Fernando Luiz Pereira de Oliveira

NONPARAMETRIC INTENSITY BOUNDS FOR THE DETECTION AND DELINEATION OF SPATIAL CLUSTERS

Belo Horizonte/MG - Brazil, March 2011.

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação em Estatística

NONPARAMETRIC INTENSITY BOUNDS FOR THE DETECTION AND DELINEATION OF SPATIAL CLUSTERS

Fernando Luiz Pereira de Oliveira

Tese de doutorado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Doutor em Estatística.

Área de Concentração: Estatística e Probabilidade
Orientador: Luiz Henrique Duczmal
Co-orientador: André Luiz Fernandes Cançado

Belo Horizonte/MG, Março de 2011

Dedicatória

Dedico este trabalho a minha mãe Benedita, ao meu pai João, meu irmão Francisco, a minha Grazinha, minha cunhada Mariana e a Deus que me acompanhou e acompanha em todos os momentos. Obrigado por me apoiarem cada um com sua forma.

Amo Vocês!

Agradecimentos

Agradeço a CAPES e FAPEMIG. Agradeço a todos que diretamente ou indiretamente vieram a contribuir para o desenvolvimento desta Tese. Em especial agradeço ao Professor Orientador Amigo Luiz Henrique Duczmal, André Luiz Fernandes Cançado e Anderson Ribeiro Duarte.

Obrigado!

Abstract

There is considerable uncertainty in the disease rate estimation for aggregated area maps, especially for small population areas. As a consequence the delineation of local clustering is subject to substantial variation. Consider the most likely disease cluster produced by any given method, like SaTScan Kulldorff [2006], for the detection and inference of spatial clusters in a map divided into areas; if this cluster 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? Do all the areas inside the cluster have the same importance from a practitioner perspective?

We propose a criterion to measure the plausibility of each area being part of a possible localized anomaly in the map. In this work we assess the problem of 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) shall be considered as just one of the possible realizations of the random variable vector with an unknown expected number of cases. In our methodology we perform m Monte Carlo replications: we consider 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 is detected and the corresponding m likelihood values obtained by means of the m replications are ranked. For each area, we determine the maximum likelihood value obtained among the most likely clusters containing that area. Thus, we construct the *intensity function* associated to each area's ranking of its respective likelihood value among the m obtained values.

The method is tested in numerical simulations and applied for three different real data maps for sharply and diffusely delineated clusters. The intensity bounds found by the method reflect the geographic dispersion of the detected clusters.

The proposed technique is able to detect irregularly shaped and multiple clusters, making use of simple tools like the circular scan. Intensity bounds for the delineation of spatial clusters are obtained and indicate the plausibility of each area belonging to the real cluster. This tool employs simple mathematical concepts and interpreting the intensity function is very intuitive in terms of the importance of each area in delineating the possible anomalies of the map of rates. The Monte Carlo simulation requires an effort similar to the circular scan algorithm, and therefore it is quite fast. We hope that this tool should be useful in public health decision making of which areas should be prioritized.

***K*eywords:** Error bounds of spatial cluster; Spatial disease cluster; Spatial scan statistics.

List of Figures

Contents

Apresentação

Motivação

Existe uma incerteza considerável na estimativa de taxas de doenças para mapas de área, especialmente para áreas de população pequena. Como conseqüência, a delimitação do agrupamento local é sujeito a variações substanciais. Considere um cluster detectado por um determinado método, como SaTScan, para a detecção e inferência de conglomerados espaciais em um mapa dividido em áreas. Se este cluster é considerado estatisticamente significativo, o que poderia ser dito das áreas externas adjacentes ao cluster? Não temos informações suficientes para exclui-las de um programa de prevenção? Será que todas as áreas dentro do cluster têm a mesma importância do ponto de vista do usuário?

O problema de detecção de clusters espaciais encontra-se presente em diversas situações, sendo importante determinar modelos satisfatórios para a execução de procedimentos para detecção e avaliação destes clusters que considerem diversos fatores inclusive os citados acima.

Principais contribuições

Nesta Tese desenvolvemos um novo conceito para a detecção e representação de clusters em mapas, descrevendo seus limites de erro. Tratamos um dos principais problemas em detecção de clusters, a medição da incerteza da definição das áreas que pertencem a um cluster detectado. A técnica desenvolvida pode potencialmente ajudar em uma limitação existente ao utilizar o scan circular, que é a não discriminação entre os grupos que são mais homogêneos daqueles que são mais irregulares ou em forma de anel. O método proposto supera várias limitações em relação à estatística espacial scan: (i) conseguimos interpretar e delinear clusters diferentes do cluster primário; (ii) fornecemos uma interpretação para a incerteza de áreas que podem pertencer ao cluster. Além disso, esse método é computacionalmente muito rápido. Outra característica importante se refere à interpretação intuitiva desta nova metodologia, tornando o conceito fácil de ser compreendido para os usuários. Esperamos que a utilização desta nova metodologia seja utilizada por diversos profissionais de saúde pública que fazem uso de busca de clusters geográficos para definir melhor suas prioridades.

Organizacão da Tese

Esta Tese está organizada da seguinte forma: no capítulo 1 apresenta-se uma introdução sobre trabalhos encontrados na literatura que abordam temas relacionados com a motivação da metodologia desenvolvida, assim como a descrição de técnicas utilizadas para visualização e detecção de clusters geográficos. No capítulo 2 descreve-se todas as metodologias que foram implementadas computacionalmente nesta Tese, como os métodos de detecção de clusters scan circular e genético, e um método de suavização de taxas muito

utilizado na literatura. Neste capítulo apresentamos também o desenvolvimento da nova metodologia proposta nesta Tese, que chamamos de função intensidade. No capítulo 3 apresentamos um estudo numérico através de simulações em diversos tipos de mapas para testarmos a eficiência da nossa metodologia proposta. No capítulo 4 apresentamos a aplicação da metodologia proposta em três estudos de casos, utilizando como método de detecção de cluster o scan circular. Nos capítulos 5 e 6 utilizamos simulações para observar o comportamento da função intensidade com um método de detecção de cluster irregular em situações computacionais mais complexas. Finalmente no capítulo 7 fazemos as considerações finais desta Tese.

Chapter 1

Introduction

There are many methods for the detection and inference of geographic clusters Cressie [1993], Elliott et al. [1995], Kulldorff [1999], Moore and Carpenter [1999], Waller and Jacquez [2000], Lawson et al. [1999], Glaz et al. [2001], Lawson [2001], Balakrishnan and V [2002], Buckeridge et al. [2005]. A large number of methods rely on the Spatial Scan Statistic (Kulldorff [1997]), a development of the Naus spatial scan statistic (Naus [1965]). Based on this statistic, several extensions were proposed, modifying the shape of the circular window used in the circular scan statistic (Kulldorff and Nagarwalla [1995]) to include irregular shapes (Duczmal and Assunção [2004], Patil and Taillie [2004], Tango and Takahashi [2005], Kulldorff [2006], Duczmal et al. [2006, 2007], Yiannakoulias et al. [2005]), see Duczmal et al. [2009] for a recent review. However, those methods generally do not discuss the possible uncertainty in the delineation of the most likely cluster found. There exists nowadays a crescent demand of interactive software for the visualization of spatial clusters (Hardisty and Conley [2008]).

A technique was developed in Boscoe et al. [2003] to visualize relative risk and statistical significance simultaneously. Given a map of k areas, with

their respective centroids, the procedure builds a grid of equidistant points between all combinations of two, three and four adjacent area centroids. For each grid point the distances to the areas centroids are computed and sorted. These distances are used to define almost circular groupings of areas, with their respective cumulative numbers of observed and expected cases. The relative risk and the likelihood ratio are then calculated for each circular grouping. The likelihood ratio values are compared to the results of a Monte Carlo simulation under the null hypothesis that there are no clusters and the cases are uniformly distributed in the population, such that the expected number of cases in each area is proportional to its population. Groupings with likelihood ratios values exceeding 95% of those obtained from the simulation are stored and stratified into ten levels of relative risk. Within each risk level, the grouping with largest likelihood ratio is then mapped. Circular groupings with lower likelihood ratio are also mapped if they did not overlap any grouping previously mapped. The final result is a ten color shaded map of areas with statistically significant relative risks, providing a very effective visualization tool to grasp these two concepts.

A visual tool was developed in Chen et al. [2008] to find circular clusters using SaTScan, repeating the search for a set of S different values for the maximum cluster size parameter. The reliability of an area a_i is defined as the number of times this area is part of a significant circular cluster found by SaTScan, divided by the number S. A typical value of S is 8, with maximum-sizes ranging from 5% to 49%, as given in the paper Chen et al. [2008]. This approach allows the interactive visual identification the so-called "core clusters", which are loosely defined as those clusters which appear more consistently through the S multiple runs varying the maximum-size parameters. This method reveals additional information about the cluster structure,

although restricted to the circular shape delineation imposed by formalism of the circular scan.

The program SaTScan detects a spatial cluster in aggregated-area maps and compute its significance based on Monte Carlo simulations. This approach allows the characterization of a potential map anomaly, dividing the map into two areas, the cluster and the area outside it. In this work proposed this thesis we are interested in pursuing further questions regarding the properties of individual areas inside and outside the detected cluster. We would like to assess the relative importance of individual areas within the cluster. We would also like to verify if the areas outside the cluster and adjacent to it could be indeed excluded from the suspected anomaly region in the map. These questions are important from a public health practitioner perspective. How to access quantitatively the risk of those areas, given that the information we have (cases count) is also subject to variation in our statistical modeling? A few papers have tackled these questions recently. For example Rosychuk [2006] produces confidence intervals for the risk in every area, which are compared to the risks inside the most likely cluster.

Geographic variability studies of disease rates are essential tools in etiology (Lawson [2009]). Maximum Likelihood Estimate Bayesian methods have been proposed to obtain unbiased rates, especially for rare diseases occurring in small population areas (Efron and Morris [1973]), thus providing more precise results than the usual maximum likelihood estimators (see Marshall [1991]). This approach includes information from adjacent areas to estimate locally the risk, consequently reducing the quadratic mean error of the estimated rates. In Manton et al. [1981, 1987], Stone [1988] approaches adjust the test significance levels for geographic risk excess. Clayton and Kaldor [1987] proposed an empirical Bayes method employing Poisson like-

lihood with gamma prior distribution in disease mapping. The authors also presented a non-parametric estimation for the prior using a method which is based on a spatial autoregressive procedure to model the prior distribution parameter devised by Laird [1978].

In this thesis, a different approach is proposed to delineate the "intensity bounds" associated to the most likely cluster, by running Monte Carlo simulations. The number of cases for each area is now considered as a random variable with mean equal to the observed rate, or to some smoothing function which takes into account its first order neighborhood. We will introduce a novel approach to assess the relative importance of individual areas in the composition of the clustering structure. The main purpose of our method is to find the error bounds for the delineation of spatial clusters in maps divided into areas, through the definition of a criterion to measure the plausibility of each area being part of the cluster. As a by-product, our method is capable of identifying irregularly shaped clusters and multiple local clustering. This method is computationally fast and relies on basic ideas about the intrinsic variation of the observed number of cases for each area. This procedure allows the quantification of the uncertainty in the delineation of spatial clusters in a very precise and intuitive way, through the definition of the intensity function.

Chapter 2

Methods

2.1 Kulldorff's Spatial Scan Statistic

Consider a map divided into k areas, with under-risk population N and C cases of an observable phenomenon. The analysis is conducted conditioned on the total number of cases so that C is considered a known constant. We define a zone as any set z of connected areas. Any circular window over the study area defines a zone z formed by areas whose centroids are inside the window.

Let Z be the set of all possible zones obtained by circular windows with varying radio and centered along each of the k areas centroids. The test proposed by Kulldorff [1997] is based on the maximization of the likelihood ratio. The parameters set is $(z; p; q)$ in which z denotes a zone in Z, p is the probability of an individual in z to be a case and q is the probability of an individual outside z to be a case. Such probabilities are constant for all individuals. Considering that there are no clusters within the map (null hypothesis), the number of cases in each area follows a Poisson distribution, with expected value proportional to its population. Define $L(z)$ as the like-

lihood under the hypothesis that the zone z is a cluster ($H_A : p > q$), and L_0 the likelihood under the null hypothesis ($H_0 : p = q$). Let $n(z)$ and $c(z)$ be, respectively, the population and cases inside z, and $\mu(z) = \frac{n(z)}{N}C$ the expected number of cases inside z under the null hypothesis. For the Poisson model the likelihood function(Kulldorff [1997]) is:

$$L(z,p,q) = \frac{e^{-pn(z)-q(N-n(z))}}{C!}p^{c(z)}q^{C-c(z)}\prod_{j=1}^{m} n(j) \tag{2.1}$$

The likelihood ratio, λ, can be written as

$$\lambda = \frac{Sup_{H_A}\{L(z)\}}{Sup_{H_0}\{L(z)\}} = \frac{Sup_{z\in Z, p>q}\{L(z,p,q)\}}{Sup_{p=q}\{L(z,p,q)\}} = \frac{L(\hat{z})}{L_0} \tag{2.2}$$

By definition, $L_0 = \frac{e^{-C}}{C!}\left(\frac{C}{N}\right)^C \prod_{j=1}^{m} n(j)$.

Hence, likelihood ratio is expressed by

$$\lambda = \begin{cases} Sup_{z\in Z}\dfrac{\left(\frac{c(z)}{n(z)}\right)^{c(z)}\left(\frac{C-c(z)}{N-n(z)}\right)^{C-c(z)}}{\left(\frac{C}{N}\right)^C} & \text{, if } \frac{c(z)}{n(z)} > \frac{C-c(z)}{N-n(z)} \\ 1 & \text{, otherwise} \end{cases}$$

The distribution of $(\lambda \mid C)$ must be obtained by a Monte Carlo simulation process (Kulldorff and Nagarwalla [1995]), since the distribution of λ depends on the population distribution, what makes it almost impossible to be obtained analytically, and the usual assintotic aproximation via Chi-square distribution, since the transformation $-2log\lambda$ is not valid because regularity conditions are not satisfied.

A simplified form for the likelihood ratio is obtained considering

$I(z) = \frac{c(z)}{\mu(z)}$ and $O(z) = \frac{C-c(z)}{C-\mu(z)}$, respectively the relative risk inside and outside z:

$$LR(z) = \frac{L(z)}{L_0} = \begin{cases} I(z)^{c(z)} O(z)^{C-c(z)} & \text{, if } I(z) > 1 \\ \\ 1 & \text{, otherwise} \end{cases}$$

The most likely cluster is the zone $\hat{z}$ that maximizes $LR(z)$ ($LR(\hat{z}) \geq LR(z)\ \forall z \in Z$). Since the logarithm is a strictly increasing function and $LR(z)$ increases very quickly, it is more convenient to maximize $LLR(z) = log\{LR(z)\}$.

Alternatively we could detect a cluster simply considering the incidence of cases in each zone, that is, the ratio between the number of observed cases and the population, or even the relative risk given by the number of observed cases divided by the expected number of cases. However, these measures do not take into account that, a low populated zone will most likely present low significance, even if it presents high relative risk. The test based on the LLR (Kulldorff [1997]) bypass this problem since it also considers not only the relative number but also the absolute number of cases.

The statistical significance of the most likely cluster of observed cases is computed through a Monte Carlo simulation, according to 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 many times, and the obtained distribution of the values is compared with the LLR of the most likely cluster of observed cases, producing an estimate of its p-value.

2.2 Single-Objective Genetic Algorithms

Conley et al. [2005] proposed a genetic algorithm to explore a configuration space of multiple agglomerations of ellipses for point data sets. The method

employed a strategy to "clean-up" the best configuration found in order to geometrically simplify the cluster. A genetic algorithm was used to find clusters in point data sets, shaped as the intersections of circles with different sizes and centers (see Sahajpal et al. [2004]). In order to use the procedures mentioned above , it is necessary to use some heuristic optimizer. Among the possible heuristics to be used in the detection spatial clusters problem, genetic algorithm was implemented for the detection of clusters and inference in Duczmal et al. [2007] using the objective maximized the test Kulldorff's Scan statistic. The algorithm parts of an initial population of possible solutions in order to build a sequence of generations. In the generations, three operators are used: *crossover* and *mutation* serve to increase the variability of the population of solutions and the *selection* operator chooses who will be part of the next generation, directing the search and maintaining a fixed population size within a generation. The crossover operator creates new individuals (new zones), combining the features of two individuals (zones) were randomly chosen and named by parents A and B. Several new individuals are produced which are intermediate zones between the two extreme zones A and B. The mutation operator introduces random perturbations in the characteristics of an individual zone (adding or removing one random region) thus increasing the variability of the population. The selection operator classifies the zones according to the value of the objective function, in this case of the Spatial Scan statistic, choosing those which will be part of the next generation. It is expected to find individuals (zones) with higher values for the objective function as the generations evolve. A geometric compactness penalty function is employed to avoid excessive irregularity of the cluster geometric shape. This algorithm is an order of magnitude faster and exhibits less variance compared to other algorithms (see Duczmal et al. [2007]), such

as the Simulated Annealing Scan presented in Duczmal and Assunção [2004], and it is more flexible than the Elliptic Scan. It has about the same power of detection as the Simulated Annealing Scan for mildly irregular clusters and it is superior for the very irregular ones.

2.3 Multi-objective Genetic Algorithms

Genetic algorithms are widely used for optimization problems in multi-objective, assessing the development of possible solutions, simultaneously evaluating two or more objectives as in Fonseca and Fleming [1995], Takahashi et al. [2003]. In Duczmal et al. [2007] it is suggested the use of Compactness Geometric penalty for a multi-objective Scan algorithm. In this proposal the penalty would be one of the objective functions, while the likelihood ratio $LLR(z)$ would be another objective function.

The pairs (LLR_i, K_i), representing the logarithm of the Scan statistic value and Compactness (or other penalty function) computed for each individual i (connected set of regions in the map) in the genetic population, are plotted in the Cartesian plane. The selection operator uses the concept of dominance: a point is called *dominated* if it is worse than another point in at least one objective, while not being better than that point in any other objective (see Chankong and Haimes [1983]). The *non-dominated set* consists of all solutions which are not dominated by any other solution.

The construction of the initial population and the operators of crossover and mutation are identical to those used in the single-objective genetic algorithm (see Duczmal et al. [2007] for a detailed description of those operators). 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 crossover operator. The next generation list, initially empty, stores the individuals that will survive for the next generation. We compute the set of non-dominated solutions P_0 of the current generation list, which is transferred to the initially empty next generation list; the same set P_0 is also removed from the current generation list. A new set P_1 of the remaining individuals is computed, and the procedure is repeated until the new generation list has grown to contain M individuals, where M is the number of regions of the original map and corresponds to the population size that will be held constant along the generations. After a number of steps, say l, the set P_l will eventually not be totally added to the next generation list, because this would cause the list to contain more than M individuals. In such cases, the individuals of P_l are transferred randomly, one by one, until the next generation list contains exactly M individuals. This procedure is known as *non-dominated sorting* (see Deb et al. [2002]).

In the context of irregularly shaped clusters, the first of the competing objectives (regularity of shape) could not be considered appropriate if it was the only objective of the search. If so, we would inevitably obtain a circularly shaped, but possibly meaningless, solution. Conversely, consider the complementary situation, when the maximization of the likelihood ratio, irrespective of shape, is the only objective: as we have seen in the introduction, this would also produce solutions which are not useful from a geographic perspective. The maximization of shape regularity only makes sense when coupled with the maximization of likelihood ratio, as developed in the multi-objective methodology. Isolated, neither objective is sufficient to guide the search for the most likely clusters, when we have the freedom to choose among clusters of arbitrary shape. A rather regularly shaped cluster usually has many

neighborhood connections with its adjacent regions compared to the number of component regions within the cluster due to the fact that its compactness is high. Otherwise, an irregularly shaped cluster is probably "tree-like" in the sense that the number of connections with adjacent regions is small compared to the number of component regions. In a situation where two clusters have the same LLR and one is more regularly shaped than the other, the former is preferred: the compactness of a cluster is generally related to the strength with which its component regions connect to each other. In this regard, compactness is considered as a measure of stability of the cluster, as a solid geographic entity: we probably can remove a few regions from a regularly shaped cluster without breaking it apart, but a similar operation may not be possible for a highly irregularly shaped cluster.

2.4 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 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 *intensityfunction*

$f : \{1, \ldots, m\} \longrightarrow \mathbb{R}$ by $f(j) = LLR_{(j)}$, $j = 1, \ldots, m$.

For each area a_i, let:

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

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.

2.5 Rate correction using empirical Bayesian estimator

We shall consider a variation of the procedure described in the previous section. Instead of using the observed number of cases, this variant uses Marshall's smoothed estimates of the number of cases based on the information of first order neighborhood of each area. We then compute the intensity function in those two situations, employing the raw number of cases and Marshall's estimates.

Empirical Bayes methods were employed by Marshall [1991] and Yasui et al. [2000]. Studies involving disease rates to show the geographical variability are common in epidemiological approaches. For this kind of approach it is important to assess the problem of obtaining unbiased estimates. Some Bayesian methods have been proposed in the literature for estimation of risks in small areas. These methods are based on information from other areas that comprise the region of study. One consequence of using these methods is the decreasing of the total mean square error of the estimates Efron and Morris [1973]. That is, relative risks are estimated more accurately by Bayesian methods than by using maximum likelihood estimation. Authors like Marshall [1991] and Yasui et al. [2000] address this issue.

Efron and Morris [1973] were among the first to work with this approach using empirical Bayes methods. Clayton and Kaldor [1987] proposed a procedure for empirical Bayes estimation using a Poisson likelihood and a gamma priori distribution. One approach was suggested by Stone [1988] to adjust the significance levels in testing for geographical risks in excess, as well as in Manton et al. [1981] and Manton et al. [1987]. Clayton and Kaldor [1987] also suggested a non-parametric estimate for the prior distribution using a

method proposed by Laird [1978] who proposed a procedure to model the parameters of a priori distribution using a spatial autoregressive method.

Using Bayesian methods in the estimation of spatial phenomena have the extra advantage of allowing the incorporation of spatial similarities between adjacent areas in risk estimates. Adding this information to the estimation of risk can lead to maps with more stable estimates and more precise differentiation between what is a true high (or very low) risk and what is indeed a random fluctuation caused by small populations. Moreover, it is expected that the estimates reproduce the spatial pattern of the real risks.

In this thesis we use the estimation procedure proposed by Marshall [1991] to obtain estimates of relative risks. We use local empirical Bayesian estimators, because it is often reasonable to consider adjacent areas whose rates are similar because they are likely to be similar in other aspects. We use the first order neighbors of the area for which we want to get the estimated rate. The methodology developed by Marshall proposed an empirical Bayesian estimator for the risk of rare diseases, where one can approximate the distribution of the number of cases by the Poisson distribution with parameter estimated by the method of moments. Consider a map divided into k areas indexed by i, $i = 1, 2, ..., k$. Suppose that events are recorded for each area in a period of time. Let θ_i be the event rate in the i-th area and assume that y_i, the number of events accumulated in the i-th area during this period, is distributed as a Poisson random variable with mean $E(y_i|\theta_i) = n_i\theta_i$, where n_i is the population at risk in the i-th area. The maximum likelihood estimator of θ_i is $t_i = y_i/n_i$. This estimator has mean and variance conditioned on θ_i given by $E(t_i|\theta_i) = \theta_i$ and $V(t_i|\theta_i) = \theta_i/n_i$, respectively. In the Bayesian approach, θ_i has *a prior* distribution with mean $m_i = E_{\theta_i}$ and variance $A_i = V_{\theta_i}$. Unconditionally, t_i has mean $m_i = E_{t_i}$ and variance $V_{t_i} = A_i + \frac{m_i}{n_i}$. Efron and

Morris [1973] showed that, given m_i and a_i, the best linear Bayes estimator for θ_i is expressed by

$$\hat{\theta}_i = w_i t_i + (1 - w_i) m_i$$

where $w_i = \frac{A_i}{(A_i + m_i/n_i)}$ is the a ratio between the a *prior* variance of θ_i and the unconditional variance of t_i. The global empirical Bayesian estimator proposed by Marshall [1991] assumes that the distribution of θ_i is the same for all areas and then replaces m_i and A_i by m and A, respectively. Using the method of moments, Marshall showed that the estimates for m and A are given, respectively, by $\tilde{m} = \frac{\sum y_i}{\sum n_i}$ and $\tilde{A} = s^2 - \frac{\tilde{m}}{\bar{n}}$, where $s^2 = \frac{\sum n_i (t_i - \tilde{m})}{\sum n_i}$, $\bar{n} = \frac{\sum n_i}{N}$ and k is the number of areas of the map. As the overall proposal is spatially invariant, i.e., independent of the performed permutation, the estimates do not change. It is necessary to change the expression of θ_i for the estimation of the a *prior* parameters set to be performed based on information from the neighboring areas of i. In this case, w_i, m, s^2 and n are replaced by W_i, M_i, s_i^2 and n_i, respectively, calculated only with data from the neighboring areas of i, and are defined as the local empirical Bayesian estimators.

Marshall's smoothing procedure is advantageous when the number of cases is very small. It will be used for the Chagas' disease map, which has a reduced number of cases, as we will see in the Results section.

Chapter 3

Results and Discussion

The methodology proposed in this thesis 2.4 was tested in numerical simulations and it was applied in three case studies.

3.1 Numerical Simulations

Three different types of "true" artificial clusters will be tested: a single circular cluster (in two maps with different relative risks), a L-shaped irregular cluster, and a double circular cluster (also in two maps with different relative risks). In all situations, the map consists of a rectangular array of 203 hexagonal cells, each cell with population 1000. The centroids of the hexagonal cells are not placed in a perfectly regular array; we introduced a slight random displacement on both x and y axes, in order to avoid ties when measuring distances between any two centroids. Cases are randomly distributed such that the cells inside the true cluster have higher probability of receiving cases than the areas outside it; the resulting maps with the randomly distributed cases are also displayed. That means that we will find clusters in "noisy" maps, where the number of cases is not homogeneously distributed

inside and outside the artificial clusters. The clusters found by the circular scan are also shown. Finally, we display the resulting maps built through the intensity function. Supposing a normal distribution of risks in the map, we consider very high relative risk clusters (the relative risk inside the cluster is 5 standard deviations above the average global risk) and moderately high relative risk clusters (the relative risk inside the cluster is 3 standard deviation above the average global risk). For a given map, the (greater than 1.0) risk is the same for all areas inside the cluster, and the risk is the same (1.0) for all areas outside the cluster.

3.2 Single Circular Cluster

Figure 3.1 shows a circularly shaped true artificial cluster with very high relative risk (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c). The intensity function is displayed in Figure 3.2(a). Finally, the intensity bounds map obtained by our method is shown in Figure 3.2(b).

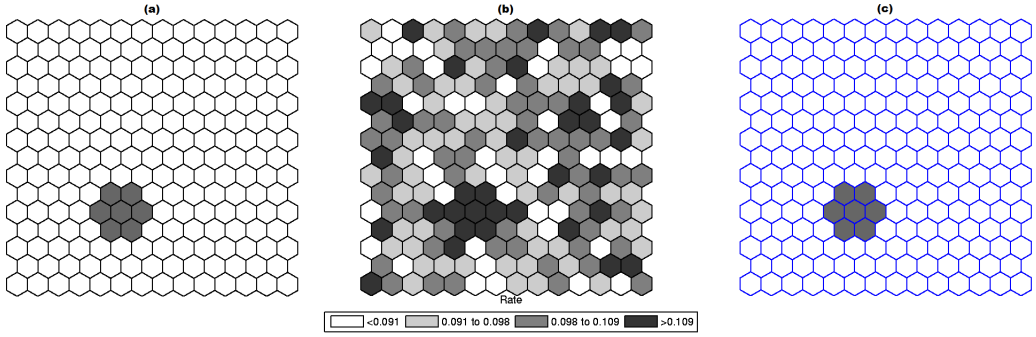

Figure 3.1: A single circularly shaped true artificial cluster with very high relative risk (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c).

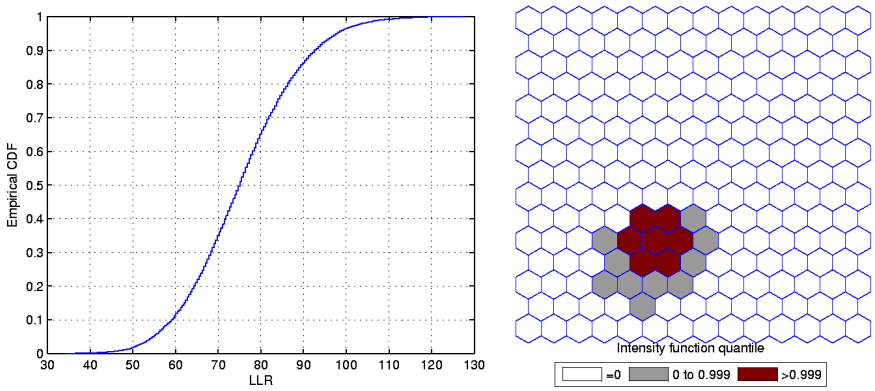

Figure 3.2: The intensity function (a) and the intensity bounds map (b) for the very high relative risk single circular cluster.

Figures 3.3 and 3.4 show the analogous results for another circularly shaped true cluster, with moderately high relative risk, for comparison.

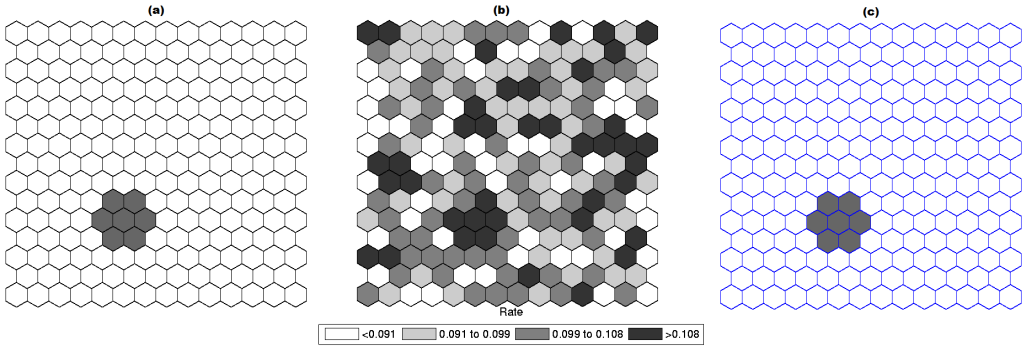

Figure 3.3: A single circularly shaped true artificial cluster with moderately high relative risk (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c).

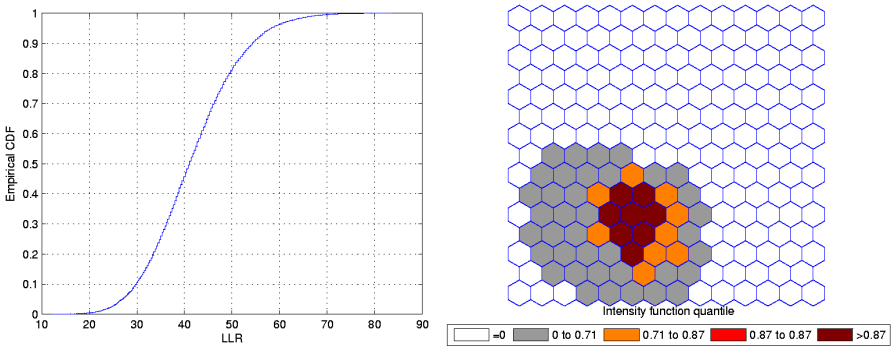

Figure 3.4: The intensity function (a) and the intensity bounds map (b) for the moderately high relative risk single circular cluster.

The intensity bounds of the very high relative risk cluster are more sharply defined than those corresponding to the moderately high relative risk cluster, as expected. Observe that in both instances the true clusters were clearly detected, as represented by the darkest shade in Figures 3.2 and 3.4.

3.3 Irregularly Shaped Cluster

Figure 3.5 shows a L-shaped true artificial cluster (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c). The intensity function is displayed in Figure 3.6(a). The intensity bounds map obtained by our method is shown in Figure 3.6(b).

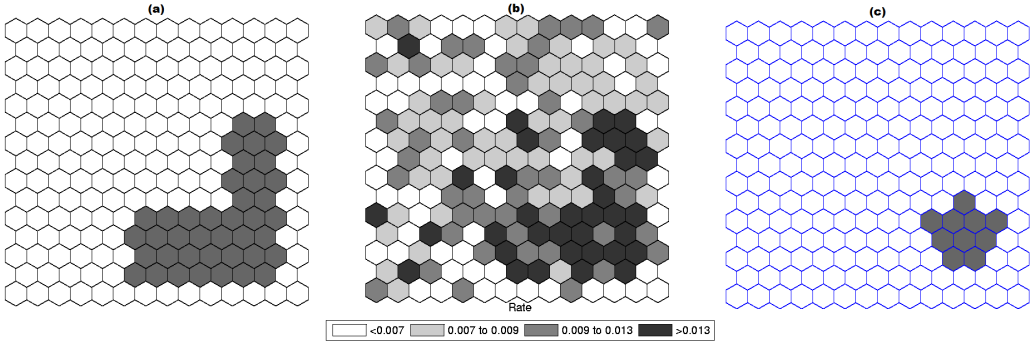

Figure 3.5: The L-shaped true artificial cluster (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c).

The circular scan detected a circular cluster centered in the angle formed by the two braces of the L-shaped cluster. However, the intensity bounds roughly delineated the L-shape, with a more intense region located around the angle of the L-shaped cluster. Sometimes the realizations of the random variable produced maps where circular clusters were found centered in the angle of the L-shaped cluster, but, very interestingly, also produced circular clusters centered along the braces of the L-shaped cluster. As a result, the overall intensity map of Figure 3.6 indicates the form of the L-shaped cluster.

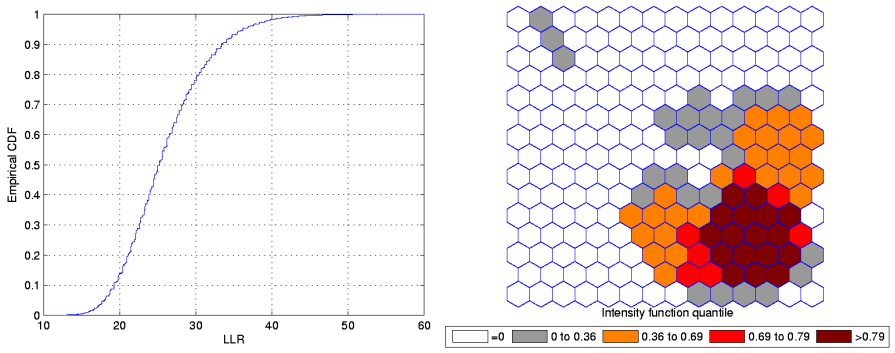

Figure 3.6: The intensity function (a) and the intensity bounds map for the L-shaped artificial cluster.

3.4 Double Circular Cluster

Figure 3.7 shows a double circularly shaped true artificial cluster with very high relative risk (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c). The intensity function is displayed in Figure 3.8(a). Finally, the intensity bounds map obtained by our method is shown in Figure 3.8(b).

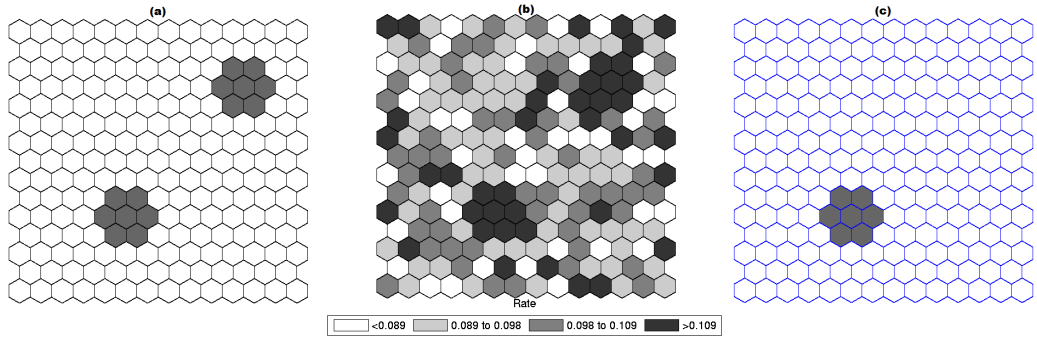

Figure 3.7: A double circularly shaped true artificial cluster with very high relative risk (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c).

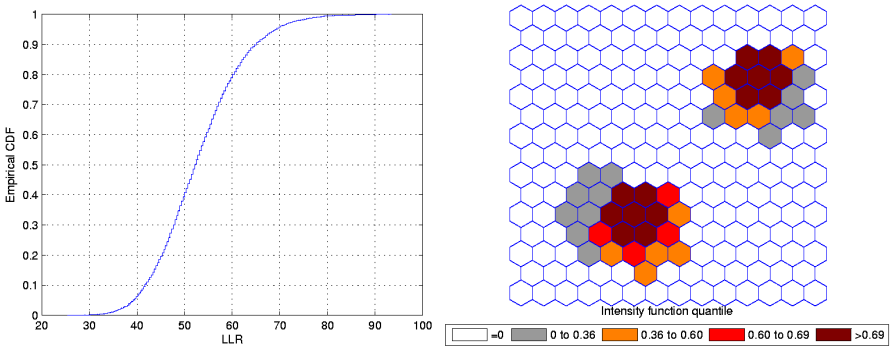

Figure 3.8: The intensity function (a) and the intensity bounds map (b) for the double circularly shaped cluster with very high relative risk.

Figures 3.9 and 3.10 show the analogous results for another double circular true cluster, with moderately high relative risk, for comparison.

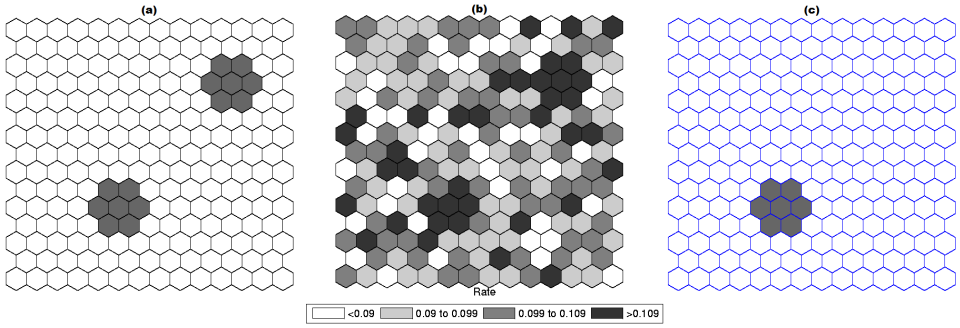

Figure 3.9: A double circularly shaped true artificial cluster with moderately high relative risk (a), the random generated cases map of rates (b), and the cluster detected by the circular scan (c).

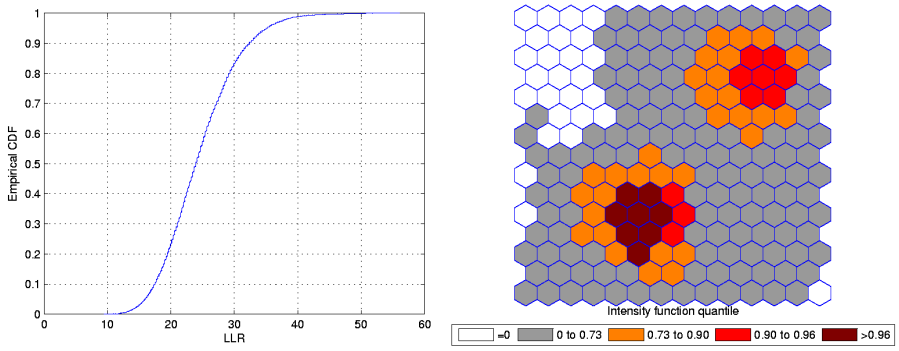

Figure 3.10: The intensity function (a) and the intensity bounds map (b) for the moderately high relative risk double circular cluster.

As displayed in Figures 3.7(b) and 3.9(b), the local rates of the two components of the double cluster are not equal, and the circular scan detected only the circular component cluster with the highest rate (Figures 3.7(c) and 3.9(c)). However, the intensity bounds delineated both circular clusters,

with a more intense region located around the highest risk circular component (Figures 3.8(b) and 3.10(b)). Sometimes the realizations of the random variable produced maps where the highest risk circular component was found, but also produced circular clusters centered in the lower risk component. As a result, the overall intensity map indicates the two components, with different intensities.

Chapter 4

Real Data Case Studies

To illustrate the method proposed in this thesis, we present three real data case studies. In the first study, with homicide cases from Minas Gerais state, Brazil, the most likely cluster is compact and very sharply delineated, with negligible geographic dispersion. The second study is a well-known benchmark of female breast cancer in the Northeast U.S. (Kulldorff [1997]), and the third case study displays Chagas' disease cases in puerperal women, also data from Minas Gerais state, Brazil. In those two last studies, the most likely clusters are not sharply delineated, presenting moderate geographic dispersion. The breast cancer study has many cases, compared to the reduced number of cases of the Chagas' disease study, allowing us to compare the performance of the map in two very different situations.

In the Chagas' disease study we used both the raw and Marshall's smoothed rates, due to the small number of cases. On the other hand, for the the other two studies we have only presented raw rates results, because there are no advantages in employing smoothed rates when the raw rates are based in a large number of cases. For all maps, each area a_i will be colored according to the quantile given by the function value $q(a_i)$, as explained in the previous

section. The choice of the quantile level representation by distinct shades of color varies in each map. We have chosen quantile levels in order to improve the visualization of the intensity function in the maps. All blank areas were never part of any cluster in the Monte Carlo simulations, corresponding to those areas a_i for which $q(a_i) = 0$. In the software, the user may choose arbitrary quantiles to represent the data. All the programming was made using Matlab 7.10 and the code is available from the authors.

4.1 Homicide Clusters

Minas Gerais state is located in Brazil's Southwest and consists of 853 municipalities, with 20,912 registered homicides from 2003 to 2007, and an estimated population of 19,150,344 in 2005. Data are available from the Brazilian Ministry of Health (http://www.datasus.gov.br) and the Brazilian Institute of Geography and Statistics (http://www.ibge.gov.br).

The raw rates map is presented in Figure 4.1(a) and the population at risk map in Figure 4.1b. The Monte Carlo procedure described in the Methodology section is performed for the raw rates, producing their respective intensity function. The intensity function for the raw rates map is displayed in Figure 4.2. Figure 4.3(a) shows the most likely cluster found by circular scan. Figure 4.3(b) show the map corresponding to the intensity function derived from the raw rates map.

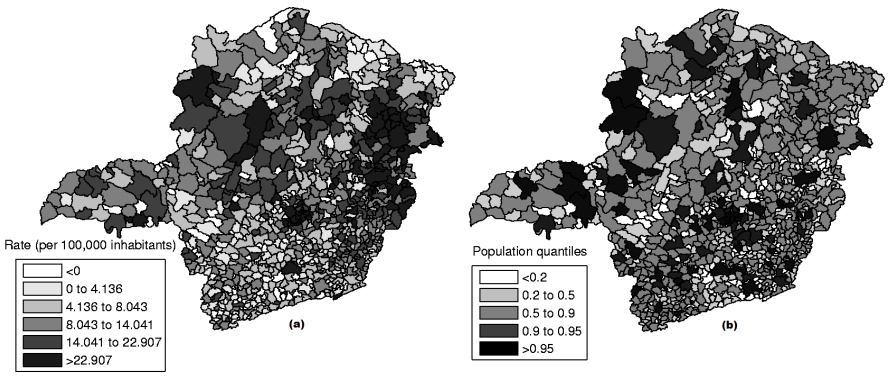

Figure 4.1: Homicide rates map (a) and population at risk map (b) in Minas Gerais State, Brazil.

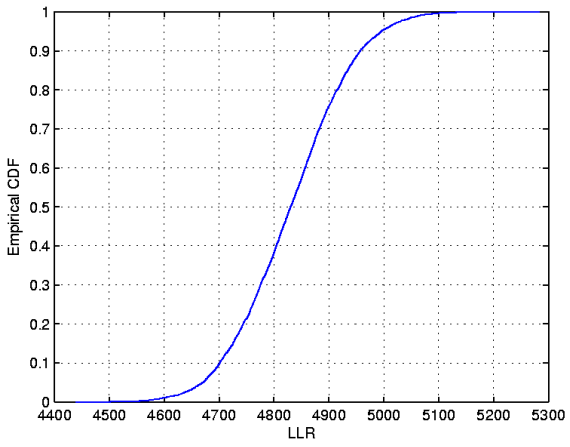

Figure 4.2: The intensity function for the homicides map.

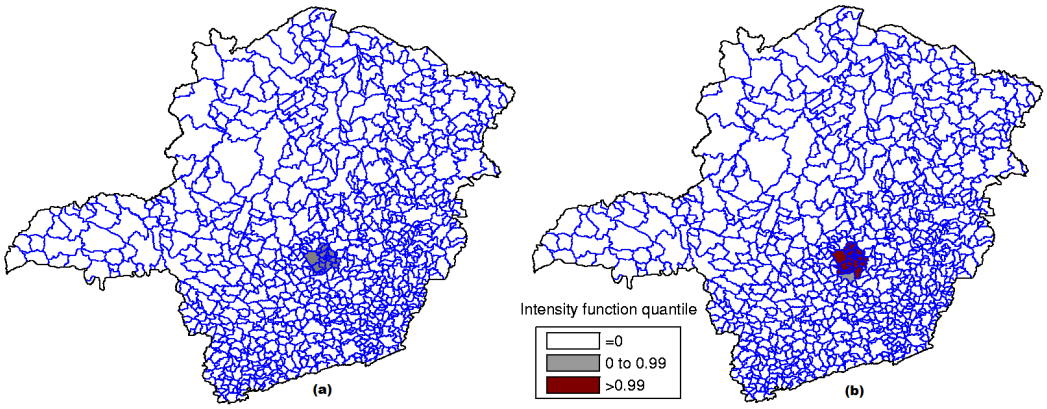

Figure 4.3: The most likely cluster found by the circular scan (a) and intensity function map (b) for the homicides map.

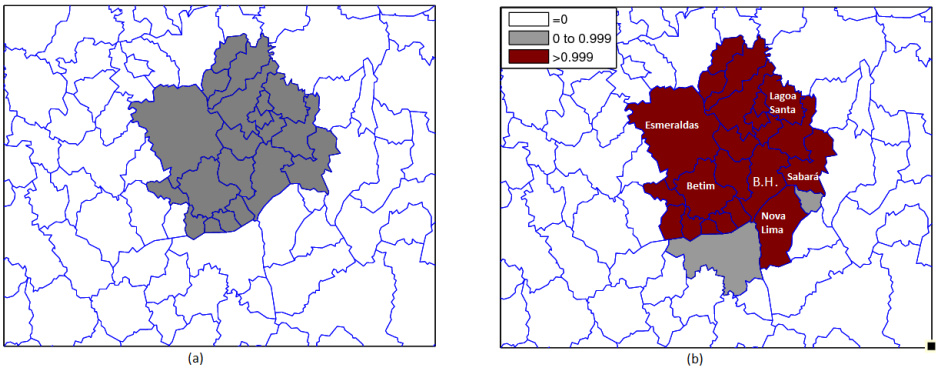

Figure 4.4: The most likely cluster found by the circular scan (a) and intensity function map (b) for the homicides map.(Zoom)

In the intensity function map, the non-blank areas attain almost the same level, meaning that the anomaly is very conspicuous. On the other hand, this anomaly is compact and coincides with the most likely cluster found by the circular scan. Although there are other places in the map where the rates are elevated, the values of the intensity function are not elevated enough to produce non-blank areas outside the anomaly in the center of the map.

4.2 The Breast Cancer Clusters in Northeastern United States

The data set of mortality from breast cancer in the Northeastern U.S. consists of age-adjusted 58,943 deaths for the period from 1988 to 1992, with the female population at risk of 29,535,210 in 1990. This map consists of 245 counties in 10 states and the District of Columbia. This dataset has been studied in detail using the circular spatial scan statistic (Kulldorff et al. [1997]) and the elliptic spatial scan statistic (Kulldorff et al. [2006]). The raw rates map is presented in Figure 4.5(a) and the population at risk map in Figure 4.5(b). The Monte Carlo procedure is performed producing its respective intensity function, displayed in Figure 4.6.

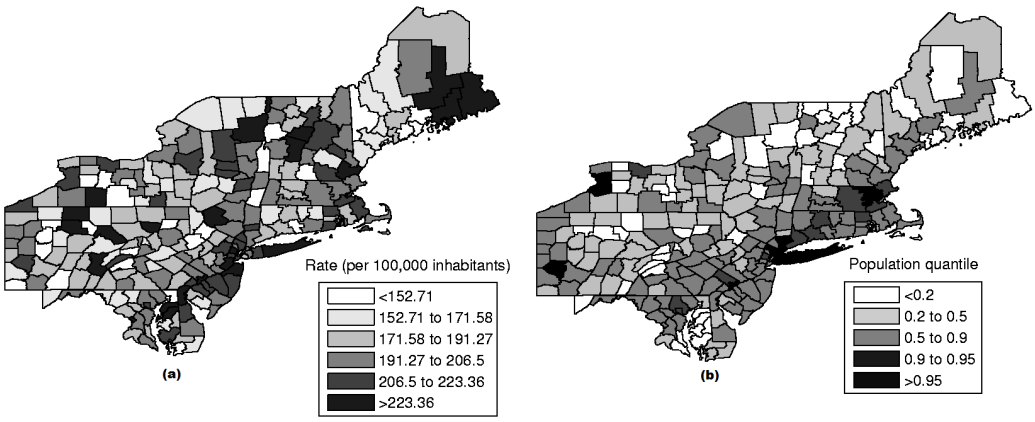

Figure 4.5: The rates map (a) and population at risk map (b) for the Northeast U.S. breast cancer data.

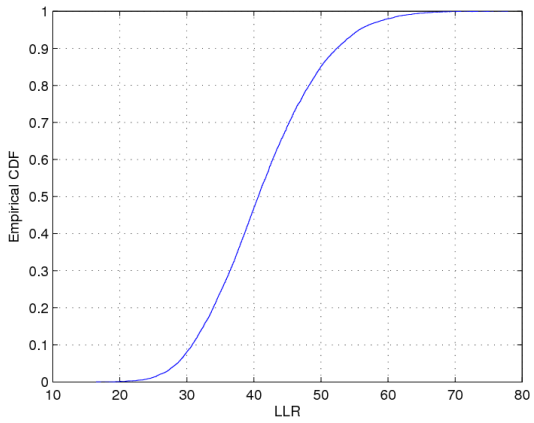

Figure 4.6: The intensity function for the Northeast U.S. breast cancer data.

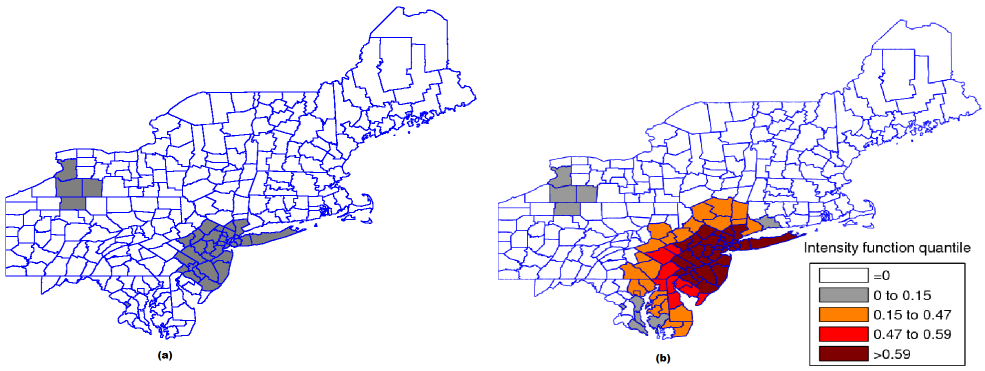

Figure 4.7: The three strongest clusters found by SaTScan Kulldorff et al. [1997] (a) and intensity function map (b) for the Northeast U.S. breast cancer data.

This case study presents a very different situation from the first example. The map derived from intensity function in Figure 4.7(b) shows the presence of various anomalies placed at different parts of the study area, indicating their geographic dispersion. We clearly observe three distinct groups of shaded areas in Figure 4.7(b), consistently matching with the three strongest clusters found by SaTScan (Kulldorff et al. [1997]), shown in Figure 4.7(a). The darkest shaded group is associated to the New York, NY-Philadelphia, PA primary cluster, with p-value 0.0001. The upper left group of four gray areas coincides exactly with the Buffalo, NY secondary cluster, with p-value 0.122. Finally the gray area at the lower center of the map corresponds to the Washington, DC secondary cluster, with p-value 0.147.

This example shows that the intensity function has the ability to delineate even the multiple and irregularly shaped potential clusters. We stress the fact that, for each Monte Carlo replication, only the primary most likely cluster was used to build the map derived from the intensity function of Figure 4.7(b).

4.3 Chagas' Disease Clusters

This subsection presents the data set of Chagas' disease cases in puerperal women in Minas Gerais state, Brazil. The population at risk consists of women that gave birth to babies in the period of July to September, 2006. The new-born babies were blood tested to detect the presence of the Chagas disease antigen, with coverage above 96%. A positive test means that the mother is infected. These tests were conducted through the project PETN-MG (Minas Gerais State Program of New-Born Screening) coordinated by the research group NUPAD-MEDICINA/UFMG from Federal University of

Minas Gerais Medical School (http://www.nupad.medicina.ufmg.br) in col-laboration with Minas Gerais State Health Secretary. The state is divided into 853 municipalities with a total population at risk of 24,969 women. Af-ter a comprehensive screening to eliminate false positives a total number of 113 cases were obtained.

The raw rates map is presented in Figure 4.8(a) and the population at risk map in Figure 4.8(b). The Monte Carlo procedure is performed for both the raw rates and Marshall's smoothed rates maps, producing their respective intensity functions. The intensity function for the raw rates map is displayed in Figure 4.9(a). The intensity function for Marshall's smoothed rates is displayed in Figure 4.9(b). Figure 4.10(a) shows the most likely cluster found by circular scan. Figures 4.10(b) and 4.10(c) show the maps corresponding to the intensity function derived from the raw rates map and the smoothed rates map, respectively.

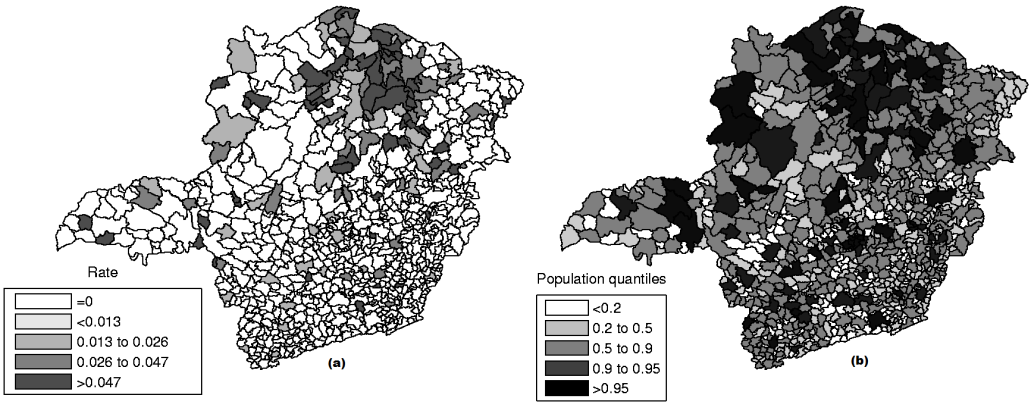

Figure 4.8: Chagas' disease rates map (a) and population at risk map (b) in Minas Gerais State, Brazil.

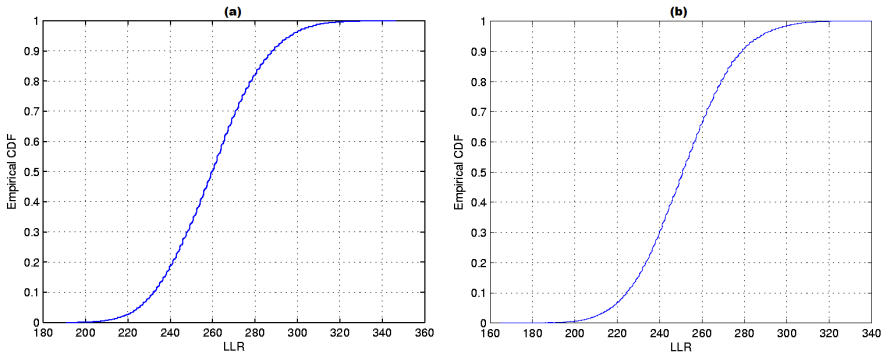

Figure 4.9: The intensity functions of the raw rates (a) and smoothed rates (b) for the Chagas' disease map.

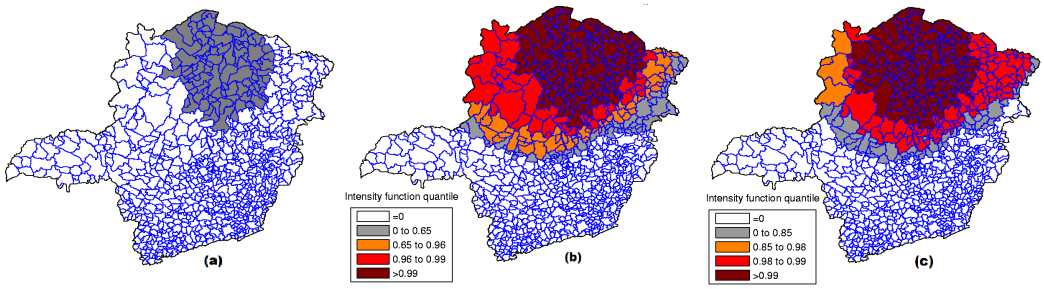

Figure 4.10: The most likely cluster found by the circular scan for the raw rates map (a), the raw rates intensity function map (b) and Marshall's smoothed rates intensity function map (c) for the Chagas' disease map.

The maps derived from the raw (Figure 4.10(b)) and smoothed (Figure 4.10(c)) intensity functions show the presence of a strong anomaly. For the map of Figure 4.10(b), the area formed by the highest intensity areas (dark colored) coincides almost perfectly with the primary cluster found by the circular scan. However, the corresponding area of Figure 4.10(c) does not match so well the primary cluster, due to the overdispersion created by Marshall's smoothing procedure. In both maps, we observe the high geographic

dispersion of the anomaly, which spreads over the northern part of the state. This example shows that the error bounds of the existing cluster were easily visualized by means of the intensity function. The application of Marshall's smoothing procedure does not contribute to improve the delineation of the anomaly, even considering that there are few cases in the study area.

Chapter 5

Irregulary shaped clusters

The circular spatial scan has several limitations, which were discussed in the literature (Duczmal et al. [2006], Kulldorff et al. [2006]). In particular the circular window is not adequate to delineate irregularly shaped clusters - either choosing a small proper subset of the cluster (underestimation) or choosing a large circle containing the cluster as a proper subset (overestimation). One important consequence is the reduction of the power of detection. In order to overcome this limitation, many algorithms were proposed in the last five years to detect irregularly shaped clusters, substituting the circularly shaped window. Usually, the only limitation in shape for those clusters is a connectivity requirement. In this section, we will analyze the impact of irregularly shaped algorithms for the application of the intensity function discusssed in the previous sections, compared to the use of the simple circular scan, which was employed as the standard method. We will present results only for the multi-objective genetic algorithm scan (Duczmal et al. [2007, 2008], Duarte et al. [2010]), adapted for the weighted non-connectivity penalty function (Cançado et al. [2010], Duarte et al. [2011]), see 8.1 in this thesis the procedure for the weighted non-connectivity penalty function.

Figure 5.1: The most likely cluster of breast cancer among woman for the period 1988-1992, occurring sround New York, and Philadelphia, Pennsylvania, as well as four secondary clusters.

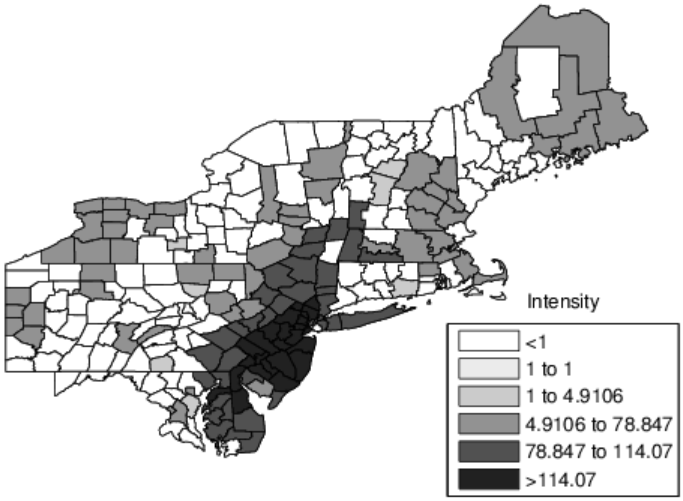

Figure 5.2: The intensity function map for the Northeast U.S. breast cancer data.

The map derived from intensity function in Figure 5.2 shows the presence of various anomalies placed at different parts of the study area, indicating their geographic focus. We observe several distinct groups of shaded areas in Figure 5.2, consistently matching with the five strongest clusters found by SaTScan (Kulldorff et al. [1997]), shown in Figure 5.1. The darkest shaded group spreads through a larger portion of the map (compared with the corresponding group found in chapter 4) and is associated to the New York, NY-Philadelphia, PA primary cluster, with p-value 0.0001. The same thing happens with the upper left group of 13 gray areas, containing the Buffalo, NY secondary cluster of four areas, with p-value 0.122. Finally the gray area at the lower center of the map corresponds to the Washington, DC secondary cluster, with p-value 0.147. The remaining two secondary clusters of Figure 5.1 have even higher p-values, and the corresponding groups in Figure 5.2 are less sharply defined. Other scattered groups also were formed through the map.

This example shows that the intensity function has the ability to delineate even more multiple and irregularly shaped potential clusters, but there is considerably more noise.

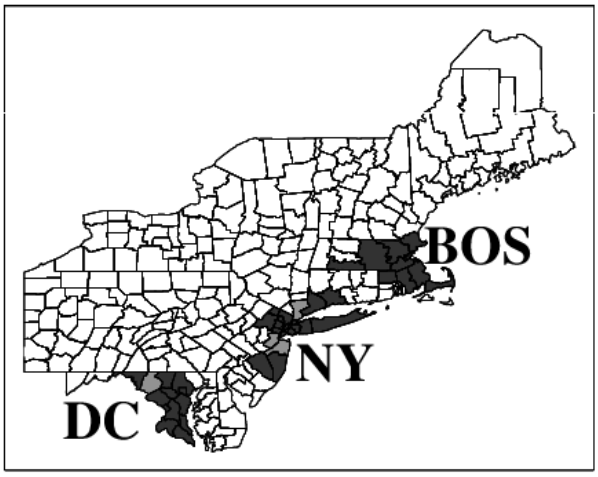

Figure 5.3: Three artificial clusters

We also present a set of simulations to illustrate the *average* behavior of the intensity function using the multiobjective genetic algorithm scan. We generated 100 Monte Carlo replications for the construction of the intensity function, for each one of the three artificial clusters shown in Figure 5.3. The intensity function map was built and then we repeated the whole process 100 times, composing the average maps shown in Figure 5.4. We stress the fact that this result is an average process, and we found a large variance in the delineation of the original clusters (Figure 5.3), as expected. Even then, the maps of Figure 5.4 show consistently the outline of the original clusters.

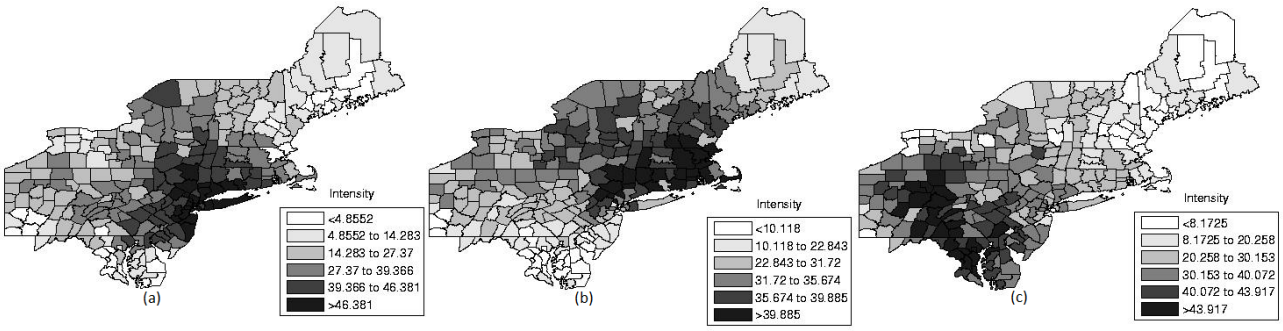

Figure 5.4: The three results for intensity function map (a) New York, (b) Boston and (c) Washington DC.

The same procedure was done for the Chagas' disease map of Minas Gerais, representing a situation where the total number of cases is small. 5.5 shows the most likely cluster of Chagas' disease in Minas Gerais found by the multi-objective genetic algorithm. 5.6 displays the combined solutions of the Pareto set, when the maximum cluster size was 5 (5.6(a)) and 10 (5.6(b)). The results show clusters considerably more sharply defined, compared to the New England map's clusters.

Figure 5.5: Most likely cluster found by genetic algorithm.

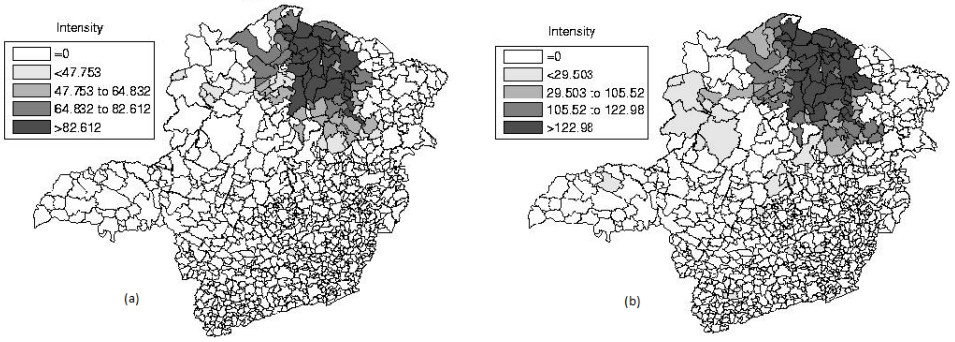

Figure 5.6: Cluster found by genetic algorith with maximum cluster size was 5(a) and cluster found by genetic algorith with maximum cluster size was 10(b).

In conclusion, our examples using artificial and real data show that there is a palpable gain when using irregularly shaped methods. This gain is translated here as a greater sensitivity to detect boundaries of the clusters, and also the capacity to detect secondary clusters. However, the informational gain is somewhat offset by the increased amount of detected noise, generated by the possibly excessive freedom of shape and/or size of the window used. Smaller, more compact windows generate less noise, but also less sensitivity. The opposite is true for larger, less penalyzed (in terms of shape) windows, which generate noiser maps with more clusters.

Our simulations seem to indicate that more complicated spatial population distributions, with several highly populated nuclei in different parts of the map, are better suited for the application of the intensity function with the circular scan; otherwise, when the population is more evenly distributed, irregularly shaped algorithms may be more useful.

It is possible to find a balance between the informational gain, but currently the most adequate parameters are not automatically chosen. A more prudent strategy, in our setting, is to evaluate several simulations with different parameter settings. Further work is needed to assess the optimal choices which could generate maps representing the adequate balance between noise and informational content. We presented results with the multi-objective genetic algorithm scan, employing the weighted non-connectivity penalty function. Other algorithms could also be used, but there is no reason to believe that the basic features should be different, when using other types of algorithms. It seems that only the range of the window size, measured as the maximum allowed population in the candidate clusters, is relevant to modify the balance of the algorithm's sensitivity to detect secondary clusters and the amount of noise in the final map.

Chapter 6

Relative Frequency Studies

One is tempted to ask if simpler criteria, aside from the intensity function definition, should suffice for the delineation of the uncertainty bounds of spatial clusters. For instance a very simple frequentist approach could be used instead: for each area a_i, consider the number m_i of Monte Carlo replications when a_i is included in a most likely cluster, divided by the total number of replications m. In Figure 6.1 we compare the intensity function map with the relative frequency map described above, for the New England breast cancer map. It could be observed that the results are almost identical. On the other hand, Figure 6.2 makes a similar comparison for the Minas Gerais Chagas' disease map. The results are considerably different; this happen because, for each area a_i, the frequentist approach does not take into account the value of the highest LLR clusters which contain the area a_i. The intensity function method, otherwise, does not produce underestimated LLR value clusters. This difference is most notable when the number of cases in the map is small, and the relative variance is larger. When the total number of cases is the map is large, both evaluations tend to produce similar clusters for each area.

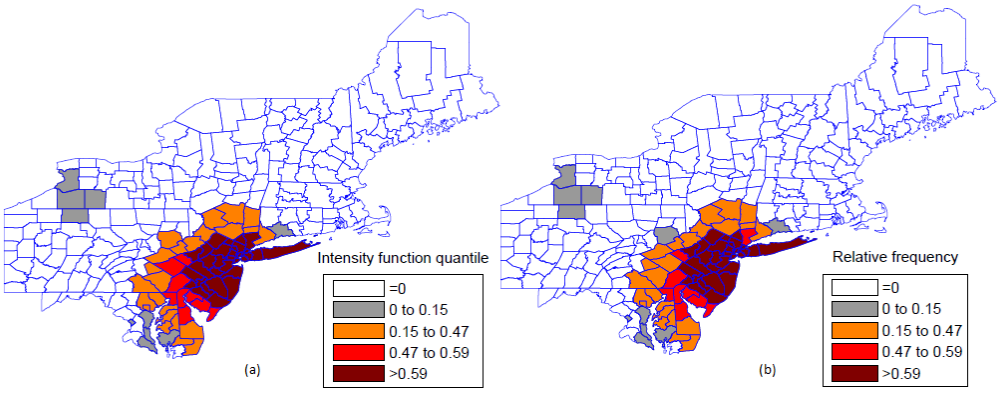

Figure 6.1: The intensity function for the raw rates map and the relative frequency map.

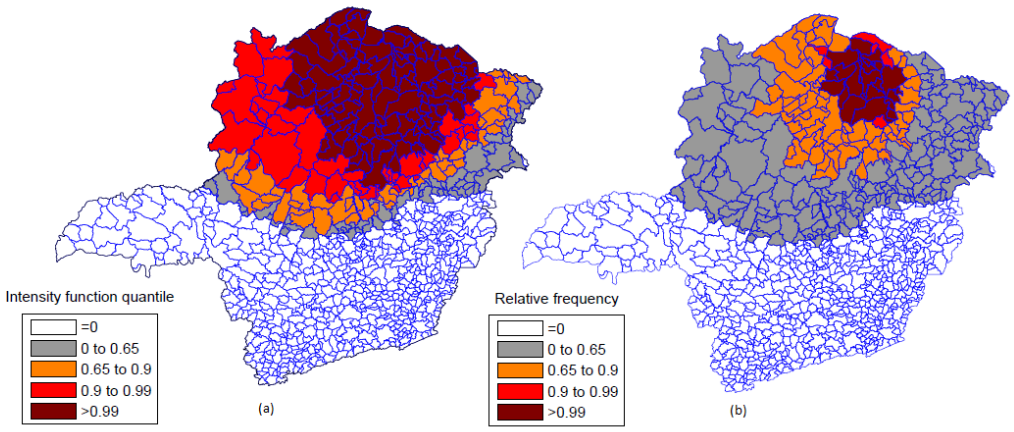

Figure 6.2: The intensity function for the raw rates map and the relative frequency map.

Chapter 7

Conclusions

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, either as a composition of neighboring areas in the map, employing Marshall's smoothing, or either considering the raw 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 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 seen in the homicides map example, 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 example low geographic dispersion occurs. However, in the other two case studies, the opposite happens. The Chagas' disease map presents 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 breast cancer, 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. It is noteworthy that the entire procedure was performed using the circular scan, and even then it identifies irregular and multiple clusters.

An analogy with our proposed method can be found in image analysis: suppose we take several short digital exposures of a very low light level scene, e.g. some deep-sky field of galaxies. Each exposure generates an image consisting of a rectangular matrix of pixels, each pixel receiving a small number of photons corresponding to the illumination of its small associated portion

of the image. The expected rate of photons is constant during all the exposures, but the number of photons received by the same pixel varies from one exposure to the other due to the stochastic nature of the process. Usually, one simply adds the values for the same pixel through all the exposures, to compose a single final image with higher sharpness (signal-to-noise). Instead, we first submit each exposure image through a filter, which in our case is the algorithm to detect the most likely cluster, and then compose all the corresponding clusters into a single "cluster image" by means of the intensity function. If the "real" cluster is very contrasting with the background noise, all exposures will produce very similar clusters, thus producing a sharply defined final cluster image. Otherwise, when the real cluster is not very conspicuous, we should observe a large variation in individual clusters, producing a poorly delineated cluster in the final image.

We presented two variants of the computation of the intensity function. The first employed the raw number of cases, and the second used Marshall's smoothed estimates of the number of cases based on the information of the first order neighborhood of each area. This was done because we were especially concerned with areas containing zero cases, which could generate biased Monte Carlo distributions of cases over the map. Marshall's smoothed estimates of cases could potentially alleviate this problem providing non-zero averages employed in the multinomial random vector. However, we have noted in all our examples that the application of Marshall's smoothed estimates produces less sharply defined intensity function maps, compared to those obtained by the use of the raw cases data. On the other hand, we could not observe any artifacts due to the use of non-smoothed raw cases data in the delineation of the anomaly. This may be explained by the simple fact that the circular spatial scan works itself as a "filter", when it joins several

areas within the circular window, thus naturally diminishing the effect of the zero cases areas in the composition of the cluster candidates. This suggests that the utilization of raw cases data does not seem to interfere with the visualization of the intensity bounds.

This tool uses simple mathematical concepts and the interpretation of the intensity function f is very intuitive in terms of the importance of each area in delineating the possible anomalies of the map of rates. The Monte Carlo simulation requires an effort similar to the circular scan algorithm, and therefore it is quite fast. Furthermore, the accuracy of the interactive construction of the map from the intensity function f increases gradually with execution time. Thus the user could stop the simulation process at any time when it is realized that the delineation of potential anomalies will converge. We therefore hope that this tool may assist in the decision process of prioritizing the areas of a map associated with potential spatial anomaly.

In this thesis we developed a new concept for detection and representation of clusters in maps, describing the their error bounds. We treat one of the principal problems in cluster detection, the uncertainty of their boundaries, measured by the plausibity of each area belonging to the real cluster. Our technique may potentially overcome one of the limitations of the spatial scan statistic which doesn't really discriminate between clusters that are homogeneous and those that are patchy or ring-like. When actually capturing this imprecision on the map, our method is able contribute with two long standing problems: first, how secondary clusters should be reported and interpreted; second, how the uncertain precision of the cluster locations should be reported and interpreted. We therefore hope that this tool may assist in the decision process of prioritizing the areas of a map associated with potential spatial anomaly.

Trabalhos Futuros

Nesta tese foi desenvolvido uma forma de delineamento da intensidade de regiões pertencerem ao cluster mais verossímil, classificando-as no mapa em questão de acordo com uma escala de intensidade, onde tratamos com dados agregados. Este procedimento permitiu um avanço em questões que antes não tinham sido abordadas como: O que pode ser dito das áreas externas adjacentes ao cluster? As áreas dentro do cluster detectado têm a mesma importância de pertencerem à anomalia? Entre outras questões. O nome dado para este procedimento foi função intensidade. Usando métodos de detecção de clusters conhecidos como scan circular e genético, com o uso da função de intensidade é possível detectar clusters irregulares e múltiplos. Este método estima a plausibilidade de todas as regiões pertencerem aos possíveis clusters existentes. O esforço computacional da função intensidade é relativamente baixo. Para trabalhos futuros desejamos estender a função intensidade para dados pontuais e para detecção de clusters espaço temporal. Assim pretendemos desenvolver ferramentas importantes para priorização de regiões que pertencem a uma determinada anomalia detectada.

Produção bibliográfica

Apresentamos as publicações que resultaram de nosso trabalho durante o doutorado. Publicações diretamente decorrentes do trabalho desenvolvido nessa tese:

Artigo publicado em periódico internacional:

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

Artigo aceito em periódico internacional:

- Oliveira, F. L. P., Duczmal, L. H., Cançado, A. L. F. (2011). Nonparametric intensity bounds for the visualization of disease clusters. *Emerging Health Threats Journal*.

Artigo submetido para publicação em periódico internacional:

- Duarte, A.R., Silva, S.B., Duczmal, L.H., Ferreira, S.J, Cancado, A.L.F., Reis, F.M and Oliveira, F.L.P. (2011). A weighted non-connectivity penalty for the detection and inference of irregular clusters. *International Journal of Geographical Information Science*.

Apresentação oral em conferência internacional:

- *9th Annual Conference International Society for Disease Surveillance* (2010), Park City, Utah.

References

N Balakrishnan and Koutras M V. *Runs and Scans with Applications*. John Wiley & Sons, London, 2002.

F P Boscoe, C McLaughlin, M J Schymura, and C L Kielb. Visualization of the spatial scan statistic using nested circles. *Health & Place*, 9:273–277, 2003.

D L Buckeridge, H Burkom, M Campbell, W R Hogan, and A W Moore. Algorithms for rapid outbreak detection: a research synthesis. *Journal of Biomedical Informatics*, 38:99–113, 2005.

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

V Chankong and Y Y Haimes. Multi-objective decision making:theory and methodology. In *North-Holland*, 1983.

J Chen, R E Roth, A T Naito, E J Lengerich, and A M MacEachren. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of u.s. cervical cancer mortality. *International Journal of Health Geographics*, 7:57, 2008.

D Clayton and J Kaldor. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681, 1987.

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

N C A Cressie. *Statistics for Spatial Data.* Wiley, New York, 1993.

K Deb, A Pratap, S Agrawal, and T Meyarivan. A fast and elitist multi-objective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6:(2):182–197, 2002.

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

A R Duarte, L H Duczmal, S B Silva, S J Ferreira, A L F Cancado, F M Reis, and F L P Oliveira. A weighted non-connectivity penalty for the detection and inference of irregular clusters. *International Journal of Geographical Information Science*, 2011. (submitted).

L Duczmal and R Assunção. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45:269–286, 2004.

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

L Duczmal, A L F Cançado, R H C Takahashi, and L F Bessegato. A genetic algorithm for irregularly shaped spatial scan statistics. *Computational Statistics & Data Analysis*, 52:43–52, 2007.

L Duczmal, A L F Cançado, and R H C Takahashi. Geographic delineation of disease clusters through multi-objective optimization. *Journal of Computational & Graphical Statistics*, 17:243–262, 2008.

L Duczmal, A R Duarte, and R Tavares. Extensions of the scan statistic for the detection and inference of spatial clusters. In N Balakrishnan and J Glaz, editors, *Scan Statistics*, pages 157–182. Birkhäuser, Boston, Basel and Berlin, 2009.

M Dwass. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28:181–187, 1957.

B Efron and C Morris. Stein's estimaton rule and its competitors - an empirical bayes approach. *Journal of the American Statistical Association*, 68: 117–130, 1973.

P Elliott, M Martuzzi, and G Shaddick. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research*, 4:137–159, 1995.

C M Fonseca and P Fleming. An overview of evolutionary algorithms in multi-objective optimization. *Evolutionary Computation*, 3:1–16, 1995.

J Glaz, J Naus, and S Wallestein. Disease mapping and risk assessment for public health. In *Springer Series in Statistics*. Springer, Berlin Heidelberg New York, 2001.

F Hardisty and J Conley. Interactive detection of spatial clusters. *Advances in Disease Surveillance*, 5:37, 2008.

M Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6):1481–1496, 1997.

M Kulldorff. Spatial scan statistics: Models, calculations, and applications. In N Balakrishnan and J Glaz, editors, *Scan Statistics and Applications*, pages 303–322. Birkhäuser, 1999.

M Kulldorff. and information management services. *Inc. SaTScan v7.0: software for the spatial and space-time scan statistics* http://www.satscan.org/, 2006.

M Kulldorff and N Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14:799–810, 1995.

M Kulldorff, E J Feuer, B A Miller, and L S Freedman. Breast cancer clusters in the northeast united states: A geographic analysis. *American Journal of Epidemiology*, 146:161–170, 1997.

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

N Laird. Non-parametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811, 1978.

A Lawson. Statistical methods in spatial epidemiology. In A Lawson, editor, *Large scale: surveillance*, pages 197–206. Wiley, London, 2001.

A Lawson, A Biggeri, D BVohning, E Lesare, J F Viel, and R Bertollini. *Disease Mapping and Risk Assessment for Public Health*. Wiley, London, 1999.

A B Lawson. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Series: Interdisciplinary Statistics.* New York, 2009.

K G Manton, M A Woodbury, and E Stallard. A variance components approach to categorical data models with heterogeneous cell populations:

analysis of spatial gradients in lung cancer mortality rates in north carolina counties. *Biometrics*, 37:259–269, 1981.

K G Manton, E Stallard, M A Woodbury, W B Riggan, J P Creason, and T J Mason. Statistically adjusted estimates of geographic mortality profiles. *J Natl Cancer Inst*, 78:805–815, 1987.

R J Marshall. Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics*, 40:283–294, 1991.

D A Moore and T E Carpenter. Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic Reviews*, 21:143–161, 1999.

J I Naus. Clustering of random points in two dimensions. *Biometrika*, 52: 263–267, 1965.

F L P Oliveira, L H Duczmal, A L F Cançado, and R Tavares. Nonparametric intensity bounds for the delineation of spatial clusters. *International Journal of Health Geographics*, 10:1, 2011.

G P Patil and C Taillie. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11:183–197, 2004.

R J Rosychuk. Identifying geographic areas with high disease rates: when do confidence intervals for rates and a disease cluster detection method agree? *International Journal of Health Geographics*, 5:46, 2006.

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

R A Stone. Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, 7:649–660, 1988.

R H C Takahashi, J A Vasconcelos, J A Ramirez, and L Krahenbuhl. A multi-objective methodology for evaluating genetic operators. *IEEE Transactions on Magnetics*, 39:(3):1321–1324, 2003.

T Tango and K Takahashi. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4:11, 2005.

L A Waller and G M Jacquez. Disease models implicit in statistical tests of disease clustering. *Epidemiology*, 6:584–590, 2000.

Y Yasui, H Liu, J Benach, and M Winget. An empirical evaluation of various priors in the empirical bayes estimation of small area disease risks. *Statistics in Medicine*, 19:2409–2420, 2000.

N Yiannakoulias, R J Rosychuk, and J Hodgson. Adaptations for finding irregularly shaped disease clusters. *International Journal of Health Geographics*, 6:28, 2005.

N Yiannakoulias, A Karosas, D P Schopflocher, L W Svenson, and M J Hodgson. Using quad trees to generate grid points for application in geographic disease surveillance. *Advances in Disease Surveillance*, 3, 2007.

Chapter 8

Annexes

8.1 Annexe A - The weighted non-connectivity penalty

8.1.1 The geometric penalty function

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 areas with high cases incidence. One way to avoid such kind of "meaningless" solution would be to use an algorithm that besides to consider the $LLR(z)$ would also use some sort of penalty for the cluster shape. One of the possible penalties that takes into 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, and 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:

$$k(z) = \frac{4\pi A(z)}{H(z)^2} \tag{8.1}$$

where $A(z)$ is the area of the zone z and $H(z)$ the perimeter of the convex hull of z. Informally, the convex hull of a planar object is the area inside a rubber band stretched around it. The compactness penalyzed scan statistic is defined as $max_{z \in Z} k(z).LLR(z)$.

8.1.2 The non-connectivity penalty function

Yiannakoulias et al. [2007] proposed a greedy algorithm to scan the set Z of all possible zones z. A 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

$$nc(z) = \frac{e(z)}{\left[3\left(v(z) - 2\right)\right].} \tag{8.2}$$

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.

8.1.3 The weighted non-connectivity penalty

We employ the multi-objective genetic algorithm, where the first objective is the logarithm of the likelihood ratio (the LLR function) and the second objective is given by our new proposal for a regularity/penalty function called *weighted non-connectivity function*.

8.1.4 Weighting the edges and nodes

The non-connectivity function proposed by Yiannakoulias et al. [2007] and given by (8.2) proved to be quite effective in the detection and inference of spatial clusters. Basically, given a zone z, this function takes into account the number of edges relative to the number of nodes of the subgraph associated with the zone z and gives the strength of the zone as a possible cluster by a measure of the connectivity between its component areas. However the non-connectivity function does not consider the population heterogeneity among the component areas. If we consider the problem of disease cluster detection in epidemiology or disease surveillance, the population heterogeneity is clearly an important feature to be included in the problem analysis. For instance, we could ask how relevant an edge is for the subgraph connectivity. If the removal of this edge breaks the graph/zone in two connected pieces, the edge relevance is different from another edge whose removal does not break the graph/zone. The original non-connectivity function does distinguish between these two situations.

But if we consider an edge whose removal breaks the graph/zone, how relevant is this edge as an element of an associated graph of a possible cluster if this edge connects two nodes corresponding to high populated areas, or to low populated areas? Our new proposal tries to answer this ques-

tion. Besides considering the associated graph connectivity structure we propose to give 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 areas 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: $P(e_{i,j}) = (pop(R_i) + pop(R_j))/2$. For a node v_i associated with the area R_i whose population is $pop(R_i)$, the weight is just the node population: $P(v_i) = pop(R_i)$.

8.1.5 Weighted non-connectivity function

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

$$w(z) = \frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} P(e_{i,j})}{3\left[\sum_{i=1}^{k} P(v_i) - 2\left(\frac{\sum_{i=1}^{k} P(v_i)}{k}\right)\right]} \tag{8.3}$$

We remark that if we consider that all areas have the same population, we recover the expression (8.2) for the non-connectivity function introduced by Yiannakoulias et al. [2007].