What is the true shape of a disease cluster? The multi-objective genetic scan

Luiz Duczmal , André L. F. Cançado , Ricardo H. C. Takahashi
Universidade Federal de Minas Gerais, Brazil

OBJECTIVE

We propose a novel approach to the delineation of irregularly shaped disease clusters, treating it as a multi-objective optimization problem. We present 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 among the many possible ones.

BACKGROUND

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 [1]. 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 [2].

METHODS

We propose quantitative criteria for choosing the best cluster solution, maximizing simultaneously two competing objectives: regularity of shape (K(z)), and scan statistic value (LLR). The *Pareto set* consists of all clusters candidates *z* such that no other cluster has both higher LLR and higher regularity than *z*. For each value of K(z), a separate empirical distribution of LLR under null-hypothesis is computed, constituting a two-dimensional p-value surface. The cluster with lowest p-value is considered the most likely cluster [2]. Instead of running a cluster finding algorithm with varying degrees of penalization, the complete set of solutions is found in parallel, through a genetic algorithm. The p-value surface is computed using Gumbel approximations [3] (Figure 1). Although different shapes are dealt with simultaneously, multiple testing does not occur, since the null hypothesis maps also produce Pareto-sets using exactly the same algorithm as the observed cases map.

RESULTS

Figure 2 shows an example for a simulated cluster. In this particular example, the p-values isolines are approximated by straight lines. A real data application for breast cancer is discussed. The method is fast, with power of detection similar to the genetic, the simulated annealing and the elliptic scans.

CONCLUSIONS

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 arbitrary parameters. The automatic selection process and the high speed of the method remove two stumbling blocks for the utilization of irregular cluster detection methods in syndromic surveillance.

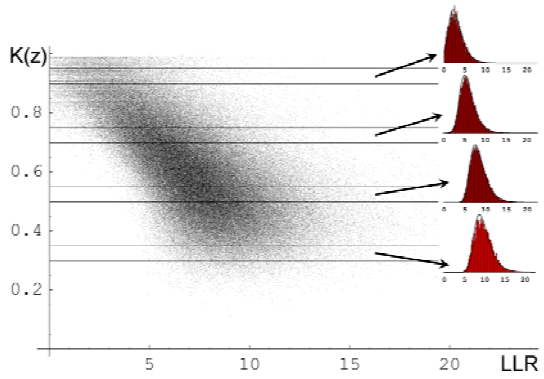

Figure 1 – The Gumbel parametric approximation of the null hypothesis empirical distribution in the LLR versus K(z) space.

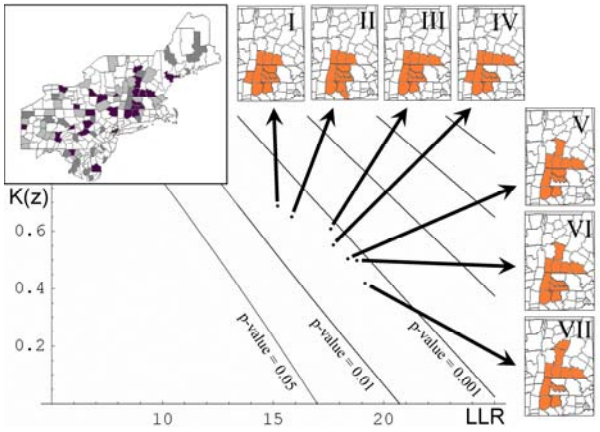

Figure 2 – Each point I-VII in the Pareto-set is a cluster solution for this simulated New England map (inset). Cluster III has the lowest p-value, and is hence considered as the most likely cluster.

REFERENCES

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

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

[3] Abrams A, Kulldorff M, Kleinman K, 2006. Empirical/Assymptotic P-values for Monte Carlo-Based Hypothesis Testing: an Application to Cluster Detection Using the Scan Statistic. *Advances in Disease Surveillance, Vol. 1.*

Corresponding author: duczmal@ufmg.br