A topological correction for a graph based spatial scan cluster detection algorithm

Luiz Duczmal, Sabino José Ferreira, Marcos da Cunha Santos

Statistics Department, Universidade Federal de Minas Gerais

duczmal@est.ufmg sabino@est.ufmg.br msantos@est.ufmg.br

ABSTRACT

Many spatial cluster finder algorithms do not have adequate procedures for controlling the shapes of the clusters found. The cluster candidates may sometimes spread through large portions of the map, making it difficult for the practitioner to assess the geographical meaning of the solution. We propose a novel scan statistic algorithm for finding irregularly shaped spatial clusters in a map divided in a finite number of regions, whose adjacency is defined by a graph structure, by means of a new penalty function, the cohesion penalty correction. Based on the graph topology, the cohesion correction was developed to avoid the excessive irregularity of the clusters. The cohesion correction is compared with the geometric concept of compactness correction, which was used previously as a penalty function. We show that the cohesion correction has advantages over the compactness correction, boosting the power to detect elongated clusters, and being less computer-intensive. A multi-objective genetic algorithm is used to compute the solutions, consisting of the Pareto-set of clusters candidates. The goal of the cluster finder algorithm is to maximize two objectives: the scan statistic and the cohesion of the graph structure. The statistical significances of the clusters in the Pareto-set are estimated through Monte Carlo simulations, using Gumbel's approximation for the scan statistic distribution, which are used as a criterion for choosing the best solution.

1. INTRODUCTION

Epidemiologists and crime analysts make use of cluster detectors as an effective tool to study the geographical patterns of diseases (Lawson et al., 1999), syndromic surveillance (Duczmal and Buckeridge 2006a; Kulldorff et al. 2005a,b, 2006a,b) and criminal activity. Softwares such as SatScan (Kulldorff, 1999) and ClusterSeer (TerraSeer, 2004), employing the spatial scan statistic (Kulldorff 1997) are increasingly popular. We are interested in the detection spatial clusters that are not restricted to circular shape. Recently, several methods were developed to detect irregularly shaped clusters (Patil and Tallie 2004, Duczmal et al. 2004, 2006a,b,c,d, Iyengar 2004, Sahajpal et al. 2004, Conley et al. 2005, Tango and Takahashi 2005, Neill et al. 2006, Assuncao et al. 2006 and Kulldorff et al.2005, 2006a,b).

We develop a novel methodology, based on a genetic multi-objective optimization algorithm, which was developed elsewhere (Duczmal et al. 2006d) for selecting the best cluster solution, among the many possible solutions found. That algorithm was developed to maximize two competing objectives, namely the cluster scan likelihood ratio, and the regularity of cluster shape. In our present paper we introduce a novel penalty function, which, instead of being based on the geometric shape of the clusters, is based on its topology. This change was motivated by the discussion in Duczmal et al. (2006b). In that paper, it was noted that the some clusters have low population regions, which should disconnect it if removed (the so called *weak links*). The presence of weak links impacts the power of detection of such cluster. This happens because the variance of the number of cases inside a weak link is elevated, due to its low population. As a consequence, we expect to find very few cases inside a weak link, thus making it difficult for the cluster detector to join the parts of the cluster separated by the weak link. In other words, weak links generates noise, in the sense that a legitimate cluster which includes a weak link should be penalized (see details in Duczmal et al. 2006b). The usual geometric compactness correction does not penalize enough clusters with weak links. As we shall see, the topological correction developed in this work actuates effectively in clusters with weak links, thus inhibiting the proliferation of such clusters, which should otherwise compete unfairly with the legitimate clusters solutions.

The paper is divided as follows. In section 2, we summarize the concepts of Kulldorff´s spatial scan statistic and review the multi-objective genetic algorithm. Section 3 introduces the novel penalty function based on the cluster's graph topology. Power tests are presented in section 4. The final remarks are discussed in section 5.

2. THE MULTI-OBJECTIVE SPATIAL SCAN GENETIC ALGORITHM

In this section we review the spatial scan statistic (Kulldorff, 1997), the geometric concept of compactness (Duczmal et al., 2006a), the usual geometric compactness correction, the multi-objective genetic algorithm and the statistical significance estimation of the clusters found (Duczmal et al., 2006b, 2006c), using the Gumbel´s approximation (Abrams et al. 2005).

2.1. THE SPATIAL SCAN STATISTIC

Given a map divided into M regions, total population N and C cases, a zone Z is defined as any set of connected regions. The number of cases in each region follows a Poisson distribution, with average proportional to its population, under the null hypothesis that there are no clusters in the map. Defining μ_Z as the expected number of cases inside Z under the null hypothesis, c_Z as the number of observed cases inside Z, c_Z/μ_Z as the relative incidence inside Z, and $(C-c_Z)/(C-\mu_Z)$ as the relative incidence outside Z, the Kulldorff´s spatial scan statistic is defined as

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

if the relative incidence inside Z is higher than 1, and 1 otherwise. The zone with the maximum likelihood is defined as *the most likely cluster*. The test statistic is $\max_z LR(z)$. This likelihood ratio, maximized over all the zones, identifies the zone that constitutes the most likely cluster. See Kulldorff (1997) for details. The test statistic can detect not only circular clusters, but we should expect lesser power for the irregular ones.

2.2. THE REGULARITY OF CLUSTER SHAPE

Highly irregularly shaped zones in the map tend to have very large scan likelihood ratio values and are almost always undesirable, because they compete unfairly with the more legitimate cluster candidates, which tend to have more regular shapes and more modest scan likelihood ratio values. In this sense, they constitute noise, against which the signal, represented by the more geographically meaningful cluster candidates, is superimposed. We now define a quantitative measure of the regularity of cluster shape. Given a planar geometric object z, define $A(z)$ as the area of z and $H(z)$ as 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. The *compactness* of z is

$K(z) = 4\pi A(z) / H(z)^2$.

This formula is equivalent to $A(z)$ divided by the area of the circle with perimeter $H(z)$. Compactness does not depend on the size of the object, only on its shape. The maximum compactness value, namely one, is attained by the circle. Compactness can be used as a penalty function acting as a filter to restrain the presence of those extremely high *LLR* valued large tree-shaped clusters, allowing the presence of the somewhat lower *LLR* valued clusters solutions with real geographic meaning that we are looking for. For details, see Duczmal et al. (2006b).

2.3. THE GENETIC ALGORITHM

A genetic algorithm was developed for spatial cluster detection and inference using the scan statistic in a map divided in regions. The algorithm aims to maximize an objective function, modifying an initial set of individuals, or population, for a number of generations. The variance of the population is increased through the *crossing-over* and *mutation* operators. The *selection* operator picks the individuals that will remain in the next generation, maintaining the population size fixed during the process. The crossing-over operator creates new children individuals, or zones, mixings the features of two randomly chosen parents at a time, which are themselves zones from the previous generation, see details in Duczmal et al. (2006c). In this manner, several children are produced, which are intermediate zones between the two extremes zones *A* and *B*. The selection operator rank the zones according to the value of the objective function, namely the compactness corrected spatial scan statistic mentioned in section 2.2. We expect to find individuals with increasingly higher values as the algorithm advance through the generations.

The statistical significance of the most likely cluster of observed cases is computed through a Monte Carlo simulation (Dwass, 1957). Under null hypothesis, simulated cases are distributed over the map and the scan statistic is computed for the most likely cluster. This procedure is repeated thousands of times, and the obtained distribution of the values is compared with that of the most likely cluster of observed cases, producing an estimate of its p-value. The algorithm has fast convergence, and good power of detection.

The genetic algorithm is suitably modified to deal simultaneously with the two quantities: the compactness $K(.)$ (section 2.2), and Kulldorff´s original spatial scan $LLR(.)$ (section 2.1), constituting the multi-objective genetic algorithm (Duczmal et al. 2006d). The pairs

(K_i, L_i), indicating the compactness and scan statistic computed for each individual *i,* are plotted in the Cartesian plane. The selection operator is now defined in terms of two objectives, maximizing the compactness and the scan statistic. This operator relies on the concept of *dominance*: a point is said to be dominated if it is worse than another point in at least one objective, while not being better than that point in any other objective (Chankong and Heimes, 1983). The *Pareto-set* is the set that does not contain any dominated solution (Takahashi et al. 2003).

The selection operator is modified as follows. At first we compute the *current generation list*, which consists of the parents set augmented several times with the addition of newly produced offspring through the crossing-over operator. Next we compute the Pareto-set of the current generation list, which is stored in the initially empty *next generation list*, and then removed from the current generation list. This procedure is repeated until the new generation list has grown to contain at least *M* individuals. The eventually excessive individuals added in the last step are removed randomly to form a new next generation list with exactly *M* individuals. The next generation population consists of the individuals of the next generation list, thus formed by selecting a set consisting of M points that are not dominated by any other point outside this set. The current generation list is finally substituted by the next generation list.

The crossing-over operator builds again new offspring, and the procedure is repeated for a number of successive generations. We observe a collective displacement of points towards generally higher values of *K* and *L*. The Pareto-set of the latest generation is considered the solution given by the algorithm. We observe that the later generations point sets become closer to their respective Pareto-sets, and that proximity could be used as a criterion for convergence.

2.4. THE MULTI-OBJECTIVE ALGORITHM

A scheme based on *multi-objective optimization* was developed elsewhere (Duczmal et al. 2006d) for selecting the best cluster solution, among the many possible solutions found. The algorithm aims to maximize two competing objectives, namely the cluster scan likelihood ratio, and the regularity of cluster shape. This is done computing the Pareto-set of the collection of all the solutions found. Following that, a Monte Carlo simulation is conducted: we assign random cases to the regions, according to a Poisson distribution under null hypothesis, where the average of cases allocated to each region is proportional to its population. The process of finding the Pareto-set is repeated hundreds of times, each time for a different allocation of random cases under the null hypothesis. Those Pareto-sets are joined, obtaining a collection of thousands of points. We then partition the strip $(0,1)\times(0,\infty)$ into a number of parallel bands $(s_i, s_{i+1})\times(0,\infty)$, where $s_i < s_{i+1}$. Next, we compute the approximate Gumbel′s distribution that fits best to the log likelihood ratio values of the points encased in each band. The integral of the tail of the Gumbel distribution furnishes our estimated p-values for the mid point $(s_i + s_{i+1})/2$. The collection of p-values estimates for all the strips are then used to compute the *interpolated p-value surface*, from which the individual p-values for the points in the observed cases Pareto-set are then computed. In the example of Figure 1, where 10,000 Pareto sets were obtained under the null hypothesis, the 0.01, 0.001 and 0.0001-*value isocline* curves are shown in green. The blue curve indicates the 0.05-*value isocline* curve.

3. THE COHESION CORRECTION OF THE CLUSTER

In this section we define the cohesion correction to be used as a measure of topological regularity of the cluster associated to a graph G. To each graph G with n nodes we associate a set of nodes or vertices which belongs to G and that are called the disconnection set $G_{disc} = \{x_1, x_2, \ldots, x_k\}$, $k \le n$. Each node $x_i \in G_{disc}$ is such that the resulting graph $G - \{x_i\}$ is not connected. In other words, each time any node $x_i \in G_{disc}$ is removed from the graph G, this removal breaks the original graph in two or more connected pieces. To each $x_i \in G_{disc}$ we define a partition G_i of the original graph G made of the node x_i and the two or more parts into which the node breaks the original graph as it is removed from it. We write an arbitrary partition G_i as $G_i = \{x_i, \hat{x}_{P_1}, \hat{x}_{P_2}, \ldots, \hat{x}_{P_L}\}$ where $\hat{x}_{P_J}$ denotes the J-th connected part of the broken graph and the number of parts L is less or equal to $n-1$. Also to each connected part $\hat{x}_{P_J}$ we associate its relative population

$$\mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_J}) = \frac{\mathrm{pop}\,(\hat{x}_{P_J})}{\mathrm{pop}\,(G)},$$

where respectively, $\mathrm{pop}\,(\hat{x}_{P_J})$ and $\mathrm{pop}\,(G)$ represent the populations of the connected part $\hat{x}_{P_J}$ and the whole graph G.

Given any partition $G_i = \{x_i, \hat{x}_{P_1}, \hat{x}_{P_2}, \ldots, \hat{x}_{P_L}\}$ of some graph G we rank the connected parts according their relative populations and rewrite the G_i partition in its ranked version as $G_i = \{x_i, \hat{x}_{P_{(1)}}, \hat{x}_{P_{(2)}}, \ldots, \hat{x}_{P_{(L)}}\}$. Then we define a cohesion function for a ranked partition as:

$$\mathrm{ch}\,(G_i) \;=\; (\,n\;\mathrm{pop}_{\mathrm{r}}(x_i)\,)^{dp} \;\times\; \left(\frac{\Delta_1 \qquad\qquad \Delta_2 \qquad\quad \cdots \qquad\qquad \Delta_L}{\mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_{(1)}}) \;\; \mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_{(2)}}) \quad \cdots \quad \mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_{(L)}})} \right) \quad (1)$$

where

$$\Delta_J = \mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_{(J)}}) \,-\, \mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_{(J+1)}}) \;\; \text{for} \qquad 1 \le J \le L-1,$$
$$\Delta_L = \mathrm{pop}_{\mathrm{r}}(\hat{x}_{P_{(L)}}) \,-\, 0$$

and the exponent dp from damping factor are defined as:

$$dp \; = \; (1 - (\, \mathrm{pop_r}(x_i) + \mathrm{pop_r}(\hat{x}_{P_{(1)}}\,)\,)\,)$$

Now we define the cohesion function of the whole graph G as:

$$\mathrm{ch}\,(G) \; = \; \min_{x_i \in \, G_{disc}} \quad \mathrm{ch}\,(G_i)$$

If the set G_{disc} is the empty set, then $\mathrm{ch}\,(G) = 1.$

The cohesion function is used as a penalization factor in the expression for the test statistic (actually it goes as an exponent of the likelihood ratio). The cohesion function is conceived with the intention to penalize graphs that are broken in “homogeneous” parts (parts that have approximately the same population). This is associated with the second term on the right side of formula (1). Also the penalization is high if the disconnecting node is “weak”, meaning that it has small population. This is accomplished by the first term on the right side of formula (1). This particular penalization effect is attenuated by the damping factor dp that takes in account if the remaining relative population, namely, one minus the relative population of the disconnecting node and the relative population of the most populated part, is still significant.

4. POWER TESTS

A benchmark dataset is constructed using real data population from the northeastern U.S. (245 regions) with one of 11 simulated irregularly shaped clusters A to K, displayed in Figure 2 (Duczmal et al. 2006b). These clusters were chosen with the purpose of testing the 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). 10,000 runs of the both the compactness (GAC) and topological (GAT) genetic algorithm statistics were done under the null hypothesis, plus 1,000 runs for each entry in the table, under the alternative hypothesis. Observe that the power is significantly higher for the elongated string-shaped clusters C, F and G. for the other clusters, the power is about the same for both algorithms. The running time was less than one second for each Monte Carlo simulation.

5. CONCLUSIONS

We developed and tested a novel penalty function, the cohesion penalty correction. Based on the graph topology, the cohesion correction was developed to avoid the excessive irregularity of the clusters. The cohesion correction was compared with the geometric concept of compactness correction through a multi-objective genetic algorithm. It was shown that the cohesion correction has higher power of detection, when used for finding elongated string-shaped clusters. For less elongated clusters, the algorithm based on the cohesion correction exhibited the same power of detection of the compactness correction algorithm. We also have shown that the cohesion correction is faster.

6. REFERENCES

Abrams A, Kulldorff M, Kleinman K, 2005. Empirical/Assymptotic P-values for Monte Carlo-Based Hypothesis Testing: an Application to Cluster Detection Using the Scan Statistic. *2005 Syndromic Surveillance Conference.*

Chankong V, Haimes YY, 1983. *Multiobjective Decision Making: Theory and Methodology.* North-Holland.

Conley J, Gahegan M, Macgill J, 2005. A genetic approach to detecting clusters in point-data sets. *Geographical Analysis*, 37, 286-314

Duczmal L, Assunção R, 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, *Comp. Stat. & Data Anal.*, 45, 269-286.

Duczmal L, Buckeridge DL, 2006a. A Workflow Spatial Scan Statistic. *Statistics in Medicine*, 25;743-754.

Duczmal L, Kulldorff M, Huang L., 2006b. Evaluation of spatial scan statistics for irregularly shaped clusters. *J. Comput. Graph. Stat.* 15;1-15.

Duczmal L, Cançado ALF, Takahashi RHC, Bessegato LF, 2006c. A genetic algorithm for irregularly shaped spatial scan statistics (*submitted to Comput. Stat. Data Anal.*).

Duczmal L, Cançado ALF, Takahashi RHC, 2006d. Delineation of Irregularly Shaped Disease Clusters through Multi-Objective Optimization (*submitted*)

Dwass M. 1957. Modified randomization tests for nonparametric hypotheses. *Ann. Math. Stat.*, 28:181-187.

Iyengar, VS, 2004. Space-time Clusters with flexible shapes. *IBM Research Report* RC23398 (W0408-068) August 13, 2004.

Kulldorff M, 1997. A Spatial Scan Statistic, *Comm. Statist. Theory Meth.,* 26(6), 1481-1496.

Kulldorff M, 1999. Spatial scan statistics: Models, calculations and applications. In *Scan Statistics and Applications*, Glaz and Balakrishnan (eds.). Boston: Birkhauser, 303-322.

Kulldorff M, Tango T, Park PJ., 2003. Power comparisons for disease clustering sets, *Comp. Stat. & Data Anal.,* 42, 665-684.

Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F, 2005. A Space-Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Medicine, Feb.15.*

Kulldorff M, Huang L, Pickle L, Duczmal L, 2006. An Elliptic Spatial Scan Statistic. *Statistics in Medicine* (to appear).

Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, Platt R., 2006b, Multivariate Scan Statistics for Disease Surveillance. *Statistics in Medicine* (to appear).

Lawson A., Biggeri A., Böhning D. *Disease mapping and risk assessment for public health*. New York, John Wiley and Sons, 1999.

Neill DB, Moore AW, Cooper GF., 2006, A Bayesian spatial scan statistic. *Advances in Neural Information Processing Systems 18* (in press).

Patil GP, Taillie C, 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Envir. Ecol. Stat.,* 11, 183-197.

Sahajpal R., Ramaraju G. V., Bhatt V. (2004) Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. *International Conference on Intelligent Sensing and Information Processing.*

Takahashi RHC, Vasconcelos JA, Ramirez JA, Krahenbuhl L, 2003. A multiobjective methodology for evaluating genetic operators. *IEEE Transactions on Magnetics*, 39(3), 1321-1324.

Tango T, Takahashi K., 2005. A flexibly shaped spatial scan statistic for detecting clusters. *Int. J. Health Geogr.*, 4:11.

TerraSeer, 2004. http://www.terraseer.com

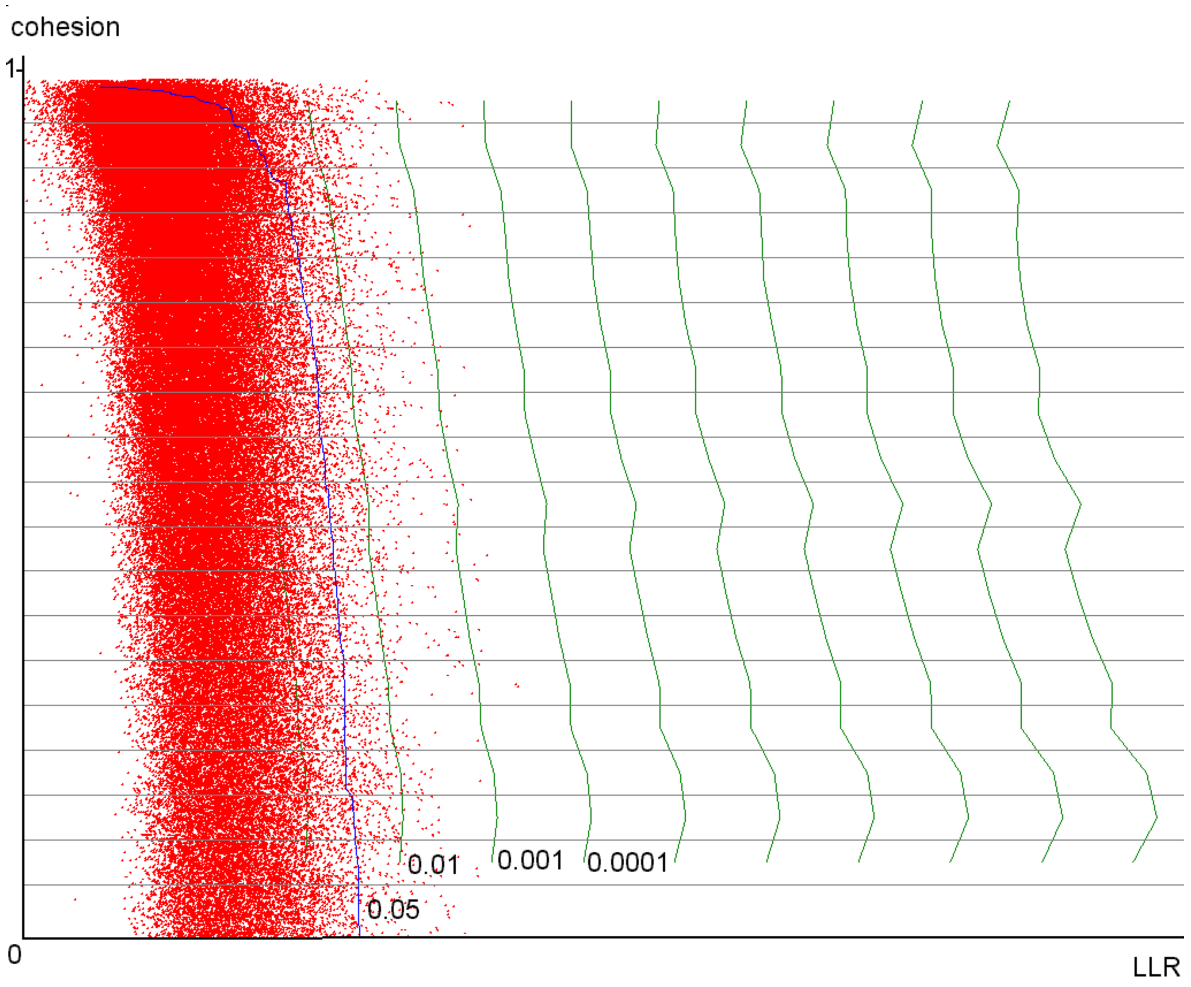

Figure 1- The p-value isocline curves for the null hypothesis Monte Carlo simulation, using 10,000 Pareto-sets.

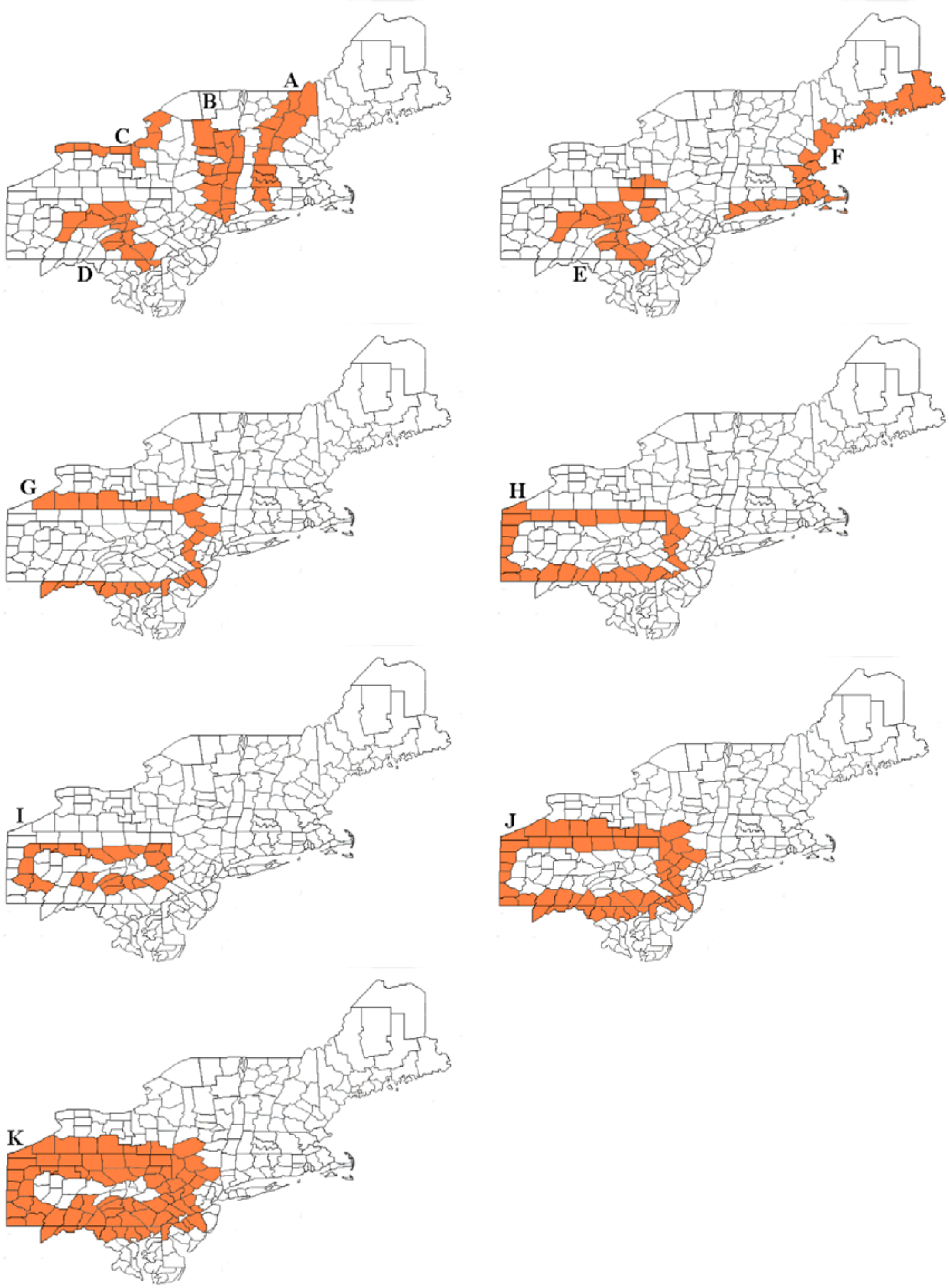

Figure 2. Simulated data clusters for the northeastern U.S. The clusters A-K, were used in the power evaluations. (Duczmal et al. 2006b).

real cluster	A	B	C	D	E	F	G	H	I	J	K
#regions	13	16	7	15	21	23	26	29	23	55	78
GAC	.86	.81	.79	.87	.77	.45	.50	.65	.62	.58	.49
GAT	.86	.79	.86	.90	.80	.56	.57	.62	.63	.58	.47

Table 1- Power comparison between the genetic algorithm employing the compactness correction (GAC) and the topological correction (GAT).