

The Cluster Detection in Spatial Health Surveillance Using the Scan Statistic

L'identificazione dei Clusters di Malattia Attraverso la Procedura Scan

Eugenia Nissi - Annalina Sarra

Dipartimento di Metodi Quantitativi e Teoria Economica

Viale Pindaro, 42 -65127 Pescara

nessi@dmqte.unich.it - asarra@dmqte.unich.it

Riassunto: I metodi per l'identificazione dei cluster di malattia assumono notevole rilevanza per gli istituti di sorveglianza della salute pubblica, incaricati di monitorare la distribuzione spaziale e temporale delle malattie. In questo lavoro, sono analizzati e confrontati recenti sviluppi della procedura proposta da Kulldorff e Nagarwalla (1995). I nuovi metodi hanno l'obiettivo di individuare clusters di forma irregolare.

Keywords: detection cluster methods, health surveillance, Monte Carlo testing, simulated annealing scan statistic, flexible scan statistic.

1. Introduction

Since the 1980's the analysis of disease clustering has generated considerable interest in statistical and epidemiological area. One can consider different aspects of the analysis of clustering. It is worth nothing the importance of detection cluster methods in public health, as part of geographical disease surveillance, whose main aim is to identify health problems as they occur and try to reduce or removed their effects on the population by appropriate public heath intervention. In this work, we review some of the cluster detection approaches. In particular, we focus on the spatial scan statistic (section 2) and its recent extensions (section 3). In section 4, we present the results of an empirical study carried out in order to evaluate the performances of the new methods in detecting clusters of irregular shapes.

2.The scan statistic: an overview

The statistical cluster detection methods indicate the likelihood that clusters exist at particular location; in other words, they provide estimates of the likely clusters geographical location and extent without previous knowledge of either how many or where they are. Among this category, we review the spatial scan statistic, introduced by Kulldorff and Nagarwalla (1995). Compared with the other previous approaches, (see Openshaw, 1988 and Besag and Newell, 1991), this method addresses some important issues: it contains a precise statement of null and alternative hypotheses and the inference is based on the likelihood ratio rather than an ad hoc test statistic. Moreover, it does not perform a separate test for each possible cluster location or for each possible cluster size. The spatial scan statistic can be used for data with exact point locations or for aggregated data. In this paper, we refer to disease incidence data available as counts from a set of geographic regions. The scan statistic has to take into account the

geometry of the area being scanned, the probability distributions generating events under the null hypothesis, the shapes and sizes of the scanning window. The study region is partitioned into n geographic sub-areas called cells. The basic form of data involves a set of counts observed $Y_1, Y_2, \dots, Y_N$ and a matching set of counts expected $E_1, E_2, \dots, E_N$ under the null hypothesis of constant risk, where the expected values are proportional to the population of each cell. The traditional scan statistic imposes a circular window on a map and lets its centre move across the study region. For any given position of centre the radius of the circle is changed continuously in size to take any value between zero and some upper limit specified by the analyst. For each distinct circular window, the method calculates the likelihood of observing the observed number of cases inside and outside the circle, assuming a Poisson (or Bernoulli) model for how the cases are generated and provide a significance value representing the detected cluster's "unusualness". So we have to calculate a likelihood ratio statistic testing the constant risk hypothesis versus the specific alternative that risk, within region having their centroid within the circle, differs from the risk in the rest of study area. We adopt the following notation: we define Y_{in} the number of cases observed in regions defining the circle of interest, E_{in} the number expected under the null hypothesis and with Y_{out} e E_{out} the corresponding cases observed and expected occurring outside the circle, respectively. The spatial scan statistic is proportional to

$$\max \left(\frac{Y_{in}}{E_{in}} \right)^{Y_{in}} \left(\frac{Y_{out}}{E_{out}} \right)^{Y_{out}} I(Y_{in} > E_{in}), \text{ where } I(.) \text{ is the indicator function.}$$

The aim is to find the area that maximizes the function L over all circles in the study region and this identifies the one that constitutes the most likely cluster. Its p -value is obtained through Monte Carlo hypothesis testing.

3. Cluster detection methods: new approaches and related tests

The major limitation of the method described in section 2 is related to the shapes of clusters: they are to be typical circular. However, in real situations, owing to social or environmental factors, we can frequently find spatial clusters with quite different shapes from circular one: such as disease concentrations along rivers, roads, etc. So, much of recent literature has been concerned with finding cluster of irregular shapes and more refined tests have been proposed as extensions of earlier method. Duczmal & Assunção (2004) introduced a simulated annealing spatial scan statistic to look for connected clusters with arbitrary shape. The theory and terminology behind simulated annealing come from statistical mechanics and it contemplates the analogy with the physical notation of temperature. The simulated annealing spatial scan statistic considers the centroids of every cell in the map as vertices of a graph, whose edges link cells with a common frontier. So the map of interconnected regions can be represented by a graph. For each vertice we know the population and the number of cases of the corresponding region. The collection of connected zones, with irregular shapes, consist of all those zones for which the corresponding subgraphs are connected. As to analyse the likelihood of $n \cdot 2$ subgraphs, which can arise from the initial graph of n vertices, is impractical, this strategy tries to visit only the most promising zone. One starts from some zone Z_0 (it is worth noting that the usual scan statistic method can also be used to provide an initial value for this algorithm), then the algorithm chooses some neighbour

Z_1 among all neighbours of Z_0 . In the next step, another Z_2 is chosen and so on. This strategy at each step chooses a new neighbour according to a temperature parameter which can take only three values: high, medium and low. In other words, there are three levels of randomness in the process of neighbour selection. With high temperature every neighbours has the same chance to be chosen. Instead, the lower is the temperature, the higher is the chance to choose a neighbour with a greater likelihood. When the temperature is low, only the neighbours with higher likelihood are selected. The temperature parameter is fixed according to the value of likelihood ratio and the number of times that the current subgraph has been visited before. The survey ends when some parameters exceed the thresholds defined within the algorithm. Tango and Takanashi (2005) introduce a flexible scan statistic imposing an irregularly shaped window on each region connecting its adjacent regions. They avoid to detect cluster of unlikely peculiar shape restricting the connected regions to a subsets of the set of region i . Their approach leaves to move the k connected regions from 1 to a prefixed maximum K . Another method to detect irregularly shaped clusters has been proposed by Patil et al. (2004). This procedure relies on the notion of “upper level set”. It is a data-driven way to select the list of candidate zone, in which the incidence rate is significantly elevated relative to the rest of region.

4. An empirical study and power evaluation: some results

In this section it will be of special interest to compare the effectiveness of simulated annealing scan statistic and the flexible scan statistic. Both these methods find the most likely cluster over the collection of all connected zone irrespectively of shape. However, as pointed out by Tango and Takanashi (2005), the simulated annealing scan statistic tends to detect a larger cluster than the true cluster by absorbing not only surrounding region where there is not-elevated risk but also faraway regions. In our work, we first illustrate the reviewed approaches to a real data set. We use a well-known data set concerning child mortality for 167 census districts, collected for a nine period for Auckland area (New Zealand). This data set is discussed in details by Bailey and Gatrell (1995).

Figure 1: *Most likely cluster using flexible spatial scan*

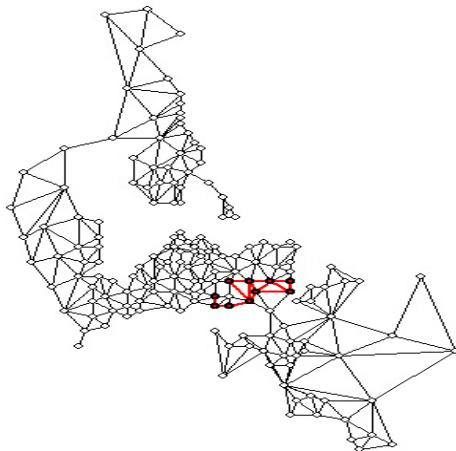
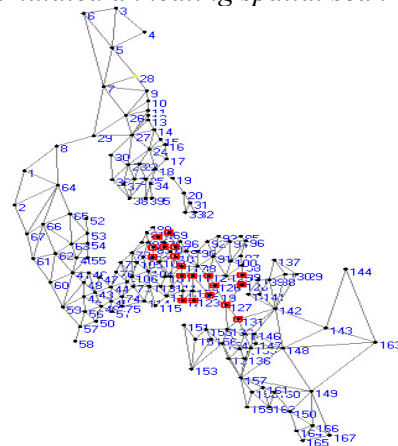


Figure 2: *Most likely cluster using the simulated annealing spatial scan*



As we expected, this application clearly shows that the simulated annealing scan statistic identifies larger clusters (as displayed in Fig.1 and Fig.2). The regions detected as the most likely cluster for the simulated annealing spatial scan are 20 whereas the census areas included in the resulting cluster of flexible scan are 10.

These results are obtained under 999 Monte Carlo replications. We note that SA spatial scan has the largest log-likelihood ratio (38.98). But for a formal comparison of test statistics it is important evaluate their power, carrying out a simulation study. In order to produce power calculation, we use benchmark dataset, based on the geography of the North-eastern United States and the female population of those counties. In this way, we take into account for uneven geographical population density. The benchmark data and how it was generated has been described in details elsewhere (Kulldorff et al. 2003). They contemplate clusters of different size and location. By using the same simulated data when comparing methods, the variance of power difference is kept to a minimum. The null data is used to estimate the critical value, which is the cut-off point for significance. Then the estimated power for an alternative multiple hot-spot cluster model is calculated as percentage of the number of simulated datasets for which the maximum LLR exceeds the critical value; the nominal level α is fixed to 0.05. Since we are making fair comparison of these tests we specified the same number of maximum cluster size to scan (we used $K=15$ for detecting small clusters). As regards the power performance, the findings show that the SA scan statistic deals better with very irregularly shaped cluster alternatives. In fact, when the simulated clusters are quite long and narrow, the SA scan statistic shows power around 0.85 while the Flexible Spatial Scan performs worst, attaining 0.76 or even 0.60. The simulation results are not displayed here.

References

- Bailey T, Gatrell A (1995) *Interactive Spatial Data Analysis*, Longman, Harlow.
- Besag J, Newell J. (1991) The detection of clusters in rare diseases, *Journal of the Royal Statistical Society, Series A*, 154, 143–155.
- Duczmal L., Assunção R. A (2004) Simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, *Computational Statistics & Data Analysis*, 45, 269–286.
- Kulldorff M. (1997) A spatial scan statistic, *Communications in Statistics*, 26, 1481–1496.
- Kulldorff M., Nagarwalla N. (1995) Spatial disease clusters: detection and inference, *Statistics in Medicine*, 14, 799–810.
- Kulldorff M., Tango T., Park P.P. (2003) Power comparison for disease clustering test, *Computational Statistics & Data Analysis*, 42, 665–684.
- Openshaw, S., Craft A.W., Charlton M., Birch J.M. (1988) Investigation of leukaemia clusters by use of a geographical analysis machine, *Lancet*, 1, 272–273.
- Patil G.P., Taillie C. (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots, *Environmental and Ecological Statistics*, 11, 183–197.
- Takahashi K., Yokoyama T., Tango T. (2004) *FleXScan: Software for the flexible spatial scan statistic*, National Institute of Public Health, Japan.
- Takahashi K., Tango T. (2005) A flexibly shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, 4, 4–11.