Statistica & Applicazioni - Vol. VIII, n. 1, 2010, pp. 87-103

DETECTING IRREGULAR SHAPED CLUSTERS VIA SCAN STATISTICS

Eugenia Nissi*
Annalina Sarra*

SUMMARY

The topic of this paper regards recent extensions of spatial scan statistics, widely used in public health research to test disease clusters and to identify their approximate locations. Despite its success, there is an important limitation associated with the traditional scan statistics: it depends on the use of circle shaped windows. As results, the identified regions are often not well localized. This limitation has motivated research aimed at developing new approaches which have the capability to detect clusters of irregular shapes. Two new techniques have been studied and compared: the spatial scan statistics, based on the graph theory, and the flexible scan statistics which imposes an irregularly shaped window. A computational study has been carried out to evaluate the effectiveness of these new approaches. A better understanding of the relative strengths and weakness of these two methods is essential to appropriate choices of methodology.

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. A wide variety of definitions can be put forward for clusters and clustering. Within spatial epidemiology, the terms cluster detection, clustering and spatial variation in risk are used to indicate different issues but the effort to distinguish them can often generate confusion (Diggle, 2000). One possible explanation is the lack of formal definition of cluster. The most basic meaning of cluster, without any assumptions about shape or form, is the following: ''Any area within the study region of significant elevated risk'' (Knox, 1989). Also the term clustering has different interpretations. We can consider different aspects of the analysis of clustering. According to Besag and Newell (1991) classification, any methods which address the issue of the location of putative clusters, are defined as specific or detection methods, whereas methods which are aimed at investigating a tendency to cluster (i.e. if cases are located close to each other no matter where they occur), are referred as general methods. From a statistical point of view, it is worth noting the importance of detection cluster methods in public health, as a part of geographical disease surveillance whose main aim is to identify health problems as

* Dipartimento di Metodi Quantitativi e Teoria Economica - Viale Pindaro, 42 - 65127 PESCARA (e-mail: nissi@dmqte.unich.it; asarra@dmqte.unich.it).

they occur and try to reduce or to remove their effects on the population by appropriate public health intervention. Furthermore, Wartengerg and Greenberg (1990) suggest that a proper classification of different clustering detection methods should consider the suspected mode of clustering in the data, which can describe two type of clusters: hotspot clusters and clinical clusters. A hotspot cluster is characterized by a uniform elevation in risk in a specific zone (for example around a point source of hazard) while an increase in incidence of disease over a small geographic region, coupled with a decline to the background disease rate across the rest of the study region, characterizes a clinical cluster. Many methods have been developed to detect spatial clusters of disease.

In this paper, we start with the popular health scan statistic method to detect a local excess of events and to test if such excess can reasonably have occurred by chance. As known, its major limitation is that it is circle-based. Research in this area is ongoing and new and refined methods have been proposed. Some recent developments of the traditional scan statistic take into account that clusters can be of any shape and cannot be captured only by circles. So, we consider and compare new approaches and their related tests, which have the capability to detect arbitrarily, shaped, hotspots. The main focus of this study is on carrying out a simulation experiment to evaluate the power performance of these new approaches in delineating irregular shaped clusters when different alternatives for non-circular cluster are taken into account. We believe that this simulation study can give useful insights on the ability of each of considered methods in detecting clusters of peculiar shape, highlighting their relative strengths and weakness.

The paper is structured as follows: in Section 2, we first review the background methods for detection clusters, we then focus on the spatial scan statistic (Section 3) and its recent extensions (Section 4). In Section 5, we first illustrate these reviewed approaches to a real data set and then present the results of simulation study carried out in order to evaluate the power of the new methods in detecting clusters of irregular shapes. Section 6 contains some concluding remarks.

2. AN OVERVIEW OF CLUSTER DETECTION APPROACHES

In this section, we turn attention to statistical cluster detection methods: they provide estimates of the likely clusters geographical location and extent without any previous knowledge of either how many or where they are. The goal for these methods is to identify areas with unusually high (or perhaps unusually low) local disease incidence rates. Generally, these techniques deal with the following situation: a region study is tessellated into cells and disease data are available in the form of nonnegative counts on cells. The cell counts are independent random variables while the cell sizes are regarded as known and fixed. As known, using disease counts arises an inferential complication. The spatial analyses of regional counts are subjects to the modifiable area unit problem (MAUP): different aggregations of individuals in different set of areas can change observed associations between variables (Openshaw, 1984).

Among this category of methods, we briefly review those superimposing a number of circular windows in the study region and determine the significance of the number of cases that fall within each circle. One can distinguish three methods which define the circles in terms of distance, in terms of number case and in terms of population size.

Openshaw, Craft, Charlton and Birch (1988) propose a method which superimposes a regular grid on the study region and circles of constant radius are drawn on the intersections of gridlines. This method, known as GAM (Geographical analysis machine), proceeds by examining a large number of overlapping circular regions and noting those with particularly high rates, that are drawn on the map because incidence proportions exceeding some user-specified threshold. Notwithstanding it is a good descriptive tool, the statistical foundation of this procedure has been long criticised. The main drawback is related to the multiple testing problem. The statistical criticism of GAM, has motivated the development of several statistical approaches seeking to provide local inference based on GAM-type operations.

So, Besag and Newell (1991) develop a method to rectify the principal problems of Openshaw's procedure. They limit attention to circles containing a constant number of cases, while GAM contemplate circles with different number of cases and different population at risk. That is, they consider the numerator of incidence rate to be fixed. The first step is to select a constant cluster size, say h. The study region is divided in a set of I areas $i = 1, 2,..I$. For each area i, they order the remaining areas by increasing distance of their centroids to the centroid of area. Then it is necessary to determine the number of areas L that must be added to area i to include the nearest h cases to the centroid of area i. A small observed value l of L indicates that an area centroid has h cases nearby and may indicate a cluster. Also this method should be intended as a screening device, as it does not use adjustment for multiple testing and the related results depend on the choice of cluster size h. Anyway, in comparison with the previous described approach, Besag and Newell's procedure has the advantage of not requiring arbitrary aggregations of areas and presents an higher probability to detect cluster in zone with low population density.

3. THE SCAN STATISTIC: THE BASIC THEORY

As previously pointed out, both the approaches described in section 2 have to be viewed as descriptive tools as they are unable to do inference on individual cluster in order to determine if the areas have an excess rate that is statistically significant or not. It would be valuable to complement them considering a procedure which not require any assumptions about crucial and subjective parameters from the researchers.

An appropriate and popular method is the spatial scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997, 1999). 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 on an *ad hoc* test statistic. Moreover,

it does not perform a separate test for each possible cluster location or each possible cluster size. The scan statistic assesses the null hypothesis of constant risk: people are likely to contract the disease regardless of location; in other terms under the null hypothesis there are no clusters of cases. The scan statistic has to take into account the geometry of area being scanned, the probability distributions generating events under the null hypothesis, the shapes and sizes of scanning window. This procedure uses a particular alternative hypothesis with an excess risk in a circular clusters. We remind that the advantage of having a well-specified alternative hypothesis is that it gives some information about the alternative for which the test can be expected to have good power. However it does not mean that it can be only used to detect such alternative hypotheses.

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 traditional scan statistic imposes a circular window on a map and lets its centre move across the study region. For aggregated data a candidate cluster is a set of sub-regions whose centroids are within each scanning circle.

For any given position of centre of the circular window, the radius of the circle is changed continuously in size. We could include all potential scanning zones of any size at every location, but we may want to specify some upper limit for the maximum cluster size, since a too large size would not be useful. The standard analytic option is that the window never includes major than 50% of total population at risk. Other choices of the maximum are also possible. The study region is partitioned into N geographic sub-divisions called cells. For each cell, we have a count y_i $(i = 1, 2, ..., N)$ the number of disease cases, and an underlying baseline e_i $(i = 1, 2, ..., N)$, which traditionally corresponds to the population at risk or may be an estimate of the expected value of the count.

Following the traditional framework in this field, under the null hypothesis we assume that the counts y_i are independent Poisson random variables

$$y_i \sim \text{Poisson}(r \cdot e_i)$$

where r represents the disease rate and e_i is the baseline. Other probability models have also been used for scan statistics. Note, if we have binary counts, for example individuals with or without some disease, we should assume a Bernoulli model for how the cases are generated. Basically, this technique consists of two steps: first to find the most significant spatial regions and then calculate the statistical significance of these regions, i.e. provide a significance value representing the detected cluster's ''unusualness''. In order to have a meaningful notion of significance, we must have a model of underlying process which have generated data. In particular, we have two models: a model for data generated under the null hypothesis of no cluster and a model for data obtained under the alternative hypotheses, each of them represents clustering in some regions. The null hypothesis signifies complete spatial randomness, with each individual equally likely to be a case: the disease rate r is uniform everywhere. Under the alternative hypothesis $H_1(S)$ r is higher inside region S than

outside the region, which corresponds to a risk of contracting the disease for people living inside zone S higher than outside the zone. Hence the system of hypotheses can be stated as follows:

$$H_0(S_i) : r_i = r$$
$$H_1(S_i) : r_i > r$$

The statistic used for finding regions where the disease rate is higher inside region S than outside is the likelihood ratio:

$$L(S) = \frac{P(Data|H_1(S))}{P(Data|H_0)}$$

Using maximum likelihood estimates of r_{in}, r_{out} and r_{all},which represent the risk inside region S, the risk outside the region S and the risk everywhere, respectively, we derive

$$L(S) = \frac{\left(\frac{y_{in}}{e_{in}}\right)^{y_{in}} \cdot \left(\frac{y_{out}}{e_{out}}\right)^{y_{out}}}{\left(\frac{y_{all}}{e_{all}}\right)^{y_{all}}} \text{ if } \frac{y_{in}}{e_{in}} > \frac{y_{out}}{e_{out}} \tag{1}$$

and $L(S) = 1$ otherwise.

Details about derivation of likelihood ratio tests, has been given elsewhere (Kulldorff, 1997). In the expression (1), we have:

$$y_{in} = \sum\nolimits_S y_i$$

$$y_{out} = \sum\nolimits_{S^c} y_i$$

$$y_{all} = \sum\nolimits_{S \cup S^c} y_i$$

and similarly

$$e_{in} = \sum\nolimits_S e_i$$

$$e_{out} = \sum\nolimits_{S^c} e_i$$

$$e_{all} = \sum\nolimits_{S \cup S^c} e_i$$

The aim is to find the region S^* that maximises the function L over all circles in the study region. The area associated with the maximum value of the likelihood ratio test statistic identifies the most likely cluster. So, we have to individuate the highest scoring region $S^* = \arg\max_S L(S)$ of the study region and its score $L^* = L(S^*)$. The statistical significance of this region is usually evaluated through Monte Carlo sampling (Dwass, 1957). As there are no known asymptotic or approximate solutions

for most disease cluster problems owing to uneven geographical population densities, the p-value is obtained in this way: random data sets are generated under the null hypothesis and the value of the scan statistic is calculated for both the real data and the simulated random data sets. Then the rank of maximum likelihood from the real data set is compared with the maximum likelihood from the random data sets.

If this rank is R, then the p-value is defined as

$$p = \frac{R}{1 + \ of \ replications \ under \ H_0}.$$

If this p-value is less than some threshold (e.g 0.05) we can conclude that the discovered region is a significant spatial cluster and not ''just a simply chance occurrence''; otherwise no significant clusters exist. Besides the most likely cluster, the scan statistics allows to identify secondary clusters in the data set. Secondary clusters are almost identical to the most likely cluster and provide little additional information but their existence means that the exact boundaries of the identified cluster remains uncertain. For this reason it is better to speak of cluster's approximation. Another issue to clarify is if the scan statistic can adjust for the multiple p-values obtained for the secondary clusters detected. The secondary clusters are irrelevant for the rejection of null hypothesis: we only need to know the likelihood of the most likely cluster of real data set and compare to those obtained from the random data sets. Furthermore, as concern the pinpointing the specific cluster causing rejection, the most likely cluster is clearly so strong to cause rejection; hence we are still concerned to reject the null hypothesis rather than doing multiple tests.

4. INNOVATIONS OF CIRCLE-BASED SPATIAL SCAN STATISTICS

The circular spatial scan statistic is routinely used in a variety of disciplines (epidemiology, medical imaging, astronomy, urban and regional planning) in which it maintains its statistical validity and meaningful interpretation. We highlight that with other suitable modifications, the scan statistic approach can be used for critical area analysis in much more fields.

Starting from its first formulation, the circular spatial scan statistic has been updated and enlarged to embrace new issues such as the introduction of an elliptic window (Kulldorff, Huang, Pickle, Duczmal, 2006), the extension to ordinal data (Jung, Kulldorff, Klassen 2007), the analysis of survival data (Huang, Kulldorff, Gregorio, 2007) and the possibility to employ it to simultaneously search cluster in more than one dataset (Kulldorff, Mostashari, Duczmal, Yih, Kleinman, Platt, 2007).

We know that the major limitation of the method is related to the shape of clusters: they are to be typically 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. (Verkasalo, 1993). So, much of recent literature has been concerned with finding cluster of irregularly shapes and more refined tests have been proposed as extensions

of earlier method. Two of these alternative approaches are discussed in detail in the next section.

4.1 *Simulated Annealing Spatial Scan Statistic*

Duczmal and Assunção (2004) introduce a simulated annealing spatial scan statistic to look for connected clusters with arbitrary shape. As known, the simulated annealing is a heuristic method used in operations research to find solutions in optimisation process avoiding to find local maximum or minimum. 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 is a strategy for looking for connected clusters with arbitrary shape using a graph algorithm. In fact 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 mathematically represented by a graph. In that graph, each region is associated to a node and if two regions are neighbours, i.e. they are connected by a segment, there is an edge in the graph linking the corresponding two nodes. For each node we know the population and the number of cases of the corresponding region. We call *zone* a connected subset of regions of map. For each zone there is a corresponding connected sub-graph of the map graph. The collection of connected zones, with irregular shapes, consists of all those zones for which the corresponding sub-graphs are connected. The goal is to identify, among all possible zones, that ones which maximise the likelihood ratio. As to analyse the likelihood of 2^n sub-graphs, 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 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. Notice that most of time this rule allows us to choose for the highest likelihood valued neighbours but it can also adopt a less deterministic decision. In order to impose limits to the choices of the procedure, we have to establish a rule by which to determine whether the current sub-graph at iteration n+1 should be accepted or rejected compared to the sub-graph at iteration n and a rule by which to change the temperature parameter. The temperature parameter is set up according to the value of likelihood ratio and the number of times that the current subgraph has been visited before.

So, for a proper selection of the successor of current sub-graph at each step of the strategy, it is crucial to define:

- high Likelihood function value, say hL: it is a flag variable indicating if it was found a neighbour with higher likelihood value at the current step;
- the number of consecutive steps, denoted by *cs,* before finding new subgraphs with hL value >1;
- the number of times the current subgraph has been visited before in the survey *(vb)*;
- the number *cv* of common vertices between the current subgraph and the highest yet valued on in the survey.

At each step the strategy checks the above parameters and the survey ends when at least one of the two parameters (*vb* and/or *cs*) exceed the thresholds defined within the algorithm. These thresholds, necessary to modify dynamically the process of selection of the successor of the current subgraph and set up a stopping criterion to the search, are defined as follows:

- the threshold for the number of consecutive steps before finding new subgraphs with hL value >1, is denoted by cs_threshold and is equal to *cv* (cs_threshold=cv);
- a second threshold for *cs* is denoted by cs_threshold_2 and is equal to cv/2 (cs_threshold_2=cv/2);
- the threshold for the number of times the current subgraph has been visited before in the survey *(vb)* is denoted by vb_threshold and it is fixed and was empirically determined and is between 6 and 10 for most situations, as shown in Duczmal and Assunção's paper (2004);
- a second threshold for *vb* is denoted by vb_threshold_2 and is equal to vb_threshold/2 (vb_threshold_2= vb_threshold/2).

The basic idea of how this procedure works can be summarised in the steps described below:

- select an initial connected subgraph G;
- find the set (say N(G))of all connected subgraphs neighbours of G;
- compute the likelihood for all new subgraphs in N(G);
- compute hL, cs, vb ,cv , cs_threshold, cs_trhreshold_2;

if (cs>cs_threshold_2) then the current subgraph G is identified through the most random strategy; otherwise if (hL=0) and (vb>vb_threshold_2) the current subgraph is identified through the strategy with medium level of temperature; otherwise if (hL=0) or (vb>vb_threshold_2) the current subgraph is chosen through the most determinist strategy. At the end we adopt a fourth strategy to use when the current subgraph is a very promising one. Such a strategy augments the set of vertices of the current subgraph near the places where there have been recent improvements of current subgraph. We will go on in this survey until cs is not bigger than cs_threshold and vb is not bigger than vb_threshold. As seen, the fourth types of strategies illustrated contemplate different levels of randomness. Summing up, we can say that:

- the most random strategy is adopted when the current subgraph has been visited many times, has a relatively low value for the likelihood and for several steps of the survey we have not an increase of the likelihood value;

- the strategy with medium level of temperature is assumed when there are very similar conditions with the previous strategy: the only exception is there has been an increase of the likelihood values for some subgraphs lately visited;
- the most deterministic strategy is applied when has been recorded an increase of the likelihood values for some recently surveyed subgraph and at least one of these conditions are true: the current subgraph has been visited many times or it has a relatively low value for the likelihood;
- the fourth strategy is finally adopted when the current subgraph has a relatively high value for the likelihood and for this reason it can be considered a ''very promising one''; it has not been visited many times and there have been an increase of the likelihood values for some recently surveyed sub-graph.

The simulated annealing is generally a computer intensive method: hundreds of thousand iterations are often necessary to find the most likely cluster. Hence the choice of initial subgraph is important. It is worth noting that the traditional scan statistic method can be used to provide an initial value for this algorithm.

4.2 *Flexible Spatial Scan Statistic*

Another recent proposal, conceived to identify irregularly shaped clusters, is due to Takanashi and Tango (2005) who introduce a flexible scan statistic. Such a strategy imposes an irregularly shaped window on each region connecting its adjacent regions and leaves to move the k connected regions from 1 to a prefixed maximum K. Notice that we need to specify a maximum number of region K to be included in the cluster (i.e. is a maximum length of cluster). In this way the anomalous aggregation is restrained to a relatively small neighbourhood of each region avoiding detecting a cluster of unlikely peculiar shape. More formally, while for any given region i the traditional spatial scan statistic consider k concentric circles, then all the windows to be scanned are included in set

$$Z_1 = \{z_{ik} | 1 \leq i \leq m,\ 1 \leq k \leq K\}$$

where m denotes the number of regions in which the entire study area is divided.

By contrast, for the flexible scan statistic all the windows to be analysed are included in the larger set

$$Z_2 = \left\{z_{ik(j)} | 1 \leq i \leq m,\ 1 \leq k \leq K,\ 1 \leq j \leq j_{ik}\right\}$$

So this new approach considers k concentric circles plus all the set of connected regions whose centroids are located within the k^{th} largest concentric circle. The underlying algorithm of this procedure can be conveniently illustrated as follows. At the beginning, we need to define a $m \times m$ matrix $\mathbf{A}$, which elements a_{ij} are equal to 1 if regions i and j are connected otherwise $a_{ij} = 0$. We also set $Z_2 = \phi$ and $i_0 = 0$. Hence we set $i_0 = i_0 + 1$ as starting region where $i_0 = \{1, 2, ..., m\}$ and create the set of $k - 1$ nearest neighbours, say W_{io}.

Then we have to consider all the set $(Z \subset W_{io})$ which include the starting region *i*. For any given such set Z we need to repeat the following steps:

1. divide Z in two disjoint set: $Z_0 = \{i_0\}$ and Z_1 which contains the other regions of Z
2. create two new sets: Z_0 which consists of regions of Z_1 that are connected to some regions of Z_0 and Z_1 which consists of the regions of Z_1 that are not connected to any regions of Z_0 and replace Z_0 and Z_1 by Z_0 and Z_1 respectively
3. repeat recursively the previous step until Z_0 or Z_1becomes null first.

At this point one can establish a decision rule like that:

''*when Z_1 becomes null first Zis said to be connected and added to the set Z_2 by contrast when Z_0 becomes null first Z is said to be disconnected and discarded''*.

Finally another method to detect irregular shaped clusters has been proposed by Patil and Tallie (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. The upper level set scan statistics allows for flexible shape of cluster and differently to the traditional scan statistic, applicable only to tessellation data, it can be also applied to responses defined on a network (stream network, highway system, water distribution network etc...) as this procedure relies on an abstract graph. Moreover Patil and Tallie (2004) suggest way to address another limitation of traditional scan statistic: response distributions have been taken as discrete (specifically binomial or Poisson). In order to correctly incorporate this aspect, their approach is illustrated for the gamma family of distributions.

5. AN EMPIRICAL STUDY AND POWER EVALUATION: SOME RESULTS

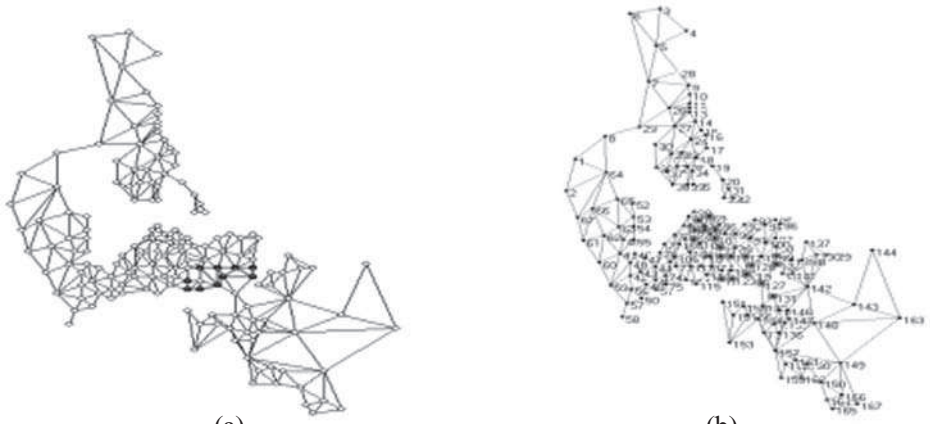
In this section it will be of special interest to compare the effectiveness of simulated annealing scan statistic and the flexible scan statistic. We aim to achieve a better understanding of properties of methods under considerations and their ability to detect true clusters. Both find the most likely 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 use a real dataset to illustrate the new reviewed approaches and related tests. The data refer to a child mortality cases among 167 census district in Auckland (New Zealand) as recorded over a nine period (1977-1985). The region involved covers approximately 5000 square kilometres. In particular this data, exploited in details by Bailey and Gatrell (1995), include the number of death in children under five years old in each small area during the nine period, together with the population of children under five as recorded in the 1981 census.

The results for the flexible spatial scan and simulated annealing spatial scan are summarized in Table 1.

TABLE 1. - *Results for flexible spatial scan and simulated annealing spatial scan*

	Flexible scan statistic	Simulated annealing scan statistic
Census areas included	10	20
Cases	126	214
LLR	19.28	38.98
Monte Carlo replications	999	999
P-value	0.01	0.01



(a) (b)

FIGURE 1. - *Most likely cluster using flexible spatial scan* (a) *and the simulated annealing spatial scan* (b)

As we expected, this application clearly shows that the simulated annealing scan statistic identifies larger clusters (as displayed in Figure 1 (a-b)). The regions detected as the most likely cluster for the simulated annealing spatial scan are 20 whereas the census area included in the resulting cluster of flexible scan are 10. These results are obtained under 999 Monte Carlo replications. We note that the simulated annealing spatial scan has the largest likelihood ratio (38.98).

For a formal comparison of test statistics it is important to evaluate their power, carrying out a simulation study. For a rigorous evaluation of statistical power of these detection methods, we use benchmark datasets, based on the geography Northeastern United States and the female population of those counties. Each of 245 counties is geographically represented by a centroid coordinate. In this way, we take into account for uneven geographical population density. Benchmark datasets, described in details in Kulldorff, Huang, Pickle, Duczmal (2006), with a random number of cases of hypothetical disease, were considered under either the null model of no clusters of cases or some alternative models, including cluster of different size and location. Simulated annealing spatial scan and flexible scan were then run on these datasets and the resulting power performances were analysed and compared. We have to

point out that we make use of the algorithm implemented in C++ code, obtained from the authors, for the simulated annealing spatial scan statistic whilst we employ the software FleXScan also provided by the authors for the flexible scan statistic.

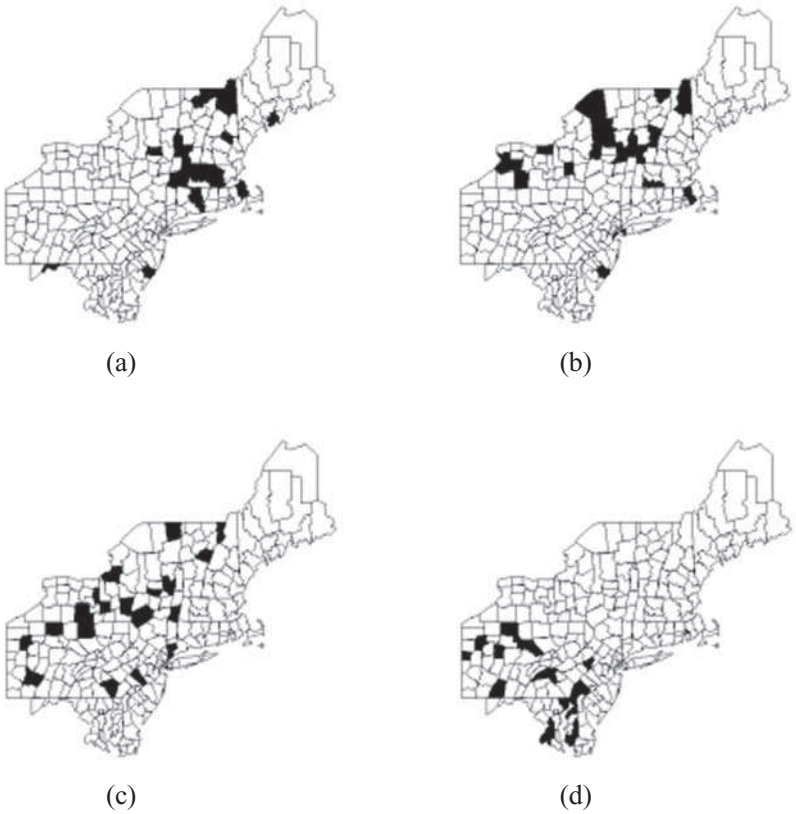
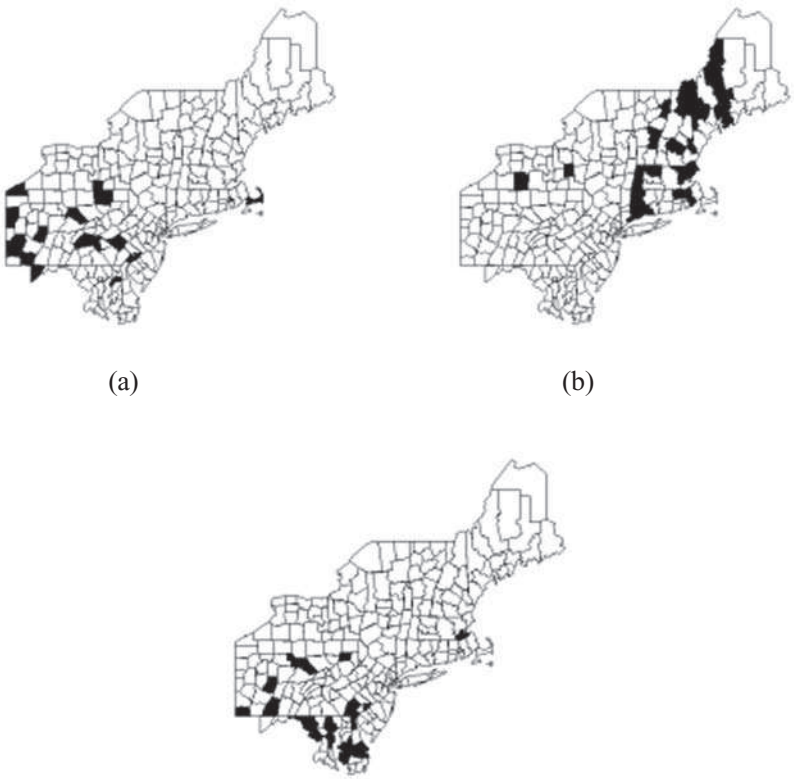


FIGURE 2. - (a) *Hotspot alternative **A**: simulated data clusters along Connecticut River;* (b) *Hotspot alternative **B**: simulated data clusters along Hudson River;* (c) *Hotspot alternative **C**: simulated data clusters along Lake Ontario Coast;* (d) *Hotspot alternative **D**: simulated data clusters along West/Lower Susquehanna River*

As known, one goal of developing benchmark datasets is to enable quick and simple comparison. Moreover, by using the same simulated data when comparing methods, the variance of power difference is kept to a minimum. The benchmark datasets include 100000 datasets generated under the null model each person living in the north-eastern of United States counties is equally to contract the disease. Accordingly, under the null model, 600 cases are randomly assigned to counties with probabilities proportional to population of areas. The benchmarks datasets also contem-

plate datasets generated under eleven alternative hypotheses. For every different cluster model 10000 datasets are available. From these hotspot alternatives, which contemplate irregularly shaped clusters, we consider only seven of eleven simulated shaped alternative models, whose geographical meaning is shown in Figure 2 (a-d) and Figure 3 (a-c). They have been chosen with the purpose of testing the two methods for some very irregular cluster shapes.



(a) (b)

(c)

FIGURE 3. - (a) *Hotspot alternative **E**: simulated data clusters along Susquehanna River;* (b) *Hotspot alternative **F**: simulated data clusters along New England Coast;* (c) *Hotspot alternative **G**: simulated data clusters along Pennsylvania External Border*

The counties within each cluster were given a higher risk than remaining counties. The relative risk is equal to one for every cell outside the cluster. The null data are used to estimate the critical value which is the cut-off point for significance. So, the critical value corresponding to a 0.05 significance level was computed by identi-

fying the 5000th highest maximum LLR from among the 99999 random data sets, generated under the null model. Then the estimated power alternative multiple hotspots was calculated as percentage of 10000 random datasets for which the maximum LLR exceeds the critical value. As specified, previously, the flexscan approach needs to specify a pre-set maximum number of regions K to be included in the cluster. Since we are making fair comparisons of these tests, we declare the same number of maximum cluster size to scan. The choice of upper bound K depends on disease map under study. Anyway this algorithm reveals suitable in detecting moderate cluster size (i.e. the length of true cluster would not be larger than 10-15% of total of number regions). We fixed $K = 15$ for detecting small clusters. The simulation results for the seven alternative models considered are provided in Table 2.

TABLE 2. - *Power evaluation for cluster alternative A-G*

Hotspot Cluster alternatives	**A**	**B**	**C**	**D**	**E**	**F**	**G**
# counties	13	16	7	15	21	23	26
Flexible scan statistic (α=0.05)	0.604	0.4892	0.7684	0.5585	0.439	0.2836	0.2896
Simulated annealing scan statistics (α=0.05)	0.86	0.84	0.84	0.90	0.85	0.61	0.63

As regards the power performance, the findings show that the simulated annealing spatial scan statistic deals better with very irregularly shaped cluster alternatives. The simulated annealing spatial scan shows power constantly above the 80% for the cluster A,B,C,D,E. The worst performance is for the cluster alternative denoted by F (0.61) and by G (0.63). Instead, the flexible scan statistic does not exhibit such high power performances. The procedure of Takahashi and Tango (2005) works well for small cluster size: for the alternative C (with 7 counties included) it attained 0.76. However in the other non circular hot-spot alternatives the power obtained is lower and it performs worse compared with the simulated annealing scan statistic. In particular, we can observe a significantly diminished power when cluster are shaped in twisted long strings as in the peculiar hot spot F e G.

6. CONCLUDING REMARKS

In this work we explored and compared two innovations of the circle based spatial scan statistic popular in health sciences. We found that the simulated annealing scan statistic showed higher power in detecting non circular hot-spot clusters considered. However, the simulated annealing scan statistic tends to detect larger cluster than the true one, particularly for very unusual cluster shapes. For this reason, there is the need to use some penalty for very irregular shapes in order to give to researchers insight of geographic cluster delineation.

Recently, Duczmal, Kulldorff and Huang (2006) and Duczmal, Cançado, Takahashi, Bessegato (2007) propose two modified versions of simulated annealing spatial scan statistic. The first modification incorporates the concept of ''non-compactness'' and penalises clusters that are very irregular in shape. The last reformulation refers to a genetic algorithm which allows to define spatial aggregations with less arbitrarily shape and requires a less computational task. As regards the flexible scan statistic, we remark its ability to detect small cluster. For larger cluster sizes we found that the method has lower power. Accordingly, we need a more efficient algorithm to detect cluster as it becomes not feasible when the size cluster increases. For future purposes, there are other issues to take into account. A detailed investigation of these cluster detection methods should use other alternative cluster models as the benchmark datasets here considered represent only a subset of potential geographical feature of disease outbreaks. For example, rather than a sudden increase in relative risk level in the cluster area, one could construct outbreak models in which the relative risk increase gradually.

Finally, we point out that a new methodology, which can be viewed as a hybrid of simulating annealing scan statistic and flexible scan statistic, is now available. Yiannakoulias, Rosychuk and Hodgson (2007) propose the Greedy Growth Search (GGS) technique which involves to set two parameters for monitoring the geometric shape of clusters and to take under control the elaboration time. So, it could be of some interest to compare the previous findings with those achievable with this hybrid procedure.

RIASSUNTO

Il lavoro si propone di confrontare alcune recenti estensioni del metodo della scan statistics che risultano particolarmente utili per l'individuazione e la localizzazione di clusters di malattia. La scan statistics, come è noto, procede con l'individuazione di clusters attraverso una finestra circolare che determina alcune limitazioni soprattutto nella fase di localizzazione della regione a rischio identificata. L'obiettivo del lavoro è quello di confrontare due metodi alternativi a quello tradizionale quali la spatial scan statistics basata sulla teoria dei grafi e la flexible scan statistics le cui potenzialità sono valutate attraverso un esperimento simulativo.

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.

Diggle P.J. (2000). Overview of statistical methods for disease mapping and its relationship to

cluster detection. In Elliott P., Wakefield J., Best N., Briggs D.J. (Eds.) *Spatial Epidemiology: Methods and Applications* (pp. 87-103), Oxford University Press.

Dwass M. (1957). Modified randomisation tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, **28**, 181-187.

Duczmal L., Assunção R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, **45**, 269-286.

Duczmal L., Kulldorff M., Huang L. (2006). Evaluation of spatial scan statistics for irregularly shaped disease clusters. *Journal of Computational and Graphical Statistics,* **15**(2), 428-442.

Duczmal L., Cançado A.L.F., Takahashi, R.H.C., Bessegato L.F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis,* **52***,* 43-52.

Glaz J., Naus J., Wallenstein S. (2001). *Scan Statistics.* Springer, New York.

Huang L., Kulldorff M., Gregorio D. (2007). A spatial scan statistic for survival data. *Biometrics*, **63**(1), 109-118.

Jung I., Kulldorff M., Klassen A. (2007). A spatial scan statistic for ordinal data. *Statistics in medicine*, **26**(7), 1594-1607.

Knox E.G. (1989). Detection of clusters. In P. Elliot (Ed.), *Methodology of enquiries into disease clustering*, Small Area Health Statistics Unit, London.

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

Kulldorff M. (1997). A spatial scan statistic. *Communications in Statistics*, **26**, 1481-1496.

Kulldorff M. (1999). *Spatial scan statistics models, calculations and applications*. In J. Glaz and N. Balakrishnan (Eds.), *Scan Statistic and Applications* (pp. 302-322). Boston, Birkhauser.

Kulldorff M., Tango T., Park P.P. (2003). Power comparison for disease clustering test. *Computational Statistics & Data Analysis*, **42**, 665-684.

Kulldorff M., Huang L., Pickle L., Duczmal L. (2006). An elliptic spatial scan statistic, *Statistics in medicine*, **25**, 3929-3943.

Kulldorff M. and Information Management services, Inc. (2006). SaTScanTM ver.7.0, Software for the spatial and space-time statistics, http://www.satscan.org.

Kulldorff M., Mostashari F., Duczmal L., Yih K., Kleinman K., Platt R. (2007). Multivariate Scan Statistics for disease surveillance. *Statistics in Medicine*, **26**(8), 1824-1833.

Openshaw S. (1984). *The Modifiable Areal Unit Problem*. Geo Book. Norwich, UK.

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., Tango T. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4-11.

Takahashi K., Yokoyama T., Tango T. (2004). FleXScan: Software for the flexible spatial scan statistic. *National Institute of Public Health*, Japan.

Verkasalo P.J. (1993). Risk of cancer in Finnish children living close to powerlines. *British Medical Journal*, **307**, 895-899.

Wartenberg D., Greenberg M. (1990). Detecting disease cluster: the importance of statistical power. *American Journal of Epidemiology*, **132**, S156-S166 (supplement).

Yiannakoulias N., Rosychuk R.J., Hodgson J. (2007). Adaptations for findings irregularly shaped disease clusters. *International Journal of Health Geographics*, 6-28.