Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)
March 20–22, Vienna, Austria *ISSN 1609-395X*
Kurt Hornik, Friedrich Leisch & Achim Zeileis (eds.)
http://www.ci.tuwien.ac.at/Conferences/DSC-2003/

Detecting Clusters of Diseases with R

V. Gómez-Rubio, J. Ferrándiz, A. López

Abstract

One of the main concerns in Public Health surveillance (Aylin, Maheswaran, Wakefield, Cockings, Jarup, Arnold, Wheeler, and Elliot, 1999) is detection of clusters of diseases, i. e., the presence of high incidence rates around a particular location, which usually means a higher risk of suffering from the disease of study.

Many methods have been proposed for cluster detection, ranging from visual inspection of disease maps to full Bayesian models estimated by using M.C.M.C.

In this paper we describe the use and implementation, as a package for R, of several methods which have been widely used in the literature, such as Openshaw's G.A.M., Stone's test and others (Wakefield, Kelsall, and Morris, 2000; Waller, Turnbull, Clarck, and Nasca, 1994).

Although some of the statistics involved in these methods have an asymptotic distribution, bootstrap will be used to estimate their actual sampling distributions.

1 Introduction

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

Beginning from the study of Snow (1854) over an outbreak of cholera in London, whose focus he found by plotting the location of those people affected, many methods have been developed to detect spatial clusters of diseases.

Some of these methods, which have repeatedly appeared in the bibliography and that have been widely used in real studies, are described in this paper. Besides, we describe the use of a new package for R called `DCluster` that implements routines to use all these methods. Although there are several packages available in R for clustering, they provide general methods and none of them is devoted to spatial clusters of diseases.

In this paper, first we will present the general structure of the data available for the problem of cluster detection, followed by the most usual statistical models used. After that, we will briefly describe methods implemented in `DCluster`, and bootstrap procedures. Finally, we explore the use of these methods using real data.

2 Data structure

Supposing that our study region is divided in n non-overlapping regions (which may be counties, provinces, municipalities, ...), data available are usually found as counts, that is, number of deaths or affected people in each region.

Let us represent by O_i the observed number of cases (usually, deaths) in region i, E_i its expected number, which may 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 over all the regions of observed cases, expected cases and population.

Usually population is stratified according to age and sex and, sometimes, a measure of deprivation or poverty. So, P_{ij} will mean people 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} are usually calculated using indirect standardisation. That is, if we have a reference population from which we know their incidence rates ($r_{ij} = O'_{ij}/P'_{ij}$) for each stratum, then we have $E_{ij} = P_{ij}r_j$.

When the reference population is the same than the population under study, standardisation is called *internal* and it holds that $O_+ = E_+$.

Finally, spatial location of regions will be done by their centroids, which mark the centre of the total area. This centroids are usually not taken as the geometrical centre, but are weighted by the actual population location within the region.

3 Statistical models for diseases

As a first approximation, we will consider O_i's to be independent and drawn from a Poisson distribution whose mean is $\theta_i E_i$, where θ_i is the relative risk, which measures the local deviation of the disease. If the relative risk is over 1 then there is an excess in risk in that region.

The maximum likelihood estimator for θ_i, which is called the *Standardised Mortality Ratio* (S.M.R.), 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 estimations arising from rare diseases or low populated areas, where the number of expected cases is really low, may lead to poor estimators.

Conditioning on O_+ leads us to a Multinomial model, in which the size is O_+ and probabilities are given by $(E_1/E_+, \ldots, E_n/E_+)$. This model is often used when performing Monte Carlo simulations to estimate distributions of different statistics (Best, Elliott, and Richardson, 2001).

Notice that this model is equivalent to distribute total observed cases at random among all the regions using E_i as weights.

Poisson model is too strict in the sense that it imposes mean and variance to be equal. When data exhibits some kind of overdispersion Poisson distribution is unlikely to be the right one.

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

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

As a consequence, O_i is distributed following a Negative Binomial with size ν and probability $\alpha/(\alpha + E_i)$. ν and α are usually estimated via Empirical Bayes.

M.L.E. for θ_i is now $(O_i + \nu)/(E_i + \alpha)$, which provides a smoothed estimator of the relatives risks. These estimators are usually used when performing a 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 they search for clusters over all the study regions or they assess the presence of a cluster just around a given region.

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

4.1 Tests for homogeneity

These methods can be used as a first approach to the problem to investigate whether relative risks are homogeneous (i.e., *equal*) along the study region. Differences between relative risks may lead to zones where they are higher (or lower) than expected and, hence, a cluster may be present.

4.1.1 Pearson's chi-square statistic

The value of this statistic is well known:

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

Test hypotheses are as follows:

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

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

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

The case in which internal standardisation is used is slightly different, in the sense that, since $O_+ = E_+$, λ must be 1 and the degrees of freedom are $n - 1$.

Notice 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 (1966a) assume that data come from a Multinomial distribution and consider the locally most powerful test for related to the next test hypotheses:

$$\begin{aligned} H_0: &\quad \theta_1 = \ldots = \theta_n = \lambda \\ H_1: &\quad \theta_i \sim Ga(\lambda^2/\sigma^2, \lambda/\sigma^2) \end{aligned}$$

Notice 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:

$$PW = E \sum \frac{O_i(O_i - 1)}{E_i}$$

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. Notice that if internal standardisation was carried out, then the hypotheses of homogeneity implies λ to be equal to 1.

4.2 Autocorrelation

Statistics presented in this section measure spatial autocorrelation of the data. Usually the quantities involved are S.M.R.s or residuals By working with the residuals we look for correlation among what wasn't explained by our primary model. When using S.M.R.s, we expect to find regions where they tend to be higher (or lower).

4.2.1 Moran's I statistic

Moran (1948) proposes a statistic, called the *I statistic*, that is very close to that of correlation coefficient

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

As mentioned before, Z_i may be either residuals $(O_i - E_i)$ or relative risks. W is a matrix which measures vicinity between regions, and it can be defined in different ways. For example, it can be 1 if the regions have a common boundary (and 0 otherwise) or the inverse of the distance between their centroids.

It has been noticed that this statistic is quite robust against changes in sampling distribution, as shown in Zoellner and Schmidtmann (1999)

4.2.2 Geary's c statistic

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

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

Notice that now 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.

It is known that these methods may fail to detect global clustering when actual clusters are small or scatter all around the study area.

4.3.1 Whittermore's statistic

The statistic proposed by Whittermore, Friend, Byron, Brown, and Holly (1987) is based on the distance between all pairs of cases, and is defined as:

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

This statistic has been heavily criticised in Tango (1999) because it only cares about the observed number of cases, and not about discrepancies between observed and expected cases.

4.3.2 Tango's statistic for general clustering

It was proposed by Tango (1999) as a modification to Whittermore's statistic, and it is defined as follows:

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

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

4.4 Scan statistics

These methods are proposed to scan small areas within the whole study region and look for clusters. Some of these methods, specially G.A.M., have been highly criticised because they perform many non-independent tests. Their defenders argue that, on the other hand, the level of the local tests can be corrected and that there's no bias in the investigation of cluster locations because data have not been explored a priori.

4.4.1 Openshaw's GAM

Probably this is the first scan method proposed (Openshaw, Charlton, Wymer, and Craft, 1987). It is based on creating a grid over the study region and building balls (i.e., circles) of a given radius centred at that points.

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

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 ball.

4.4.2 Besag & Newell

This method has been developed by Besag and Newell (1991) to detect clusters of size k, that is, regions that grouped together reach k observed cases.

Taking each case as centre of a possible cluster, the other regions are sorted according to distance to this one and the number of regions needed until k cases are found is 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 more than k cases in these l_i regions. When data come from a Poisson distribution this probability is:

$$\text{p-value} = P(L_i \geq l_i) = P(\text{N. cases} > k | \lambda = E_i^*) = 1 - \sum_{s=0}^{k-1} \frac{\exp(E_i^*)(E_i^*)^s}{s!}$$

λ represents the mean of the underlying Poisson distribution while E_i^* is the sum of the expected number of cases of the 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, they are interested in the probability of being a case inside (p) and outside (q). If p is much higher than q then the circle can be viewed as a cluster.

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

$$\begin{aligned} H_0 &: \quad p = q \\ H_1 &: \quad p > q \end{aligned}$$

They compute the maximum likelihood ratio, under the assumption of a Poisson model and conditioning to the total number of observed cases. This is equivalent to consider the next statistic:

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

where L_0 and $L(z)$ are defined this way:

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

$$L(z) = \begin{cases} \left(\frac{O_z^{O_z}(P_z - O_z)^{P_z - O_z}}{P_z^{P_z}}\right)\left(\frac{(O_+ - O_z)^{O_+ - O_z}(P_+ - P_z - (O_+ - O_z))^{P_+ - P_z - (O_+ - O_z)}}{(P_+ - P_z)^{P_+ - P_z}}\right) & \text{if } \frac{O_z}{P_z} > \frac{O_+ - O_z}{P_+ - P_z} \\ \frac{O_+^{O_+}(P_+ - O_+)^{P_+ - O_+}}{N^N} & \text{if } \frac{O_z}{P_z} \leq \frac{O_+ - O_z}{P_+ - P_z} \end{cases}$$

O_Z (P_Z) represents the sum of the observed number of cases (population at risk) of all the regions whose centroids lay within circle Z.

Pvalue can be calculated by means of bootstrap or Monte Carlo simulations.

4.5 Focused tests

Unlike scan methods, the method presented here consider a single (or just a few) region around which the hypotheses of clustering is tested. This region usually contains a pollution putative source which may be thought to affect Public Health. Examples of such sources are nuclear plants, waste deposits or incinerators.

A bias will be introduced in the study if these methods are used after data examination suggested the presence of a cluster. This is due to the fact that we try to assess whether the observed number of cases is extremely high after knowing that it is in fact high. Then, the probabilities 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 the next test:

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

which is performed with the next statistic:

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

Again, if λ is supposed to be unknown, expected number of cases must by multiplied by $\frac{O_+}{E_+}$

5 Bootstrap

Since sampling distributions of statistics used in these tests can be difficult to derive, we propose the use of bootstrap sampling to estimate them. The idea is to choose a suitable model or distribution for the data and to simulate the observed number of cases at every region. For each of this simulations, the value of the statistic being used is calculated.

After a number of simulations have been computed, we have an approximation of the sampling distribution of this statistic and pvalues can be easily calculated.

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

The first three are based on models explained in section 3. Notice that they may be used depending on whether our data exhibit extra-variation or not.

If it is not found in a preliminary study over the data, then Poisson bootstrap may be used (and even Multinomial). If we think overdispersion may be related to our data, then perhaps it is better to use Negative Binomial sampling.

Notice that, if O_+ is high compared to the number of regions (n), then Multinomial and Poisson bootstrap will produce almost the same results, since Multinomial distribution can be obtained from Poisson framework by conditioning on O_+, and small variations in O_+ will not affect Multinomial distribution strongly.

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 Lawson (2001) Zoellner and Schmidtmann (1999) by means of spatial autocorrelation.

6 `DCluster` overview

In the first place, 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). Some functions also need a distance or closeness matrix, which must be squared $n \times n$.

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

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

Once bootstrap is performed, a 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 is also drawn. 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 the general Openshaw's G.A.M. 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 have 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 deeply explained in package documentation). The object returned by these functions is a list containing the coordinates (`x` and `y`) of the point marked as cluster, the value of the statistic (`statistic`) and the associated pvalue (`pvalue`). No *boot* object is returned this time since they are used in local calculations only.

`opgam` returns all this information for the points that resulted to be significant according to the significance level choose by the user. For those points that were not clusters nothing is returned.

It is worth saying that for Besag & Newell's statistic exact p values are calculated when sampling from Multinomial, Poisson of Negative Binomial distributions. In the future exact calculation of p value for Stone's Test will be added too.

7 Example

In order to illustrate the use of package `DCluster` a brief example using real data is provided below. Data employed are 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 reformated to accomplish `DCluster` requirements. Population at risk is the number of births, while the expected number of cases have been calculated by $P_i \frac{O_+}{P_+}$. Furthermore, a matrix containing distances between centroids have been created.

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

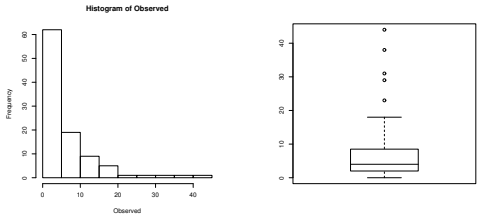

Figure 1: Boxplot and histogram of SIDS data.

In order to choose a suitable sampling model, a likelihood ratio test have been performed between a fitted Negative Binomial and Poisson models, resulting on the first one to fit better the data (pvalue 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 bootstraping.

Estimated parameters for the prior Gamma distribution are $\widehat{\nu} = 4.630689$ and $\widehat{\alpha} = 4.395678$. This means that smoother relative risks will not be strongly changed.

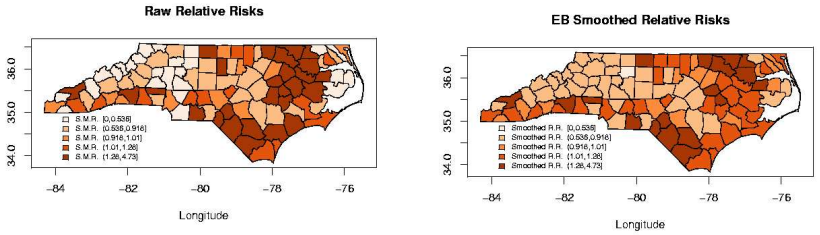

Figure 2: Relative Risks and Smoothed Relative Risks Estimators (Poisson-Gamma model).

Figure 2 shows relative risks 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 that maps to the south and northeast.

Under the assumption that data come from a Negative Binomial distribution, with the Empirical Bayes estimated Gamma parameters, pvalues related to each region have been plot in Figure 3. 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.

As shown in Figure 4 [1] both methods (Pearson's Chi-square and Potthoff - Whittinghill) used to test homogeneity in the data show that we can't reject null

[1] For both statistics it can be seen a histogram of simulated values of the statistic togehter with its observed value (dashed line), and a normal qq-plot.

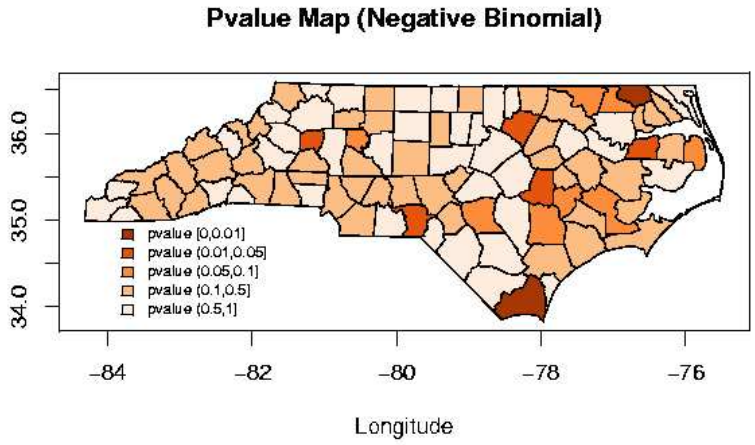

Figure 3: Pvalues calculated for each area according to the hypotheses that data are drawn from a Negative Binomial.

hypotheses of homogeneity. Negative Binomial is more variable than Poisson distribution, so it happen that more wide values are allowed and, hence, more extreme value of these statistics are also high probable in the simulations.

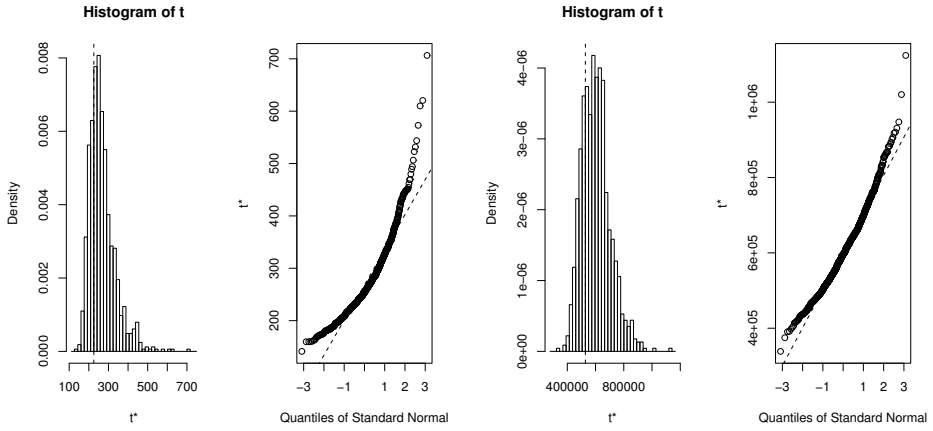

Figure 4: Chi-square Test and Potthoff-Whittinghill's Test.

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

This means that there will appear zones were the number of observed cases may be high and, hence, clusters may also appear. Notice that it is also possible that

correlated areas are those with low risks.

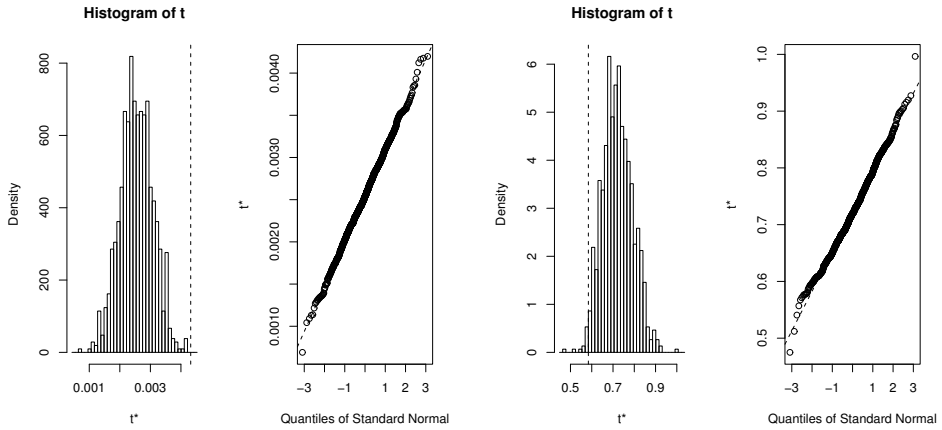

Figure 5: Moran's I Statistic and Geary's c Statistic.

General clustering statistics (Whittermore's and Tango's), as seen in Figure 6, do not show any evidence of general clustering because observed values of statistics fall in highly probability regions of sampling distributions. This fact can be explained by considering that really significant regions, according to the Negative Binomial distribution, are locally found and that there is no global tendency to cluster among them.

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

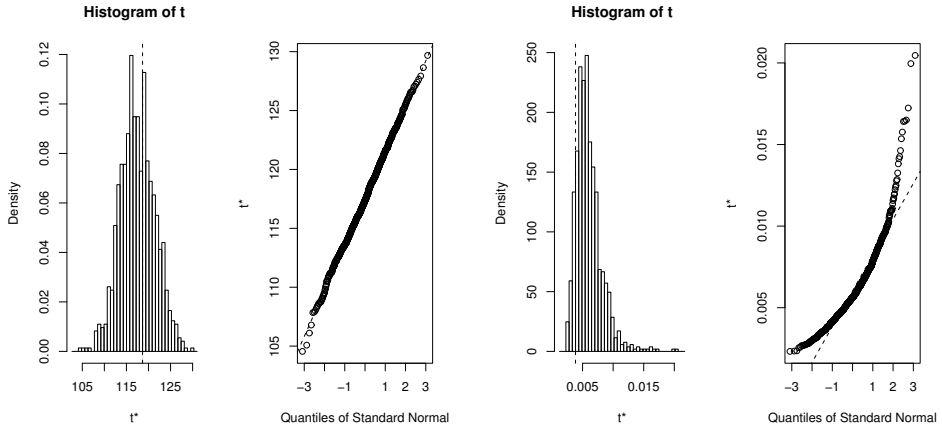

Figure 6: Whittermore' Statistic and Tango's Statistic.

Scan methods described before were also employed, and table in Figure 7 summarises the parameters used. Notice that for G.A.M. and Besag and Newell's

method no bootstrap is performed, since the exact critical value is calculated using the Negative Binomial distribution.

Method	**Grid**	**Radius**	**Sig. Level**
G.A.M.	step=radius/5	10 miles	0.002
B. & N.	centroids	20 cases	0.05
K. & N.	centroids	$\leq$0.2 tot. pop.	0.05

Figure 7: Arguments of the different scan methods used.

Significance level has been set to 0.002 for G.A.M, which is the one proposed by Openshaw et al. (1987). Concerning the two other methods, significance has been set to 0.05, as proposed by Kulldorff and Nagarwalla (1995). In this paper they also compare their method to Besag and Newell's, so we think 0.05 is a suitable significance level to show differences between both methods.

G.A.M. clearly marks just one area as cluster, which corresponds to county 4, the one with the highest S.M.R.

Kulldorff and Nagarwalla's method marked 21 counties as clusters, which can be found grouped in two zones, to the south and northeast. None of them is the one pointed out by G.A.M. Notice that these counties have high S.M.R.s as shown in Figure 2.

Besag & Newell's was tested with cluster size 20, which is over three times the mean of the observed number of cases. This method didn't marked any of the centroids as a significant centre of a cluster (of size 20). More tests should be done by varying the size of the cluster because when don't know the actual size of the clusters, if present, in the study area.

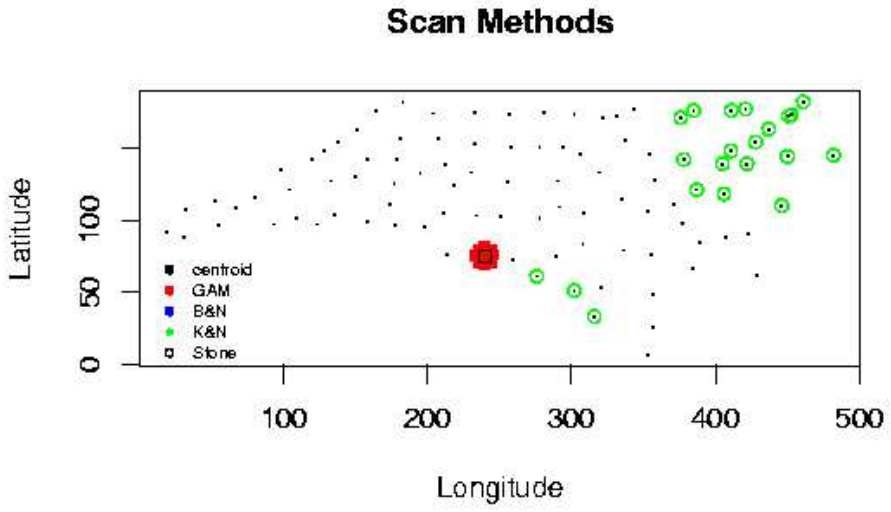

Figure 8: Results from several scan methods. Coordinates are in U.T.M. to show real distances between centroids.

Cressie and Chan (1989) mention that they removed county 4, which is the one that has been considered a cluster by G.A.M., from their study because of its high residual,

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 county 4. The result is shown in Figure 9, which clearly suggests that there is a cluster in that region.

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

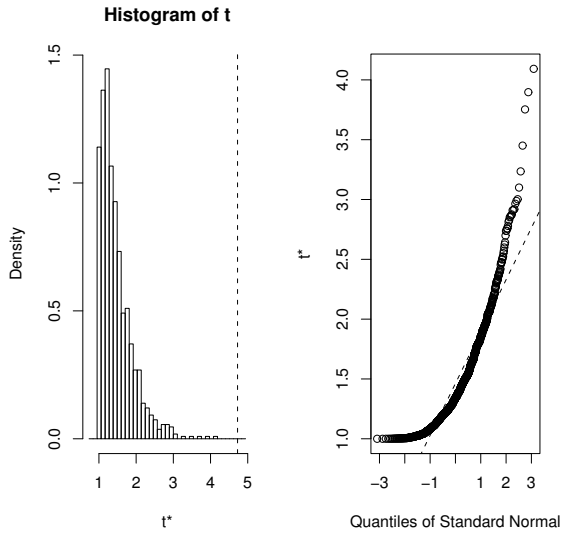

Figure 9: Stone's Test Results.

8 Concluding remarks

In this paper we have presented different methods used for exploratory analysis of epidemiological data and detection of spatial clusters, and the implementation we have done in R of all of them. A suitable bootstrap sampling has been proposed to estimate distributions of the statistics involved in the analysis.

Furthermore, an example using North Carolina SIDS data has also been discussed. In the future 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 specially useful when working with real data, since we don't know its actual distribution.

Other methods will be added to this package in the future.

Acknowledgements

This work has been partly funded by projects:

- *EUROHEIS: An European Environment and Health Information System for Exposure and Disease Mapping and Risk Assessment* - codes SI2.132454 (99CVF2-606), SI2.291820 (2000CVG2-605) and SI2.329122 (2001CVG2-604).
- *Collaboration Agreement between Consellería de Sanitat (Valencian Health Authority) and Universitat of València*

References

P. Aylin, R. Maheswaran, J. Wakefield, S. Cockings, L. Jarup, R. Arnold, G. Wheeler, and P. Elliot. A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: The u.k. small area health statistics unit. *Journal of Public Health Medicine*, 21 (3):289–298, 1999.

J. Besag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A*, 154:143–155, 1991.

Nicky Best, Paul Elliott, and Sylvia Richardson. Spatial epidemiology. short course. *http://stats.ma.ic.ac.uk/ ngb30/*, 2001.

David Clayton and John Kaldor. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681, 1987.

Noel Cressie and Ngai H. Chan. Spatial modeling of regional variables. *Journal of the American Statistical Association*, 1989.

Noel A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., 1993.

C. B. Dean. Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992.

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

Martin Kulldorff and Neville Nagarwalla. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14:799–810, 1995.

Andrew B. Lawson. *Statisticial Methods in Spatial Epidemiology*. Wiley, 2001.

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

S. E. Morris and J. C. Wakefield. Assessment of disease risk in relation to a pre-specified source. In P. Elliot, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology: Methods and Applications*, pages 153–184. Oxford University Press, 2000.

S. Openshaw, M. Charlton, C. Wymer, and A. W. Craft. A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1:335–358, 1987.

R. F. Potthoff and M. Whittinghill. Testing for homogeinity: II. The Poisson distribution. *Biometrika*, 53:183–190, 1966a.

R. F. Potthoff and M. Whittinghill. Testing for homogeneity: I. The Binomial and Multinomial distributions. *Biometrika*, 53:167–182, 1966b.

John Snow. On the mode of communication of cholera. *Churchill Livingstone*, 1854.

R. A. Stone. Investigating of excess environmental risks around putative sources: Statistical problems and a proposed test. *Statistics in Medicine*, 7:649–660, 1988.

Toshiro Tango. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, 14:2323–2334, 1995.

Toshiro Tango. Comparison of general tests for spatial clustering. In A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, and J-F. Viel, editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 8, pages 111–117. John Wiley & Sons Ltd., 1999.

J. C. Wakefield, J. E. Kelsall, and S. E. Morris. Clustering, cluster detection and spatial variation in risk. In P. Elliot, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology. Methods and Applications.*, chapter 8, pages 128–152. Oxford University Press, 2000.

Lance A. Waller, Bruce W. Turnbull, Larry C. Clarck, and Philip Nasca. Spatial pattern analyses to detect rare disease clusters. In Nicholas Lange, Louise Ryan, Lynne Billard, David Brillinger, Loveday Conquest, and Joel Greenhouse, editors, *Case Studies in Biometry*, chapter 1, pages 3–23. John Wiley & Sons, Inc., 1994.

A. S. Whittermore, N. Friend, W. Byron, J. R. Brown, and E. A. Holly. A test to detect clusters of disease. *Biometrika*, 74:631–635, 1987.

Iris K. Zoellner and Irene M. Schmidtmann. Empirical studies of cluster detection - different cluster tests in application to german cancer maps. In A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, and J-F. Viel, editors, *Disease Mapping and Risk Assessment for Public Health*, chapter 12, pages 169–178. John Wiley & Sons Ltd., 1999.

Corresponding author

Virgilio Gómez-Rubio
Departament d'Estadística i Investigació Operativa
Facultat de Matemàtiques
C/ Dr. Moliner 50
46100 Burjassot (València)
Spain