Arbitrarily shaped multiple spatial cluster detection for case event data

Christophe Demattei, Nicolas Molinari, Jean-Pierre Daurès

Arbitrarily shaped multiple spatial cluster detection for case event data.

Christophe DEMATTEI [a,*], Nicolas MOLINARI [a], Jean-Pierre DAURES [a]

[a] *Laboratoire de biostatistique, d'épidémiologie et de santé publique, UFR Médecine Site NORD UPM/IURC, 640 avenue du Doyen Gaston Giraud, 34295 Montpellier Cedex 5, France.*

Abstract

An original method is proposed for spatial cluster detection of case event data. A selection order and the distance from the nearest neighbour are attributed to each point, once pre-selected points have been taken into account. This distance is weighted by the expected distance under the uniform distribution hypothesis. Potential clusters are located by modelling the multiple structural change of the distances on the selection order and the best model (containing one or several potential clusters) is selected using the double maximum test. Finally a p-value is obtained for each potential cluster. With this method multiple clusters of any shape can be detected.

Key words: Spatial cluster detection test, Distance from nearest neighbour, Regression, Arbitrarily shaped clusters, Case event data

* Corresponding author: Laboratoire de biostatistique, d'épidémiologie et de santé publique, UFR Médecine Site NORD UPM/IURC, 640 avenue du Doyen Gaston Giraud, 34295 Montpellier Cedex 5, France. Tel.: +33 467 415 921; Fax.: +33 467 542 731.
Email address: demattei@iurc.montp.inserm.fr (Christophe DEMATTEI).

1 Introduction

The term "cluster" is an unusual aggregation, real or perceived, of events that are grouped together in time and/or space. Clusters of health events, such as chronic disease, injuries, and birth defects, are often reported to health agencies. When the etiology of a disease has not yet been established, it is sometimes required to examine data for obtaining evidence of temporal or spatial clustering and to establish an etiologic link with exposure. Spatial cluster detection affects severals fields: medicine, cosmology with spatial clustering of galaxies (Szalay et al., 2002), social sciences and criminology (Vinson and Baldry, 1999), agronomy and more.

The question of whether events are clustered in space has received considerable attention in the literature. The numerous tests proposed in all the above mentioned fields can be classified according to their purpose. The aim of global (or general) clustering tests (Ripley, 1977; Whittemore et al., 1987; Cuzick and Edwards, 1990; Besag and Newell, 1991; Tango, 1995, 2000) is to analyse the overall clustering tendency of disease incidence in a study region, without paying attention to cluster location. With cluster detection tests (Turnbull et al., 1990; Kulldorff and Nagarwalla, 1995; Kulldorff, 1997, 1999), potential clusters can not only be located but their significance also tested. Finally, focused tests (Cuzick and Edwards, 1990; Besag and Newell, 1991; Diggle et al., 1999) are used when a pre-specified focus is supposed to be related to disease incidence. Kulldorff (2002) has recently proposed a general framework which a high majority of tests for spatial randomness can be put into. Using this framework, a mathematical definition can be given to each of the three classes of tests. Given that among the great variety of existing tests, many of these are identical, it was important to have criteria to distinguish the different tests. This is achieved by the general framework in which each (unique) test can be precisely identified by the definition of several elements: a set of centroids (each case and/or population location for example), areas around them (through their type, shape, size and distance), a measure to be calculated for each area (excess number of cases or likelihood function for example), a summary quantification of the measures corresponding to areas of different size but sharing the same centroid, and areas centered around different centroids (raw or weighted summation, or maximum). Moreover, this framework is able to create new tests by choosing a new combination of these elements. As mentioned by the author, not all tests proposed in the literature fit this framework, just as the test that we will present further.

Among the cluster detection methods, the spatial scan statistic (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) has become the most popular one. The aim of this method is to scan the study area using windows of a predefined shape (generally circles) and to determine the one that groups together an abnormally high number of cases using the log-likelihood ratio test. These windows (candidate zones for the

most likely cluster) represent the reduced parameter space denoted by Ω_0.

The spatial scan statistic is very powerful (Kulldorff et al., 2003; Ozonoff et al., 2005), notably when the real cluster has a circular shape. However, this method has a low power for detecting irregularly shaped clusters due to the use of circular scanning windows (the SatScan software, now widely used and freely available at the URL *http://www.satscan.org/*, is being extended to detect elliptical clusters using the general framework). Three recent works (Patil and Taillie, 2004; Duczmal and Assunção, 2004; Tango and Takahashi, 2005) on arbitrarily shaped cluster detection have been proposed in order to overcome this limitation. The basic principle of the three methods is the same: to apply the spatial scan statistic to another reduced parameter space Ω_0 that is not restricted to regularly shaped clusters. In all cases cluster significance is obtained via Monte Carlo simulations.

The method of Patil and Taillie (2004), known as the ULS scan statistic, suggests a new approach for reducing the list of candidate zones Z. For this purpose, $\Omega_0 = \Omega_{ULS}$ consists of all connected components of all upper level sets (ULS) of a piecewise constant surface defined over the tesselation by the adjusted rate (case number / cell population). Ω_{ULS} has the structure of a tree and is data-dependent, which implies recomputing it for each replicate data set when simulating null distributions. The size of Ω_{ULS} corresponds to the number of nodes in the tree and does not exceed the number of cells in the tesselation.

Duczmal and Assunção (2004) have developed another method to be called the SA scan statistic, which does not restrict the cluster to a fixed geometric shape. The parameter space Ω_0 allows the potential clusters to be any subset of adjacent areas. In order to analyse only the most promising subsets, a simulating annealing (SA) strategy is used. The basic survey routine is repeated several times (until 99% of all cells have been visited at least once) with different starting subsets: the cluster found by the Kulldorff method and several randomly chosen one-cell subsets.

Finally, the flexible spatial scan statistic of Tango and Takahashi (2005) modifies the set of the windows to be scanned by adding connected regions to the circular scanning windows. This test has higher power than the spatial scan statistic when the true cluster is non-circular. As mentionned by the authors, the flexible spatial scan statistic can only be applied to count data.

All the methods presented above have been conceived for count data, that is data in which cases are available on an aggregated level, defined by cell counts. Even if the circular based scan statistic is also applicable to case event data, it is not the case for the recent scan statistics for arbitrarily shaped clusters since the criterion for cell adjacency is not defined for this type of data. As explained in Lawson (2001), there are significant advantages and disadvantages in using case event data. On the one

hand, the relationship between an exact location and the disease aetiology may be uncertain (work-related disease, or case moving). The exact location, often given by an address, is not always available due to possible problems of confidentiality. On the other hand, this type of data provides detailed spatial information which could be lost when counts are used. In that sense, count data are an approximation of case event data. Hence, if case event data are available, the author recommends that spatial information should not be lost by aggregation into counts, and to analyse this level of resolution.

This paper deals with precise events within $\mathbb{R}^2$, such as spatial coordinates for the occurrence of disease cases or the geographical positions of individuals. A new method of detection and inference for multiple spatial clusters is thus presented. This approach, based on transformation of the data set and a regression model, is an extension of the method presented in Molinari et al. (2001) for multiple temporal clusters. This new test belongs to the class of detection tests for case event data.

The following section describes the different stages of the method. It begins with data tranformation by determining a trajectory and the weighting of distance. The ordered weighted distances are then used in the cluster location and detection stages. In the third section, we apply the method to both simulated and real data and compare the results with relevant existing methods. We also present the results of a power study. The paper is concluded by a discussion.

2 Method

2.1 Data transformation

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables which denote the spatial coordinates of the occurrence of n events in A, a bounded set of $\mathbb{R}^2$. Initial data are constituted by the n point coordinates. The set of the n points is included into the underlying population of size N.

We introduce two variables constructed from initial data and labelled "distance" and "order". The first represents the distance from one point to its nearest neighbour, once pre-selected points have been taken into account. The "Order" variable can be seen as an order of selection for the points. It defines a trajectory through A and will allow us to establish the regression of distances on this order. The null hypothesis is the uniform distribution of points in A. Under H_0, no cluster is detected. The general idea of the method is based on the assumption that points included in a cluster have consecutive selection orders and that their associated distances are lower than those

of points outside the cluster (because the density of points is higher within the cluster).

An example with simulated data is presented in the Figure 1(a). In $A = [0, 100]^2$, a 70-point sample is simulated following a mixture of three uniform point processes $\frac{5}{7} \times \mathcal{U}([0, 100]^2) + \frac{1}{7} \times \mathcal{U}([10, 30] \times [60, 75]) + \frac{1}{7} \times \mathcal{U}([65, 80] \times [20, 40])$. The underlying population regroups $N = 1000$ individuals and has been simulated uniformly (homogeneity situation). In cluster simulation zones, outlined by dotted lines, the density is about five times higher than the density in the whole study area.

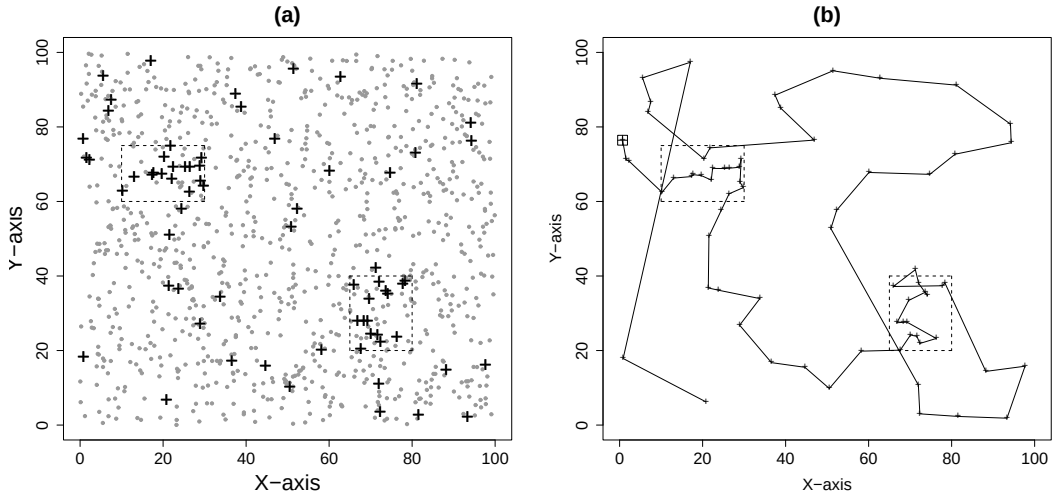

Fig. 1. **(a)** Simulated data ($n = 70$). Cases are represented by crosses, and population individuals by grey points. **(b)** Trajectory followed according to the selection order of points. The square point is the first selected. The rectangular areas in dotted lines represent cluster simulation zones.

2.2 Trajectory

For $k = 1, \ldots, n$, let be x_k a realization of X_k and $x_{(k)}$ be the point to which the selection order k is attributed. The order variable is constructed using a recursive algorithm initiated from the first order point $x_{(1)}$. Point $x_{(k+1)}$ is determined by the knowledge of points $x_{(1)}, \ldots, x_{(k)}$.

The choice of $x_{(1)}$ is arbitrary: we decided to take the point nearest to the study area border. This choice will be debated in section 2.7. Then, given $x_{(1)}, \ldots, x_{(k)}$, point $x_{(k+1)}$ is the nearest point from $x_{(k)}$ among the $n - k$ points not yet selected.

A trajectory that successively links each point to the next order point as shown in Figure 1(b) is thus defined.

2.3 Distance weighting

The elimination process of pre-selected points decreases the number of potential candidate points in the search for the nearest neighbour. Thus, the distance to the nearest neighbour observed is greater for the points selected later. Distance weighting is therefore necessary. As we will see, this weighting also makes it possible to adjust for inhomogeneity in the underlying population density. The expected distance from one point to its nearest neighbour under the uniform distribution hypothesis is given in Bickel and Breiman (1983). In what follows, we have adapted this reasoning to our particular case.

Let $x_{(1)}, \ldots, x_{(n)}$ be a realization of $X_{(1)}, \ldots, X_{(n)}$. These n points are sampled independently from an underlying density $h(x)$. As previously defined, $x_{(k)}$ is the k-order point. For $k = 1, \ldots, n-1$, let us define $D_k = d(X_{(k)}, X_{(k+1)})$ the distance from $X_{(k)}$ to $X_{(k+1)}$, with a density function g_k and distribution function G_k. Thus, $d_k = d(x_{(k)}, x_{(k+1)})$, a realization of D_k, is the distance observed from $x_{(k)}$ to $x_{(k+1)}$. The weighted distance is then defined as the ratio of the distance observed and its expectation under the hypothesis of uniform distribution :

$$d_k^w = d_k \times E_{H_0}\left[D_k \mid X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right]^{-1}. \tag{1}$$

A weighted distance greater (respectively less) than 1 means that the distance observed is greater (respectively less) than its expectation. This means that the hypothesis of uniform distribution will not be rejected if the weighted distance is statistically close to 1.

Given that D_k, with density and distribution functions g_k and G_k, is positive and bounded (since A is bounded), an integration by parts allows us to write the expectation above as

$$\int_0^a r g_k(r) dr = \int_0^a \left(1 - G_k(r)\right) dr,$$

in which

$$a = \max_{(u,v)\in A^2} d(u, v)$$

is the diameter of A . Hence

$$\begin{aligned} &E\left[D_k \mid X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right] \\ &= \int_0^a P\left(D_k > r \mid X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right) dr. \end{aligned} \tag{2}$$

Let $S(x, r)$ be the sphere with its center at x and radius r. The set $\{D_k > r \mid X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\}$ is equal to the event that none of $X_{(k+1)}, \ldots, X_{(n)}$ fall within

$S(x_{(k)}, r)$. Hence

$$\begin{aligned} & P\left(D_k > r \mid X_{(1)} = x_{(1)}, \ldots, X_{(k)} = x_{(k)}\right) \\ &= \prod_{i=k+1}^{n} \left[1 - P\left(x_{(i)} \in A_{k-1} \bigcap S(x_{(k)}, r) \mid x_{(i)} \in A_{k-1}\right)\right] \\ &= \left[1 - \frac{\int_{A_{k-1} \bigcap S(x_{(k)}, r)} h(x)dx}{\int_{A_{k-1}} h(x)dx}\right]^{n-k}, \end{aligned} \tag{3}$$

in which

$$A_k = A \setminus \left\{\bigcup_{i=1}^{k} S(x_{(i)}, d_i)\right\}$$

is the whole study area deprived of the trajectory already made as far as the k-order point. By convention, $A_0 = A$.

d_k^w is thus given by (1), (2) and (3). Its real computation involves numerical approximations. We decided to use the trapezoidal rule for the numerical integration in (2). The density integrals in (3) can be estimated using the underlying population. Let Z be this population constituted by N individuals $\{z_i : i = 1, \ldots, N\}$ in which $N \gg n$. For any set $B \subset A$, $\int_B h(x)dx$ can be approximated by $\#\{i/z_i \in B\}/N$. Hence, a potential inhomogeneity in the underlying population density will be taken into account in the computation of d_k^w.

Figure 2 illustrates the weighting computation for $k = 21$. The trajectory already made as far as $x_{(20)}$, $\{\cup_{i=1}^{20} S(x_{(i)}, d_i)\} \cap A$, is represented in white. The grey area (including the shaded portion of the disc) is A_{20} and the shaded portion of the disc is $A_{20} \cap S(x_{(21)}, r)$. Thus,

$$P\left(D_{21} > r \mid X_{(1)} = x_{(1)}, \ldots, X_{(21)} = x_{(21)}\right) = \left[1 - \frac{\#\{i/z_i \in \text{shaded area}\}}{\#\{i/z_i \in \text{grey area}\}}\right]^{n-21},$$

and $E\left[D_{(21)} \mid X_{(1)} = x_{(1)}, \ldots, X_{(21)} = x_{(21)}\right]$ can be obtained by computing the last quantity for a discrete set of values for r from 0 to a, using the trapezoidal rule.

The ordered weighted distances $\{d_k^w : k = 1, \ldots, n-1\}$ obtained now make it possible to locate potential clusters, while adjusting for underlying inhomogeneity.

2.4 Cluster location

Consider the data set $(k, d_k^w)_{k=1,\ldots,T}$ in which $T = n - 1$. In order to determine potential cluster bounds, we took the weighted distance regression on the selection

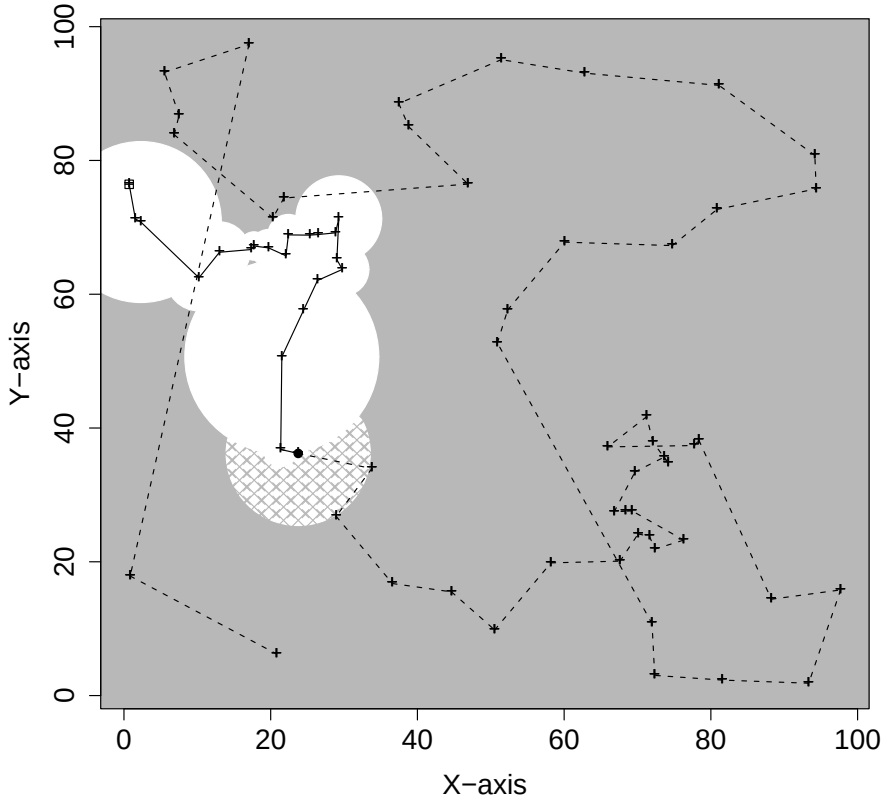

Fig. 2. Illustration of the weighting computation. The round black point is $x_{(21)} \approx (24; 36)$. The shaded disc portion, $A_{20} \cap S(x_{(21)}, r)$, is represented in the particular case of $r = d_{21}$.

order. Under the no cluster hypothesis, an appropriate regression function to use would be the constant one

$$f(t) = \overline{d} = \frac{1}{T} \sum_{k=1}^{T} d_k^w \text{ for } t = 1, \ldots, T.$$

In what follows, the values of the variable t are $1, \ldots, T$. If one cluster is present, a constant regression function with 2 breaks would be more appropriate and

$$f(t) = \overline{d}_{[1;T_1]} \times I_{[1;T_1]}(t) + \overline{d}_{[T_1+1;T_2]} \times I_{[T_1+1;T_2]}(t) + \overline{d}_{[T_2+1;T]} \times I_{[T_2+1;T]}(t) \ ,$$

in which T_1 and T_2 are the breaks, the notation $\overline{d}_{[i;j]}$ $(1 \leq i < j \leq T)$ indicates the mean of d_t^w for t in $[i; j]$, $I_{[i;j]}(t) = 1$ if $t \in [i; j]$ and 0 if not. The potential cluster is the portion between 2 breaks with the lower mean distance. Generally this will be $[T_1 + 1; T_2]$. However it may also be $[1; T_1]$ if the cluster is at the beginning of the trajectory or $[T_2 + 1; T]$ if at the end. In these cases, a one-break model would be

preferable with

$$f(t) = \overline{d}_{[1;T_1]} \times I_{[1;T_1]}(t) + \overline{d}_{[T_1+1;T]} \times I_{[T_1+1;T]}(t),$$

and the potential cluster is $[1;T_1]$ or $[T_1+1;T]$ according to the mean distance on the two portions.

More generally, to determine the presence of m breaks ($m+1$ regimes), the regression function taken into consideration is:

$$f(t) = \sum_{j=1}^{m+1} \overline{d}_{[T_{j-1}+1;T_j]} \times I_{[T_{j-1}+1;T_j]}(t)$$

with the convention $T_0 = 0$ and $T_{m+1} = T$.

Taking into account the fact that x_{k+1} is chosen among the unyet selected point sub-group, the f function will be around 1 as long as the distribution is uniform, and become distant from 1 as soon as the distribution is not uniform, that is when a cluster appears. Therefore, graphically (see Figure 3), the f staircase function will be close to 1 on the steps where there is no cluster.

In order to carry out an asymptotic analysis in the detection stage, it is necessary to impose certain restrictions on the possible values of the breaks. Indeed, at this stage, we need to test the hypothesis in the presence of parameters (break locations) which enter the model only under the alternative. This problem has been largely dealt within the literature (Davies, 1987; Andrews, 1993; Owen, 1991). Note that, in the scan statistic, the no-cluster model (same rate for all cells) is the limit of the alternative (higher rate inside a zone Z than outside) as the rate inside Z tends towards the rate outside, and the parameter Z is not identifiable within the limit. However the problem is avoided in this particular case since the statistical distribution is approximated using simulations. Recently, the method of Bai and Perron (1998) has taken this problem into account for modelling multiple structural changes. In particular, each break must be asymptotically distinct and different from the sample boundaries. The set of possible partitions is defined as follows: for some arbitrary positive number $\epsilon \in [0;1]$, $\Delta_\epsilon = \{(T_1, \ldots, T_m)\ ;\ \forall i = 1, \ldots, m+1,\ card\,([T_{i-1}+1;T_i]) \geq |T\epsilon|\}$. For example, an ϵ of 0.1 means that the number of points between two breaks is imposed as being at least 10% of the total number of points.

Breaks (cluster bounds) are estimated by resolving the constrained least square problem

$$\min_{(T_1,\ldots,T_m)\in\Delta_\epsilon} \sum_{t=1}^{T} \left(d_t^w - f(t)\right)^2 \ .$$

We note $(\hat{T}_1, \ldots, \hat{T}_m)$ the solution.

A method, based on dynamic algorithm programming for computing these estimates efficiently is presented by Bai and Perron (2003a).

In order to visualize the cluster zone, we then determined a cluster wrap as follows: we surrounded each point located within the cluster by a disc with a population inside the disc equal to the total population divided by the sample size (the number of cases). This disc can be seen as the influence zone of a point in the uniform case. If the population density is not homogeneous, the radius is not the same for all the discs since it depends on the population density around each case. The wrap is then defined as the union of all the discs. Thus, all points included in the wrap can be considered as part of the cluster.

Another way to build a wrap would be to use Voronoï tesselation, a description of which is given by Allard and Fraley (1997). This natural division of the space into disjointed areas allows us to define an area of influence for each point. Such defined areas are data-dependent and not obtained in the uniform distribution hypothesis. This can lead to large areas for points on the edges of the cluster.

These two wrap definitions can be seen as complementary. One can choose between the two or take their intersection. For the example illustrating the method and for the first point influence simulation, the disc-based definition was used.

Figure 3 displays the results obtained using models with one, four and eight breaks with the same data as the example shown in Figure 1. The union of the grey discs represents the cluster wrap. The radiuses of discs are almost the same since the underlying population is homogeneous.

2.5 Model selection

We must first consider the no break test versus a fixed number $m = k$ of breaks. The test statistic proposed by Bai and Perron (1998) is

$$F_T(\hat{T}_1, \ldots, \hat{T}_k) = \left(\frac{T - (k+1)}{k} \right) \hat{\delta} R' (R \hat{V}(\hat{\delta}) R')^{-1} R \hat{\delta} = \sup_{(T_1, \ldots, T_k) \in \Delta_\epsilon} F_T(T_1, \ldots, T_k)$$

in which $\hat{\delta} = (\hat{\delta}_1, \ldots, \hat{\delta}_{k+1})' = (\overline{d}_{[1;\hat{T}_1]}, \ldots, \overline{d}_{[\hat{T}_k+1;T]})'$ and R is a $k \times (k+1)$ matrix so that $R\hat{\delta} = (\hat{\delta}_1 - \hat{\delta}_2, \ldots, \hat{\delta}_k - \hat{\delta}_{k+1})'$. $\hat{V}(\hat{\delta})$ is an estimate of the $(k+1) \times (k+1)$ variance covariance matrix of $\hat{\delta}$:

$$\hat{V}(\hat{\delta})[i,i] = \hat{\sigma}_i^2 \frac{T}{\hat{T}_{i+1} - \hat{T}_i} \text{ for } i = 1, \ldots, k+1 \text{ and } \hat{V}(\hat{\delta})[i,j] = 0 \text{ for } i \neq j \ .$$

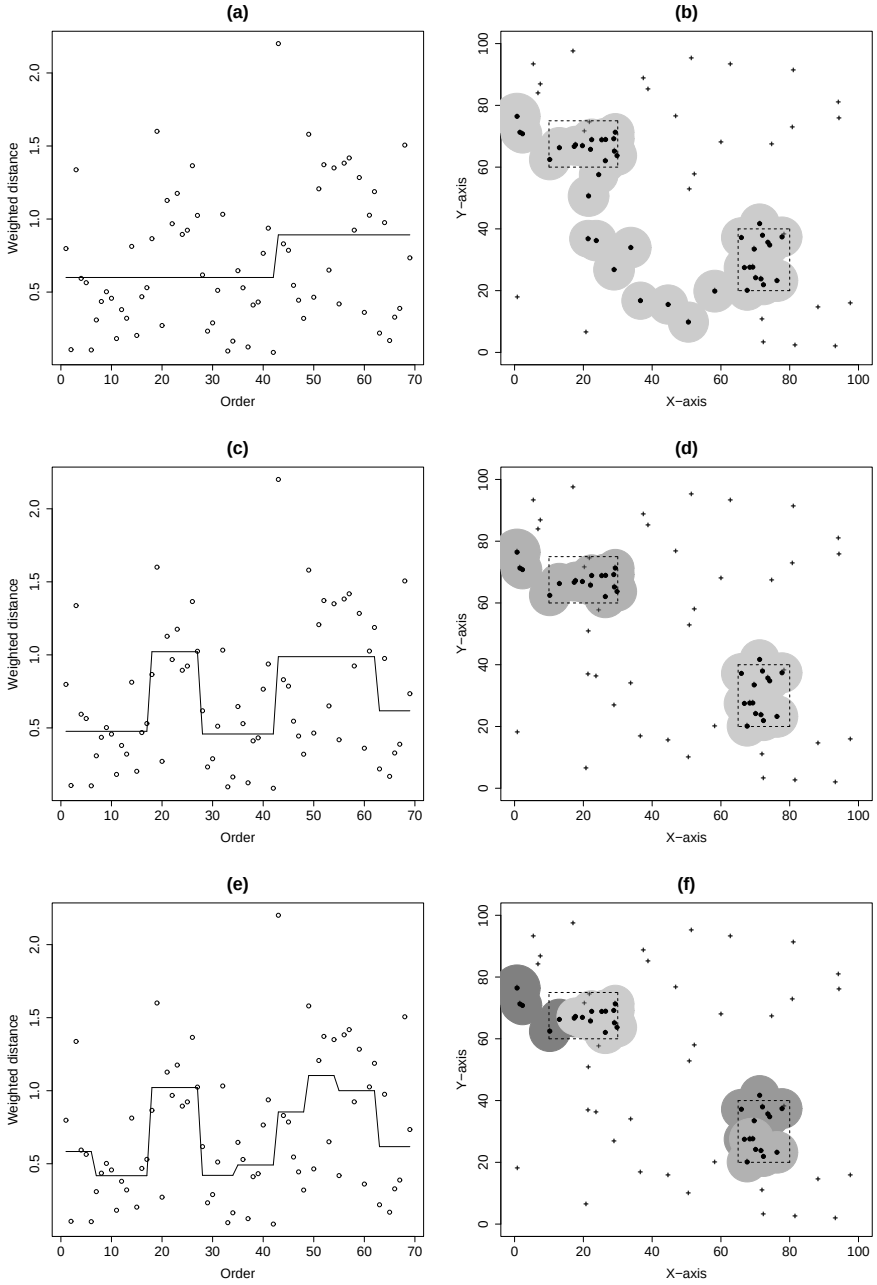

Fig. 3. Results for three models on Figure 1 simulated data: regression of distance on the order and representation of the cluster(s) located by **(a)** and **(b)** the one-cluster model with one break, **(c)** and **(d)** the two-cluster model with 4 breaks and **(e)** and **(f)** the two-cluster model with 8 breaks. Points located in the cluster(s) are round points, surrounded by a grey disc. Grey levels are used for the different portions, if necessary.

$$\hat{\sigma}_i^2 = \frac{1}{\hat{T}_{i+1} - \hat{T}_i} \sum_{j=\hat{T}_i+1}^{\hat{T}_{i+1}} \left(d_j^w - \overline{d}_{[\hat{T}_i+1;\hat{T}_{i+1}]} \right)^2 \text{ for } i = 1, \ldots, k+1.$$

F_T is the statistic for testing $\hat{\delta}_1 = \ldots = \hat{\delta}_{k+1}$ against $\hat{\delta}_i \neq \hat{\delta}_{i+1}$ for a certain i. A high value of F_T means a shift away from no break hypothesis. Critical values for this test statistic are given by Bai and Perron (2003b) for values of ϵ between 0.05 and 0.25.

In the example in Figure 3, the F values are 6.39, 9.71 and 5.60 for the 1, 4 and 8 break models respectively. Only the first one is not significant since the critical values are 9.1, 6.84 and 3.58 respectively. The models with 3, 5, 6 and 7 breaks are also significant (values not displayed). Thus the 1-break model represented in Figures 3(a) and 3(b) is not significant whereas the 4 and 8-break models in Figures 3(c) and 3(d) and Figures 3(e) and 3(f) are significant.

The best model must now be selected and the number of breaks determined, whilst taking into account the multiple testing problem. The double maximum test, defined in Bai and Perron (1998), allows us to test the null hypothesis of no break against an unknown number of breaks given a certain upper bound M. Let $c(\alpha, m)$ denote the asymptotic critical value of the test $F_T(\hat{T}_1, \ldots, \hat{T}_m)$ for a significance level α. The test is denoted:

$$WD\ max\ F_T(M) = \max_{1 \leq m \leq M} \frac{c(\alpha, 1)}{c(\alpha, m)} F_T(T_1, \ldots, T_m)\ .$$

Critical values for this corrected test statistic are given by Bai and Perron (2003b) for values of ϵ between 0.05 and 0.25 and $M \leq 9$. As ϵ represents the minimum size on potential clusters, the choice of its value depends on clinical problems. The speciality should help the statisticians to choose ϵ because the minimum cluster size has a specific interpretation. Moreover, we recommend not to choose a value greater than 0.2 in order to obtain a type I error rate near from 0.05. In what follows, we chose $\epsilon = 0.1$. With this value, our simulations gave a type I error rate of 0.062.

The number of breaks is then chosen as the $argmax$ of the $WD\ max$ statistic.

In the example in Figure 3, the $argmax$ is 8 breaks. The value of the statistic is 16.60 and is greater than the critical value for $M = 8$ (10.39). The 8-break model, corresponding to 2 clusters that each groups together 2 portions, is thus significant.

2.6 Cluster detection

The best selected model contains one or several portions (potential clusters). If the best model has a significant $WD\ max$ statisticic, the detection step consists in computing a p-value associated with each portions. To achieve this, we compute the case density in the wrap of each portion. This density is the ratio number of cases in the

wrap / number of population indivuals in the wrap. We also simulate 9999 density values under the no-cluster hypothesis (for each of the 9999 samples, we retain the higher density among the densities of the portions of the model selected) and determine, for each portion, the rank of the density in the ordered 9999 density vector. The p-value is then obtained by dividing this rank by 10000. If two (or more) portions have a non empty wrap intersection, the case density is computed for the union of the two (or more) wraps.

In the example presented in Figure 3, the best model contains 4 portions that must be grouped together in two distinct wraps (one in the top left corner and the other in the bottom right side corner). The wrap of the portion represented in the top left side corner (two grey levels) groups together 20 cases and 50 individuals (density of 0.40) and $p = 0.0003$. The wrap in the bottom right side corner (two grey levels) groups together 16 cases and 46 individuals (density of 0.35) and $p = 0.0032$. In this example, the two simulated clusters are significant.

2.7 First point influence

In order to study the influence of the first order point on the cluster location, 70-point data with a 20-point cluster was simulated. The underlying population was a 32×32 regular grid. For each of the 70 points, the cluster location method was applied, with each point as the first point of the trajectory. Only significant clusters with the WD max test were retained. The WD max test was not significant only twice out of the 70 times. Thus, we counted the number of times where the grid points fell into one of the 68 significant cluster wraps. Figure 4 shows the results of this study and confirms the robustness of the method concerning the choice of the first order point.

The method can thus be applied by choosing an arbitrary point as first point of the trajectory. A determination rule of such a point can be by example the closest point from the study area bounds, or the closest point from one of the corners.

3 Results

3.1 Power study and simulations

We simulated data samples of 100 points with different situations for clustering. Two cluster simulation zones are defined $c_1 = [20; 60] \times [75; 85]$ and $c_2 = [70; 80] \times [20; 60]$.

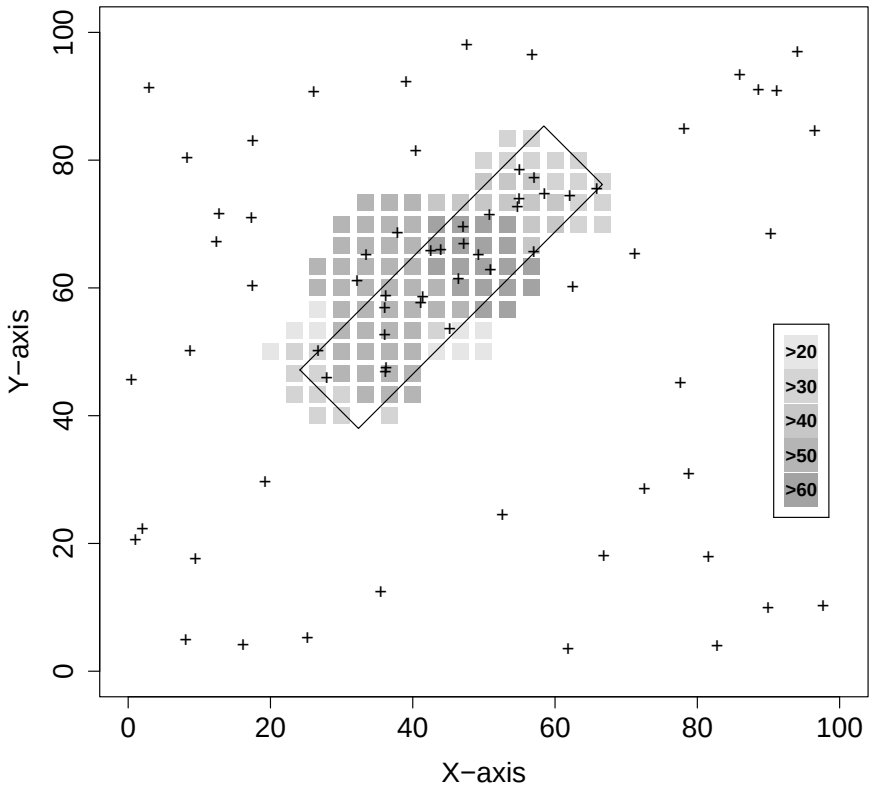

Fig. 4. Illustration of the influence on first order point choice on simulated data ($n = 70$) following a mixture of two uniform point processes $\frac{5}{7} \times \mathcal{U}([0, 100]^2) + \frac{2}{7} \times \mathcal{U}(C)$ in which C is the rectangle. Each of the n points were taken successively as the first order point. The number of times that a point is located in a significant cluster is represented by square points in grey levels. For exemple, "> 40" means that points of this color were located in the cluster between 41 and 50 times on the 70 modelizations.

c_0 denotes the whole study area ($[0; 100] \times [0; 100]$) deprived of $c_1 \cup c_2$. For $i = 0, 1, 2$, γ_i denotes the case density in c_i, that is to say the ratio number of cases in c_i / number of population individuals in c_i. The different clustering situations are: no cluster ($\gamma_0 = \gamma_1 = \gamma_2$), one cluster simulated in c_1 with a case density k times higher inside c_1, with k successively equal to 3, 6 and 10 ($\gamma_1 = k \times \gamma_0$ and $\gamma_2 = \gamma_0$), and two clusters simulated in c_1 and c_2 with a case density 6 times higher inside c_1 and c_2 ($\gamma_1 = \gamma_2 = 6 \times \gamma_0$). The underlying population was a 32×32 regular grid. 1000 samples were simulated for the different cluterization situations. For each sample, the method was applied, and the double maximum test statistic for $M = 8$ and $\epsilon = 0.1$, the density and the p-value for each portion were computed. Then, for each of the 5 simulation situations, we computed the number of classification errors (false positive and false negative) and proportion of samples in which one or more clusters were significant ($p < 0.05$). The number of errors was counted both in the total number

of replicates and on the samples with significant cluster detected. We also applied the spatial scan statistic in order to determine the same quantities and compare the two methods.

In the no-cluster situation, the results gave a type I error rate of 6.2% for our method and 6% for the spatial scan statistic. The number of classifications errors can not be computed in this case since no cluster simulation zone is defined. The results for the other clustering situations are presented in Table 1. The spatial scan statistic is more powerfull than our method. However, the spatial scan statistic locates the simulated cluster less accurately than our method since the cluster simulation zone is lengthened. The circular based statistic can not fit the lengthened rectangular zone. Indeed, globally, the number of errors is higher with the scan statistic. This is due to a twice higher number of false positives. The number of false negatives is slightly lower with the scan statistic. Those results indicate that the circular window maximizing the likelihood contains the whole rectangular simulation zone, which implies an high number of false positives and a low number of false negatives. Moreover, the difference of number of errors between the two methods is stressed when computed on significant clusters. This means that when our method detects a significant cluster, it locates it accurately.

In order to illustrate the flexibility of our method, we simulated a 70-point samples with an "L"-shaped 30-point cluster. The cluster simulation zone represents 6% of the total area. We attributed to the sample a 27×27 regular grid to represent the underlying population. The cluster location result is presented in Figure 5 with both kinds of wraps. The $WD\ max$ statistic value was 25.98, greater than the critical value for $M = 8$ and $\epsilon = 0.1$. The no-cluster hypothesis was rejected and the model with 8 breaks was selected. The four portions are significant ($p < 0.05$) and are grouped together to form a "L"-shaped located cluster wich is also significant ($p = 0.0002$). The Kulldorff circle represented in the figure is significant ($p < 0.001$). Note that a large quarter of the Kulldorff circle was empty of points which illustrates that the method is not adapted to the location of clusters with a shape very different from a circle.

3.2 Pharmacy clusters in Montpellier

One application was to check the uniform distribution of pharmacies in Montpellier, France. Indeed, the geographic location of pharmacies is supposed to depend on the surrounding population.

The 99 pharmacies in Montpellier were located using the global positioning system

Relative case density $\frac{\gamma_1}{\gamma_0}$ in c_1	1	3	6	10	6	
Relative case density $\frac{\gamma_2}{\gamma_0}$ in c_2	1	1	1	1	6	
Power						
Us	0.062	0.376	0.908	0.987	0.936	
Kulldorff	0.060	0.485	0.995	1	/	
Errors						
Us	/	11.671	10.479	9.131	11.391	11.166
Kulldorff	/	15.377	11.321	10.377	/	
False positive						
Us	/	1.206	3.297	3.112	2.926	3.093
Kulldorff	/	6.626	7.142	6.541	/	
False negative						
Us	/	10.466	7.182	6.019	8.465	8.073
Kulldorff	/	8.751	4.419	3.836	/	
Errors on significant clusters						
Us	/	7.785	8.774	8.834	10.479	10.359
Kulldorff	/	17.035	11.229	10.370	/	
False positive on significant clusters						
Us	/	3.205	3.631	3.153	3.126	3.304
Kulldorff	/	13.662	7.178	6.541	/	
False negative on significant clusters						
Us	/	4.580	5.143	5.681	7.353	7.054
Kulldorff	/	3.373	4.051	3.836	/	

Table 1
Results for the power study.

(GPS). In order to take the varying underlying population density into account, we used the 30 IRIS Montpellier divisional system. IRIS were defined by the INSEE (National Institute for Statistics and Economic Studies) in 2000. The Montpellier population as defined by IRIS was taken from the French 1999 population census. The population density in each IRIS is represented in Figure 6 (a). In concern for

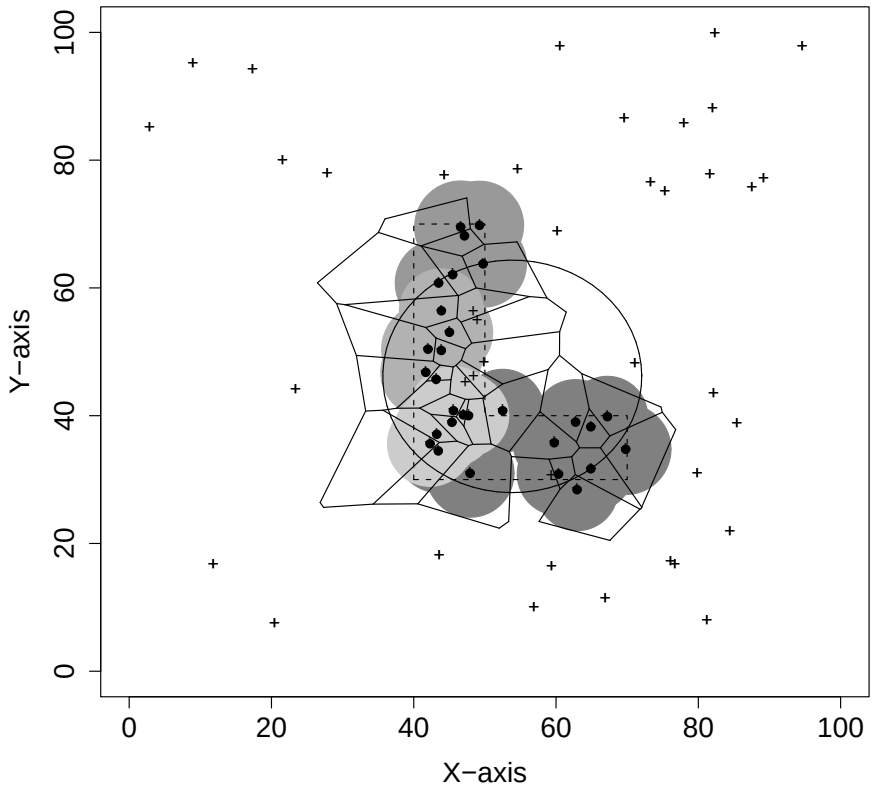

Fig. 5. 70-point sample with an "L"-shaped 30-point cluster. The cluster simulation zones are represented by dotted lines. Points located in the clusters (8-break model) are represented by a black point, surrounded by a grey disc and a Voronoi polygon. The union of grey discs represents the disc-based cluster wrap. The union of polygons represents the Voronoi-based cluster wrap. The disc represents the most likely cluster located by the spatial scan statistic.

clarity of result presentation, we wrote the identification number of each IRIS in its centroid. The pharmacy distribution is represented in the Figure 6 (b). The rate of pharmacies in the whole study area is 0.44 per 1000 inhabitants.

The case data (pharmacies) and the background data (population) are not at the same aggregation level. The typical approach is to aggregate both data to the same level as we will do for SA and ULS scan statistics. In order to apply our method, we built the underlying population by simulating a uniform point process in each IRIS with a size proportional to the IRIS population. Our method detected a significant cluster of pharmacies in the town center, shown in the Figure 6 (c) in dark grey. The 2-break (one cluster) model was selected, the $WD\ max$ statistic value was 40.5 and $p = 0.0002$. The disc-based wrap of this cluster groups together 14 pharmacies for a population of 7000 individuals (rate of 2 pharmacies for 1000 individuals). This result

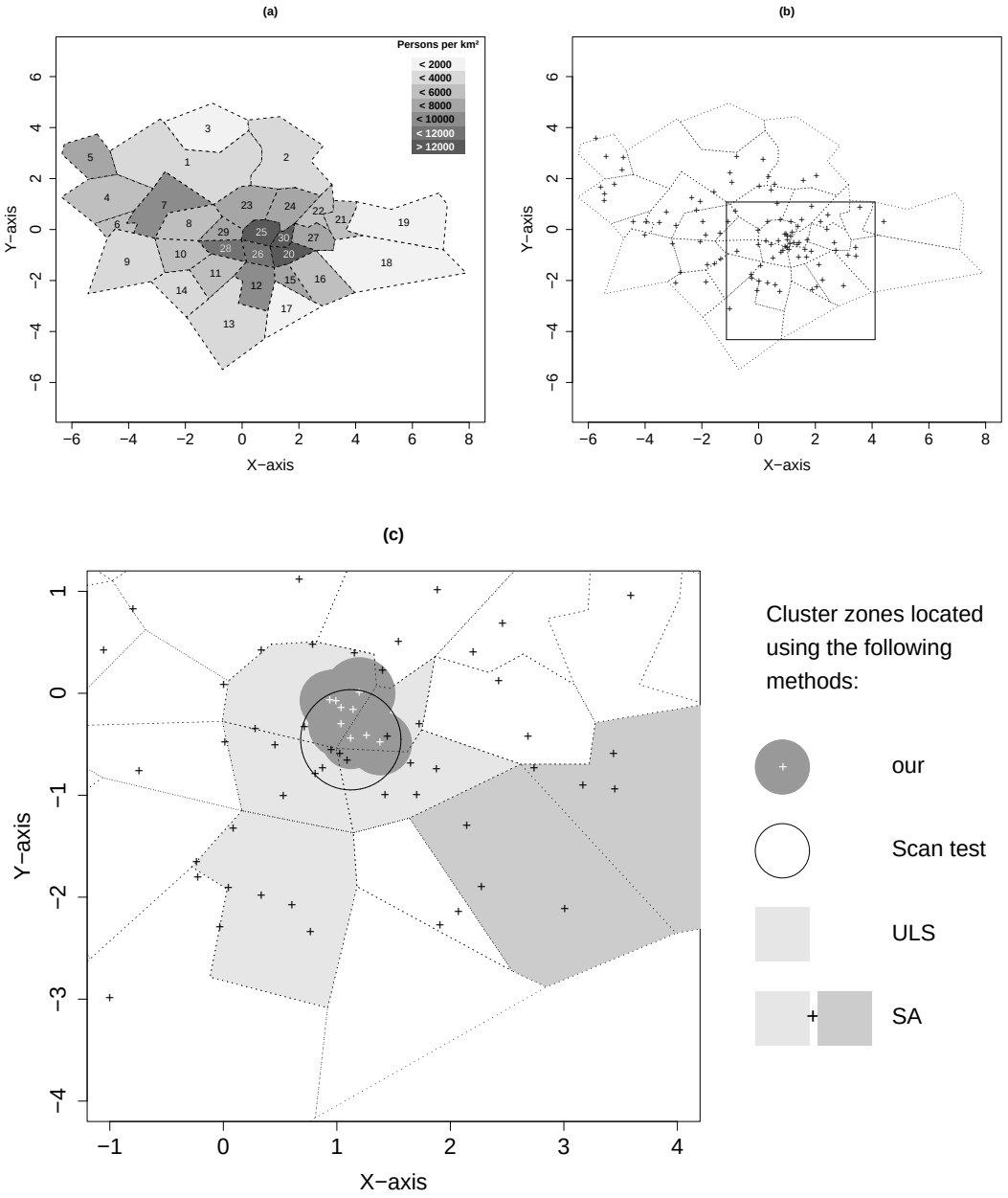

Fig. 6. **(a)** Density map of Montpellier, France. Density in each IRIS is represented by grey levels. For each IRIS, the identification number is written in its centroid. **(b)** Pharmacy distribution in Montpellier. Each cross represents a pharmacy. **(c)** Enlargement of the squared zone in figure (b) with the cluster zones located using the different methods. Axis unit is the kilometer.

can be explained by the huge number of people in the center during the day, which considerably increases the population in this area and necessitates the presence of more pharmacies than in the outskirts. Note also that the radiuses of the wrap discs are almost the same since the underlying population is homogeneous in the 2 IRIS

(25 and 30) that include the pharmacies located in the cluster.

We also applied the ULS scan statistic (Patil and Taillie, 2004) and the SA method (Duczmal and Assunção, 2004). As point data are not adapted to these methods, we computed the number of pharmacies in each cell (IRIS) in order to obtain grouped data. For these two methods, the criterion for adjacency between two cells was "their common boundary has a positive length".

For the ULS method, since no reliable software is available for this programming-intensive method, we applied the tree-based procedure for locating the most likely cluster for the data only, and not for the replicates (Ω_{ULS} must be recalculated for each replicate and efficient algorithms are needed for this calculation). Hence, we cannot provide the p-value for the most likely cluster found on the data. The most likely cluster is shown in Figure 6 (c) in pale grey (it is composed of the IRIS numbers 12, 20, 25, 26 and 30). It groups together 33 pharmacies for a population of 38800 individuals (rate of 0.85/1000). Note that this most likely cluster is the same as the one detected by the spatial scan statistic applied to count data. The latter is significant ($p = 0.007$).

For the SA method, we used the algorithm implemented in C++ code obtained from the authors. The most likely cluster is shown in Figure 6 (c) in two shades of pale grey (it is composed of the IRIS numbers 12, 20, 25, 26, 30, 16 and 18 - this last IRIS is shown only partially). It corresponds to the ULS one plus two cells on the right, and groups together 40 pharmacies for a population of 51600 individuals (rate of 0.78/1000). This most likely cluster is not significant ($p = 0.2$, obtained with 999 Monte Carlo replications). This non significant result is an illustration of the relative power of the SA and Kulldorff methods when a real cluster has a circular shape, which seems to be the case here. This issue is mentionned by Duczmal and Assunção (2004) in their conclusion.

Finally, we chose to apply the circular based spatial scan to case event data. The Poisson model was used with the simulated underlying population. The cluster located by our method seems to have a circular shape which justifies the use of this method. The most likely cluster located by the Kulldorff method is represented by the disc in Figure 6 (c). This cluster is significant ($p = 0.002$) and groups together 16 pharmacies for a population of 5400 individuals (rate of 2.97/1000).

This example allows us to compare methods for individual data with methods for count data. The clusters located by our method and Kulldorff's are included in the ULS and SA most likely clusters, and this is coherent since, by definition, these tests cannot detect a cluster with a lower resolution than the cells. Moreover, among the 5 or 7 cells which form respectively the ULS and SA most likely clusters, the two methods for individual data have found the zones (not corresponding to a whole cell)

with the highest rates of pharmacies. For example, our cluster includes only a small part of IRIS 25 in which the majority of pharmacies are concentrated. This example also illustrates, as one might expect, the advantage of the Kulldorff method when the real cluster has a circular shape. The rate of the circular zone is higher than the rate of our cluster (2.97 versus 2). However, this remark must be attenuated since the contour of the circular zone goes through a case - the furthest pharmacy from the disc center - whereas with our method it is impossible for the contour of the cluster wrap to pass through the cases.

3.3 Childhood leukaemia and lymphoma in North Humberside

The proposed method was applied on a "case/control" data set previously analysed by Cuzick and Edwards (1990). 62 cases of childhood leukaemia and lymphoma were diagnosed between 1974 and 1986 in North Humberside (England) and 141 controls were selected at random from entries on the birth registers. Their spatial distributions are shown in the Figure 7.

Potential clusters located by our method and the spatial scan statistic were not significant. The $WD\ max$ statistic value was 12.72, greater than the critical value for $M = 8$ and $\epsilon = 0.1$ (10.39). The 2-break model (one cluster) was selected. However this selected cluster were not significant ($p = 0.092$). The log likelihood ratio was 4.84 for the spatial scan statistic ($p = 0.676$). These most likely clusters are represented by grey areas in Figure 7. For both methods, the 203 cases and controls were used for the underlying population. The spatial scan statistic was applied using the Bernouilli model.

4 Discussion

The method presented here has the advantage of being very flexible. Firstly, it can be used to detect and locate several clusters, with no need to adjust for the multiple testing problem. Secondly, since the method does not need the definition of a predefined shape for potential clusters, the clusters detected can be of any shape.

This method has been conceived for case event data only. As noted in the introduction of this paper, with case event data, if available, detailed spatial information is not lost. Moreover, our method is free from map partition. As explained by Duczmal and Assunção (2004), the choice of count cell size can affect the results of cluster location and detection. This is illustrated in the pharmacy example. Indeed, cases are often

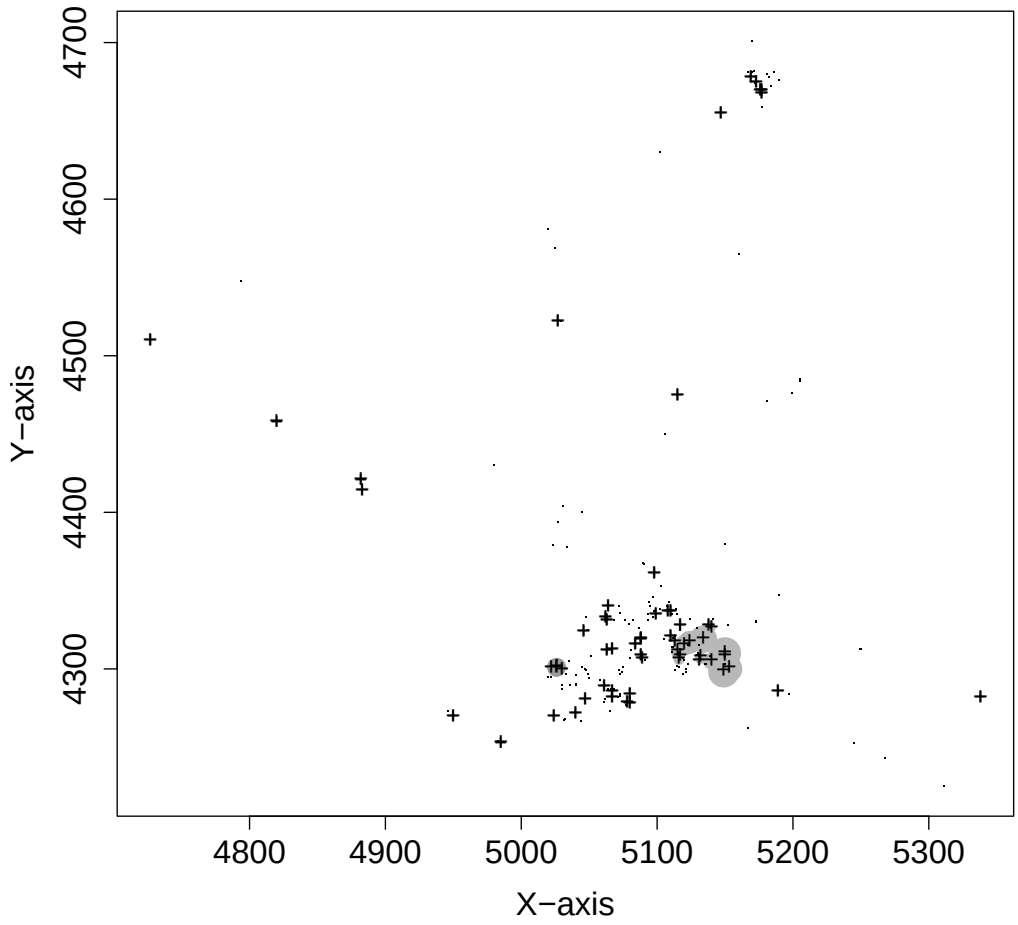

Fig. 7. Distribution of cases of childhood leukaemia and lymphoma in North Humberside. Cases are represented by a cross and controls by a point. The grey area represents the disc-based wrap of the most likely cluster located by our method. The little grey disc centered around (5020,4300) represents the most likely cluster located by the spatial scan statistic.

located near a frontier between two cells since these frontiers often correspond to a main thoroughfare in a city or to a natural frontier, such as a river, in a study on a higher scale. This can lead to ambiguities in attributing a case to a particular cell, which was noted by Turnbull et al. (1990) with regard to the leukemia incidence data in upstate New York. Thus, a slight shift or ambiguity in case location may greatly affect the cell counts and change the cluster location and detection results. Therefore we believe that the method presented here would be preferable than count data methods when case event data are available.

Although our method falls into the category of tests for spatial randomness that ad-

just for an underlying inhomogeneity, it does not fit the general framework proposed by Kulldorff (2002). The most evident departure of this method from that framework is data transformation conditionally to the trajectory (the value attributed to a point, in this case a distance ratio, depends on pre-selected points). We are aware that many tests already exist (Kulldorff, 2002). However, this method is innovative and can be seen as a way of analysing spatial data which is complementary to commonly used powerful methods.

The first limiting point is the possibility that the trajectory leaves the cluster before going through all the cluster points. This is generally a false problem. Indeed, the remaining cluster points will be detected as a second cluster and the proximity analysis of these 2 clusters by specialists could allow them to build a new bigger cluster as the union of the 2 clusters detected.

The second limiting point concerns the availability of case event data. Indeed, this type of data is not always easy to collect and count data are often preferred. In such a case, our method cannot by applied. However, as noted by Bailey (2001), health information systems are steadily improving and there is thus an increasing demand for methods that can be used with case event data.

The extension of this method to an n dimensional spatial process ($n > 2$) is immediate, by replacing the distance used here by the euclidian distance in $\mathbb{R}^n$. Once the data has been transformed, the method is roughly the same. Another extension of this method, less simple, is to adapt it to space-time cluster detection. In this issue, the main difficulty is the different roles played by the time dimension and the spatial ones. Possible applications for $\mathbb{R}^{3+1}$ are cluster detection in functional Magnetic Resonance Imaging or meteorologic data.

Another issue that should be investigated is the covariate adjustment. The method proposed here only adjusts for an underlying population inhomogeneity. The adjustment for covariates such as age or gender is not yet possible.

All the computations conducted and all the figures shown in this paper were produced using R software (version 1.9.0). For users who are interested in applying this method, we can provide upon request its implementation in an R package, named SPATCLUS.

Acknowledgements

The authors wish to express their gratitude to Martin Kulldorff for its helpful comments, to Luiz Duczmal for transmitting to us the C++ code of its algorithm, and to Teresa Sawyers for its comments that improved the presentation of this article.

References

Allard, D. and Fraley, C., 1997. Non parametric maximum likelihood estimation of features in spatial point processes using Voronoï tesselation. Journal of American Statistical Association, 92, 1485–1493.

Andrews, D., 1993. Tests for parameter instability and structural change with unknown change point. Econometrica, 61, 821–856.

Bai, J. and Perron, P., 1998. Estimating and testing linear models with multiple structural changes. Econometrica, 66, 47–78.

Bai, J. and Perron, P., 2003a. Computation and analysis of multiple structural change models. Journal of Applied Econometrics, 18, 1–22.

Bai, J. and Perron, P., 2003b. Critical values for multiple structural change tests. Econometrics Journal, 6, 72–78.

Bailey, T., 2001. Spatial statistical methods in health. Cadernos de Saúde Pública, 17, 1083–1098.

Besag, J. and Newell, J., 1991. The detection of clusters in rare diseases. Journal of the Royal Statistical Society A, 154, 143–155.

Bickel, P. and Breiman, L., 1983. Sums of functions of nearest neighbour distances, moment bounds, limit theorems and a goodness of fit test. Annals of Probability, 11, 185–214.

Cuzick, J. and Edwards, R., 1990. Spatial clustering for inhomogeneous populations. Journal of the Royal Statistical Society B, 52, 73–104.

Davies, R., 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika, 74, 33–43.

Diggle, P., Morris, S. and Morton-Jones, T., 1999. Case-control isotonic regression for investigation of elevation in risk around a point source. Statistics in Medicine, 18, 1605–1613.

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

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics - Theory and Methods, 26, 1481–1496.

Kulldorff, M., 1999. An isotonic spatial scan statistic for geographical disease surveillance. Journal of the National Institute of Public Health, 48, 94–101.

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

Kulldorff, M., 2002. Tests for spatial randomness adjusted for an inhomogeneity: A general framework. Technical report, Harvard Medical School.

Kulldorff, M., Tango, T. and Park, P.J., 2003. Power comparisons for disease clustering tests. Computational Statistics and Data Analysis, 42, 665–684.

Lawson, A., 2001. Statistical Methods in Spatial Epidemiology. Wiley, Chichester.

Molinari, N., Bonaldi, C. and Daurès, J. P., 2001. Multiple temporal cluster detection. Biometrics, 57, 577–583.

Owen, A., 1991. Comment on "multivariate adaptative regression splines" by friedman j.h. The Annals of Statistics, 19, 102–112.

Ozonoff, A., Bonetti, M., Forsberg, L. and Pagano, M., 2003. Power comparisons for an improved disease clustering test. Computational Statistics and Data Analysis, 48, 679–684.

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

Ripley, B. D., 1977. Modelling spatial patterns. Journal of the Royal Statistical Society B, 39, 172–192.

Szalay, A. S., Budavári, T., Connolly, A., Gray, J., Matsubara, T., Pope, A. and Szapudi, I., 2002. Spatial clustering of galaxies in large datasets. Technical Report MSR-TR-2002-86, Microsoft Research.

Tango, T., 1995. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statistics in Medicine, 14, 2323–2334.

Tango, T., 2000. A test for spatial disease clustering adjusted for multiple testing. Statistics in Medicine, 19, 191–204.

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

Turnbull, B. W., Iwano, E., Burnett, W., Howe, H. and Clark, L., 1990. Monitoring for clusters of disease: Application to leukemia incidence in upstate new york. American Journal of Epidemiology, 132, 136–143.

Vinson, T. and Baldry, E., 1999. The spatial clustering of child maltreatment: Are micro-social environments involved? Australian Institute of Criminology, Trends and Issues, 119.

Whittemore, A., Friend, N., Brown, B. and Holly, E., 1987. A test to detect clusters of disease. Biometrika, 74, 631–635.