(to appear in Journal of Computational and Graphical Statistics)

Evaluation of spatial scan statistics for irregularly shaped disease clusters

Luiz Duczmal[1], Martin Kulldorff[2], Lan Huang[3]

[1] Statistics Department, Universidade Federal de Minas Gerais, Brazil

[2] Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care

[3] Division of Cancer Control and Population Sciences, National Cancer Institute

ABSTRACT

Spatial scan statistics are commonly used for geographic disease cluster detection and evaluation. We propose and implement a modified version of the simulated annealing spatial scan statistic that incorporates the concept of "non-compactness" in order to penalize clusters that are very irregular in shape. We evaluate its power for the simulated annealing scan and compare it with the circular and elliptic spatial scan statistics. We observe that, with the non-compactness penalty, the simulated annealing method is competitive with the circular and elliptic scan statistic, and both have good power performance. The elliptic scan statistic is computationally faster and is well suited for mildly irregular clusters, but the simulated annealing method deals better with highly irregular cluster shapes. The new method is applied to breast cancer mortality data from northeastern United States.

Keywords: Disease clusters, Elliptic scan, Non-Compactness penalty, Power evaluation, Simulated annealing scan, Spatial cluster detection.

1. Introduction

In disease surveillance, the spatial scan statistic is often used for the detection and evaluation of geographic clusters (Kulldorff 1997). The objective is to identify spatial concentrations of individuals with a particular disease condition (Hsu et al. 2004; Sheehan et al. 2004; Odoi et al. 2004; Heffernan et al. 2004; Washington et al. 2004; Andrade et al. 2004). The methods assume that we have at our disposal a map of locations, each with a defined risk population and a certain number of observed cases. The spatial scan statistic is a maximum likelihood ratio test statistic, which finds the maximum likelihood ratio when sweeping over all zones circumscribed by a window with variable size and shape (Kulldorff 1997). For each zone, the likelihood ratio is computed by counting the observed number of cases inside and outside that zone. The zone that maximizes the likelihood ratio defines the most likely cluster. This window may be of any shape, including circular (Kulldorff et al. 1995), elliptic (Kulldorff et al. 2005) or irregular (Duczmal et al. 2004, Patil et al. 2004a,b, Tango et al. 2005). Although many clusters may be approximately circular, it is very common to find irregularly shaped clusters arising in real situations, such as disease concentrations along rivers, roads or shores. A pollutant plume or the extent of underwater contamination may assume an irregular geographic shape (Biggeri et al. 1996; Katsouyanni et al. 1991; Xu et al. 1989).

[1] Corresponding author: duczmal@est.ufmg.br, Universidade Federal de Minas Gerais, Statistics Dept., Campus Pampulha, Belo Horizonte, MG 31270-901 Brazil. Phone:55-31-3499-5900 fax:55-31-3499-5924

The simulated annealing scan statistic (Duczmal et al. 2004) tries to find the most likely cluster over the collection of all connected zones, irrespectively of shape. The unlimited geometric freedom of cluster shapes could lead to low power of cluster detection. This happens because the algorithm tends to find tree-shaped clusters that merely link the highest likelihood ratio cells of the map. These large and extremely high-value tree-shaped clusters obfuscate the somewhat lower-value clusters that contain the real geographic meaning we are seeking. These latter clusters are in general less irregularly shaped than the tree-shaped ones. In this paper, we propose and implement a modification of the simulated annealing scan statistic that incorporates the concept of geometric "non-compactness." It uses a modified maximum likelihood function (Good et al. 1971, Ogata et al. 1988) that penalizes irregular shaped clusters, generalizing an idea that was used for the special case of ellipses (Kulldorff et al. 2005). We make power evaluations for this new simulated annealing scan statistic, and compare it with the circular and elliptic scans. The simulated annealing technique using non-compactness correction is applied to breast cancer mortality data in the northeastern United States for 1988-1992 (Kulldorff et al. 1997). The spatial scan statistic is presented in section 2, and we describe its variations briefly for the elliptic scan and the simulated annealing scan. In section 3, we define the non-compactness penalty criterion for irregularly shaped spatial clusters. We show an application for breast cancer mortality in section 4. The power evaluations are presented in section 5. Section 6 contains some final remarks. The software in C language is available as freeware from the corresponding author.

2. The Spatial Scan Statistic

Spatial scan statistics can be used for data with exact point locations or for aggregated data. In this paper, we only consider the latter. A study region is partitioned into m cells, with a total of C disease cases, and total population N. A zone is *connected* when there is a transitive neighborhood relationship between any two cells belonging to it; in other words, this zone can be described as a one-piece geometric object. The zone z may be formed by all the cells whose centroids are within a circle of specified center and radius, or within some ellipse with given size, position, eccentricity and orientation, or simply a connected set of cells with irregular geometric format. We assume that under the null hypothesis, there is no cluster in the map, and the number of cases in each cell is Poisson distributed, with expected values proportional to its population size. We define $L(z)$ as the likelihood, under the alternative hypothesis, that there is a cluster in the zone z, and L_0 the likelihood under the null-hypothesis.

The zone z with the maximum likelihood ratio is defined as *the most likely cluster*. If μ_Z is the expected number of cases inside the zone z under the null hypothesis, and c_Z is the number of cases inside z, it can be shown that

$$LR(z) = \frac{L(z)}{L_0} = \left(\frac{c_Z}{\mu_Z}\right)^{c_Z} \left(\frac{C - c_Z}{C - \mu_Z}\right)^{C - c_Z}$$

when $c_Z > \mu_Z$, and one otherwise (Kulldorff 1997). This likelihood ratio test, being maximized over all the zones, identifies the zone that constitutes the most likely cluster (Kulldorff 1997). The test statistic is $\max_Z LR(z)$, where the maximization is done over some collection of zones. Usually we refer to the logarithm of *LR(z)*, or *LLR(z)*. The collection of zones determines when one uses a circular, elliptic, irregular or some other type of spatial scan statistic. A Monte Carlo procedure is used to evaluate the statistical significance of the cluster found (Dwass 1957).

2.1. The Elliptic Scan Statistic

The elliptic spatial scan statistic has been described in detail elsewhere (Kulldorff et al. 2005). An ellipse is defined by the x and y coordinates of its centroid, and its size, shape and angle of the inclination of its longest axis. The shape is defined as the ratio of the length and width of the ellipse. For a given map, we define a finite collection of ellipses E as follows. For computational reasons, the shapes s in E are restricted to 1, 1.5, 2, 3, 4, 5, 6, 8, 10, 15 and 20. A finite set of angles is chosen such that we have an overlapping of about 70 percent for neighboring ellipses with the same shape, size and centroid. The ellipses´ centroids are set identical to the cells´ centroids in the map. We choose a finite number of ellipses whose sizes define uniquely all the possible zones z formed by the cells in that map whose centroids lie within some ellipse of the subset. The collection E is thus formed grouping together all these subsets, for each cells´ centroid, shape and angle. We further define $E(s)$ as the subset of E that includes all the shapes listed above in this section up to and including s. The choice of the collection E and its associated collection of zones is done beforehand and only once for a given map. The spatial scan statistic is thus applied to the collection of zones defined by E.

2.2. The Simulated Annealing Scan Statistics

It is useful to treat the centroids of every cell in the map as vertices of a graph, whose edges link cells with a common boundary. For the simulated annealing (SA) spatial scan statistic, the collection of connected irregularly shaped zones consists of all those zones for which the corresponding subgraphs are connected. This collection is very large, and it is impractical to calculate the likelihood for all of them. Instead we shall try to visit only the most promising zones, as follows (see Duczmal et al.(2004) for details). The zones z and w are *neighbors* when only one of the two sets $w-z$ or $z-w$ consists of a single cell. Starting from some zone z_0, the algorithm chooses some neighbor z_1 among all the neighbors of z_0. In the next step, another neighbor z_2 is chosen among the neighbors of z_1, and so on. Thus, at each step we build a new zone adding or excluding one cell from the zone in the previous step. It is only required that there is a maximum size for the number of cells in each zone (usually half of the total number of cells). Instead of always choosing the highest LR neighbor at every step, the SA algorithm evaluates if there has been little or no LR improvement during the latest steps; in that case, the algorithm opts for choosing a random neighbor. This is done while trying to avoid getting stuck at LR local maxima. We restart the search many times, each time using each individual cell of the map as the initial zone. Thus, the effect of this strategy is to keep the program openly exploring the most promising zones in the configuration space and abandoning the directions that seems uninteresting. The best solution found by the program is called a *quasi-optimal solution*, and for our purposes, it is a compromise due to computer time restraints for the identification of the geographical location of the clusters.

3. Simulated Annealing Scan Statistic with Non-Compactness Penalty

With the simulated annealing scan statistic, we frequently end up with a zone that is nothing more than the collection of the highest incidence cells in the map, linked together forming a "tree-shaped" cluster spread through the map; the associated subgraph resembles a tree, except possibly for some few additional edges. Figure 1 shows an example of a tree-shaped cluster found by the "pure" SA algorithm. This cluster has 122 counties spread across the 245 counties of New England, with random allocated cases under the null hypothesis. Clearly this cluster does not add new information with regard to its special geographical significance in the map. One easy way to avoid that problem is simply to set a smaller upper bound to the maximum number of cells within a zone. As we shall see later in the numerical simulation results section, this approach is only effective when cluster size is rather small (i.e., for detecting those clusters occupying roughly up to 10% of the cells of the map). For larger upper bounds in size, the increased geometric freedom favors the occurrence of very irregularly shaped tree-like clusters similar to the one in Figure 1, thus impacting the power of detection.

We will now explore another elegant and effective way to deal with this problem. Instead of allowing shapes with unlimited freedom, as is the case for the tree-shaped clusters, we need to have some shape control for the zones that are being analyzed. We would like to penalize the zones in the map that are highly irregularly shaped. For this purpose, we define the *compactness* of a planar geometric object z as

$$K(z) = \frac{4\pi A(z)}{H(z)^2},$$

where $A(z)$ is the area of z and $H(z)$ is the perimeter of the convex hull of z. Intuitively, the convex hull of a planar object is the cell inside a rubber band stretched around it. A circle is the shape with maximum compactness, namely 1. Compactness is dependent on the shape of the object, but not on its size. Compactness also penalizes a shape that has small area compared to the area of its convex hull. Rewriting the definition above as

$$K(z) = A(z) \Big/ \pi \left(\frac{H(z)}{2\pi} \right)^2$$

we can interpret $K(z)$ as the area of z divided by the area of the circle with perimeter $H(z)$. A square has compactness $\pi/4 = .785$. A $a \times 1$ rectangle has compactness $\pi a/(1+a)^2$. The compactness for some $a \times 1$ ellipses and other geometric shapes are given in Figure 2. A rounded 2×1 rectangle (i.e., a square with two semi-circles at the tips) has compactness 0.849. Finally, observe that a cluster formed by the cells whose centroids lie within a circle may have compactness less than one, because its cells may form a zone whose shape is not circular.

The idea of using a penalty function for spatial cluster detection, based on the irregularity of its shape, was first used for ellipses (Kulldorff et al. 2005), although a different formula was employed. We would like to set the likelihood ratio $LR(z)$ of a zone z to 1 when the compactness of z is close to zero, and leave the likelihood ratio unaltered when the compactness of z is 1. We then substitute the likelihood ratio $LR(z)$ with the expression $LR(z)^{K(z)}$.

3.1. Approximating Area and Perimeter

Two technical issues are related to the computation of the compactness of a zone z, namely the computation of $A(z)$ and $H(z)$. The area $A(z)$ of each cell is generally available to the practitioner. The perimeter $H(z)$ of the convex hull of z is computed as follows. Define C_0 as the set of centroids of the cells in z, and C_1 as the set C_0 augmented with the centroids of the neighbors of the cells in z. Let $H_0(z)$ and $H_1(z)$ be respectively the perimeter of the convex hull of C_0 and C_1. We now define $H(z)$ as the average of $H_0(z)$ and $H_1(z)$, as exemplified in Figure 3. The convex hull computation for C_0 and C_1 is made using the Quickhull algorithm (Schneider et al. 2003). Numerical tests suggest that the complete procedure for the SA method with the computation of $K(z)$ consumes only 25% more time than the "pure" SA, thus making it feasible for practical use.

3.2. A General Definition of Compactness

The compactness concept may be generalized, using the formula $K(z)^a$, with $a > 0$, instead of $K(z)$. The expression $LR(z)^{K(z)^a}$ may be used as the likelihood test function instead of the likelihood ratio $LR(z)$. With $K(z)^a$ and $a < 1$, (medium compactness correction) those shapes that depart more from a disk are less penalized than using the usual definition of $K(z)$. When $a \to 0$, the corresponding test function $LR(z)^{K(z)^a}$ tends to the original test function $LR(z)$. Values of a greater than 1 penalize strongly any irregular shape. When $a \to \infty$, only circular shapes are admitted.

4. The Breast Cancer Clusters in Northeastern United States

An application of the compactness corrected SA scan statistics described in the previous sections is done using 1988-1992 breast cancer mortality data from northeastern United States. This data set has been studied using the circular spatial scan statistic (Kulldorff et al. 1997) and the elliptic spatial scan statistic (Kulldorff et al. 2005). The northeastern U.S. map contains 245 cells (counties and county equivalents) in ten states and the District of Columbia, with a total population of 29,535,210 women with 58,943 breast cancer deaths. The mortality rate was 39.91 per 100,000 women per year. A Poisson model was used for the observed number of deaths in each cell. The analyses are adjusted for age applying indirect standardization with 18 distinct five-year age groups: 0-4, 5-9,…, 80-84, and 85+. Figure 4 displays the top 10% highest *LLR* and the top 10% highest incidence isolated cells of the map.

The most likely clusters found by the SA scan statistic are displayed in Figure 5A-D, for the adopted values ***a***=0 (no compactness correction), ***a***=0.25 and ***a***=0.5 (medium compactness correction), and ***a***=1 (full compactness correction). We used an upper limit of 13% (32 counties) for the number of cells in the clusters. It was chosen as a compromise between the extremes used in the power evaluation of section 5, and the clusters obtained seem to exhibit just enough flexibility using this upper limit. We also display in Figure 5E-F, for comparison, the clusters obtained by the circular scan algorithm (Kulldorff et al. 1997) and the elliptic scan algorithm (Kulldorff et al. 2005). Table 1 summarizes these results. It can be observed that for the highest values of ***a***, the most likely clusters assume a less irregular shape. The relatively compact most likely cluster of Figure 5D is formed by joining the highest *LLR* cells present within the circular cluster of Figure 5E. The irregularly shaped most likely clusters of Figure 5A-C seem to be formed by trying to gather some of the highest *LLR* isolated cells, and is more extensive for the lowest values of ***a*** (compare with Figure 4 at left). This gathering process occurs chiefly for the highest *LLR* cells inside the cluster of figure 5E. The p-value of the most likely clusters found were in every case 0.001, from 999 Monte Carlo replications of the null-hypothesis model for each value of ***a***.

5. Power Evaluation

In this section we build the alternative cluster model for the execution of the power evaluations, and compare the power for the circular, elliptic and simulated annealing methods.

5.1. Irregularly Shaped Cluster Alternatives

We compare the performance of the circular, elliptic and SA scan statistics. A benchmark dataset is constructed using real data population from the northeastern U.S. with one of 11 simulated irregularly shaped clusters A to K, displayed in Figure 6. Those irregularly shaped hot-spot clusters are briefly described in the accompanying table, which lists its geographic meaning. The third column in the table within Figure 6 shows the values of $K(z)$ for the simulated clusters in the study region. These clusters were chosen with the purpose of testing the circular, elliptic and SA algorithms for some very irregular cluster shapes. From now on, the clusters A to K will be denoted *real* clusters, in contrast to the *detected* clusters found by the scan statistics. For each simulation of data under these eleven alternative hypotheses, 600 cases are distributed randomly according to a Poisson model using a single cluster; we set a relative risk equal to one for every cell outside the real cluster, and greater than one and identical in each cell within the cluster. The relative risks are defined such that if the exact location of the real cluster was known in advance, the power to detect it should be 0.999 (Kulldorff et al. 2003). For each upper limit of the detected cluster size and each compactness correction parameter ***a***, 100,000 runs of the SA statistic were done under the null hypothesis, plus 10,000 runs for each entry in the table, under the alternative hypothesis.

5.2. Power Evaluation Results

The estimated power is presented in Table 2, for the circular, elliptic and the SA scan statistics and for all 11 alternative cluster models.

5.2.1 Power Results for the Circular and Elliptic Scans

The optimal power of the circular and elliptic scan was above 0.83 for clusters A-E, I and K, and below 0.75 for the remaining data with clusters F, G, H and J. The performance was very poor on simulated data with cluster G, and the optimal power achieved was only 0.61, using the maximum shape parameter 20. Similar comments apply for clusters F, H and J, with optimal power about 0.70 and maximum shape parameters 4 and 8. Better power was not achieved when we increased the maximum elliptic shape to 20 for these data. When clusters are shaped as twisted long strings, the elliptic scan tended to detect only straight pieces within them: this phenomenon was observed in clusters F, G, H and J, resulting in diminished power. Otherwise, when a cluster fits well within some ellipse of the set, best power results were attained, as observed for the remaining clusters.

5.2.2 Power Results for the Simulated Annealing Scan

As expected, the power is low when the upper limit for the detected cluster size is set to 50% of the number of cells, without compactness correction (SA(122) and ***a***=0 in Table 2). The maximum power for each real cluster is attained when the maximum allowed size for the detected clusters roughly matches the size of the real cluster in the map, as can be seen in the lines of Table 2 when ***a***=0.

When ***a***=1 (full compactness correction), we can compare the values in the third line of each SA group and verify that for the higher upper bounds in size, the power values roughly increase when the real cluster compactness increases. It performs significantly better compared with the pure SA scan (***a***=0) with the same upper bound for size.

In some instances, the subgraph associated with a cluster may become disconnected when we remove one of its vertices. If the cell associated with this vertex is significantly less populated than the average, it is denoted a *weak link cell.* The string-like clusters F, G, H, I and J have weak links cells. With 600 cases distributed randomly through the map, it happens that the weak link cells end up with zero cases very often. This occurs 35% of the times with the string-like shaped cluster F, and more than 90% of times with cluster G. As a result, the SA method may not detect the whole cluster because, with zero cases in a weak link cell, it is difficult to bind the two isolated pieces of the split cluster, thus diminishing the power. Most cells within this cluster are connected to only two neighbors, and the presence of a weak link impacts considerably the cluster detection. To a lesser extent, similar comments also apply to the clusters H, I and J. Examples like these, exhibiting at the same time shape irregularity and unevenly distributed populations, represent very severe instances of characteristics that could be attributed to clusters, and they are presented here to test the limits of the power of detection of the algorithms. Clusters F and G scored the worst power performance, respectively 0.61 and 0.63. Clusters H, I and J attained nearly 0.70, and cluster K attained 0.75. For the remaining clusters, the SA scan statistic showed power consistently above 0.85.

When ***a***=0.5 (medium compactness correction), the power results are situated between the full and pure algorithms, as can be seen in the second lines for each SA group. Additional simulations were done using ***a***=0.25, but the power values are virtually the same as those obtained using ***a***=0, and are not shown. Values of ***a*** greater than one were also tested but found to be very shape restrictive. Even values as low as ***a***=1.5 produce clusters that are not very different from the circular ones, and they are not displayed here.

5.3. Power Comparison between the Circular, Elliptic and Simulated Annealing Scans

The power values for the statistics analyzed here are not very different. The power performance was good on all statistics for clusters A-E and K. When a large fraction of a cluster fits within some ellipse, the elliptic scan works well. The circular scan works well, except for cluster G. Generally, when there are no weak links the SA scan also works well. The worst comparative performances for the SA method, against the other methods, occurred in clusters F and I, according to the previous section. Each one of these two clusters fits reasonably well within some ellipse, and both have weak links. Clusters G, H and J perform poorly on the elliptic and SA scans because they have weak links and also do not fit well within any ellipse.

6.Conclusions

We compared different spatial scan statistics regarding their power to detect irregular shaped clusters. The power performance of the new compactness corrected SA scan is significantly better than the previous uncorrected SA scan. The elliptic scan is well suited for those clusters that fit well within some ellipse. The SA compactness-corrected scan performs well when there are no weak links. One possible solution for this problem is to pre-process the entire map in advance, aggregating very dissimilarly populated neighbor cells according to some appropriate criteria, thus minimizing the occurrence of weak links. The circular, elliptic and SA scans have similar power in general. The elliptic scan method is computationally faster and is well suited for mildly irregular-shaped cluster detection, but the non-compactness-corrected SA scan detects clusters with every possible shape, including the highly irregular ones. The choice of the statistic depends on the initial assumptions about the degree of shape irregularity to allow, and also on the availability of computer time.

The power analyses show that the circular scan statistic works very well even for non-circular clusters, so most public health departments may not want to bother with the more computer-intensive spatial scan statistics for routine daily surveillance activities. While multiple smaller circular clusters from the circular spatial scan statistic could be used to approximate an irregularly shaped cluster, it is not clear which of those circles should be used, as it is not clear which collection provides the maximum likelihood or penalized maximum likelihood. For more in-depth epidemiological research about the geographical distribution of a specific disease, the elliptic and/or simulated annealing-based scan statistics have an important purpose, as they do a better job at detecting and delineating irregular shaped clusters. It is important though to use some form of penalty on very irregular shapes. For the SA scan statistic, we recommend using a penalty with $a = 1$ when we want to have the ability to detect large as well as small clusters. If we are only interested in detecting very small clusters using a very restrictive upper limit on the cluster size, then the use of this penalty function is much less important.

Acknowledgements

The authors gratefully acknowledge the comments from the editor and the reviewers. We also thank Ms. Judith Swan (NHI/NCI) for the grammatical revision of the text. This research was supported in part by NCI grant number ROICA095979.

References

1. Glaz J., Naus J. I., Wallenstein S. (2001) Scan Statistics. *Springer Verlag: New York*.
2. Kulldorff M.(1997) A spatial scan statistic. *Comm. Stat. - Theo. Meth.*, 26**:**1481-1496.
3. Hsu C., Jacobson H., Soto Mas F. (2004) Evaluating the disparity of female breast cancer mortality among racial groups - a spatiotemporal analysis, *lnt. J. of Health Geographics*, 3:4
4. Sheehan T. J., DeChello L. M., Kulldorff M., Gregorio D. I., Gershman S., Mroszczyk M. (2004) The geographic distribution of breast cancer incidence in Massachusetts 1988 to 1997, adjusted for covariates. *Int. J. of Health Geographics*, 3:17
5. Odoi A., Martin S.W., Michel P., Middleton D., Holt J., Wilson J. (2004) Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *Int. J. of Health Geographics*, 3:11
6. Heffernan R., Mostashari F., Das D., Karpati A., Kulldorff M., Weiss D. (2004) Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases*, 10:858
7. Washington C.H., Radday J., Streit T.G., Boyd H.A., Beach M.J., Addiss D.G., Lovince R., Lovegrove M.C., Lafontant J.G., Lammie P.J., Hightower A.W. (2004) Spatial clustering of filarial transmission before and after a Mass Drug Administration in a setting of low infection prevalence. *Filaria J* May 5, 3:3
8. Andrade L.S.S., Silva S.A., Martelli C.M.T., Oliveira R.M., Morais Neto O.L., Siqueira Júnior J.B., Melo L.K., Di Fábio J.L. (2004) Population-based surveillance of pediatric pneumonia: use of spatial analysis in an urban area of Central Brazil. *Cadernos de Saúde Pública,* 20:2, 411-421
9. Kulldorff M., Nagarwalla N. (1995) Spatial disease clusters: detection and inference. *Statistics in Medicine,* 14:779-810.
10. Kulldorff M. (1999) Spatial scan statistics: Models, calculations and applications. In *Scan Statistics and Applications*, Glaz and Balakrishnan (eds.). Boston: Birkhauser, 303-322.
11. Kulldorff M., Huang L., Pickle L., Duczmal L. (2005) An elliptic spatial scan statistic. S*ubmitted.*
12. Good I.J., Gaskins R.A. (1971) Nonparametric roughness penalties for probability densities, *Biometrika*, 58:255-257.
13. Ogata Y., Katsura K. (1988) Likelihood analysis of spatial inhomogeneity for marked point patterns, *Annals of the Institute of Statistical Mathematics*, 40:29-29
14. Duczmal L., Assunção R. (2004) A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis*, 45:269-286.
15. Patil, G.P., Taillie, C. (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics,* 11, 183-197.
16. Patil, G.P., Bishop, J.A., Myers, W.L., Taillie, C., Vraney, R., Wardrop. D. (2004) Detection and delineation of critical areas using echelons and spatial scan statistics with synoptic cellular data. *Environmental and Ecological Statistics,* 11, 139-164.
17. Tango T., Takahashi K. (2005) A flexibly shaped spatial scan statistic for detecting clusters. International Journal of Health Geographics, 4:11.

18. Biggeri A., Barbone F., Lagazio C., Bovenzi M., and Stanta G. (1996) Air pollution and lung cancer in Trieste: spatial analysis of risk as a function of distance from sources. *Environmental Health Perspectives*, 104:750-754.
19. Katsouyanni K., Trichopoulos D., Kalandidi A., Tomos P., Riboli E. (1991) A case-control study of air pollution and tobacco smoking in lung cancer among women in Athens. *Preventive Medicine*, 20:271-280.
20. Xu Z.Y., Blot W.J., and Xiao H.P. (1989) Smoking air pollution and the high rates of lung cancer in Shenyang, China. *Journal of the National Institute of Cancer*, 81: 1800-1806.
21. Kulldorff M., Feuer E.J., Miller B.A., Freedman L.S. (1997) Breast cancer clusters in the Northeast United States: a geographic analysis. *American Journal of Epidemiology*, 146:161-170.
22. Dwass M. (1957) Modified randomization tests for nonparametric hypotheses. *Ann. Math. Stat.*, 28:181-187.
23. Schneider P.J., Eberly D.H. (2003) *Geometric Tools for Computer Graphics*, Elsevier Science.
24. Kulldorff M., Tango T., Park P.J. (2003) Power comparisons for disease clustering tests. *Computational Statistics and Data Analysis*, 42:665-684.

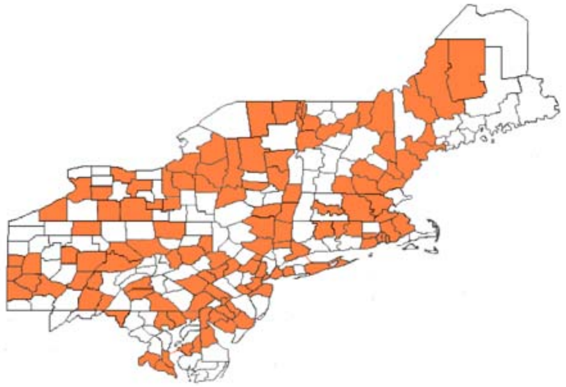


Figure 1. A very irregularly shaped connected cluster (with 122 counties, or 50% of the study region) found by the simulated annealing algorithm without the compactness penalty. The simulated cases were distributed randomly in the study region under the null hypothesis.

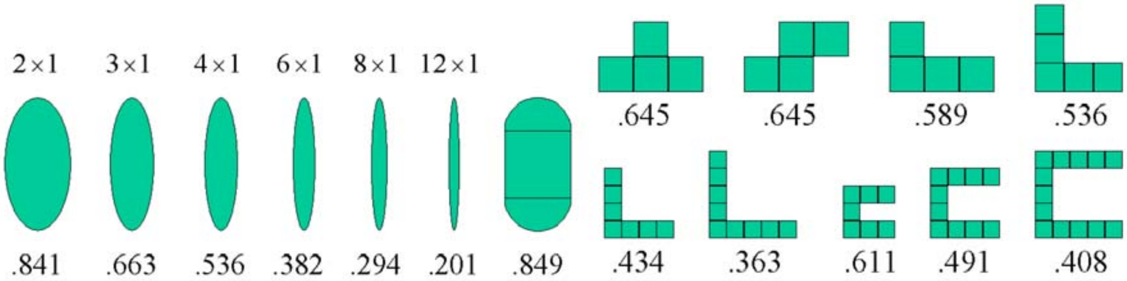


Figure 2. Compactness values for some simple geometric forms. From the left, are depicted ellipses with relative proportions 2×1, 3×1, 4×1, 6×1, 8×1 and 12×1, a rounded 2×1 rectangle (i.e., a square with two semi-circles at the tips), and some non-convex geometric shapes.

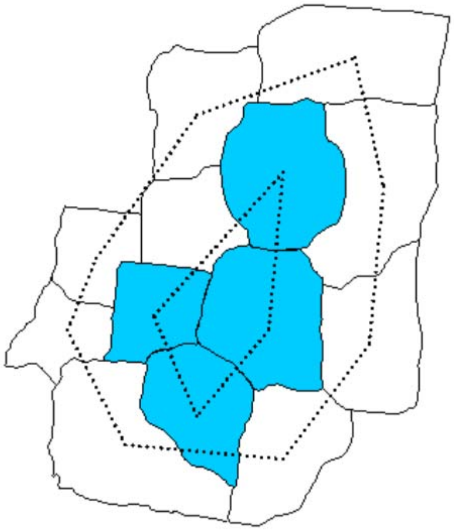


Figure 3. Approximating the perimeter of the convex hull of the shaded zone z as the average of $H_0(z)$ and $H_1(z)$ (shown as dotted polygons).

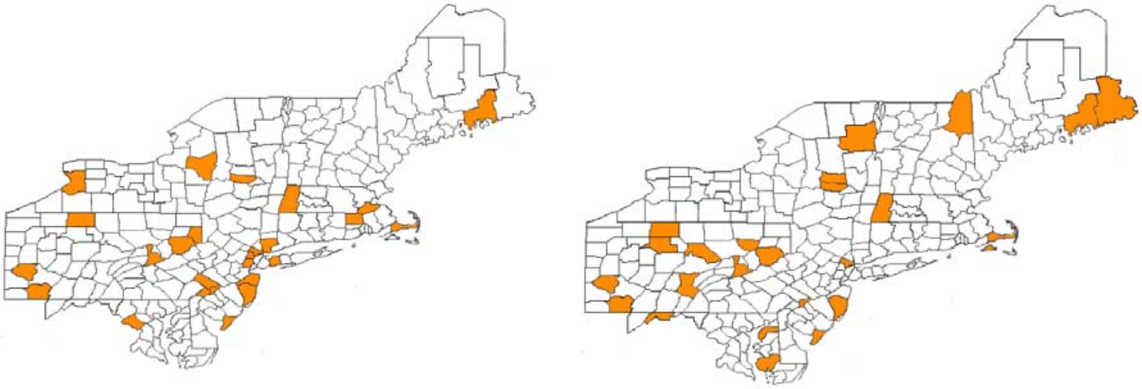
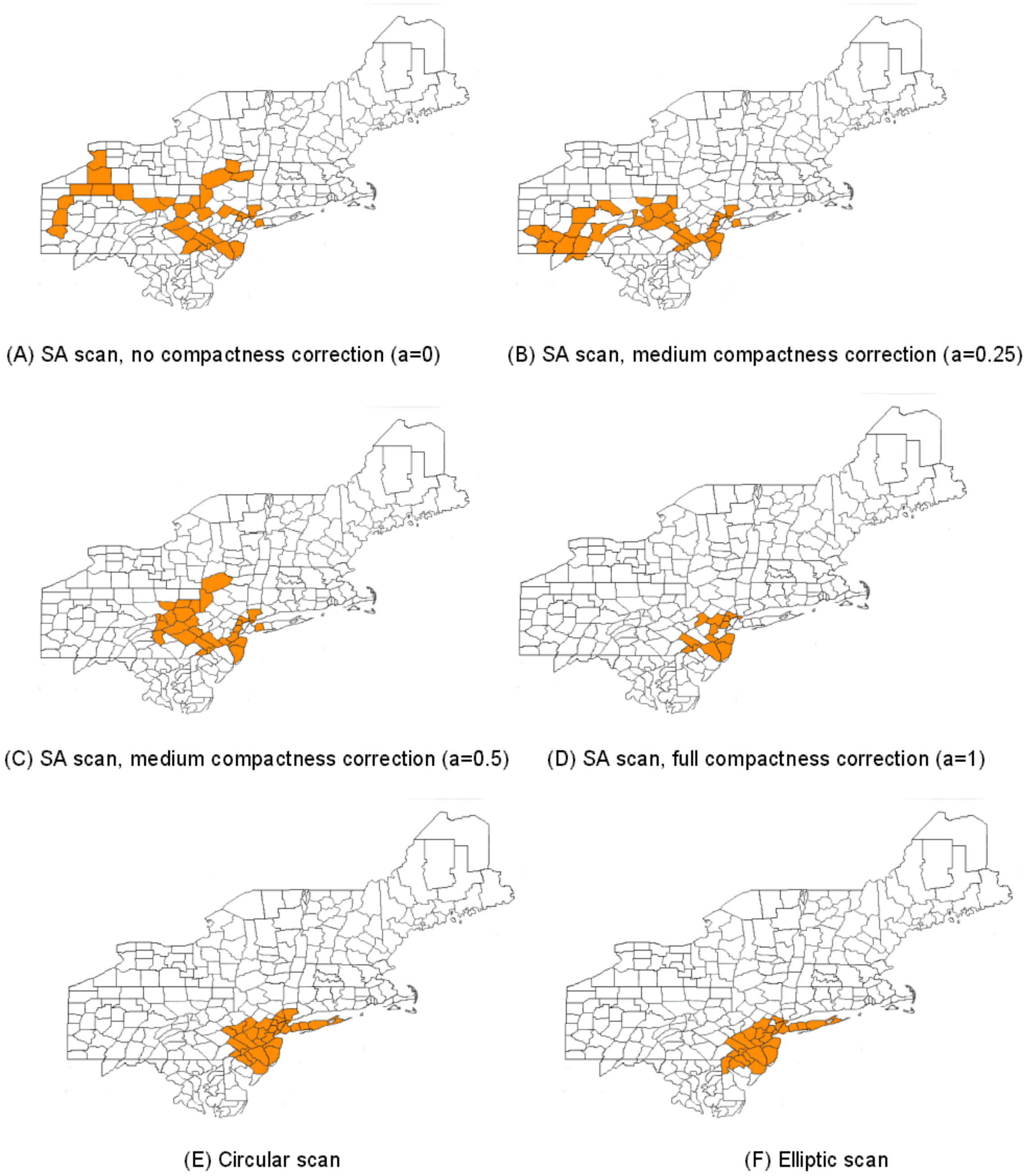


Figure 4 - Top 10% highest likelihood ratio (left) and the top 10% highest incidence (right) isolated cells of observed deaths of breast cancer in the northeastern U.S., 1988-1992.



(A) SA scan, no compactness correction (a=0)

(B) SA scan, medium compactness correction (a=0.25)

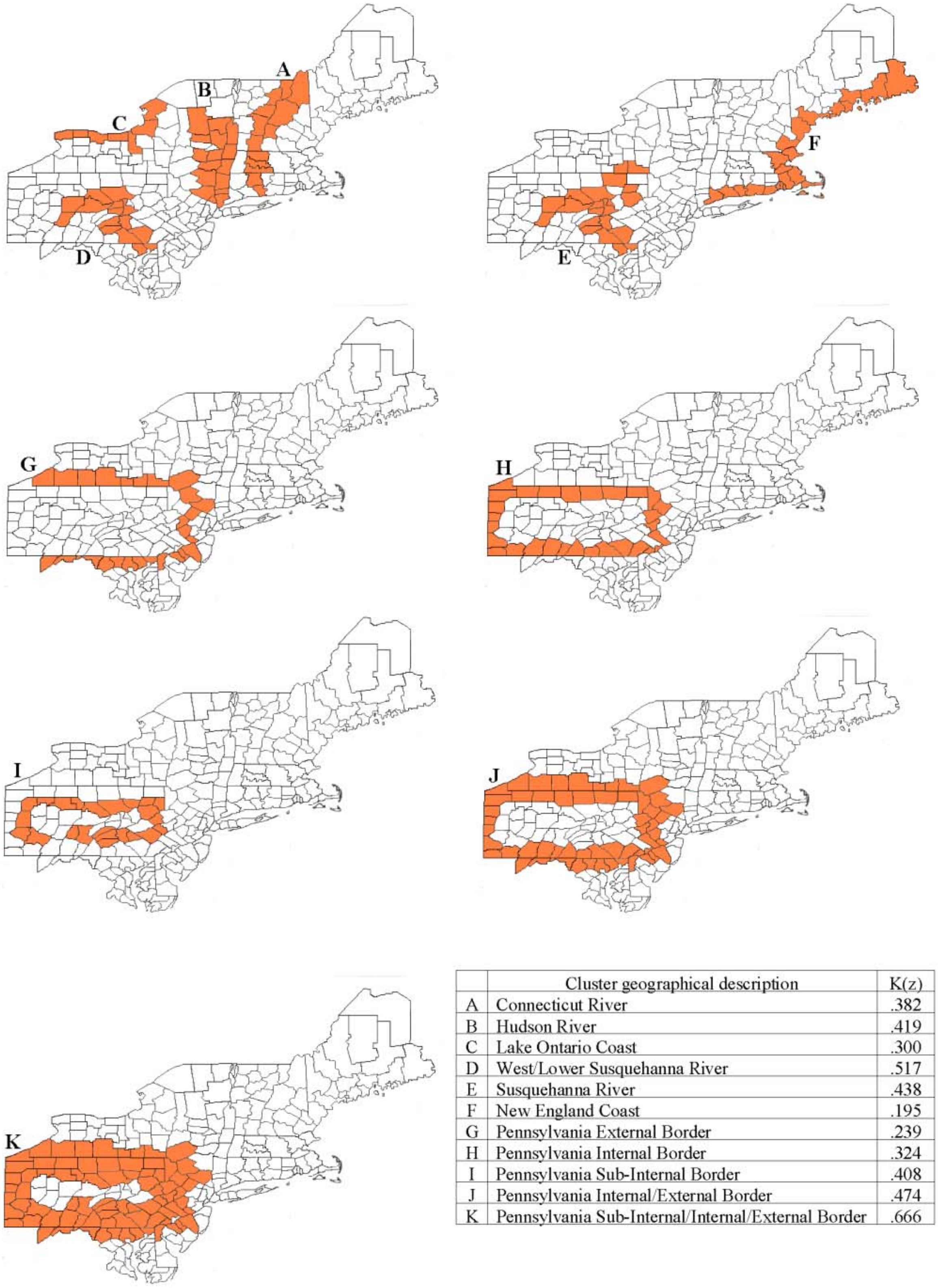
(C) SA scan, medium compactness correction (a=0.5)

(D) SA scan, full compactness correction (a=1)

(E) Circular scan

(F) Elliptic scan

Figure 5 – Clusters found by the compactness corrected simulated annealing scan (A-D), the circular scan (E) and the elliptic scan statistics (F) of observed deaths of breast cancer in the northeastern U.S., 1988-1992.



	Cluster geographical description	K(z)
A	Connecticut River	.382
B	Hudson River	.419
C	Lake Ontario Coast	.300
D	West/Lower Susquehanna River	.517
E	Susquehanna River	.438
F	New England Coast	.195
G	Pennsylvania External Border	.239
H	Pennsylvania Internal Border	.324
I	Pennsylvania Sub-Internal Border	.408
J	Pennsylvania Internal/External Border	.474
K	Pennsylvania Sub-Internal/Internal/External Border	.666

Figure 6. Simulated data clusters for the northeastern U.S. The clusters A-K, whose geographical meaning and compactness are given on the table, were used in the power evaluations.

cluster	***a***	cells	relative risk	$K(z)$	LLR	p-value
5A	0.0	32	1.144	0.26	177.3	0.001
5B	0.25	32	1.150	0.28	157.8	0.001
5C	0.5	27	1.138	0.43	101.4	0.001
5D	1.0	13	1.129	0.57	51.4	0.001
5E	-	30	1.044	0.59	35.7	0.001
5F	-	26	1.080	0.57	40.5	0.001

Table 1 – The clusters of observed deaths of breast cancer in figure 5.

real cluster		A	B	C	D	E	F	G	H	I	J	K
#counties		13	16	7	15	21	23	26	29	23	55	78
$K(z)$		.38	.42	.30	.52	.44	.20	.24	.32	.41	.47	.67
Circular	E(1)	.85	.79	.88	.86	.81	.70	.46	.66	.77	.68	.80
Elliptic	E(2)	.88	.83	.89	.89	.85	.72	.52	.69	**.83**	.71	**.84**
Elliptic	E(4)	**.89**	**.84**	**.90**	**.90**	**.86**	**.75**	.56	.71	.83	.72	.82
Elliptic	E(8)	.89	.82	.90	.90	.85	.75	.59	**.72**	.82	**.72**	.81
Elliptic	E(20)	.88	.80	.90	.88	.83	.73	**.61**	.71	.81	.70	.78
SA(8)	a=0	**.87**	.83	**.87**	.89	.82	.58	.61	.69	.65	.60	.51
	a=.5	.86	.82	.83	.89	.80	.50	.56	.68	.64	.59	.50
	a=1	.86	.78	.79	.85	.72	.44	.49	.62	.59	.54	.43
SA(12)	a=0	.86	.84	.84	.90	.85	**.61**	**.63**	**.70**	.67	.66	.60
	a=.5	.86	.85	.81	**.91**	.85	.52	.59	.70	.65	.65	.60
	a=1	.86	.84	.79	.89	.81	.45	.52	.67	.64	.63	.55
SA(20)	a=0	.79	.81	.77	.88	**.87**	.59	.62	.69	**.69**	.69	.68
	a=.5	.82	.84	.73	.90	.87	.51	.57	.68	.67	.68	.68
	a=1	.84	**.86**	.74	.90	.86	.46	.52	.67	.66	.67	.66
SA(30)	a=0	.66	.74	.65	.81	.84	.51	.59	.67	.67	**.70**	.72
	a=.5	.72	.79	.64	.85	.86	.47	.54	.63	.65	.67	.71
	a=1	.79	.83	.65	.87	.86	.44	.50	.64	.65	.67	.71
SA(61)	a=0	.30	.40	.31	.45	.56	.28	.42	.47	.45	.58	.67
	a=.5	.43	.54	.37	.60	.68	.31	.42	.47	.53	.58	.70
	a=1	.61	.69	.46	.74	.77	.35	.40	.52	.59	.60	.72
SA(122)	a=0	.09	.14	.09	.15	.22	.12	.25	.25	.20	.38	.53
	a=.5	.20	.27	.16	.31	.40	.16	.31	.32	.35	.48	.68
	a=1	.39	.47	.26	.52	.56	.21	.32	.39	.47	.54	**.75**

Table 2 – Power evaluation for the clusters A-K of figure 6. The first three lines display real cluster letter, size and compactness. Lines 4 to 8 show the power for the elliptic scan using each ellipse sets $E(1)$, $E(2)$, $E(4)$, $E(8)$ and $E(20)$. The last seven groups of three lines display the power for the Simulated Annealing (SA) scan with a=0 (no compactness correction), a=0.5 (medium compactness correction), and a=1 (full compactness correction), using for each group different upper bounds for the detected clusters sizes. The groups labels SA(8), SA(12), SA(20), SA(30), SA(40), SA(61) and SA(122) on the left of Table 2 refer to the different upper bounds tested (3%, 5%, 8%, 12.5%, 16%, 25% and 50% of the 245 regions, respectively). Bold type values represent the highest power obtained for each real cluster for each method.