A NON-CONNECTIVITY WEIGHTED PENALTY FUNCTION FOR IRREGULAR-SHAPED CLUSTER DETECTION

Anderson Ribeiro DUARTE[1]
Spencer Barbosa da SILVA[1]
Fernando Luiz Pereira de OLIVEIRA[1]
Marcelo Carlos RIBEIRO[2]
André Luiz Fernandes CANÇADO[3]
Flávio dos Reis MOURA[1]

▪ *ABSTRACT: Methods for the detection and inference of irregularly shaped geographic clusters with count data are important tools in disease surveillance and epidemiology. Recently, several methods were developed which combine Kulldorff's Spatial Scan Statistic with some penalty function to control the excessive freedom of shape of spatial clusters. Different penalty functions were conceived based on the cluster geometric shape or on the adjacency structure and non-connectivity of the cluster associated graph. Those penalty function were also implemented using the framework of multi-objective optimization methods. In particular, the non-connectivity penalty was shown to be very effective in cluster detection. Basically, the non-connectivity penalty function relies on the adjacency structure of the cluster's associated graph but it does not take into account the population distribution within the cluster. Here we introduce a modification of the non-connectivity penalty function, introducing weights in the components of the penalty function according to the cluster population distribution. Our methods is able to identify multiple clusters in the study area. We show through numerical simulations that our weighted non-connectivity penalty function outperforms the original non-connectivity function in terms of power of detection, sensitivity and positive predictive value, also being computationally fast. Both single-objective and multi-objective versions of the algorithm are implemented and compared.*

[1] Universidade Federal de Ouro Preto - UFOP, Departmento de Estatística, Ouro Preto, MG, Brazil. E-mail: *duarte.andersonr@gmail.com*; *spencerbars@gmail.com*; *fernandoluizest@gmail.com*; *prof.flaviomoura@gmail.com*

[2] Universidade Federal de Viçosa - UFV, Departmento de Estatística, Viçosa, MG, Brazil. E-mail: *prof.marcelocarlosribeiro@gmail.com*

[3] Universidade de Brasília - UnB, Departmento de Estatística, Brasília, DF, Brazil. E-mail: *cancado@gmail.com*

▪ *KEYWORDS: Spatial Scan statistic; irregular clusters; multi-objective algorithms; compactness Function; non-connectivity function; weighted non-connectivity function.*

1 Introduction

Consider some study area represented by a map divided in regions where some kind of occurences (disease, crimes, etc.) are distributed among the regions. We define a spatial cluster as a connected set of regions where the risk of some occurence is anomalously high or low compared with the map remainder regions. The delineation of geographic clusters is a valuable tool in epidemiology (see LAWSON *et al.*, 1999; BALAKRISHNAN; KOUTRAS, 2002; BUCKERIDGE *et al.*, 2005; LAWSON, 2009 and LAWSON, 2010). One method for the detection and inference of spatial clusters is the Circular Scan (see KULLDORFF; NAGARWALLA, 1995), a particular case of Kulldorff's Spatial Scan Statistic (see KULLDORFF, 1997). The circular scan is efficient when the cluster has a regular shape. However, disease clusters with arbitrary shapes occur along traffic ways, plumes of air pollution or geographical features such as rivers, shores or valleys. Several methods for detecting irregularly shaped clusters have been developed recently. A recent review may be found in Duczmal *et al.* (2009).

A irregularly shaped spatial cluster detection algorithms frequently may end up with a cluster solution that is merely a collection of the high incidence regions, linked together forming a "tree-shaped" zone spread out through the map. In general it is hard to give a geographical meaning for this kind of cluster, because this solution does not add any new information with regard to its special location in the map. This in turn motivates the use a penalty function as a mechanism to prevent excessive freedom on the shape of a possible cluster solution. Such penalty functions, also called regularity functions, are combined with the spatial scan statistic producing a penalized spatial scan statistic. Among those penalty functions, we mention the geometric penalty function presented in (see KULLDORFF *et al.*, 2006; DUCZMAL *et al.*, 2006; DUCZMAL *et al.*, 2007; DUCZMAL *et al.*, 2008; DUARTE *et al.*, 2010) and the non-connectivity penalty function presented in Yiannakoulias *et al.* (2007).

In the present paper we introduce a modification of the non-connectivity penalty function. The original non-connectivity penalty function as proposed in Yiannakoulias *et al.* (2007) takes in account only the cluster associated graph connectivity structure. Our proposal, denominated weighted non-connectivity penalty function, incorporates weights in the components of the original non-connectivity function according to the cluster population distribution. Our methods is able to identify multiple clusters in the study area. Details will be presented in Section 2.

Duczmal *et al.*, 2008 and Duarte *et al.*, 2010 uses a multi-objective genetic algorithm to the problem of detecting irregular clusters. Instead of just one objective, namely, to search among all the possible clusters the one which maximizes the penalized scan statistic, this new method seeks to maximize two objectives, namely the spatial scan statistic and some penalty function. The multi-objective algorithm presents a major advantage: all clusters are considered to be a potential one without a classification according to the values of the penalty. So the rating on the quality of possible solutions is performed only after all candidates are evaluated.

We compare three multi-objective methods for clusters detection using the spatial scan statistic as the first objective and the geometric compactness, the non-connectivity and our new proposal the weighted non-connectivity as the second objective. Those methods are compared with the corresponding single-objective penalized likelihood methods. We use numerical simulations to study the power to detect irregularly shaped clusters, sensitivity and positive predict value of those methods. The rest of this paper is organized as follows. In the next section, we present a brief review of the literature on spatial statistics and genetic algorithms. Section 2 we introduced our proposed of the weighted non-connectivity function. Then, some results from datasets analysis are presented and discussed. Final remarks and topics for future research in the area close this paper.

1.1 The spatial Scan statistics

Consider the study map A, divided into M regions, with total population N and total number of cases C. A non-oriented graph G_A is associated with the study map A. In the associated graph G_A there are M nodes each one corresponding to one region of the map and edges connecting nodes associated with adjacent regions, so this is the graph associated with the adjacency matrix. A zone is any collection of connected regions and corresponds to a sub-graph of the associated graph. Whereas a cluster is a z zone whose the rate of occurrence of the study phenomenon is discrepant, the null hypothesis states that there are no clusters in the map. Under the null hypothesis the number of cases (occurrences of the phenomenon of interest) in each region is Poisson distributed proportionally to its population. For each zone z, the number of observed cases is c_z, the population is n_z and the expected number of cases under the null hypothesis is $\mu_z = C(n_z/N)$. The relative risk of a zone z is $I(z) = c_z/\mu_z$ while the relative risk outside the zone z (complement of z) is given by $O(z) = (C - c_z)/(C - \mu_z)$. Denoting L_0 as the likelihood function under the null hypothesis and $L(z)$ as the likelihood function under the alternative hypothesis that there is a cluster on the map in the study it can be shown (see for details) that the logarithm of the likelihood ratio is given by:

$$LLR(z) = \begin{cases} c_z \log(I(z)) + (C - c_z)\log(O(z)) & \text{if } I(z) > 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

A set Z of zones is chosen according to some restrictive criteria in order to avoid a exhaustive search over all possible zones but instead only on a set of promising ones, identifying the zone that maximizes the likelihood ratio function called the most likely cluster. A widely used choice for the set Z is when it is composed of zones defined by circular windows of different radii and centers. In such case

$\max_{z \in Z} LLR(z)$ is the circular scan statistic proposed by Kulldorff and Nagarwalla (1995). The statistical significance of the most likely cluster is obtained through a Monte Carlo simulation (see DWASS, 1957). Under the null hypothesis and conditioned on the total number of observed cases, simulated cases are distributed among the regions of the map under study. Given the simulated cases distribution the scan statistic is calculated for the most likely cluster. This procedure is repeated thousands of times, and an empirical distribution of values for the likelihood ratio is obtained. An estimated p-value for the most likely cluster of observed cases is obtained when its likelihood ratio value is compared with the empirical distribution.

As previously mentioned, algorithms for detecting spatial clusters making an unrestricted search can eventually choose a cluster that spreads across the whole map just connecting regions with high relative risk. One way to avoid such kind of "meaningless" solution uses an algorithm that considers the $LLR(z)$ together with some sort of penalty for the possible cluster shape. In order to detect clusters using the spatial scan statistic combined with some penalty function, we can maximize the product of the $LLR(z)$ with the penalty function or try to simultaneously maximize the penalty function and the $LLR(z)$.

1.2 The non-connectivity penalty function

Yiannakoulias *et al.* (2007) proposed a greedy algorithm to scan the set Z of all possible zones z. A penalty function called non-connectivity was proposed. It was based on the ratio of the number of nodes $v(z)$ to the number of edges $e(z)$ of the subgraph associated with the zone z. The non-connectivity penalty function of a zone z is defined by:

$$nc(z) = \frac{e(z)}{[3\,(v(z) - 2)]} \tag{2}$$

the expression in the denominator represents the maximum number of edges of a planar graph given its number of vertices. The most penalized zones are the ones with tree-like associated graphs, meaning that they have a small number of nodes compared with the number of edges.

1.3 Genetic algorithms

Many stochastic optimization methods were proposed for the detection of irregularly spatial clusters (see DUCZMAL; ASSUNÇÃO, 2004; CONLEY *et al.*, 2005; PEI *et al.*, 2011; WAN *et al.*, 2012; COSTA; KULLDORFF, 2014). In this work we use the genetic algorithm proposed in Duczmal *et al.*, (2007) and later adapted for the multi-objective framework in Duczmal *et al.*, (2008).

Genetic algorithms (GAs) are powerful tools mainly used to approximate the solution of complex optimization problems. The GA starts with a set of solutions randomly chosen in the search space. This set is called "population". The GA then performs a series of operations over the population, generating a new population that is expected to be more “adapted” than the first one. A typical GA uses at least the following operators:

- crossover: new individuals are generated combining the information of two or more individuals of the current population;
- mutation: new individuals are generated applying random perturbations over individuals of the current population;
- selection: choice of individuals that will compose the population in the subsequent generation.

There are many ways to implement each of these operators and many other operators have been proposed. These operations are then performed over and over again, generating a sequence of populations. In the end of the process, the last population is expected to contain the most adapted individuals, i.e., the solutions that optimize the objective function.

In the context of spatial clusters, given a map of regions with count data, the algorithm objective is to identify the zone that maximizes Kulldorff’s spatial scan statistic (eqn. 1). A detailed description of the algorithm and its operators can be found in Duczmal *et al.*, (2007).

In the multi-objective approach to the cluster detection problem, the best cluster solutions are found by maximizing simultaneously two competing objectives, namely Kulldorff’s logarithm of the likelihood ratio $LLR(z)$ and some penalty or regularity function. In this approach the regularity function is no longer used as a penalty correction to the $LLR(z)$ function but, instead, as another objective function. GAs are quite efficient tools for dealing with multi-objective optimization problems because they are capable of evolving the whole population in parallel towards a set of optimal solutions in the objective space (see FONSECA; FLEMING, 1995; TAKAHASHI *et al.*, 2003).

The construction of the initial population, the crossover and the mutation operators are identical to those used in the single-objective genetic algorithm. The selection operator uses the concept of dominance: a solution x is *dominated* if there is a solution y such that x is worse than y in at least one objective, while not being better than y in any other objective (see CHANKONG; HAIMES, 1983). The *non-dominated set* consists of all solutions that are not dominated by any other solution. The non-dominated solutions are also called *efficient* solutions.

Such non-dominated set is then computed for the observed data and for each of the replications under the null hypothesis. Let x be a solution in the non-dominated set obtained for the observed data. If, for a specific simulation under the null hypothesis, the obtained non-dominated set contains at least one solution that dominates x, then we say that this non-dominated set *attains* x. To assess the significance of the solution x we can then compute the proportion of non-dominated sets obtained under the null hypothesis containing at least one solution that dominates x, that is, the proportion of non-dominated sets that attain x. More details on the computation of the attainment function can be seen on Fonseca *et al.* (2005) and Duczmal *et al.*, (2008).

2 Weighted non-connectivity function

The non-connectivity penalty function (eqn. 2) proved to be quite eficient in the detection and inference of spatial clusters (see YIANNAKOULIAS *et al.*, 2007). However it does not consider the population heterogeneity among the component areas. In epidemiology and disease surveillance, the population heterogeneity is clearly an important feature to be included in cluster analysis. In this context, we could ask how relevant an edge is for the subgraph connectivity. We will modify the non-connectivity penalty function in order to distinguish the relative importance of the edges, by considering the populations.

It was observed, through numerical experiments that irregularly shaped clusters detection algorithms improve their power of detection by using penalty functions (see DUCZMAL *et al.*, 2006). 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.

Besides considering the associated graph connectivity structure we propose to assign weights to the edges and the nodes according to their associated areas' populations. For an edge $e_{i,j}$ connecting the nodes v_i and v_j associated with regions R_i and R_j with populations $pop(R_i)$ and $pop(R_j)$, we used the average population of the two connected nodes as the weight: $P(e_{i,j}) = (pop(R_i) + pop(R_j))\,/2$, when the nodes i and j are not connected $P(e_{i,j}) = 0$. For a node v_i associated with the area R_i whose population is $pop(R_i)$, the weight is just the node population: $P(v_i) = pop(R_i)$.

2.1 The weighted non-connectivity function

Given a zone z composed of k connected regions, we formally define our novel proposal for a penalty function called weighted non-connectivity function and denoted by $w(z)$ as:

$$w(z) = \frac{\displaystyle\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} P(e_{i,j})}{3\left[\displaystyle\sum_{i=1}^{k} P(v_i) - 2\left(\frac{\displaystyle\sum_{i=1}^{k} P(v_i)}{k}\right)\right]} \tag{3}$$

Figure 1 present a hypothetical zone with two different population distributions.

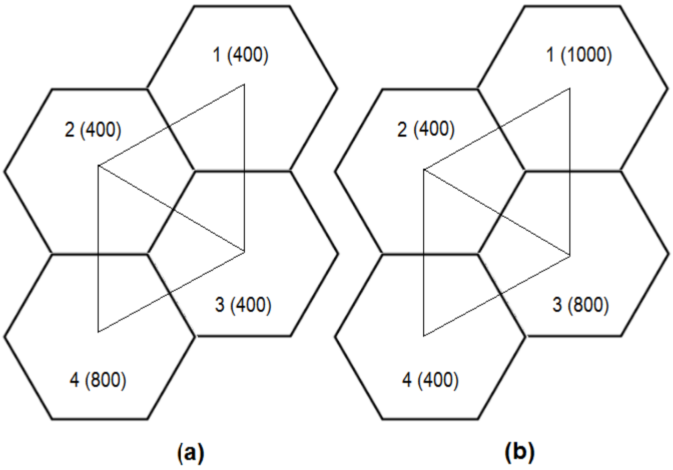

Figure 1 - Hypothetical zone with two different population distributions. The numbers inside parenthesis are the regions populations.

If we consider the different population distributions, we notice that the non-connectivity function and the weighted non-connectivity function takes distinct values.

$$nc(z) = \frac{5}{3\,(4-2)} = 0.833$$

In contrast with the non-connectivity function $nc(z)$ that does not distinguish between the two different population distributions, the weighted non-connectivity function $w(z)$ takes in account the areas' populations indicated inside parentheses in Figure 1 and assumes, respectively, the following values $w(z) = 0.769$ (left) and $w(z) = 0.897$ (right).

In the first population distribution of Figure 1 the most populated regions (R_1 and R_4) are not connected and the value (0.769) of the weighted non-connectivity function decreases compared to value assumed by the non-connectivity function (0.833). In the second example of Figure 1 the most populated regions (R_2 and R_3) are connected and the value (0.897) of the weighted non-connectivity function increases compared to the value assumed by the non-connectivity function (0.833).

3 Results and discussion

3.1 Power, sensitivity and PPV tests

We use a benchmark dataset with the map of the Northeastern US (see DUCZMAL *et al.*, 2006), consisting of 245 counties in 10 states and the District of Columbia, with a total population at risk for breast cancer of $29,535,210$ women. The alternative models in the benchmark are the nine simulated irregularly shaped clusters A to F, NYC, BOS and D.C., displayed in the three maps of Figure 2. These clusters were built with the purpose of testing the limits of the algorithms for very irregular cluster shapes.

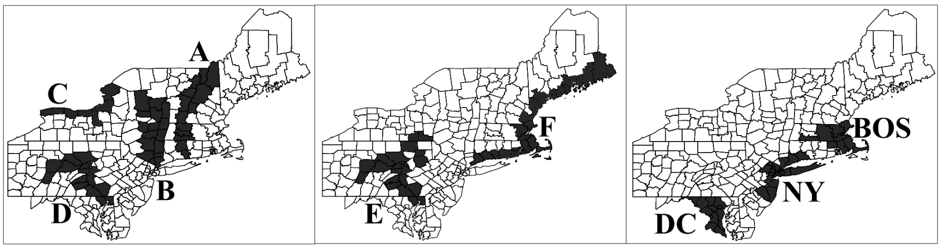

Figure 2 - Artificial clusters generated in the Northeastern US.

Data under the null hypothesis model is replicated $10,000$ so that for the mono-objective approaches $10,000$ values of the test statistic are available under the null hypothesis, while for the multi-objective approaches $10,000$ non dominated sets are available. For each scenario A–F $5,000$ runs of the scans are performed, thus producing $5,000$ sets of efficient solutions. To generate cases under alternative hypotheses we compute the relative risk for regions inside and outside the cluster such that an abnormal concentration of cases would be formed inside the cluster with probability 0.999 considering an ordinary binomial test. Details on the computation of relative risks under the alternative hypothesis can be found in Kulldorff *et al.* (2003).

The following single-objective methods were compared: the geometric compactness scan (SGC), the non-connectivity scan (SNC), and the weighted non-connectivity scan (SWN). We also compare the corresponding multi-objective scans: the multi-objective geometric compactness scan (MGC), the multi-objective non-connectivity scan (MNC), and the multi-objective weighted non-connectivity scan (MWN). All were evaluated according to their power of detection, sensitivity and positive predictive value (PPV).

For the three single-objective scans, the power, sensitivity and PPV were computed for the most likely cluster in each replicate (see Table 1). For the multi-objective scans, the power, sensitivity an PPV were computed for the cluster within the non-dominated set which has the lowest estimated p-value, according to

Table 1 - Power, positive predictive value and sensitivity comparisons for the single-objective algorithms

	Power			PPV			Sensitivity		
cluster	SGC	SNC	SWN	SGC	SNC	SWN	SGC	SNC	SWN
A	0.822	0.881	**0.900**	0.578	**0.665**	0.629	0.551	0.792	**0.818**
B	0.843	**0.926**	0.915	0.691	0.786	**0.792**	0.598	**0.784**	0.768
C	0.814	0.826	**0.864**	0.344	0.659	**0.685**	0.360	0.796	**0.819**
D	0.840	0.922	**0.937**	0.616	0.771	**0.788**	0.506	0.713	**0.752**
E	0.778	0.885	**0.893**	0.633	0.762	**0.768**	0.414	0.544	**0.562**
F	0.433	**0.585**	0.583	0.314	**0.650**	0.624	0.170	0.523	**0.535**
NY	0.747	0.819	**0.826**	0.621	0.929	**0.934**	0.364	0.650	**0.664**
BOS	0.834	0.864	**0.909**	0.389	0.827	**0.841**	0.295	0.806	**0.843**
D.C.	**0.903**	0.877	0.887	0.518	0.865	**0.885**	0.426	0.791	**0.818**
C-BOS	0.686	0.742	**0.791**	0.399	0.782	**0.805**	0.331	0.466	**0.500**

Table 2 - Power, positive predictive value and sensitivity comparisons for the multi-objective algorithms

	Power			PPV			Sensitivity		
cluster	MGC	MNC	MWN	MGC	MNC	MWN	MGC	MNC	MWN
A	0.950	0.942	**0.957**	**0.803**	0.711	0.663	0.732	0.748	**0.806**
B	0.954	**0.969**	0.959	0.781	0.821	**0.824**	0.702	**0.767**	0.728
C	0.933	0.915	**0.935**	0.716	0.734	**0.745**	0.735	0.749	**0.800**
D	0.962	0.965	**0.971**	0.751	0.803	**0.809**	0.629	0.656	**0.730**
E	0.947	0.946	**0.958**	0.760	0.785	**0.787**	0.514	0.507	**0.548**
F	0.746	0.743	**0.841**	0.710	**0.729**	0.717	0.519	0.524	**0.562**
NY	0.888	**0.909**	0.900	0.918	**0.942**	**0.942**	0.572	0.638	**0.664**
BOS	0.918	0.928	**0.950**	**0.891**	0.854	0.857	0.692	0.743	**0.834**
D.C.	**0.955**	0.936	0.939	**0.931**	0.882	0.893	0.748	0.756	**0.793**
C-BOS	0.897	0.890	**0.919**	**0.799**	0.733	0.763	0.498	0.701	**0.716**

the previous subsection (see Table 2).

Tables 1 and 2 present the average power, sensitivity and PPV for 5,000 replications of each of the nine alternative hypotheses for all the six scans. The best performances for both single-objective scans and multi-objective scans are presented in bold type.

When comparing the tables 1 and 2, we observe consistently better performance, regarding to power and PPV, when compared with the single-objective scans. The results for sensitivity for the single objective SWN scan and the multi-objective MWN scan (which are respectively the best in their groups) were about the same.

It is important to note that the original proposal by Yiannakoulias *et al.* (2007) is a single-objective formulation and uses an algorithm less robust than the genetic algorithm NSGA-II, thus the fair comparison between the new and original proposal would be using respectively the algorithms MWN and SNC. In this comparison it is clear the better performance by the new proposal.

Among the single-objective scans, the SWN scan presented better results for most clusters, as can be seen in Table 1. Moreover, Table 2 shows that the power, sensitivity and PPV values of the MWN scan were consistently higher those of the MGC and MNC scans. Particularly, the performance of the MWN scan for the highly irregular cluster F was significantly better, compared to the other scans. That result reinforces the notion that the MWN performed uniformly well for all the analized alternative models.

We provided an example of this situation in the ninth alternative hypothesis model, consisting of the double (disconnected) cluster C and BOS (see Table 2). The relative risk was chosen similarly to the previous eight alternative models; it means that we are considering the region at risk as consisting of the union of both C and BOS clusters; we stress that we are not considering separately each zone C and BOS in the relative risk computation, which would give a strong risk for both zones, thus inducing an unrealistic stronger signal to noise for both components C and BOS. Even then, the power of detection (0.919) was almost as high as in the other single zone alternative models in the table. The sensitivity (0.716) and PPV (0.763) were also consistent, indicating that the methodology works well enough for multiple clusters. Using an *Intel(R) Core i7* processor with 3.33 GHz desktop, 1000 benchmark executions took 290 seconds for the MNC, and 293 seconds for the MWN, compared to the much slower MGC scan, which took 1285 seconds. That result shows that the added computation time required for the calculation of the weighted edges in the MWN scan was negligible.

Conclusions

The purpose of the original non-connectivity penalty function was to penalize the cluster candidates which were not strongly connected; this was achieved by counting the vertices and edges of the adjacency subgraph associated to the cluster candidate, and penalizing those clusters which have relatively few edges compared with the number of vertices. Thus, less connected clusters with associated subgraphs like trees were the most penalized, and strongly connected planar graphs were the less penalized. This strategy presented two significant advantages: (i) it was relatively computationally inexpensive, and (ii) it was very efficient. Properties (i) and (ii) were discussed in detail in Cançado *et al.* (2010), where it was shown through numerical simulations that the non-connectivity scan statistic attains the highest performance with low cost for moderately irregularly shaped clusters, compared to other penalty functions of the literature.

However, the non-connectivity penalty function did not consider the population heterogeneity within the component areas of the cluster candidates. We proposed in this paper the weighted non-connectivity penalty function, a modification of the non-connectivity scan statistic for the detection and inference of spatial clusters in aggregated data maps, considering this added feature. Our strategy was to consider weighted edges in the adjacency subgraph, defined in such a way that the presence of links between relatively populated areas reinforced the cluster candidate. In order to mantain the consistency of the new definition, relatively to the previous non-connectivity penalty function, both penalties must give the same values for maps with homogeneously populated areas. The introduced changes were very simple: (i) the areas' populations substituted the counts of vertices; and (ii) each edge was weighted taking the average of the two neighboring areas' populations. That straightforward modifications endowed the weighted non-connectivity penalty function with two good properties, as verified by our numerical simulations: (i) the very low additional cost to compute the weight of the edges; and (ii) the improved performance in the detection of highly irregularly shaped clusters, without reducing the performance to detect moderately irregularly shaped clusters.

In the study we made with artificial clusters, we obtain significantly improvement in the power of detection for the irregular clusters (C, D, E and F, in Figure 2). Our simulations show that our proposed methods is still able to identify multiple clusters without loss in power performance. All scans were tested according to their power of detection, sensitivity and positive predictive value.

We also show that the multi-objective version of the scan employing the weighted non-connectivity penalty function have significantly better performance compared with the corresponding single-objective scan, confirming similar results already obtained for other penalty functions, as extensively studied in Cançado *et al.* (2010).

As future research proposals have the possibility to evaluate the new penalty using as weights the population density associated with each of the regions on one geographical map. Studies producing new functions using other regular or optimization strategies can be developed.

Acknowledgments

The authors have been partially funded by the Brazilian agencies, CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico of the Ministry for Science and Technology) via project 459535/2014-5, CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais).

DUARTE, A. R.; SILVA, S. B.; OLIVEIRA, F. L. P.; RIBEIRO, M. C.; CANÇADO, A. L. F.; MOURA, F. R. Uma penalização por não-conectividade ponderada para a detecção de clusters irregulares. *Rev. Bras. Biom.,* Lavras, v.35, n.1, p.160-173, 2017.

- RESUMO: Métodos para a detecção e inferência de clusters irregulares com dados de contagem são ferramentas importantes na vigilância da doenças e em epidemiologia. Recentemente, vários métodos foram desenvolvidos utilizando a estatística espacial Scan de Kulldorff em conjunto com alguma função de penalidade para controlar a excessiva liberdade de forma dos clusters. Penalizações distintas foram concebidos com base na forma geométrica cluster ou sobre a estrutura de vizinhança e conectividade do grafo associado. As diversas funções de penalização foram implementados utilizando métodos de otimização multi-objetivos. Em particular, a penalização por não-conectividade não mostrou ser suficientemente eficaz no procedimento de detecção. Basicamente, a função de não-conectividade baseia-se na estrutura de adjacência do grafo associado ao cluster, mas não leva em conta a distribuição populacional dentro do cluster. Aqui, busca-se introduzir uma modificação da função de não-conectividade, introduzindo pesos nos componentes da função de acordo com a distribuição da população do cluster. O método proposto é capaz de identificar vários clusters na área de estudo. Mostra-se através de simulações numéricas que a função de não-conectividade ponderada supera a função não-conectividade original em termos de poder de detecção, sensibilidade e valor preditivo positivo, sendo também computacionalmente ágil como a anterior. Ambas as versões mono-objectivo e multi-objetivo do algoritmo são implementados e comparadas.

- PALAVRAS-CHAVE: Estatística espacial Scan; conglomerados irregulares; algoritmos multi-objetivo; função de compacidade; função de não-conectividade; função de não-conectividade ponderada.

References

BALAKRISHNAN, N; KOUTRAS, M.V. *Runs and Scans with Applications.* London: John Wiley & Sons, 2002. 488p.

BUCKERIDGE, D.L.; BURKOM, H.; CAMPBELL, M.; HOGAN, W.R.; MOORE, A.W. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, v.38, p.99-113, 2005.

CANÇADO, A. L. F; DUARTE, A.R.; DUCZMAL, L.; FERREIRA, S. J.; FONSECA, C. M.; GONTIJO, E. C. M. D. Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters. *International Journal of Health Geographics*, v.9, p.55, 2010.

CHANKONG, V.; HAIMES, Y.Y. *Multi-objective Decision Making Theory and Methodology.* New York: Elsevier Science, 1983. 406p.

CONLEY, J.; GAHEGAN, M.; MACGILL, J. A Genetic Approach to Detecting Clusters in Point Data Sets. *Geographical Analysis*, v.37, p.3, p.286-314, 2005.

COSTA, M.; KULLDORFF, M. Maximum linkage space-time permutation scan statistics for disease outbreak detection. *International Journal of Health Geographics*, v.13, n.1, 2014.

DUARTE, A. R.; DUCZMAL, L. FERREIRA, S. J.; CANÇADO, A. L. F. Internal cohesion and geometric shape of spatial clusters. *Environmental and Ecological Statistics*, v.17, p.203-229, 2010.

DUCZMAL, L.; ASSUNÇÃO, R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, v.45, p.269-286, 2004.

DUCZMAL, L.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C.; BESSEGATO, L. F. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, v.52, p.43-52, 2007.

DUCZMAL, L.; CANÇADO, A. L. F.; TAKAHASHI, R. H. C. Geographic delineation of disease clusters through multi-objective optimization. *Journal of Computational & Graphical Statistics*, v.17, p.243-262, 2008.

DUCZMAL, L.; DUARTE, A. R.; TAVARES, R. Extensions of the scan statistic for the detection and inference of spatial clusters. In: BALAKRISHNAN, N.; GLAZ, J. (Ed.). *Scan Statistics*, Boston, Basel and Berlin: Birkhäuser, 2009. p.157-182.

DUCZMAL, L.; KULLDORFF, M; HUANG, L. Evaluation of spatial scan statistics for irregularly shaped disease clusters. *Journal of Computational & Graphical Statistics*, v.15, p.428-442, 2006.

DWASS, M. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, v.28, p.181-187, 1957.

FONSECA, C. M.; FLEMING, P. An overview of evolutionary algorithms in multi-objective optimization. *Evolutionary Computation*, v.3, p.1-16, 1995.

FONSECA, C. M.; DA FONSECA V. G.; PAQUETE, L. Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function. *Evolutionary Multi-Criterion Optimization. Third International Conference, EMO*, vol. 3410 of Lecture Notes in Computer Science, p.250-264, Berlin: Springer, 2005.

KULLDORFF, M. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, v.26, n.6, p.1481-1496, 1997.

KULLDORFF, M.; HUANG, L.; PICKLE, L.; DUCZMAL, L. An elliptic spatial scan statistic. *Statistics in Medicine*, v.25, p.3929-3943, 2006.

KULLDORFF, M.; NAGARWALLA, N. Spatial disease clusters: detection and inference. *Statistics in Medicine*, v.14, p.799-810, 1995.

KULLDORFF, M.; TANGO, T.; PARK, P. Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*. v.42, p.665-684, 2003.

LAWSON, A. B.; BIGGERI, A.; BÖHNING, D.; LESAFFRE, E.; VIEL J. F.; BERTOLLINI, R. *Disease mapping and risk assessment for public health*. Chichester: John Wiley & Sons, 1999. 482p.

LAWSON, A. B. *Bayesian Disease Mapping : hierarchical modeling in Spatial epidemiology*. Boca Raton: CRC Press, 2009. 378p.

LAWSON, A. B. Hot-spot detection and clustering: ways and means. *Environmental and Ecological Statistics*, v.17, p.231-245, 2010.

PEI, T.; WAN, Y.; JIANG, Y.; QU, C.; ZHOU, C.; QIAO, Y. Detecting arbitrarily shaped clusters using ant colony optimization. *International Journal of Geographical Information Science*, v.25, n.10, p. 1575-1595, 2011.

TAKAHASHI, R. H. C.; VASCONCELOS, J.A.; RAMIREZ, J. A.; KRAHENBUHL, L. A multi-objective methodology for evaluating genetic operators. *IEEE Transactions on Magnetics*, v.39, n.3, p.1321-1324, 2003.

WAN, Y.; PEI, T.; ZHOU, C.; JIANG, Y.; QU, C.; QIAO, Y. ACOMCD: A multiple cluster detection algorithm based on the spatial scan statistic and ant colony optimization. *Computational Statistics & Data Analysis*. v.56, n.2, p.283-296, 2012.

YIANNAKOULIAS, N.; ROSYCHUK, R. J.; HODGSON, J. Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics*, v.6, n.28, 2007.

Received in 10.05.2016.

Approved after revised in 24.06.2016.