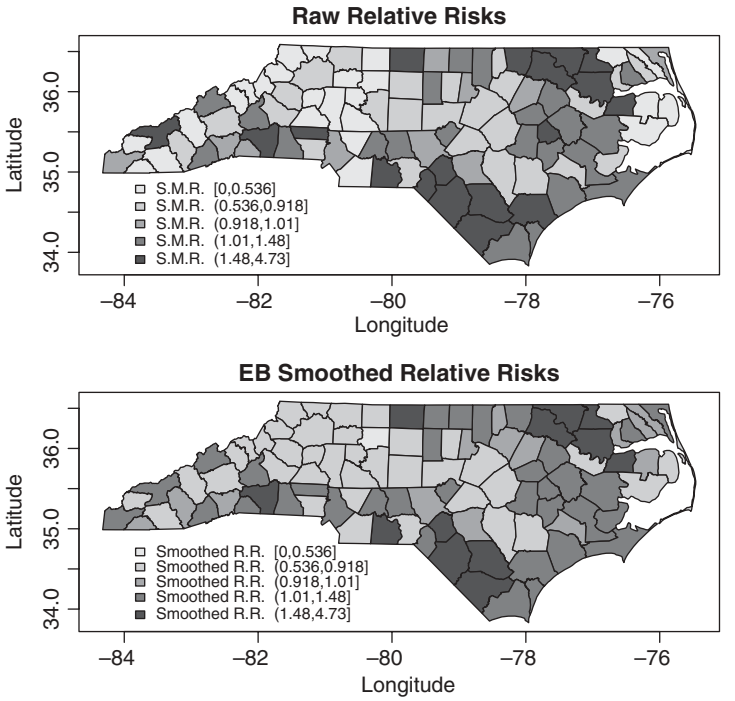

Fig. 2 Relative risks estimators (SMRs) and smoothed relative risks estimators (Poisson-Gamma model)

Figure 3 shows results for tests of homogeneity (Pearson's Chi-square and Potthoff-Whittinghill's) under the null hypothesis of data drawn from a Poisson distribution. A histogram of simulated values of both statistics, together with their observed value (dashed line), and a normal qq-plot are displayed. Since observed values are significant for both methods we reject null hypotheses of homogeneity and believe that data are overdispersed.

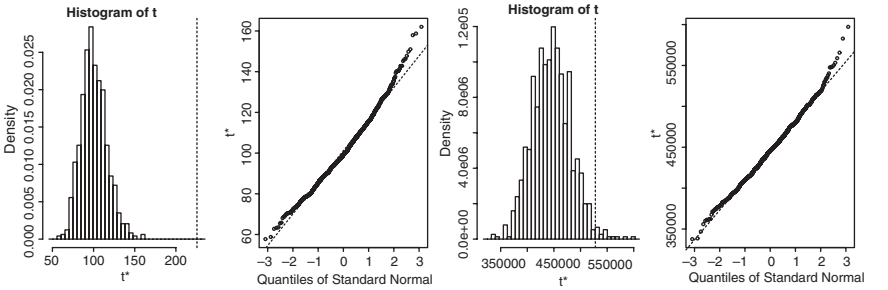

Fig. 3 Chi-square Test and Potthoff-Whittinghill's Test (Poisson model)

These results agree with those provided by likelihood ratio test and Dean's test and thus, we are supported in using a Negative Binomial distribution when sampling.

Estimated parameters for the prior Gamma distribution are $\widehat{v} = 4.630689$ and $\widehat{\alpha} = 4.395678$.

Under the assumption that data come from a Negative Binomial distribution, with the Gamma parameters obtained from an Empirical Bayes estimation using equations provided by Clayton and Kaldor (1987), probability values related to observed number of cases have been plotted (Choynowski 1959) in Fig. 4. It shows that just a few isolated areas have been marked as significant, which means that with this distribution, data apparently do not cluster around any location.

Autocorrelation measures calculated for residuals are shown in Fig. 5. The weights used were 1 if counties share boundaries, and 0 if otherwise. It is clear that data exhibit some kind of spatial correlation because observed values are found in the tails of sampling distributions. In this case permutation bootstrap was used, instead of sampling from a Negative Binomial distribution.

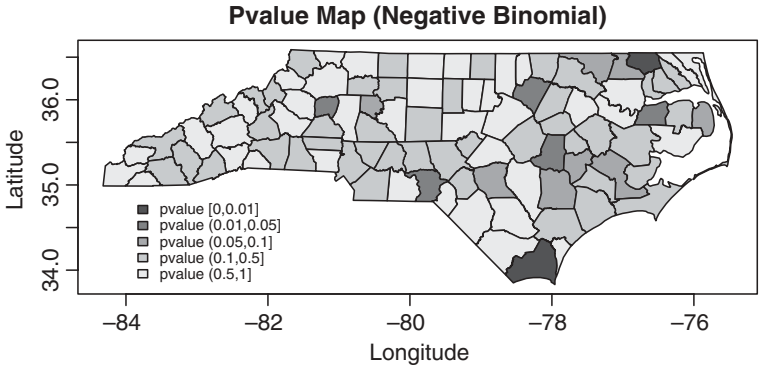

Fig. 4 Probability values calculated for each area according to the hypothesis that data are drawn from a Negative Binomial distribution

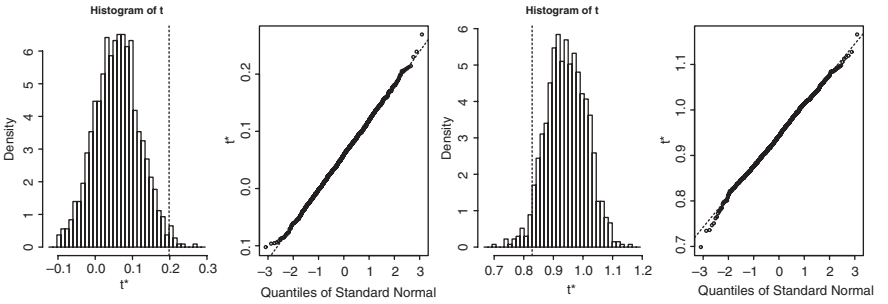

Fig. 5 Moran's I Statistic and Geary's c Statistic (Poisson-Gamma model)

These results mean that there are zones in the study area where residuals are quite similar, be they high or low. When they are significantly high, these regions constitute a cluster. Note that it is also possible that spatial autocorrelation is due to groups of areas that have low risks (Cliff and Ord 1981), but we will not investigate this case.

General clustering statistics (Whittermore's and Tango's) are shown in Fig. 6. They do not show any evidence of general clustering because observed values of these statistics fall in highly probable regions of their sampling distributions. This fact can be explained by considering that really significant regions, under the null hypothesis of data distributed according to a Negative Binomial distribution, are isolated and, hence, there is no global tendency to cluster.

Since these methods are designed to detect global trends, clusters that are small or weak will not be detected by them, which is probably the case here.

Scan methods described before were also employed, and Table 1 summarises parameters used. The significance level has been set to 0.002 for GAM, as proposed by Openshaw et al. (1987). With regard to the other two methods, significance has been set as 0.05, as proposed by their respective authors (Besag and Newell, 1991; Kulldorff and Nagarwalla, 1995). Results are difficult to compare, since only Kulldorff and Nagarwalla's method controls the overall significance level.

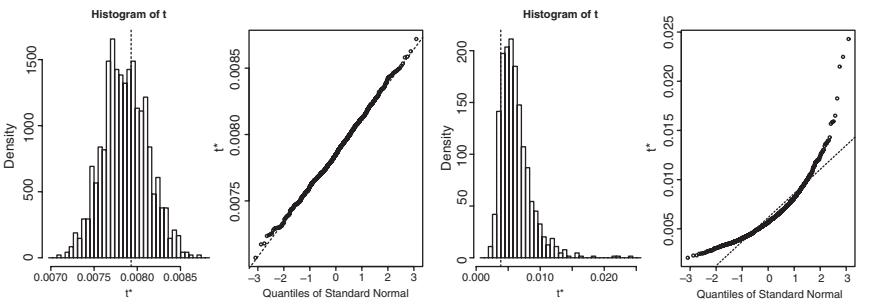

Fig. 6 Whittermore's Statistic and Tango's Statistic (Poisson-Gamma model)

Table 1 Parameters used when testing scan methods

Method	Grid	Radius	Sig. level
GAM	Step = radius/5	50 km	0.002
B. & N.	Centroids	20 cases	0.05
K. & N.	Centroids	≤ 0.2 tot. pop.	0.05

GAM clearly marks just one area as a cluster in the northeast. The results of this method are clearly sensitive to the grid step and radius selected, so it is often useful to try different values of these parameters.

Kulldorff and Nagarwalla's method marked 20 counties as being part of a cluster, which can be found grouped in two zones (i.e., two real clusters), to the south and northeast. Notice that these counties have high SMRs as shown in Fig. 2. Some of the regions in the latter cluster have been pointed out by GAM too.

Besag and Newell's method was tested with cluster size 20 ($k = 19$), which is over three times the mean of the observed number of cases in North Carolina. This method has marked two regions to the northeast and another to the south as significant centres of clusters. More tests should be performed by varying the size of the cluster since it may happen that a cluster exists but its size is different from 20. This analysis is outside the illustrative scope of this paper.

Cressie and Chan (1989) mention that they removed Anson County from their study because of its high residual. This county was considered a cluster by Kulldorff and Nagarwalla's statistic.

Since this high residual may be due to an unknown risk factor in the county which may be the responsible of the appearance of a cluster, Stone's Test was carried out over Anson County.

Note that this test must be performed **before** examining the data, since a bias is produced by trying to apply Stone's Test on those regions with the highest relative risks. This will produce an increment in the probability of being significant.

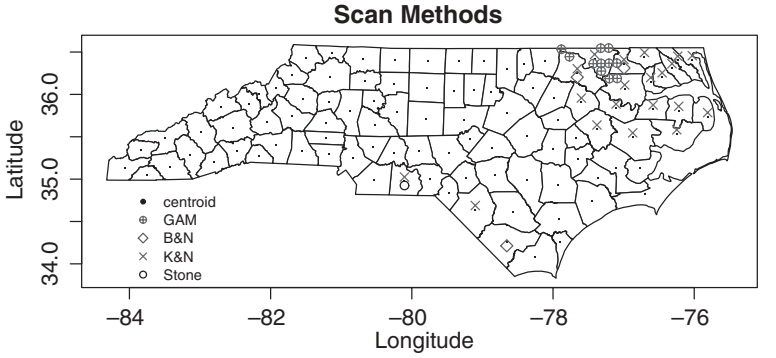

Fig. 7 Results from several scan methods and Stone's Test (Poisson-Gamma model)

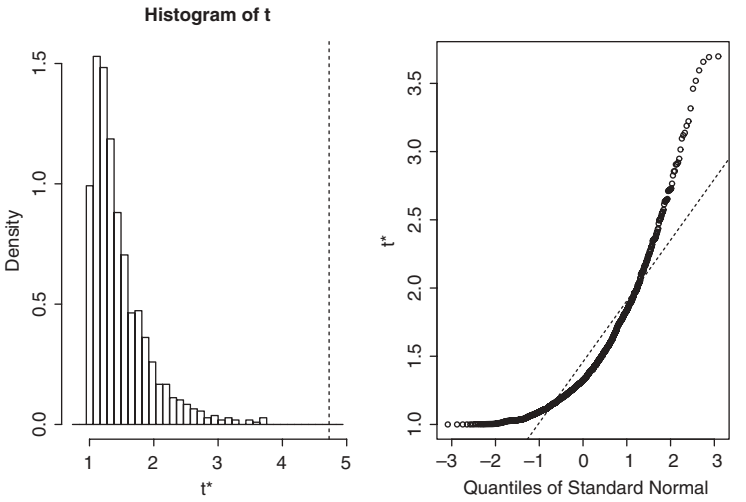

Fig. 8 Stone's Test Results (Poisson-Gamma model) around Anson County

Therefore, the test is only to illustrate the method since the cluster is specified after seeing the data. The result is shown in Fig. 8, which clearly suggests that there is a cluster around Anson County.

8 Concluding remarks

In this paper we have presented different methods used for exploratory analysis of epidemiological data and detection of spatial clusters of disease. We have implemented all these methods and have developed a package called DCluster for the R programming language which is freely available. A suitable bootstrap has also been proposed to estimate sampling distributions of statistics involved in the analysis.

Moreover, an example using North Carolina SIDS data has been discussed. We plan to compare the behaviour of all these methods under the different bootstrap samplings in order to see which methods are more robust. This is especially useful when working with real data, since their distribution will remain unknown.

We also expect to add other methods to this package in the future.

References

Assunçao RM, Reis EA (1999) A new proposal to adjust Moran's I for population density. Stat Med 18:2147–2162

Aylin P, Maheswaran R, Wakefield J, Cockings S, Jarup L, Arnold R, Wheeler G, Elliot P (1999) A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: The UK Small Area Health Statistics Unit. J Public Health Med 21(3):289–298

Besag J, Newell J (1991) The detection of clusters in rare diseases. J Roy Stat Soc Series A 154:143–155
Best N, Elliott P, Richardson S (2001) Spatial epidemiology. Short course. http://stats.-ma.ic.ac.uk/∼ngb30/ [Last access: 19-07-2004]
Choynowski M (1959) Map based on probabilities. J Am Stat Soc 54(286):385–388
Clayton D, Kaldor J (1987) Empirical bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671–681
Cliff A, Ord JK (1981) Spatial processes: models and applications. Pion Limited, London
Cressie N (1993) Statistics for spatial data. Wiley, New York
Cressie N, Chan NH (1989) Spatial modeling of regional variables. J Am Stat Assoc 84:393–401
Dean CB (1992) Testing for overdispersion in poisson and binomial regression models. J Am Stat Assoc 87(418):451–457
Diggle P, Elliott P, Morris S, Shaddick G (1997) Regression modelling of disease risk in relation to point sources. J Roy Stat Soc, Series A 160 (3):491–505
Ferrándiz J, Abellán JJ, Gómez-Rubio V, López-Quílez A, Sanmartín P, Abellán C, Martínez-Beneito MA, Melchor I, Vanaclocha H, Zurriaga O, Ballester F, Gil JM, Pérez-Hoyos S, Ocaña R (2004) Spatial analysis of the relationship between cardiovascular mortality and drinking water hardness. Environ Health Perspect 112(9):1037–1044
Geary RC (1954) The contiguity ratio and statistical mapping. Incorporated Stat 5:115–145
Gómez-Rubio V, López-Quílez A (2005) RArcInfo: Using GIS data with R. Computers & Geosciences (in press)
Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comput. Graphical Stat 5(3):299–314
Jenicek M, Cléroux R (1982) Epidemiologie. Principes, techniques, applications, 2nd ed. Edisem Inc.
Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. Stat Med 14:799–810
McCullagh P, Nelder J (1989) Generalized linear models, 2nd ed. Chapman and Hall, London
Moran PAP (1948) The interpretation of statistical maps. J Roy Stat Soc, Series B 10:243–251
Morris SE, Wakefield JC (2000) Assessment of disease risk in relation to a pre-specified source. In: Elliot P, Wakefield JC, Best NG, Briggs DJ (eds) Spatial epidemiology: methods and applications. Oxford University Press, New York, pp 153–184
Oden N (1995) Adjusting Moran's I for population density. Stat Med 14:17–26
Openshaw S, Charlton M, Wymer C, Craft AW (1987) A Mark I geographical analysis machine for the automated analysis of point data sets. Int J Geographical Inf Syst 1:335–358
Pekkanen J, Pukkala E, Vahteristo M, Vatiainen T (1995) Studying cancer incidence around an oil refinery as an example of a small area study based on map coordinates. Environ Res 71(2):128–134
Potthoff RF, Whittinghill M (1966a) Testing for homogeneity: I. The Binomial and Multinomial distributions. Biometrika 53:167–182
Potthoff RF, Whittinghill M (1966b) Testing for homogeneity: II. The Poisson distribution. Biometrika 53:183–190
Snow J (1854) On the mode of communication of cholera. Churchill Livingstone, London
Stone RA (1988) Investigating of excess environmental risks around putative sources: Statistical problems and a proposed test. Stat Med 7:649–660
Tango T (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med 14:2323–2334
Tango T (1999) Comparison of general tests for spatial clustering. In: Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF (eds) Disease mapping and risk assessment for public health. Wiley, New York, pp 111–117
Wakefield JC, Best NG, Waller L (2000a) Bayesian approaches to disease mapping. In: Elliot P, Wakefield JC, Best NG, Briggs DJ (eds) Spatial epidemiology: methods and applications. Oxford University Press, New York, pp 104–127
Wakefield JC, Kelsall JE, Morris SE (2000b) Clustering, cluster detection and spatial variation in risk. In: Elliot P, Wakefield JC, Best NG, Briggs DJ (eds) Spatial epi-

demiology. methods and applications. Oxford University Press, New York, pp 128–152
Waldhör T (1996) The spatial autocorrelation coefficient Moran's I under heterocedasticity. Stat Med 15:887–892
Walter SD (1993) Assessing spatial patterns in disease rates. Stat Med 12:1885–1894
Whittermore AS, Friend N, Byron W, Brown JR, Holly EA (1987) A test to detect clusters of disease. Biometrika 74:631–635
Zoellner IK, Schmidtmann IM (1999) Empirical studies of cluster detection- different cluster tests in application to German cancer maps. In: Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF (eds) Disease mapping and risk assessment for public health. Wiley, New York, pp 169–178