3

Assessing the Outline Uncertainty of Spatial Disease Clusters

Fernando L. P. Oliveira[1], André L. F. Cançado[2], Luiz H. Duczmal[3] and Anderson R. Duarte[1]
[1]Departament of Mathematics, Universidade Federal de Ouro Preto
[2]Departament of Statistics, Universidade de Brasília
[3]Departament of Statistics, Universidade Federal de Minas Gerais
Brazil

1. Introduction

The spatial analysis of disease incidence is a fundamental tool in public health monitoring (Lawson et al., 1999). Suppose that a geographic study area is divided into administrative areas, with known populations at risk and observed cases of disease within a certain period of time. An interesting question is the possible existence of spatial anomalies in the study area: are there localized regions within the map for which the relative concentration of cases among the population at risk is significantly higher than would be expected if the cases were distributed at random? Such anomalies, known as *spatial clusters*, are inherently difficult to delineate, for several reasons (Cancado et al., 2010; Lawson, 2009). Due to the stochastic nature of the number of observed cases of disease, the uncertainty may be elevated in the disease rate estimation for aggregated area maps, especially for small population areas. Thus the most likely disease cluster produced by any given method for the detection and inference of spatial clusters (like SaTScan (Kulldorff, 1999) or any other irregularly shaped scan) is subject to a lot of variation. If it is found to be statistically significant, what could be said of the external areas adjacent to the cluster? Do we have enough information to exclude them from a health program of prevention?

A criterion was proposed (Goovaerts, 2006) to measure the uncertainty of each area being part of a possible localized anomaly in the map, finding error bounds for the delineation of spatial clusters in maps of areas with known populations and observed number of cases. A given map with the vector of real data (the number of observed cases for each area) was considered as just one of the possible realizations of the random variable vector with an unknown expected number of cases. In this methodology, m Monte Carlo replications were performed, considering that the simulated number of cases for each area is the realization of a random variable with average equal to the observed number of cases of the original map. Then the most likely cluster for each replicated map was detected. Finally, to each area a_i it was assigned the number of simulations that a_i was included in a most likely cluster. If an area belonged to the most likely cluster on all the m replications, it was colored as black;

otherwise, if it never was part of a most likely cluster, then it was colored as white, with intermediate shades of gray in-between. A Bayesian variant along these lines, to detect and represent spatial clusters, was also proposed recently Neill (2011).

Another approach to represent the uncertainty in the delineation of spatial clusters appeared recently (Oliveira et al., 2011), employing a ranking based scheme known as *intensity function*. That procedure uses the circular spatial scan statistic (Kulldorff, 1999) to find the circularly shaped most likely cluster for each replicated map. The corresponding m likelihood values (obtained by means of the m Monte Carlo replications) are ranked. For each area a_i , the maximum likelihood value, obtained among the most likely clusters containing the area a_i, is determined. Finally, the intensity function associated to each areaŠs ranking of its respective likelihood value among the m obtained values is constructed. The latest procedure generally produce less biased results when compared with the two previous schemes.

However, the circular spatial scan has several limitations, which were discussed in the literature (Duczmal et al., 2006; Kulldorff et al., 2006). Particularly, the circular window is not suitable to make the correct delineation of irregularly shaped clusters because it either chooses a proper subset of the true cluster (underestimation) or chooses a large circle containing the cluster as a proper subset (overestimation). One important consequence is the reduction of the power of detection (Duczmal et al., 2006). In order to overcome this limitation, many algorithms were recently proposed to detect irregularly shaped clusters, replacing the circularly shaped window scheme for any strategy of finding irregularly shaped solutions. Usually, the only limitation in shape for those clusters is a connectivity requirement. In this work, we will analyze the utilization of irregularly shaped algorithms for the application of the intensity function (Oliveira et al., 2011), compared to the use of the simple circular scan, which was employed as the standard method. Due to the regular shape of the most likely cluster found, a question was left, at least in part unanswered: do all the areas inside the cluster have the same importance from a practitioner perspective? In this work is proposed an application of the intensity function for irregularly shaped algorithms, thus avoiding a potential problem inherent in the use of the circular spatial scan, which may described as the lack of resolution inside the circular cluster. As a consequence, it may be difficult or impossible to distinguish the relative importance of the areas inside the detected circular cluster. As we shall see, this problem does not occur when using irregularly shaped scans. Besides, the maximum allowed size for the most likely cluster has a large influence in the result of the cluster search (Chen J, 2008).

In this work novel results are presented, applying the multi-objective genetic algorithm scan (Duarte et al., 2010; Duczmal et al., 2008; 2007), adapted for the weighted non-connectivity penalty function (Cancado et al., 2010). Also, by allowing several different maximum sizes for the most likely cluster, the possible anomaly could be identified with greater precision. As will be demonstrated in the following sections, much better delineated cluster maps of the intensity function will be generated, as compared with the previous version using the simpler circular scan. As a consequence, the relative importance of individual regions composing the spatial anomalies may be assessed, and several interesting phenomena related to the geographical distribution of chronic and acute diseases may be visualized.

2. The intensity function

In this section we define a criterion to measure the plausibility of each area being part of a possible localized anomaly in the map. Following Oliveira et al. (2011), instead of finding the most likely cluster in the original map with the observed number of cases for each area, we consider maps where the number of cases are replications of a vector of random variables, whose averages are defined based on the observed number of cases of the original map. We formalize this procedure in the following.

The original map has c_i observed cases in the area a_i, $i = 1, \ldots, K$. Now we construct a Monte Carlo replication distributing randomly the $C = \sum_{i=1}^{K} c_i$ cases among the K areas $a_1, \ldots, a_K$ according to a multinomial distribution where the probability associated to the area a_i is c_i/C. Let $V = (s_1, \ldots, s_K)$ the realization of the multinomial random vector where s_i is the number of simulated cases in the area a_i, $i = 1, \ldots, K$, where $\sum_{i=1}^{K} s_i = C$. The cluster finder algorithm (in our setting we use the circular scan or we use the elliptic scan) now finds the most likely cluster MLC_1 with likelihood ratio value LLR_1. The Monte Carlo procedure above is repeated m times, generating a set of m likelihood ratio values $\{LLR_1, \ldots, LLR_m\}$ corresponding to the most likely clusters $\{MLC_1, \ldots, MLC_m\}$. The likelihood ratio values are sorted in increasing order as $\{LLR_{(1)}, \ldots, LLR_{(m)}\}$ for the corresponding most likely clusters found $\{MLC_{(1)}, \ldots, MLC_{(m)}\}$. We now define the $intensity function$ $f : \{1, \ldots, m\} \longrightarrow \mathbb{R}$ by $f(j) = LLR_{(j)}$, $j = 1, \ldots, m$.

For each area a_i, let:

$$q(a_i) = \frac{1}{m} \ \arg \max_{1 \le j \le m, a_i \in MLC_{(j)}} f(j), i = 1, \ldots, K$$

If the area a_i does not belong to any of the sets $MLC_{(1)}, \ldots, MLC_{(m)}$ then we set $q(a_i) = 0$. The value $q(a_i)$ represents the quantile of the highest likelihood ratio among the ranked values of the likelihood ratios of the most likely clusters found in the m Monte Carlo replications, which take into account the variability of the number of cases in each area. In this sense, the value $q(a_i)$ may be interpreted as the relative importance of the area a_i as part of the anomaly of the map, where the value $f(a_i)$ represents the maximum likelihood ratio found for the most likely clusters which contain the area a_i. This concept gives more information about the anomaly than the clear-cut division between cluster and non-cluster areas, as given by the usual process of finding the most likely cluster in the original map. See Oliveira et al. (2011) for further details.

3. Genetic algorithm for spatial cluster finding

3.1 Introduction

Genetic algorithms (GA's) constitute an important class of optimization methods. Its importance comes from the fact that the GA's are robust algorithms, in the sense that they are able to treat a wide variety of problems. While some optimization methods require certain assumptions about the problem to be solved, without which these methods fail, the GA's do not require any assumption of continuity, convexity, differentiability and unimodality. In fact, the only assumption a GA requires is that the function to be optimized presents a "global trend" that can be captured or learned by the algorithm. Of course, not making any kind of assumption and, consequently, not using these characteristics in favor, GA'a tend to be computationally intensive, so its usage is justified for difficult problems.

When looking for a most likely cluster, one faces a challenging otimization problem: given a set R of n regions in a map, some of which are neighbors, find the connected subset S of R that assumes the highest LLR value. By "connected" we mean that, starting from any region in S there's always a path to any other region of S formed by a chain of neighbors, all of them inside S.

Solving this problem exactly means that we would have to look at all of the 2^n subsets of R, test which ones are connected, evaluate their LLR values and pick up the most likely one. For maps with just a few dozens of regions this problem is already intractable. So we need another strategy to find such optimal solution. GA's showed to be a good alternative for the spatial cluster finding problem (Duczmal et al., 2008; 2007).

3.2 The genetic algorithm

The natural evolution of living beings can be compared to an optimization process. In fact, if individuals who are best adapted survive - in the sense of transmitting their genetic information - while less adapted individuals tend to disappear, it is expected that after a number of generations the population is composed of individuals who are generally better adapted than those of earlier generations. This is also the idea behind a genetic algorithm. It tries to simulate the mechanisms of random variation and selection of adaptive evolution. The mechanisms (or genetic operators) that form the basis of a genetic algorithm are:

- crossover operator, which combines the information of two or more individuals - called parents - generating new individuals - called children;
- mutation operator, which applies a random perturbation to the information of an individual, generating a new one;
- selection operator, which defines the probability of an individual to transmit its genetic information (generate children) based on its adaptation level.

In this context, an individual is a candidate-solution to the optimization problem and a population is a set of individuals. For the spatial cluster detection problem a solution - or individual - is a set of connected regions of the map (the candidate cluster). So, the population is a set of lists, each list being a set of regions that form the solution.

Starting with an initial population the GA forms a sequence of generations. At each iteration it applies the selection, crossover and mutation operators to the current population, generating a new population. The GA used in this work was primarily described in Duczmal et al. (2007) and its biobjective versions were used by Duczmal et al. (2008), Cancado et al. (2010) and Duarte et al. (2010).

3.2.1 Generating the initial population

The initial population is generated by a greedy procedure. Given a map with n regions we generate a population of n individuals, each of which is generated from one region of the map. So, starting with a region, the solution incorporates more regions, choosing at each step to aggregate, among all the regions that are neighbors of some region in the actual solution, the one that makes the LLR value to increase the most when added to the solution. The individual grows until it reaches a maximum size set by the user.

3.2.2 The selection operator

Each solution is evaluated by means of its LLR value and this is the adaptation indicator: higher LLR-valued individuals are more adapted. The selection operator will then give more chances to the more adapted individuals to generate offspring. This is done through a mechanism called binary tournment. For each tournment two individuals are chosen from the current population, each individual having the same probability of being chosed. Then we compare the two solutions and the one with higher LLR value is selected. This procedure is repeated n times, thus producing a set of n selected individuals.

3.2.3 The crossover operator

Now, selected individuals have the chance to trasmit their genetic information to new individuals by generating offspring. Crossover is applied to pairs of parents randomly chosen from the list of selected individuals. The offspring is generated in a way that the children inherit characteristis from both parents. In addition, it is well known that GA's particularly designed for a specific problem perform much better than multiple-purpose generic GA's. Thus, it is highly recommended that operators are designed so that they can take advantage of the intrinsic structure of the problem. For example, in our case we would discard any disconnected cluster candidate because it is an infeasible solution. While a generic crossover operator could, most of the time, generate infeasible clusters, we chose to use a crossover operator that ensures that every generated solution is feasible.

The crossover operator described by Duczmal et al. (2007) presents all these features, being capable of efficiently generating feasible offspring having characteristics of both parents. The operator is implemented in sucha a way that it is only possible to perform a crossover between two parents if they share a nonempty intersection. Once this condition is verified, the offspring is generated. Figure 1 shows an example with two parents (A and B) and the generated offspring 1-5. Note that the offspring constitutes a "path" from one parent to another, with child 1 being more like parent A, while child 5 is almost like parent B. In the middle of the figure we can see parents inside the map with the two intersection regions (in gray).

3.2.4 The mutation operator

Each individual generated by the crossover process has a probability of suffering a mutation. Mutation consistis in introducing a random perturbation in the genetic code of the individual. In our case, the mutation consists of adding to or removing from the cluster a randomly chosen region, provided the cluster's connectivity.

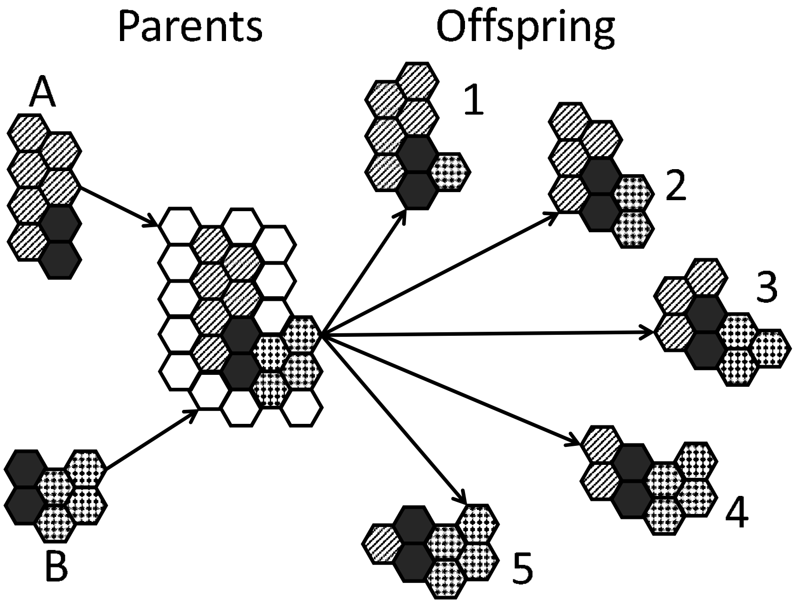


Fig. 1. A splitted vision of parents *A* and *B* (left), parents *A* and *B* inside the map (middle) and offspring (right).

3.3 The biobjective genetic algorithm

Many times one wants to find a solution that simultaneously optimizes two or more functionals. For example, a costumer may want to buy a car which is powerful and cheap. Of course, it is very unlikely that, say, the most powerful is also the cheaper car, because these two criteria are conflicting. Based on these two criteria, a whole set of cars can be of interest for this costumer: powerful (but expensive) cars and cheap (but underpowered) cars. Of course, a costumer (again, based on just these criteria) will reject cars that cost too much and are low powered.

Following the same reasoning, a biobjective GA was proposed (Cancado et al., 2010; Duarte et al., 2010; Duczmal et al., 2008) to deal with the problem of spatial cluster detection. Using the LLR as the only objective to be maximized would lead to geographically meaningless tree-shaped solutions and it is necessary to consider some shape regularity measure, such as geometric compactness (Duczmal et al., 2008) or topological corrections (Cancado et al., 2010; Yiannakoulias et al., 2007). This regularity measure works as a second objective to be maximized. As in the power/price car example, LLR and regularity are conflicting objectives, because high values of LLR are associated to very irregular clusters, while regular solutions tend to

Instead of an optimal solution, a biobjective maximization problem will lead, in general, to a set of optimal solutions, called the Pareto-set. This set is composed by all solutions that are not worse than any other solution in both objectives simulteanously. Such solution is called nondominated solution. Because GA's work with a population of candidate-solutions they can find the Pareto-set in one execution with almost the same effort spent by its mono-objective version. Figure 2 illustrates a set of solutions in the objectives space. Nondominated solutions are indicated by black dots.

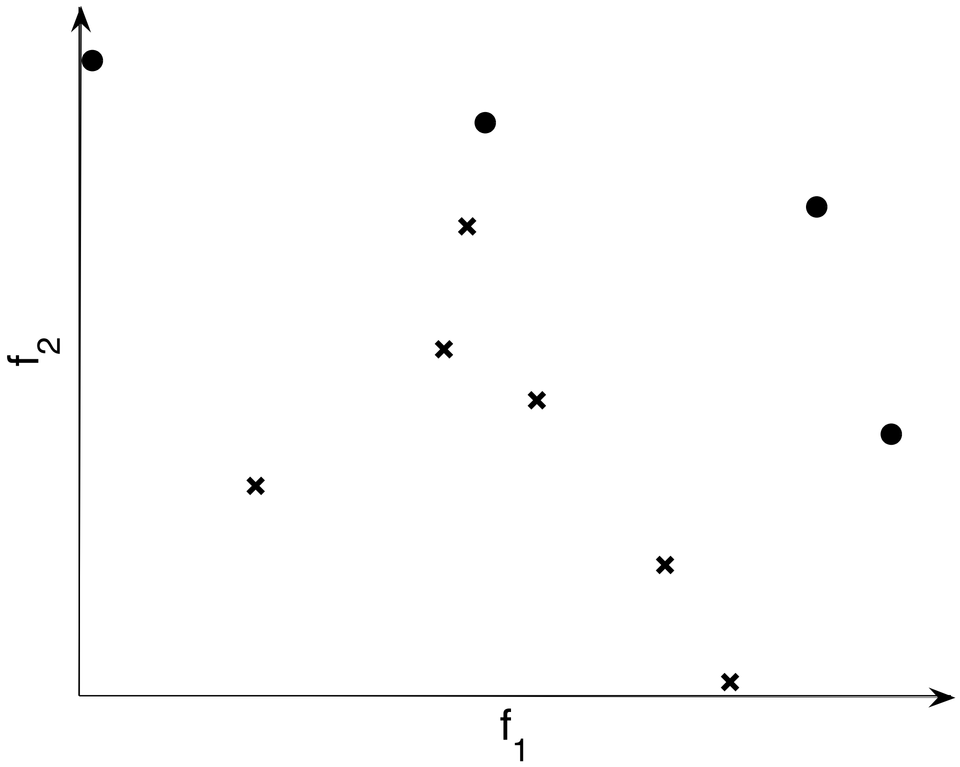


Fig. 2. A set of solutions in the objectives space: dominated solutions ($\times$) and Pareto-set ($\bullet$).

3.4 Inference and the attainment function

Once the most likely cluster is identified, we want to check its significace. This will allow the practitioner to verify if the cluster can be considered a disease outbreak or if the disease cases are randomly spreaded over the map. Since the distribution of LLR under H_0 is not known we must perform a Monte Carlo simulation. For the mono-objective case, the LLR value is calculated for the most likely cluster in each Monte Carlo replication under H_0 and the p-value is computed comparing the value of LLR for the observed data and the empirical distribution obtained through the Monte Carlo procedure.

For the biobjective case, we consider the attainment function (da Fonseca et al., 2001; Fonseca et al., 2005), as also used by Cancado et al. (2010). A single run of the biobjective GA would produce a Pareto-set, defining two distinct regions in the objectives space: points that are dominated by the Pareto-set and points that are not dominated by it. Then, for inference purposes we can consider, for each point of the Pareto-set obtained for the observed data, the proportion times that the point is dominated by the Pareto-sets under H_0. This is exactly the p-value for that point.

3.5 The geometric penalty function

One of the possible penalties that takes in account the cluster geometric shape is the called compactness geometric penalty function. This penalty function introduced in Duczmal et al.

(2006) aims to penalize zones in the map that have very irregular shape. The compactness geometric function $k(z)$ of a zone z is given by the area of z divided by the area of a circle with the same perimeter as the convex hull of z. The compactness geometric function takes values between zero and one, the circle has the most compact shape ($k(z) = 1$). Compactness depends on the shape of the zone, but not on its size. The expression for $k(z)$ is given by:

$$k(z) = \frac{4\pi A(z)}{H(z)^2} \tag{1}$$

where $A(z)$ is the area of the zone z and $H(z)$ the perimeter of the convex hull of z. The compactness penalyzed scan statistic is defined as $max_{z \in Z} k(z).LLR(z)$.

3.6 The non-connectivity penalty function

Yiannakoulias et al. (2007) proposed a greedy algorithm to scan the set Z of all possible zones z. A new 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 was used as a multiplier for the $LLR(z)$. 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 whith tree-like associated graphs, meaning that they have a small number of nodes compared with the number of edges. Although there is some similarity between the non-connectivity penalty to the geometric compactness penalty, there is an important difference: the non-connectivity penalty does not rely on the geometric shape of the candidate cluster, which could be an interesting feature when searching for real clusters which are highly irregularly shaped, but present good connectivity properties.

3.7 Evaluation of the candidate solutions

Differently from the previous procedure employing the circular scan, each run of the multiobjective genetic scan produces a set of several non-dominated solutions.

In the circular scan, the scan statistic value for the most likely cluster was assigned to each area of the solution cluster, and later the maximum value of this quantity was obtained for all the executions. However, in the multiobjective procedure, the scan statistic value will be assigned for each component area of each solution cluster of the non-dominated solution set. In the event that a given area belongs to more than one solution cluster, the largest scan statistic value is assigned to the area. The remaining of the process is identical to the usual procedure using the circular scan, obtaining the maximum value of this quantity for all the executions, and building the intensity function as usual.

4. Results and discussion

Epidemiological surveillance is essential to monitoring possible changes in the geographical distribution pattern of both acute and chronic diseases. To illustrate the techniques presented in this chapter, four diseases (dengue fever, tuberculosis, diabetes and hypertension) are analyzed. Those four diseases are currently among the most serious health threats to the Brazilian population. Our studies were concentrated in the Minas Gerais state in southeast Brazil, with 853 municipalities and total population of 19,597,330 (2010 census). For each disease, only the specific population at risk at each municipality was considered. Population data was available at Instituto Brasileiro de Geografia e Estatística (www.ibge.gov.br), and disease data was obtained through DATASUS, the Brazilian Ministry of Health's central data system (www.datasus.gov.br). Dengue fever data was collected by SINAN/MS system from the Brazilian Ministry of Health (www.sinam.org.br). During the period 2007-2010, 349.005 cases were registered, and the population at risk was assumed to be the total population of the 2010 census. Tuberculosis disease cases, using SINAN/MS data, were considered for the 2001-2010 period for the following age groups (years): 15-19, 20-39 and 40-59, making a total of 41,824 cases for a population at risk of 12,892,744. Hypertension data was obtained through the Hiperdia program of Brazilian Ministry of Health from January 2002 to January 2011. Data was available to the following age groups (years): 50-59, 60-69, 70-79 and 80+, with a total population at risk of 4,365,352 individuals and 941,710 cases. Diabetes types 1 and 2 data were also obtained through the Hiperdia program from January 2002 to May 2011. The age groups were: 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 and 80+ years, with 28.039 cases.

Diabetes mellitus and hypertension are considered chronic diseases and their control and treatment depend on the individuals behavior in relation to their lifestyle: healthy eating, physical activity, and weight control. These diseases are responsible for high rates of hospital expenses, so the investment in shares of health promotion and prevention is potentially very cost effective. The importance of dengue in our study lies in the fact that it is an infectious disease and even in regions with previous low incidence rates are subject to outbreaks. This disease is subject to major public health campaigns in Brazil. The report on the epidemiology of dengue published by the Secretariat of Health Surveillance in 2010 indicates Minas Gerais state as one of the critical states in need of stricter monitoring. Hypertension and diabetes are very common chronic diseases, and hypertension is a major public health problem in Brazil. Tuberculosis has become relevant to this study due to its high incidence, and its early diagnosis and effective treatment are of great importance to public health. The biggest challenge for public health professionals has been to promote action to encourage compliance and continuity of care, since many individuals do not join or do not follow the prescribed treatment.

4.1 Real data case studies

In what follows, we present the obtained sets of intensity function maps for dengue fever, tuberculosis, diabetes and hypertension in Minas Gerais state (Figures 3, 4, 5 and 6, respectively). North is up for all the maps. For each disease set we present six maps: (a) the quantiles of population at risk, (b) the quantiles of disease rate and the intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively). The population at risk was different for each disease in our study.

As can be noted on all four disease sets, the probability that each area belongs to the "‘true"’ cluster decreases as the maximum cluster size increases from 10 to 40. For instance, in the dengue fever set, the dark brown areas have probability of belonging to the "‘true cluster"’ greater than 94%, 88%, 83% and 76%, as the maximum cluster size increases from 10, 20, 30 and 40, respectively. It means that the intensity function maps produced with the smaller maximum cluster sizes (10 and 20)indicate inner "‘core"’ regions within the "‘true cluster"’. On the other hand, the intensity function maps produced with the larger maximum cluster sizes (30 and 40)indicate "‘borderline"’ regions with respect of the "‘true cluster"’.

Another important feature is the complexity of the shapes displayed in the sequence of intensity function maps as the maximum cluster size increases.

4.1.1 Dengue fever

In Figure 3c, the maximum size 10 inner core region of dengue fever includes the municipalities arround the state capital Belo Horizonte urban area (population 4 million) in the central part of the state. The maximum sizes 20 and 30 intensity function maps (Figures 3d and 3e respectively) show the anomaly spreading northward following the São Francisco river basin, a region with elevated humidity and high mosquito incidence. Finally, the larger maximum size 40 anomaly (Figure 3f) spreads along the highway joining the cities of Ipatinga, Valadares and Teofilo Otoni in the eastern part of the state.

4.1.2 Tuberculosis

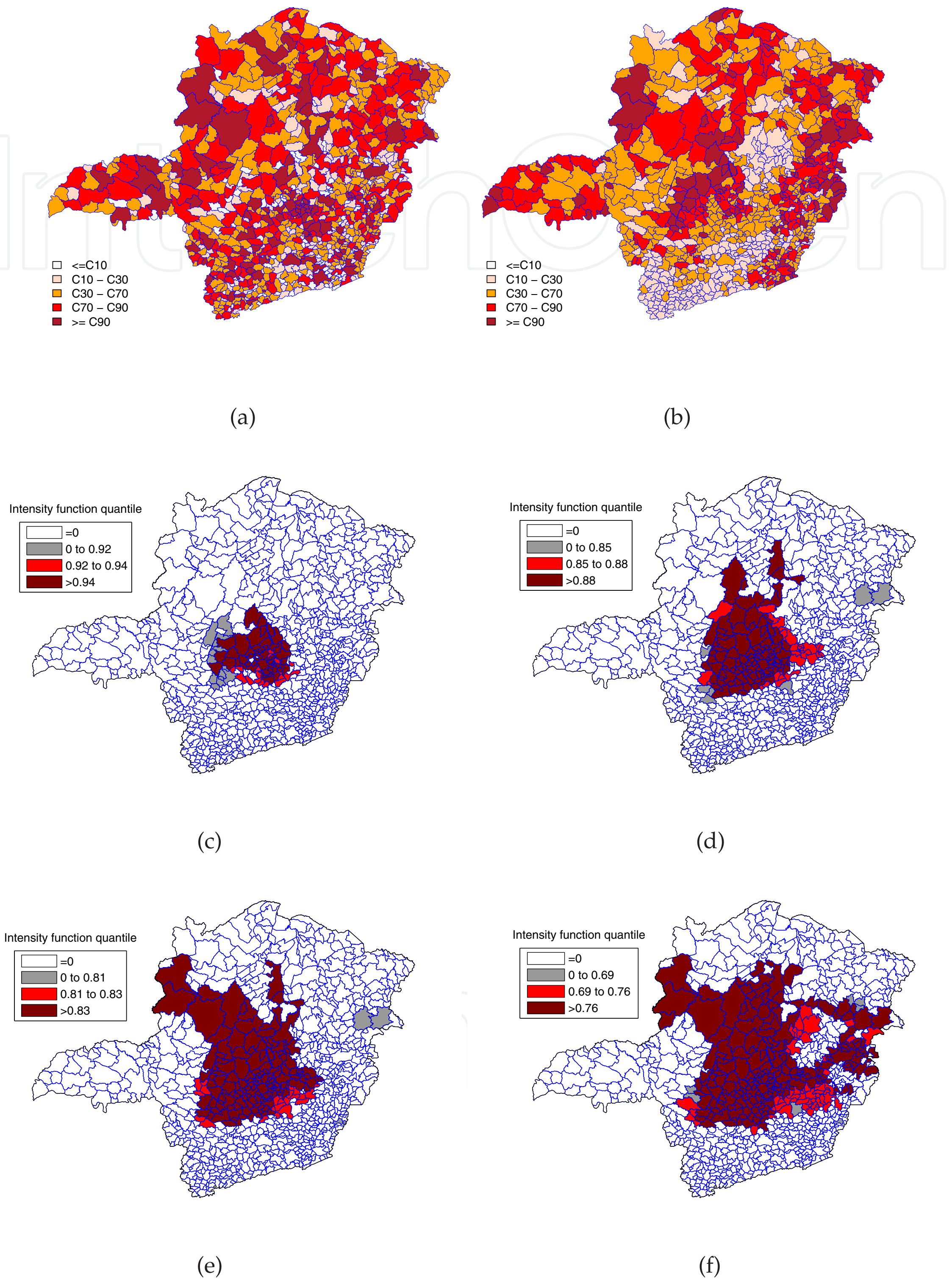
In Figure 4c, the maximum size 10 inner core region of tuberculosis includes the predominantly urban area of Belo Horizonte (in the central part of the state) and two weaker urban regions: (i) the highway joining the cities of Ipatinga, Valadares and Teofilo Otoni in the eastern part of the state, and (ii) the areas surrounding the city of Juiz de Fora, the second largest city of the state in the south. As the maximum cluster size increases (Figures 4d, 4e and 4f), the tuberculosis anomaly is reinforced to include the surrounding municipalities, and also the neighbors of the populous Montes Claros city in the northern part of the state.

4.1.3 Diabetes

In Figure 5c, the maximum size 10 inner core region of diabetes includes the southwest part of the state and the weaker urban region of Valadares city in the east. As the maximum cluster size increases (Figures 5d, 5e and 5f), the diabetes anomaly is reinforced to include the surrounding municipalities.

4.1.4 Hypertension

In Figure 6c, the maximum size 10 inner core region of hypertension includes several scattered regions in the center and mid southeast parts of the state. As the maximum cluster size increases (Figures 6d, 6e and 6f), the hypertension anomaly is reinforced to include the surrounding municipalities.

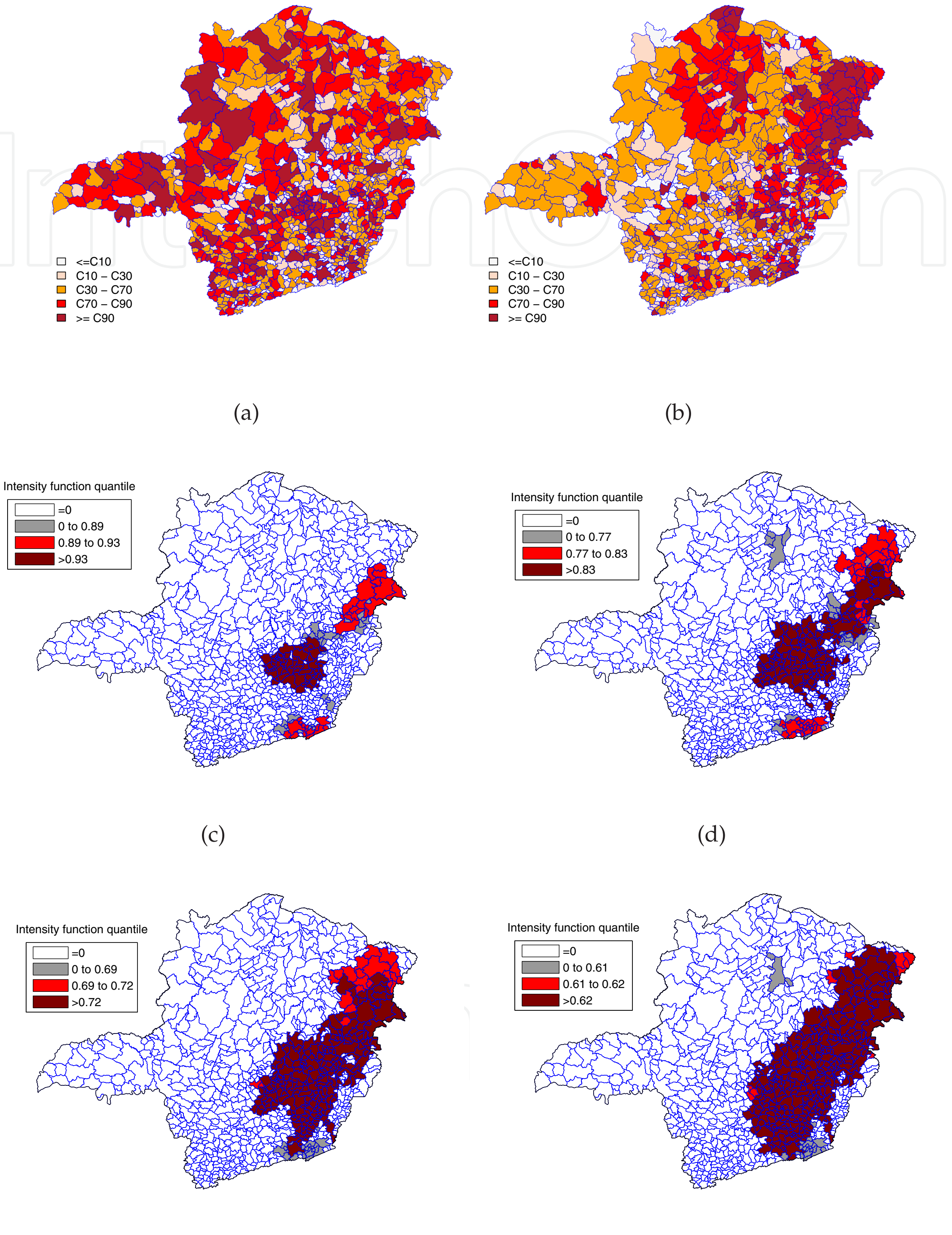


(a) (b)

(c) (d)

(e) (f)

Fig. 3. Population at risk quantiles (a), dengue fever rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

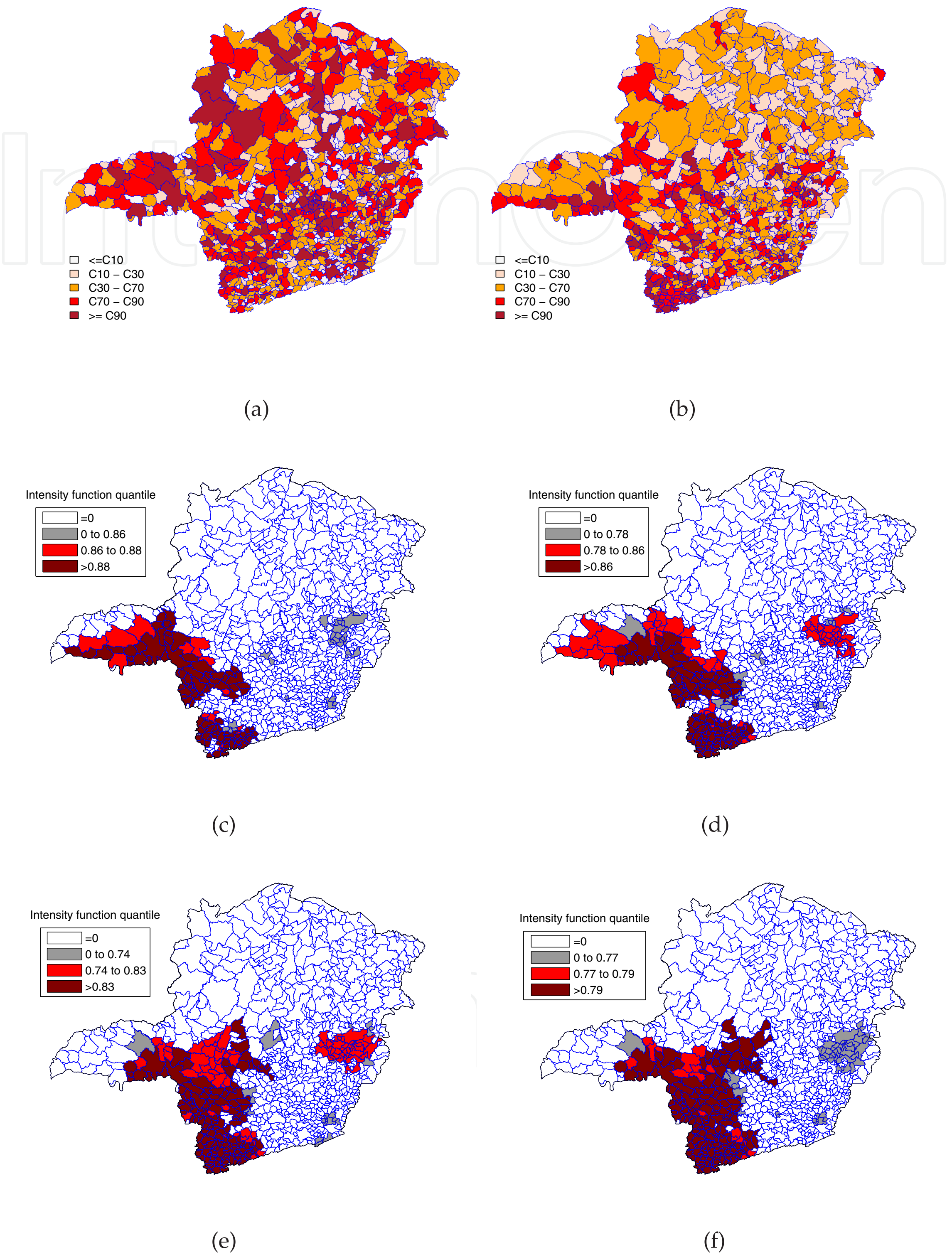


(a) (b)

(c) (d)

(e) (f)

Fig. 4. Population at risk quantiles (a), tuberculosis rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

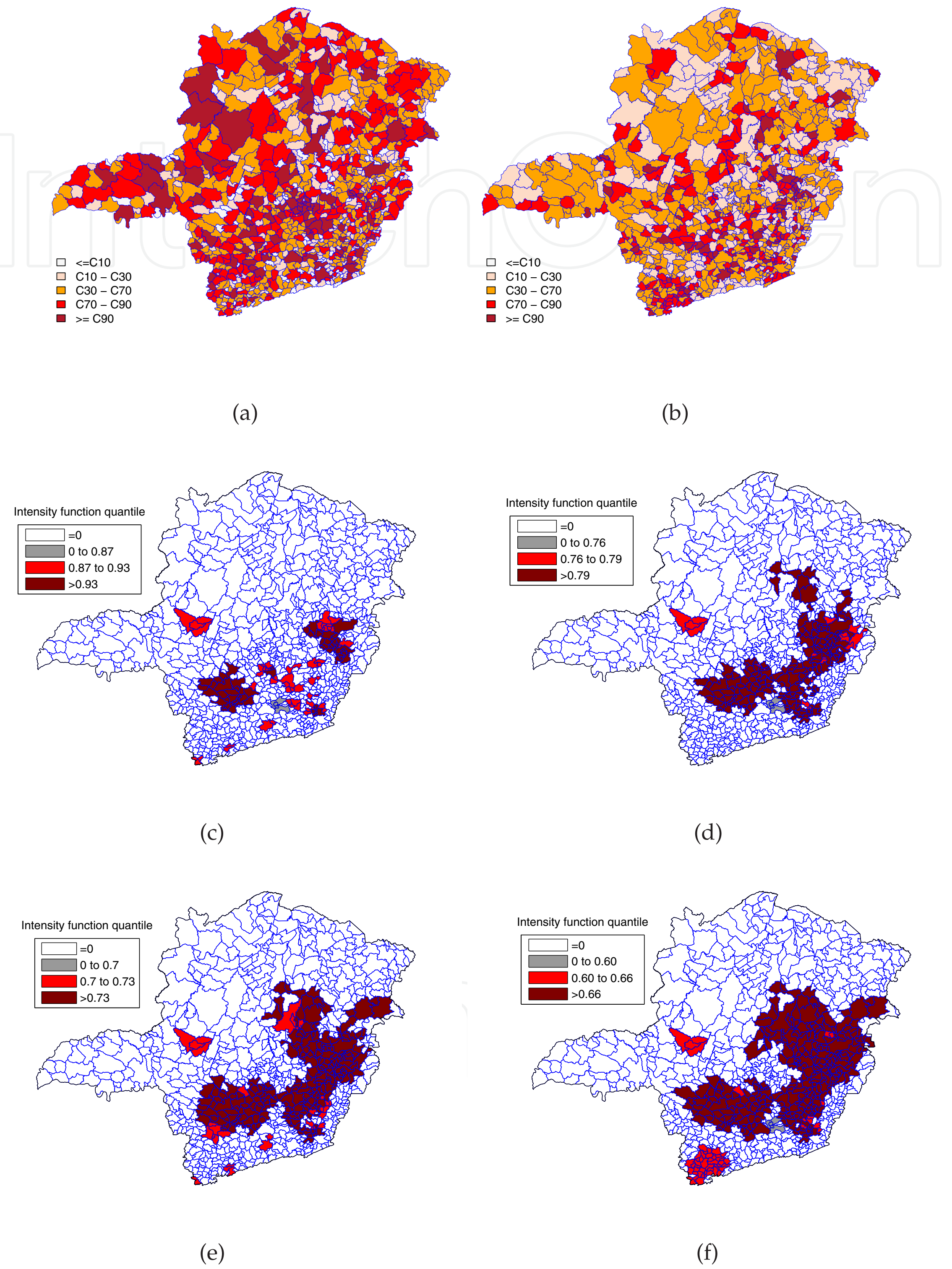


(a) (b)

(c) (d)

(e) (f)

Fig. 5. Population at risk quantiles (a), diabetes rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)



(a) (b)

(c) (d)

(e) (f)

Fig. 6. Population at risk quantiles (a), hypertension rates (b), and intensity function maps based on the genetic multi-objective algorithm for maximum clusters of sizes 10, 20, 30 and 40 (c, d, e and f respectively)

5. Conclusion

Our methodology takes into account the variability in the observed number of disease cases on area aggregated maps to nonparametrically infer the uncertainty in the delineation of spatial clusters. A given real data map is regarded as just one possible realization of an unknown random variable vector with expected number of cases. The real data vector of the number of observed cases in each area is used to construct a new vector of expected values of random variables, considering the count of cases as the average of the random variables. This vector is now an estimate of the unknown random variable vector with expected number of cases. Our methodology performs m Monte Carlo replications based on this estimated vector of averages. The most likely cluster of each replicated map is detected and the m corresponding likelihood values obtained in the replications are ranked. For each area we determine the maximum likelihood value among the most likely clusters containing that area. Thus, we obtain the intensity function associated to each area's ranking of their respective likelihood value among the m values. The intensity of each area can be interpreted as the importance of that area in the delineation of the possibly existing anomaly on the map, considering only the initially given information of the observed number of cases. This procedure, based on the empirical distribution, takes into account the intrinsic variability of the observed number of cases, which generally is not considered directly in the existing algorithms used to detect spatial clusters.

In our case studies we could see different situations with respect to the intrinsic variability of the existing spatial anomaly. When the most likely cluster is quite prominent, as in the diabetes case study, the intensity function is such that almost all areas associated with the most likely clusters found in the m replications coincides with those areas composing the most likely cluster detected for the original observed cases. In this situation the geographic anomaly is highly focused. However, in a different scenario, a disease map may present an intrinsically wide variability of data. Many areas near or adjacent to the most likely cluster have values of the intensity function close to the values corresponding to areas of the most likely cluster. In the case study of hypertension, this intrinsic variability produces a map with clearly unrelated areas, but with rather close probability ranking, indicating a situation of multiplicity of clusters, i. e., the most likely cluster is clearly poorly delineated.

In this work we included two new features that extended the original ideas of the previous paper Oliveira et al. (2011). First, instead of the circular scan, we have used an irregularly shaped cluster finder based on a multiobjective genetic algorithm. It allowed a much better delineation of the complex shapes found in the real data clusters. As a consequence, several new phenomena could be distinguished in the spatial distribution of disease, which could not be observed with the simples spatial scan. The second modification was the sequential execution of runs with different sizes for the maximum allowed cluster to composing the intensity function maps. With this modified procedure, instead of only one map, it was obtained a sequence of intensity function maps: as the maximum cluster size increased, larger anomalies of lesser intensity were displayed. This allowed the identification of "‘core"’ and "‘borderline"’ regions, with different levels of uncertainty.

The visualization tool developed in this work may serve as a support for the decision making process to prioritize areas of public health intervention, in a more precise manner than provided by ordinary methods of cluster finding.

6. References

Cancado, A. L. F., Duarte, A. R., Duczmal, L., Ferreira, S. J., Fonseca, C. M. & Gontijo, E. C. D. M. (2010). Penalized likelihood and multi-objective spatial scans for the detection and inference of irregular clusters, *International Journal of Health Geographics* 55(9).

Chen J, Roth RE, N. A. L. E. M. A. (2008). Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of u.s. cervical cancer mortality, *International Journal of Health Geographics* 7(57).

da Fonseca, V. G., Fonseca, C. M. & Hall, A. O. (2001). Inferential performance assessment of stochastic optimisers and the attainment function, *Proceedings of the First International Conference on Evolutionary Multi-Criterion Optimization, Lecture Notes In Computer Science*, Vol. 1993, Springer-Verlag, Berlin, pp. 213–225.

Duarte, A. R., Duczmal, L. H., Ferreira, S. J. & Cancado, A. L. F. (2010). Internal cohesion and geometric shape of spatial clusters, *Environmental and Ecological Statistics* 17: 203–229.

Duczmal, L., Cancado, A. L. F. & Takahashi, R. H. C. (2008). Delineation of irregularly shaped disease clusters through multiobjective optimization, *Journal of Computational and Graphical Statistics* 17(2): 243–262.

Duczmal, L., Cancado, A. L. F., Takahashi, R. H. C. & Bessegato, L. F. (2007). A genetic algorithm for irregularly shaped spatial scan statistics, *Computational Statistics and Data Analysis* 52: 43–52. DOI:10.1016/j.csda.2007.01.016.

Duczmal, L., Kulldorff, M. & Huang, L. (2006). Evaluation of spatial scan statistics for irregularly shaped clusters, *Journal of Computational and Graphical Statistics* 15(2): 428–442.

Fonseca, C. M., da Fonseca, V. G. & Paquete, L. (2005). Exploring the performance of stochastic multiobjective optimisers with the second-order attainment function, *Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization, Lecture Notes In Computer Science*, Vol. 3410, Springer-Verlag, Berlin, pp. 250–264.

Goovaerts, P. (2006). Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using poisson kriging and p-field simulation, *International Journal of Health Geographics* 5(7).

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations and applications, *in* J. Glaz & N. Balakrishnan (eds), *Scan Statistics and Applications*, Springer Netherlands, pp. 303–322.

Kulldorff, M., Huang, L., Pickle, L. & Duczmal, L. (2006). An elliptic spatial scan statistic, *Statistics in Medicine* 25: 3929–3943.

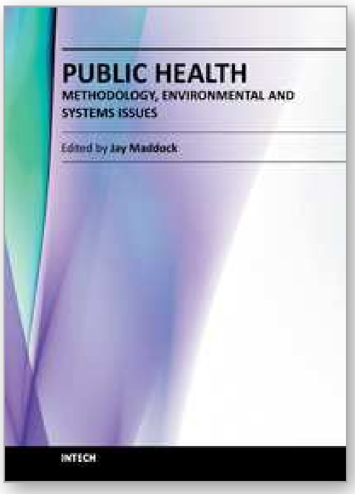
Lawson, A. (2009). *Bayesian Disease mapping*, CRC Press.

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

Neill, D. B. (2011). Fast bayesian scan statistics for multivariate event detection and visualization, *Statistics in Medicine* 30(28): 455–469.

Oliveira, F. L. P., Duczmal, L., Cancado, A. L. F. & Tavares, R. (2011). Nonparametric intensity bounds for the delineation of spatial clusters, *International Journal of Health Geographics* 1(10).

Yiannakoulias, N., Rosychuk, R. J. & Hodgson, J. (2007). Adaptations for finding irregularly shaped disease clusters, *International Journal of Health Geographics* 6(28).



PUBLIC HEALTH
METHODOLOGY, ENVIRONMENTAL AND SYSTEMS ISSUES
Edited by Jay Maddock
INTECH