Geographic Delineation of Disease Clusters through Multi-Objective Optimization

Luiz Duczmal[1,*] , André L. F. Cançado[2] , Ricardo H. C. Takahashi[3]

ABSTRACT

Irregularly shaped spatial disease clusters occur commonly in epidemiological studies, but their geographic delineation is poorly defined. Most current spatial scan software usually displays only one of the many possible cluster solutions with different shapes, from the most compact round cluster to the most irregularly shaped one, corresponding to varying degrees of penalization parameters imposed to the freedom of shape. Even when a fairly complete set of solutions is available, the choice of the most appropriate parameter setting is left to the practitioner, whose decision is often subjective. We propose quantitative criteria for choosing the best cluster solution, through multi-objective optimization, by finding the Pareto-set in the solution space. Two competing objectives are involved in the search: regularity of shape, and scan statistic value. Instead of running sequentially a cluster finding algorithm with varying degrees of penalization, the complete set of solutions is found in parallel, employing a genetic algorithm. The cluster significance concept is extended for this set in a natural and unbiased way, being employed as a decision criterion for choosing the optimal solution. The Gumbel distribution is used to approximate the empiric scan statistic distribution, speeding up the significance estimation. The method is fast, with good power of detection. An application to breast cancer clusters is discussed.

Keywords**:** Spatial scan statistic, disease clusters, geometric compactness penalty correction, Pareto-sets, multi-objective optimization, Gumbel distribution, genetic algorithm.

1. INTRODUCTION

Epidemiology and disease surveillance make intensive use of techniques for the detection and inference of spatial clusters. The geographic delineation of clusters is an important tool in etiological studies (Lawson, 1999) and in the early warning of intentional and non-intentional infectious diseases outbreaks (Duczmal and Buckeridge 2005, 2006a; Kulldorff et al. 2005, 2006a,b). The spatial scan statistic (Kulldorff 1997), employed by the softwares SatScan (Kulldorff, 1999) and ClusterSeer (TerraSeer, 2004) is currently used by many health departments to detect circularly shaped disease clusters (Kulldorff and Nagarwalla 1995). In many scenarios, however, we are interested in the detection spatial clusters that are not restricted to circular shape. Diseases may be concentrated along rivers, ocean and lake shores, transport ways, or plumes of air pollution. The SatScan approach was extended to detect elliptic shaped clusters (Kulldorff et al.2006a), increasing the geometric versatility of the original SatScan. Recently, other methods were proposed to detect connected clusters of irregular shape (Duczmal et al., 2004, 2006b, Iyengar, 2004, Tango and Takahashi, 2005, Assuncao et al., 2006, Neill et al., 2005, Patil and Tallie, 2004). Contley et al. (2005) proposed a point dataset genetic algorithm to explore a configuration space of multiple agglomerations of ellipses. Sahajpal et al., 2004 also used a genetic algorithm to find clusters shaped as intersections of circles of different sizes and centers. Duczmal et al.(2006c) presented a genetic algorithm for finding arbitrarily shaped clusters in a map divided into a finite number of regions, maximizing the scan statistic with a penalty function correction for the highly irregular shapes.

[1] Statistics Department, Universidade Federal de Minas Gerais, Brazil.
[2] Electrical Engineering Department, Universidade Federal de Minas Gerais, Brazil.
[3] Mathematics Department, Universidade Federal de Minas Gerais, Brazil.
*corresponding author: duczmal@est.ufmg.br

The geographic delineation of irregularly shaped clusters presents some difficulties. The unlimited geometric freedom of cluster shapes diminishes the power of detection (Duczmal et al., 2006b). This happens because the collection of all connected zones, irrespectively of shape, is very large. The maximum value of the objective function is likely to be associated with "tree-shaped" clusters, which merely link the highest likelihood ratio cells of the map, without contributing to the discovery of geographically meaningful solutions that delineate correctly the "true" cluster. In other words, there is much "noise", against which the legitimate solutions cannot be distinguished. That problem occurs in every irregularly shaped cluster detector. It may be mitigated in part by limiting the maximum number of regions that should constitute the cluster. A more elegant solution consists in applying a penalty function, through the concept of geometric "non-compactness" (Duczmal et al.2006b), diminishing the value of the scan statistic according to the irregularity of cluster shape, generalizing an idea that was used for the special case of ellipses (Kulldorff et al., 2006a).

By varying the amount of penalization to the freedom of shape, several cluster solutions may be found, ranging from the circular to the very irregularly shaped cluster. Most current spatial scan algorithms do not allow appropriate control of the freedom of shape, and generally only one solution is displayed. Even when a fairly complete set of solutions is available, by running the algorithm through several parameter settings, as in Duczmal et al.(2006b,c), the choice of the most appropriate parameter setting is left to the practitioner, whose decision is often subjective.

In this paper, we develop and implement a novel multi-objective algorithm for cluster detection and inference, based on a genetic algorithm. Two competing objectives are involved in the search for clusters: regularity of shape and scan statistic value. We propose quantitative criteria for choosing the best cluster solution, through multi-objective optimization, by finding the Pareto-set in the solution space, followed by the application of a decision criterion that is based on the maximization of solution significance over that set. In this way, the arbitrary and subjective choice of a "best solution" that might be performed in the former methodologies is replaced by a systematic and theoretically-founded methodology for finding such solution. The concept of optimal solution becomes well-defined in the context of irregularly-shaped cluster detection. As a by-product of the proposed approach, a whole set of alternative solutions (the Pareto-set) becomes available for the user, for the purpose of comparison and of analyzing the problem intrinsic structure. These ideas are quite new in the context of cluster detection, keeping some similarity with the recent idea of viewing other machine learning problems in a multi-objective framework, as proposed in the references (Teixeira et al., 2000, Nepomuceno et al., 2003).

The proposed multi-objective genetic algorithm, instead of running sequentially a cluster finding algorithm with varying degrees of penalization, finds the complete set of solutions in parallel. This intrinsic parallel multi-solution search is the main reason that makes the multi-objective genetic algorithms particularly efficient for solving multi-objective problems (Fonseca and Fleming, 1995). In addition, the genetic algorithm search procedure allows escaping from locally optimal solutions, making the multi-objective genetic algorithm a natural choice of optimization machinery for the problem of cluster detection (Duczmal et al., 2006b). Using Pareto-sets, the concept of cluster significance is extended in a natural and unbiased way, by means of the Gumbel approximation. The multi-objective algorithm presents a new insight into the geographic meaning of the cluster solution set, providing a quantitative approach to the problem of selecting the most appropriate solution. The method was implemented in C language. Please contact the corresponding author for a freeware copy of the program.

We also extend the remarkable recent result of Abrams et al. (2005) for the parametric approximation of the likelihood ratio scan statistic distribution for circular clusters, using the Gumbel distribution. In this work, we verify through numerical experiments that this approximation also seems to be valid for the genetic based irregularly shaped clusters scan statistic. That technique does not only speed up the p-value estimation, but allows the significance estimation of clusters to achieve better accuracy.

The paper is divided as follows. In section 2, we review Kulldorff's spatial scan statistic, the non-compactness penalty function, and the mono-objective genetic algorithm. We introduce the Gumbel approximation for the penalized genetic algorithm scan statistic in section 3. The multi-objective genetic algorithm is presented in section 4. Power evaluations are described in section 5. We present an application for breast cancer clusters in North Eastern US in section 6. We conclude with the final remarks in section 7.

2. THE PENALIZED SPATIAL SCAN GENETIC ALGORITHM

In this section we review the spatial scan statistic (Kulldorff, 1997), its geometric penalized extension (Duczmal et al., 2006b), and its implementation through a genetic algorithm (Duczmal et al., 2006c).

2.1. THE SPATIAL SCAN STATISTIC

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

$LR(Z) = I(Z)^{c_Z} O(Z)^{C - c_Z}$ if $I(Z) > 1$,

and 1 otherwise. When derived as a likelihood ratio test, it is based on a set of alternative hypotheses. See Kulldorff(1997) for details. The zone Z that maximizes $LR(.)$ is defined as *the most likely cluster*. The logarithm of the likelihood ratio (*LLR*) is commonly used. The test statistic can detect not only circular clusters, but we should expect lesser power for the irregular ones. 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 values are compared with that of the most likely cluster of observed cases, producing an estimate of its p-value.

2.2. THE COMPACTNESS PENALTY CORRECTION

In order to penalize the highly irregularly shaped zones in the map, 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

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

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 (Duczmal et al., 2006b). The maximum compactness value, one, is attained by the circle. The expression $LR(z)^{K(z)}$ is employed instead of the likelihood ratio $LR(z)$. A parameter $a > 0$ is used here as a penalty factor, varying the strength of the compactness measure, through the formula $K(z)^a$, replacing $K(z)$. The generalized expression $LR(z)^{K(z)^a}$ is thus employed as the corrected likelihood test function instead of $LR(z)$.

The penalization is weaker when $a < 1$, and stronger when $a > 1$. It was observed, through numerical experiments, that irregularly shaped cluster detection algorithms benefit from the application of the non-compactness penalty correction (Duczmal et al., 2006b). The penalty correction acts 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. These last clusters are in general less irregularly shaped than the tree-shaped ones, and should otherwise be obfuscated by them.

2.3. THE MONO-OBJECTIVE GENETIC ALGORITHM

Genetic algorithms (GA's) are optimization algorithms that employ a set of candidate solutions (called the *population*) that are spread over a region of the space of optimization variable vectors, for performing the search for function optima. The GA's evolve these candidate solutions (each such candidate solution is called an *individual*) using rules that mimic the process of biological evolution, in its process of species evolution through natural selection. A *crossover* operator randomly mixes the features of randomly chosen parent individuals, leading to offspring individuals, and a *mutation* operator introduces random perturbations in the features of an individual, increasing the variance of the population. The *selection* operator decides which individuals should continue in the population or should be discarded in the next generation, based on the objective function evaluation of all the individuals of the current generation and using a stochastic procedure. This process is repeated for a number of generations, and we expect to find individuals with increasing higher values for the objective function in the later generations. This kind of algorithm is becoming increasingly important, in the context of the optimization of objective functions that present multiple local optima and irregular landscapes.

A genetic algorithm was developed for spatial cluster detection and inference using the scan statistic. Given a map with M regions, the mono-objective optimization algorithm starts with an initial population consisting of M individuals, or zones. Every region in the map belongs to at least one of the zones of the initial population, which are formed by an aggregation process. The core of the algorithm is the routine that builds the offspring resulting from the crossover of two non-disjoint given parents *A* and *B*. Each parent and each offspring is thus a set of connected regions in the map, or zone. The offspring is constructed based on a random numbering of two trees associated to the parents *A* and *B,* see details in Duczmal et al. (2006c). Several children are produced, which are intermediate zones between the two extremes zones *A* and *B*. The computation of the objective function (the compactness corrected spatial scan statistic of section 2.2) is very inexpensive, and each individual child is already known to be connected, as a property of the special crossover operator that was developed. The algorithm has fast convergence, and has good power of detection.

3. THE GUMBEL DISTRIBUTION APPROXIMATION

Through extensive numerical tests, Abrams et al. (2005) showed that under the null hypothesis, the empirical distribution of values of the Kulldorff's scan statistic for circular clusters is approximated by the well-known Gumbel distribution

$f(x) = \theta^{-1} \exp\{-\exp[(x-\mu)/\theta] - (x-\mu)/\theta\}$,

with parameters μ *(mode)* and θ *(scale).*

In this section we extend these findings for the empirical distribution of the penalized genetic algorithm scan statistic under null hypothesis. Numerical tests suggest that it is also reasonably well approximated by the Gumbel distribution. We do not attempt here to give a rigorous proof of this result. The rationale follows the same argument used for the circular scan: the penalized genetic algorithm scan statistic is also an extreme value distribution. We are interested in computing only small p-values. We will adjust the tail of the Gumbel function, giving less importance to the smaller values of *LLR* of the empirical distribution, instead of simply computing the usual parameters μ_0 and θ_0 using the whole set of N values of the empirical distribution. Sorting those N values by decreasing order of *LLR,* we now look for the optimal parameters μ_{opt} and θ_{opt} which minimize the functional

$$\sum_{i=0}^{i=i_{\max}} \left| \mu - \theta \log\left(\log\left(\frac{1}{1-p_i} \right) \right) - x_i \right|$$

where $p_i = \dfrac{2^i}{N}$ and x_i belongs to the list of empirical values $x_0,\ldots,x_{i_{\max}}$, located at the positions $2^0,\ldots,2^{i_{\max}}$, respectively. The integer $i_{\max}$ is chosen to be the largest number such that $2^{i_{\max}} < N$. Each term in the summation above measures the absolute difference between the empirical values x_i at the selected positions of the list and the values estimated by the Gumbel distribution at the corresponding quantiles p_i. The values $x_0,\ldots,x_{i_{\max}}$ are more concentrated toward the larger values of the empirical distribution. Thus, the fitting of the tail is privileged, as compared to the standard fitting procedure that considers all the empirical values as equally important. The optimal parameters μ_{opt} and θ_{opt} are found by a standard bissection procedure in two variables, employing the usual parameters μ_0 and θ_0 as initial entries. In section 4.3 we present numerical simulations results showing the adequacy of the parametric Gumbel approximation.

4. MULTI-OBJECTIVE OPTIMIZATION

Genetic algorithms have been found to be particularly well suited for dealing with multi-objective optimization problems, since they evolve a whole set of tentative solutions towards the Pareto-set. This allows finding an entire estimate of this set in a single run of the algorithm (Fonseca and Fleming, 1995). Examples of multi-objective genetic algorithms developed for different application contexts are reported in (Ramos et al., 2003, Takahashi et al., 2004, Carrano et al., 2006). The references (Takahashi et al., 2003, Carrano et al., 2006) present illustrations of how a Pareto-set can be employed for the *a posteriori* analysis of the problem in such a way that no mono-objective algorithm could perform.

We now describe the novel multi-objective optimization approach to the problem of finding spatial clusters. The genetic algorithm described in section 2.3 will be modified to deal simultaneously with the two quantities: the compactness $K(.)$ (section 2.2), and Kulldorff′s original spatial scan $LLR(.)$ (section 2.1). The compactness $K(.)$ will not be used anymore as a penalty correction, but instead as a new variable. That approach simplifies the problem and allows a stronger grasp of the question of finding the "best" cluster solution.

4.1. PARETO-SETS

The pairs (L_i, K_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 Haimes, 1983). The *Pareto-set* is the set that does not contain any dominated solution, thus consisting of points that are not simultaneously worse than any other point in both objectives.

4.2. THE MULTI-OBJECTIVE OPTIMIZATION GENETIC ALGORITHM

The initial population construction and the crossing-over and mutation operators are identical to those employed in the genetic algorithm of section 2.3. The difference relies in the selection operator for the multi-objective optimization genetic algorithm, as follows.

At the beginning of each generation, we compute the *current generation list*, which consists of the set of parent individuals augmented several times with the addition of newly produced offspring through the crossing-over operator. The *next generation list*, initially empty, stores the individuals that will survive for the next generation. We compute the Pareto-set of the current generation list, which is transferred to the initially empty *next generation list*; the same Pareto-set is also removed from the current generation list. A new Pareto-set of the remaining individuals is computed, and the 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 current generation list is finally substituted by the next generation list. This procedure is similar to the "Non-Dominated Sorting" selection operator, which is employed in some of the most efficient multi-objective genetic algorithms that are available up to now (Deb et al., 2002). The crossing-over operator builds again new offspring, and the procedure of this paragraph is repeated for a number of successive generations.

We present an example using the Northeast US map benchmark (Duczmal et al. 2006b). This map has 245 counties, population of 29,535,210 women, and real data of 58,943 breast cancer deaths, during the period of 1988-1992 (Kulldorff et al.1997). The analyses are adjusted for age applying indirect standardization with 18 distinct five years age groups: 0-4, 5-9,…, 80-84, and 85+. The annual mortality rate was 39.91 per 100,000 women. Figure 1 displays a sequence of *LLR* versus compactness graphs obtained through the multi-objective genetic algorithm for the initial population, and for generations 2, 4, 11, 14 and 20. We observe a collective displacement of points towards generally higher values of *K* and *L*. Observe that the convergence is very fast for points with high compactness. In that example, an isolated point with high *LLR* appears in generation 14, being followed by newly generated offspring with even higher *LLR* values in generation 20. There is a pronounced change in the first generations, followed by more subtle population changes in the later generations. The population may contain multiple copies of some individuals, especially in the later generations. The Pareto-set of the latest generation is considered the solution given by the genetic algorithm. In that example, no further changes occur within the Pareto-set during the subsequent generations. The later generations point sets become closer to their respective Pareto-sets, and that proximity could be used as a criterion for convergence.

4.3. COMPUTING THE CLUSTERS SIGNIFICANCE

In this section we show how to compute the statistical significance of the points in the Pareto-set solution for the map of observed cases, which is compared to the hundreds of Pareto-set solutions computed for each one of the simulated cases maps under the null hypothesis.

A Monte Carlo simulation is conducted: the algorithm of section 4.2 is executed hundreds of times for maps containing random cases spread 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 for each different allocation of random cases. Those Pareto-sets are joined, obtaining a collection of thousands of points distributed in the $LLR \times compactness$ space, namely the strip $(0,\infty)\times(0,1]$ (see Figure 2). Let $D(l,c)$ be the "true" bivariate distribution of an arbitrarily large collection of cluster points in $(0,\infty)\times(0,1]$. Our goal now is to obtain a good approximation for $D(l,c)$. Similarly to section 3, extensive numerical tests suggest that the marginal distribution of $D(l,c)$ on the variable l is approximated by a Gumbel distribution, as follows. The strip $(0,\infty)\times(0,1]$ is partitioned into a number of parallel strips

$(0,\infty)\times(s_j,s_{j+1})$, $s_j < s_{j+1}$. The average and variance of the *LLR* values of the points contained in each strip are used to compute the parameters of the best fitting Gumbel distribution for that particular strip. Let G_j be the Gumbel distribution for the strip $(0,\infty)\times(s_j,s_{j+1})$. The values $s_j < s_{j+1}$ are thus chosen close enough such that the marginal distribution of D does not change much in the interval (s_j,s_{j+1}) and also there are sufficient points inside the strip to evaluate appropriately the parameters of G_j. Let $P_0 = (l_0,c_0)$ be a point belonging to the observed cases Pareto-set. The Gumbel distribution G_j for the strip $(0,\infty)\times(s_j,s_{j+1})$ containing the point P_0 is then used to compute the estimated p-value for P_0 as $\int_{l_0}^{\infty} G_j(t)dt$.

The example of Figure 2 displays the Gumbel approximations for the four strips selected, with the corresponding histograms of *LLR* empirical values, using data from the Northeast US map benchmark, now with 20,000 sets of 600 cases randomly distributed under the null hypothesis. The 20,000 Pareto-sets found by the multi-objective algorithm constitute a total of about 305,000 points. The many thousands of points with *K*=1 (corresponding to clusters containing only one region) were removed from this figure, for clarity.

We obtain a more accurate significance computation, using the p-values estimates for all the strips to build the *interpolated p-value surface*. The individual p-values for each one of the points in the observed cases Pareto-set are thus computed by means of the *p-values isoclines*. Figure 3 displays several isoclines, with the p-values indicated, along with the computed points used in their interpolation. The same data of Figure 2 was used here. P-values isoclines as low as 10^{-27} may be traced in this example, although the accuracy is diminished. Arbitrary precision arithmetic may be necessary to compute those parameters for those extremely small p-values.

One could be concerned about the occurrence of multiple testing, since we are dealing simultaneously with various shapes in the Pareto-sets of the map of observed cases. This in fact does not happen since the null hypothesis maps also produce Pareto-sets associated with the simulated cluster solutions of various shapes, using exactly the same algorithm used for finding the Pareto-sets corresponding to the clusters of observed cases.

We conduct a simple experiment to check the adequacy of the Gumbel distribution approximation. Table 1 displays comparative results of the 20,000 Pareto-sets (about 305,000 points) simulation under null hypothesis for the Northeast US map for some strips at different quantiles (0.1, 0.05, 0.01, 0.001 and 0.0001). The method of section 3 was used to adjust the tail of the Gumbel distributions. The expected number of values within each quantile semi-infinite interval, according to the Gumbel tail-adjusted distribution, is compared to the number of values that have occurred in the simulation, in parenthesis. The last columns show the parameters for the original Gumbel distributions (μ_0,θ_0) and for the tail-adjusted ones (μ_{opt},θ_{opt}). The good agreement between the estimated and observed number of values for each quantile suggests the plausibility of employing the Gumbel distributions approximations.

4.4 A SIMULATED CLUSTER EXAMPLE

We pick up a set of random allocations of 600 cases in the artificial cluster *A* alternative hypothesis model from the Northeast US benchmark (see Duczmal et al., 2006b). It is an elongated cluster located along the Connecticut River, just to the right of the center in the Northeast US inset map of Figure 4. This shaded map displays the incidence rates for this particular random allocation of cases. Figure 4 shows the Pareto-set solution found by the multi-objective genetic algorithm, consisting of 7 cluster points within the p-values isoclines graph, along with the corresponding clusters I-VII, displayed as 7 detailed inset maps. Table 2 shows *LLR* value, compactness, cluster size (number of component regions), population, number of cases, incidence rate and computed p-value for each cluster. Cluster I is relatively round. Adding and subtracting one region, we obtain cluster II, decreasing the compactness and increasing the *LLR* value. When we pass to cluster III, again exchanging one region, there is a significant increase in the *LLR* value, and the compactness is only slightly reduced. The increase in the p-value is clearly seen, as the cluster point moves closer to the 0.001-value isocline. The next exchange of regions, producing cluster IV, decreases the compactness without improving appreciably the *LLR* value. As expected, the cluster point moves away from the 0.001-value isocline, reducing its p-value. Clusters V and VI have increased *LLR* values and decreased compactness, and the p-values increase. The last region exchange, producing cluster VII, is not a good deal because the *LLR* value increase is not compensated by the much lower compactness, thus reducing the p-value. The best solution, measured by the p-value criterion is thus cluster III, followed closely by cluster VI.

5. POWER EVALUATION

Given an alternative hypothesis model, hundreds of runs of the multi-objective algorithm produce the correspondent Pareto-sets, which are joined and compared to the 0.05 p-value isocline, obtained under null hypothesis, as described in section 4.3. The proportion of points to the right of the isocline is an estimate of the *average power* of the algorithm for that particular alternative hypothesis model. A more detailed power analysis considers the proportion of points to the right of the 0.05 isocline for each horizontal strip. Table 3 shows the power comparison between the multi-objective genetic algorithm (MGA) and the mono-objective compactness corrected genetic algorithm (GA), using the same datasets benchmark of section 3, with 6 artificial irregularly shaped clusters A-F (Duczmal et al. 2006b,c). The size and compactness are displayed for each cluster. For the MGA, we show the average power (considering all the solution points) and consider the power separately for each strip. The *maximum power* is defined as the maximum power value among all the strips, indicated by bold face values in Table 1. The non-compactness penalty correction parameter *a* was set to 1 (full correction) in the GA. For each cluster, the Monte Carlo was based on 10,000 Pareto-sets, or about 150,000 cluster points (MGA) and 10,000 most likely clusters (GA), plus the two corresponding null hypotheses simulations.

Although the MGA average power is generally lower than the GA power, the MGA maximum power was about the same as the GA power. Interestingly, the compactness of the simulated cluster always belongs to the compactness interval corresponding to the strip where maximum power was attained. In other words, the GA seems to have maximum efficiency through the compactness correction. This suggests that the compactness penalty function "normalizes" the GA search, balancing the relative importance of the clusters of different shapes.

6. AN APPLICATION TO BREAST CANCER CLUSTERS

Figure 5 displays the isoclines for several p-values, built with the methods of section 3 and 4.3, based on 20,000 Pareto-sets null hypothesis Monte Carlo simulations, along with the Pareto-set of observed cases for the breast cancer deaths real data for Northeast US map. Arbitrary precision arithmetic was necessary for the computation of the very low p-values. The corresponding clusters are described in Table 4, displaying *LLR* value, compactness, size, population, number of cases, incidence rate and approximated computed p-value. Figure 5 shows the selected clusters *a, c , f, g, o* and *r* within the Pareto-set. Observe that the upper left points correspond to low scan valued round shaped clusters, and the lower right points correspond to the high *LLR*, more irregularly shaped clusters. In this example, the most irregularly shaped clusters are the most significant, with estimated p-values smaller than 10^{-27}. Although the individual p-value accuracy is lower for those clusters, their relative positions in the isoclines graph are accurate enough for deciding which ones are the most significant. The cluster *a* is a secondary cluster, consisting of only one region (see Kulldorff et al., 2003).

7. DISCUSSION

We described a quantitative criterion for the problem of the geographic delineation of irregularly shaped spatial clusters of disease. Given a set of optimal solutions found by a cluster finder algorithm, the problem was reduced to choosing the most significant solution among them. We developed a genetic algorithm that finds the Pareto-set solution, based on the maximization of two objectives, the scan statistic and the regularity of cluster shape (compactness). This multi-objective algorithm presents a set of cluster solutions, which should be ranked according to their individual statistical significance. We extended the usual concept of significance for the multi-objective problem by means of a p-value surface in the two variables scan statistic and compactness. The p-value isoclines are computed with the help of the Gumbel approximation, allowing higher accuracy and computational speed for the significance estimation. The possibility of identifying the "best solution" through a quantitative criterion gives a new insight in the process of finding irregularly shaped clusters. The former attempts of dealing with the problem of simultaneously taking into account the maximization of LLR value and the choice of a suitable cluster geometry, in mono-objective optimization settings, necessarily lead to some arbitrary choice of the trade-off between these objectives that could not be satisfactorily accommodated in theory. The introduction of the concept of Pareto-set in this problem, followed by the choice of the most significant solution, is shown to allow a rigorous statement about what is such "best solution", without the need of any arbitrary parameter.

The numerical evaluations show that the multi-objective algorithm average power is slightly inferior to the power of the mono-objective genetic algorithm. But when we measure the power of the multi-objective algorithm separately for each strip and choose the maximum value (*maximum power*) we observe that this value is about the same as the power of the mono-objective genetic algorithm. The maximum power definition is useful here because it measures the power of the algorithm to detect a cluster when the search is restricted to clusters with compactness within a certain interval; as we have seen, this interval always contains the value of the compactness of the cluster that we are trying to detect.

The algorithm is as fast as the genetic mono-objective. One interesting question for a future work concerns the disposition of the Pareto-set against the p-value isoclines: in the simulated cluster example of this paper, the maximum significance was in the middle of the Pareto-set list, but in the real data example the maximum significance cluster was attained at the extreme of the list. As shown in the breast cancer real data example, there is a potential for finding secondary clusters within the Pareto-set, and that possibility needs further investigation. A surprising feature of that example refers to the extremely small p-values for those real-data clusters. That provides a dramatic illustration of the high sensitivity of the scan statistic. Other algorithms could also possibly be adapted for the multi-objective search, instead of using a genetic algorithm.

ACKNOWLEDGEMENTS

The authors wish to thank Martin Kulldorff for the thoughtful comments.

8. 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.*

Assuncao R, Tavares A, Costa M, Ferreira S, 2006. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine*, 25;1-21.

Carrano EG, Soares LAE, Takahashi RHC, Saldanha RR, Neto, OM, 2006. Electric distribution network multiobjective design using a problem-specific genetic algorithm. *IEEE Transactions on Power Delivery,* 21(2):995-1005.

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

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

Deb K, Pratap A, Agrawal S, Meyarivan T, 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation.* 6(2):182-197.

Duczmal L, Assuncao 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., 2005. Using modified Spatial Scan Statistic to Improve Detection of Disease Outbreak When Exposure Occurs in Workplace – Virginia, 2004. *Morbidity and Mortality Weekly Report*, Vol.54 Suppl.187.

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. (to appear).*

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

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

Fonseca CM, Fleming P. 1995. An overview of evolutionary algorithms in multiobjective optimization. *Evolutionary Computation*, 3(1):1-16.

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

Kulldorff M, Nagarwalla N, 1995. Spatial disease clusters: detection and inference. *Statistics in Medicine,* 14, 779-810.

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

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.

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, 2006a. 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. (submitted to *Statistics in Medicine*)

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

Neill DB, Moore AW, Maheshkumar RS, Daniel K, 2005. An Expectation-Based Scan Statistic for Detection of Space-Time Clusters, *2005 Syndromic Surveillance Conference*.

Nepomuceno EG, Takahashi RHC, Amaral GFV, Aguirre LA, 2003. Nonlinear identification using prior knowledge of fixed points: a multiobjective approach. *International Journal of Bifurcation and Chaos*, 13(5):1229-1246.

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

Ramos RM, Saldanha RR, Takahashi RHC, Moreira FJS, 2003. The real-biased multiobjective genetic algorithm and its application to the design of wire antennas. *IEEE Transactions on Magnetics*, 39(3):1329-1332.

Sahajpal R, Ramaraju GV, Bhatt V, 2004. Applying niching genetic algorithms for multiple cluster discovery in spatial analysis. *Int. Conf. on Intelligent Sensing and Information Processing.*

Takahashi RHC, Palhares RM, Dutra DA, Gonçalves LPS, 2004. Estimation of Pareto sets in the mixed H2/H-infinity control problems. *International Journal of Systems Science*, 35(1):55-67.

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.

Teixeira RA, Braga AP, Takahashi RHC, Saldanha RR, 2000. Improving generalization of MLPs with multi-objetive optimization. *Neurocomputing*, 35(4):189-194.

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

Figure 1

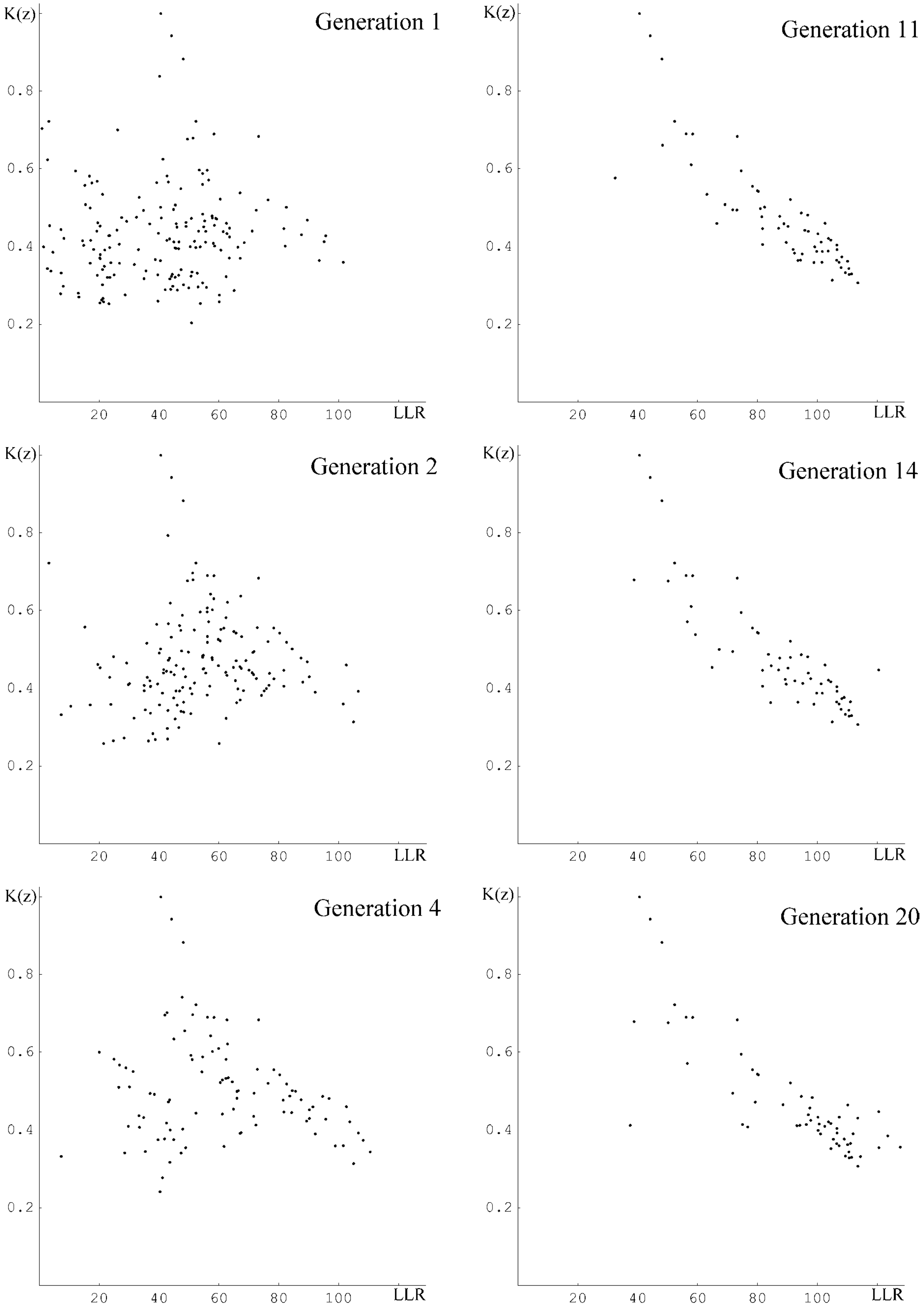
Generation 1
Generation 11
Generation 2
Generation 14
Generation 4
Generation 20
K(z)
0.8
0.6
0.4
0.2
20
40
60
80
100
LLR

Figure 2

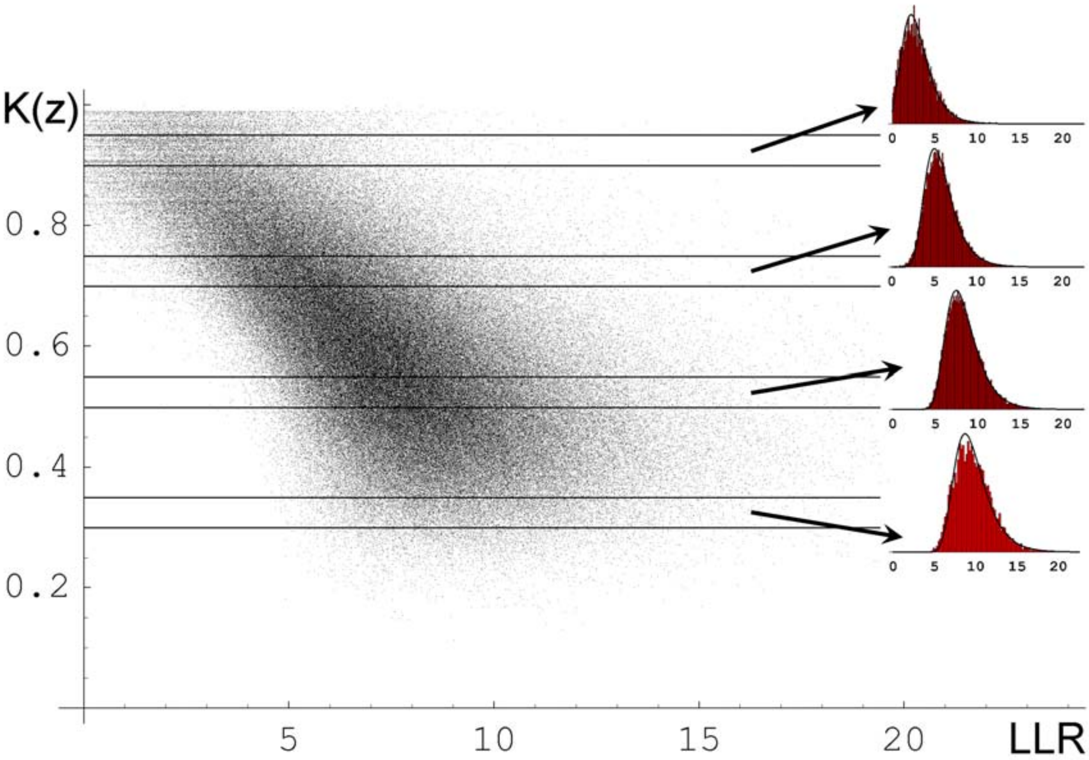
K(z)
0.8
0.6
0.4
0.2
5
10
15
20
LLR
0 5 10 15 20
0 5 10 15 20
0 5 10 15 20
0 5 10 15 20

Figure 3

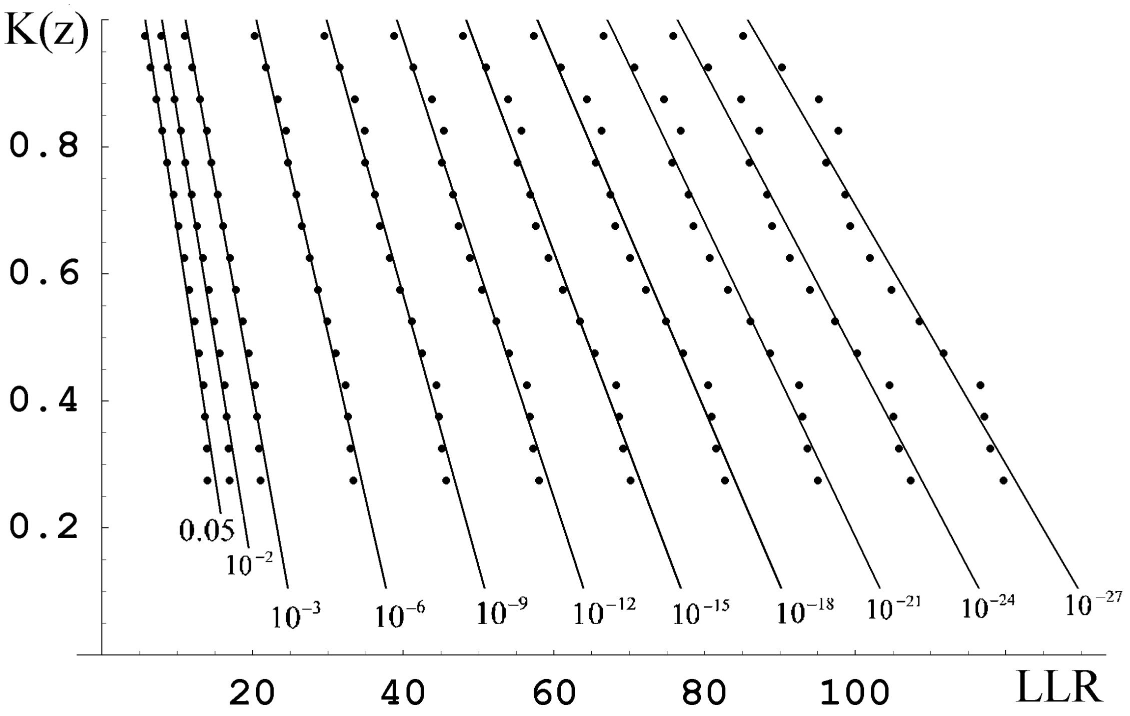
K(z)
0.8
0.6
0.4
0.2
0.05
10^{-2}
10^{-3}
10^{-6}
10^{-9}
10^{-12}
10^{-15}
10^{-18}
10^{-21}
10^{-24}
10^{-27}
20
40
60
80
100
LLR

Figure 4

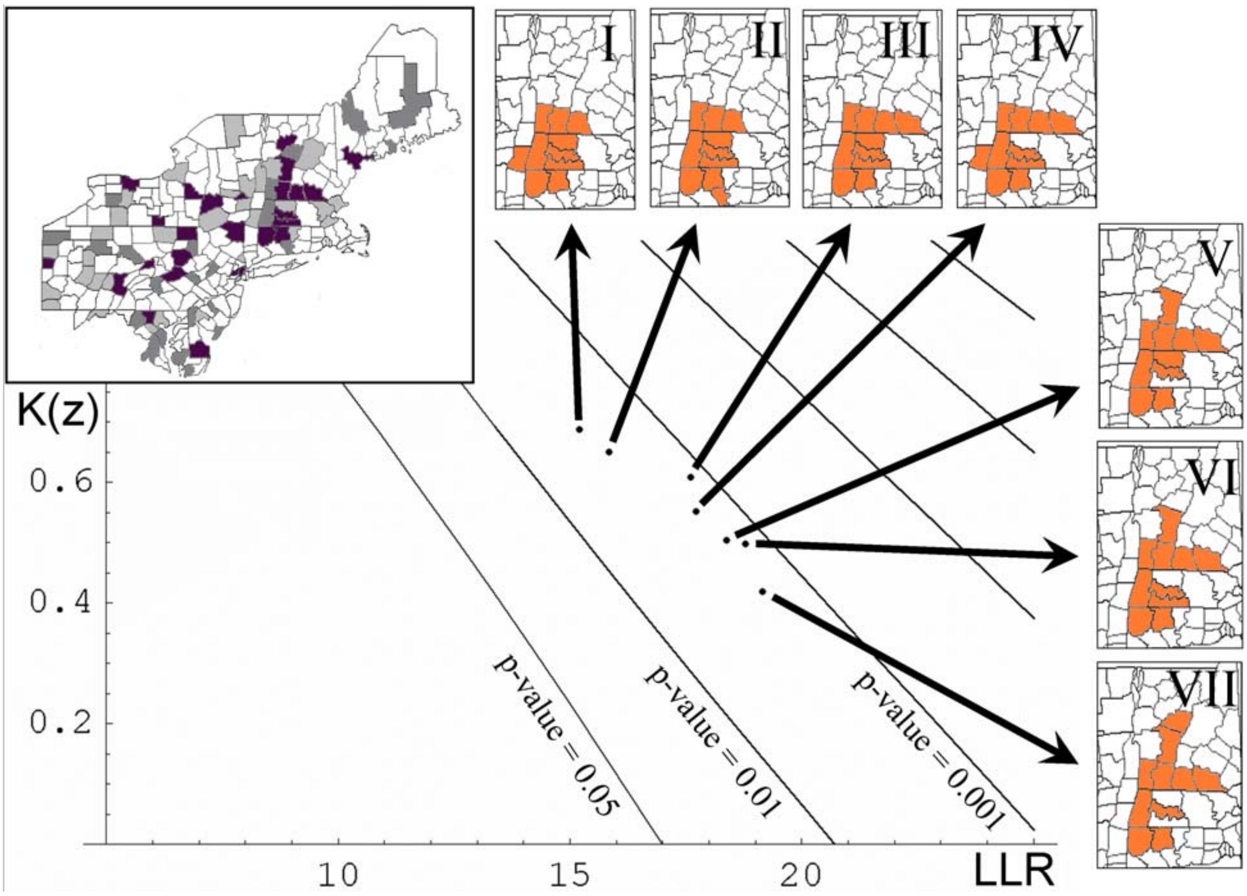
I
II
III
IV
V
VI
VII
K(z)
0.6
0.4
0.2
10
15
20
LLR
p-value = 0.05
p-value = 0.01
p-value = 0.001

Figure 5

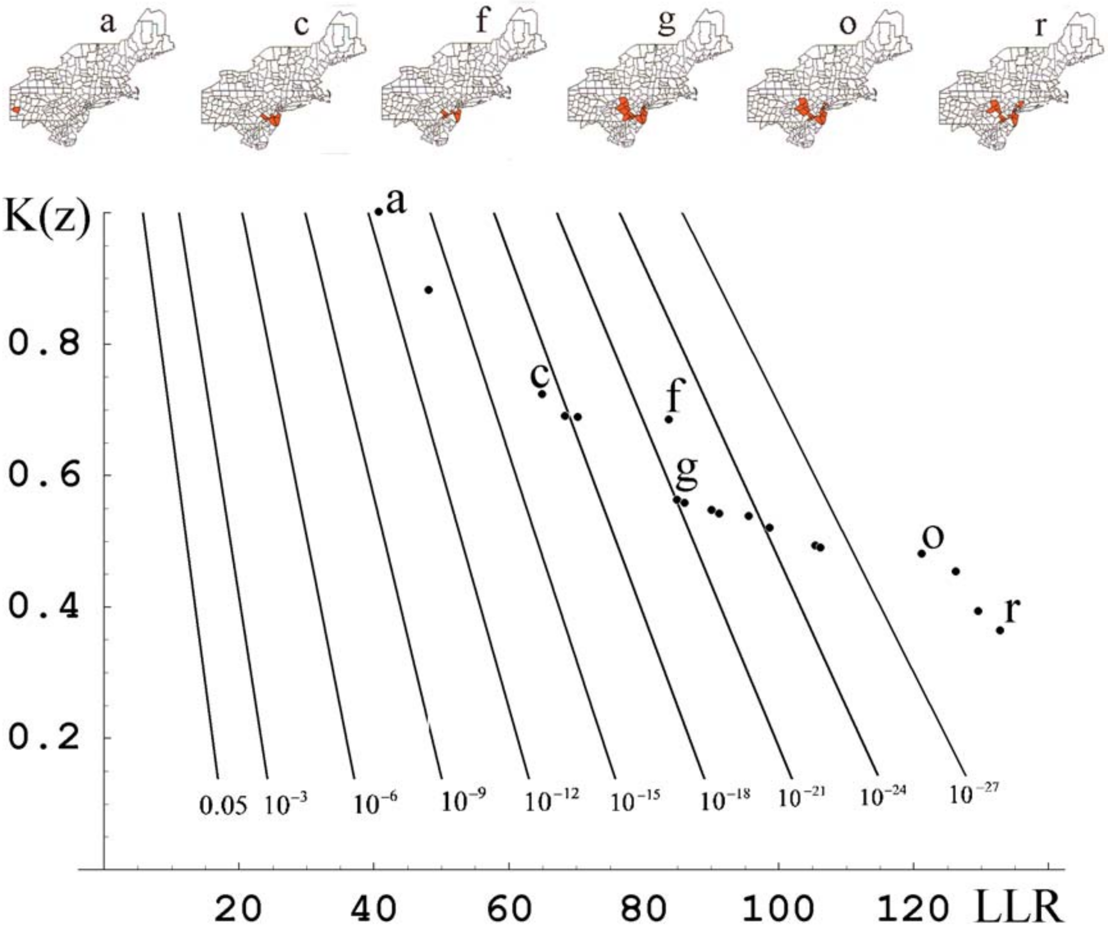
a
c
f
g
o
r
K(z)
0.8
0.6
0.4
0.2
0.05
10^{-3}
10^{-6}
10^{-9}
10^{-12}
10^{-15}
10^{-18}
10^{-21}
10^{-24}
10^{-27}
20
40
60
80
100
120
LLR

Table 1

strip	***N***	**quantiles**					**parameters**			
		0.1000	**0.05**	**0.01**	**0.001**	**0.0001**	μ_0	θ_0	μ_{opt}	θ_{opt}
.95, 1.0	8003	800.3 (838)	400.2 (441)	80.0 (77)	8.0 (6)	0.800 (1)	1.73	1.34	2.02	1.24
.90, .95	10911	1091 (1140)	545.6 (576)	109.1 (121)	10.9 (13)	1.091 (0)	2.22	1.42	2.66	1.23
.85, .90	13805	1380 (1454)	690.2 (744)	138.1 (134)	13.8 (16)	1.380 (1)	2.78	1.49	3.15	1.30
.80, .85	16874	1687 (2050)	843.7 (981)	168.7 (153)	16.9 (14)	1.687 (3)	3.49	1.52	3.36	1.45
.75, .80	19954	1995 (2135)	997.7 (1038)	199.5 (201)	20.0 (19)	1.995 (2)	4.30	1.48	4.55	1.32
.70, .75	25122	2512 (2685)	1256 (1388)	251.2 (238)	25.1 (26)	2.512 (5)	5.00	1.51	5.24	1.35
.65, .70	28826	2883 (3172)	1441 (1648)	288.3 (327)	28.8 (26)	2.883 (4)	5.70	1.51	5.95	1.34
.60, .65	32137	3214 (3378)	1607 (1724)	321.4 (314)	32.1 (34)	3.214 (1)	6.36	1.54	6.59	1.41
.55, .60	34043	3404 (3607)	1702 (1778)	340.4 (348)	34.0 (39)	3.404 (0)	6.94	1.57	7.21	1.43
.50, .55	33892	3389 (3613)	1695 (1839)	338.9 (346)	33.9 (38)	3.389 (0)	7.46	1.63	7.71	1.48
.45, .50	30870	3087 (3307)	1544 (1758)	308.7 (392)	30.9 (35)	3.087 (3)	7.93	1.67	8.43	1.42
.40, .45	23827	2383 (2358)	1191 (1221)	238.3 (279)	23.8 (24)	2.383 (1)	8.29	1.74	8.79	1.53
.35, .40	15430	1543 (1678)	771.5 (864)	154.3 (169)	15.4 (17)	1.543 (1)	8.49	1.75	8.83	1.54
.30, .35	7854	785.4 (820)	392.7 (425)	78.5 (90)	7.9 (6)	0.790 (2)	8.70	1.76	9.45	1.41
.25, .30	2615	261.5 (280)	130.8 (132)	26.2 (39)	2.6 (0)	0.262 (0)	8.74	1.78	9.02	1.62
.20, .25	558	55.8 (53)	27.9 (22)	5.6 (9)	0.6 (0)	0.056 (0)	8.66	2.00	8.66	1.99
.15, .20	94	9.4 (9)	4.7 (5)	0.9 (1)	0.1 (0)	0.009 (0)	8.73	1.57	8.65	1.41

Table 2

cluster	size	LLR	K(z)	population	cases	rate × 100,000	p-value
I	10	15.21	0.686	1,062,145	51	4.8016	0.0034
II	10	15.85	0.650	1,103,353	53	4.8035	0.0029
III	10	17.60	0.607	1,201,644	58	4.8267	0.0013
IV	10	17.72	0.551	1,197,728	58	4.8425	0.0019
V	10	18.38	0.504	990,792	52	5.2483	0.0017
VI	10	18.79	0.498	1,193,323	59	4.9442	0.0014
VII	10	19.15	0.418	967,908	52	5.3724	0.0019

Table 3

	A	**B**	**C**	**D**	**E**	**F**
K(z)	.38	.42	.30	.52	.44	.20
size	13	16	7	15	21	23
GA power	.85	.83	.79	.88	.80	.46
MGA power (average)	.70	.62	.65	.75	.68	.29
strip						
.95, 1.0	.04	.02	.10	.04	.04	.04
.90, .95	.16	.02	.18	.16	.13	.03
.85, .90	.26	.09	.25	.35	.28	.03
.80, .85	.42	.22	.37	.49	.37	.05
.75, .80	.57	.36	.47	.63	.49	.08
.70, .75	.63	.52	.54	.70	.58	.11
.65, .70	.71	.62	.61	.79	.68	.15
.60, .65	.76	.70	.62	.84	.75	.19
.55, .60	.78	.74	.68	.87	.80	.25
.50, .55	.82	.77	.73	**.89**	**.82**	.31
.45, .50	**.83**	**.78**	.78	**.89**	**.82**	.38
.40, .45	**.85**	**.79**	**.80**	.88	**.82**	.42
.35, .40	**.83**	**.77**	**.82**	.85	.80	**.45**
.30, .35	.79	.73	**.81**	.82	.77	**.46**
.25, .30	.77	.68	.79	.71	.73	**.45**
.20, .25	.74	.51	.76	.70	.69	**.45**

Table 4.

cluster	size	LLR	K(z)	pop	Cases	annual rate $\times 100{,}000$	$\log_{10}$ (p-value)
a	1	40.6	1.00	710196	1765	49.70	-12
b	5	48.1	0.88	1732559	4031	46.53	-14
c	6	52.3	0.72	2085959	4814	46.16	-17
d	8	56.2	0.69	2715157	6177	45.50	-18
e	6	58.4	0.69	2019613	4709	46.63	-18
f	6	73.3	0.68	2173282	5125	47.16	-22
g	15	74.7	0.56	4017127	9052	45.07	-21
h	14	75.9	0.55	3580296	8140	45.47	-21
i	9	80.3	0.54	3176964	7314	46.04	-22
j	15	81.4	0.54	5121469	11411	44.56	-22
k	15	86.3	0.53	3993404	9081	45.48	-24
l	14	89.8	0.52	3824703	8747	45.74	-24
m	15	97.5	0.49	4418903	10050	45.49	-26
n	15	98.3	0.48	4251402	9702	45.64	-26
o	15	114.8	0.48	4255849	9812	46.11	-29
p	15	120.5	0.45	4224819	9779	46.29	-30
q	15	124.2	0.39	4655988	10714	46.02	-30
r	15	127.7	0.36	4511453	10428	46.23	-30