Biocybernetics and Biomedical Engineering
2009, Volume 29, Number 2, pp. 19–30

Clustering and Spatial Variation in Risk

DENIS ENĂCHESCU*, CORNELIA ENĂCHESCU

Intitute for Mathematical Statistics and Applied Mathematics,
Academy of Romania, Bucharest, Romania

Motivated by recent interest in the possible spatial clustering of rare diseases, the paper presents two approaches to the assessment of spatial clustering. The first approach emphasizes estimation of the nature and physical scale of the clustering effects rather than testing for their existence. The second approach presents a scan statistic that can detect irregular shaped clusters within relatively small neighborhoods of each region. A Monte Carlo test of significance is given and the performance is examined in comparison with that of the Kulldorff's circular spatial scan statistic. An application to data on the spatial distribution of childhood leukemia and lymphoma in Nord Pas de Calais region (France) is described.

Keywords: spatial clustering, childhood leukemia, scan statistics, K-functions, complete spatial randomness, scan statistic, Monte Carlo method

1. Introduction

The question of whether disease cases are clustered in space has received considerable attention in the literature, in part prompted by increasing concerns over possible links between disease and source of environmental pollution (see, for example, [1, 2, 3] and [4]). In this paper we present two approaches to the assessment of spatial clustering based on:

- the *K-functions for labeled point processes* which quantify the departures from the hypothesis that in a realization of a stationary spatial point process consisting of events of two qualitatively different types, the disease cases are a random sample from the superposition of disease and healthy events;

* Correspondence to: Denis Enăchescu, Casa Academiei, Cal. 13 Septembrie no 13, 050711 Bucuresti, Romania, e-mail: denaches@fmi.unibuc.ro

Received 25 August 2008; Accepted 22 January 2009

- The *flexibly shaped spatial scan statistic* which test the statistical significance of the hypothesis that the observed and the expected number of cases in each region of the study area are equal.

The first approach estimates the second-moment properties of a labeled point process and the spatial clustering is assessed indirectly through the nature and physical scale of the possible clusters. The second approach detects the irregular shaped clusters by testing the null hypothesis on concentric C circles plus all the sets of connected regions (including the single current region) whose centroids are located within the C-th largest concentric circle.

The above techniques are applied to detect clusters of childhood leukemia and lymphoma in Nord-Pas-de-Calais region (France). The conclusion is that the methods are complementary and work well for small to moderate number of events and cluster sizes. For larger numbers of events and cluster sizes the methods are not feasible and more efficient algorithms are needed.

2. The Methods

2.1. K-Functions for Labeled Point Processes

Let $\mathbf{x}_i$, $i = 1,...,n_1$ denote the locations of all cases of disease in a geographical region A. A traditional starting point for the analysis of such data is to test the hypothesis of *complete spatial randomness* (CSR), whereby the $\mathbf{x}_i$ constitutes a partial realization of a homogeneous planar Poisson process (see, for example, [5]).

In an epidemiological setting, the hypothesis of CSR is implausible because of natural spatial variation in population density. A more plausible starting point is to assume an inhomogeneous Poisson process with spatially varying density $\lambda(\mathbf{x})$. In [6] is proposed to select a random sample of controls $\mathbf{x}_i$, $i = n_1 + 1,...,n_1 + n_2$ from the population at risk in A to avoid the difficulties in formulating an appropriate parametric form for $\lambda(\mathbf{x})$. Under the Poisson assumption, the cases then represent a random sample from the superposition of cases and controls. We identify this *random labeling* hypothesis H_0^K as the null hypothesis of no spatial clustering. We wish to test H_0^K and to quantify departures from H_0^K.

Note that H_0^K makes no explicit reference to an underlying Poisson process of cases or controls. Subsequently we shall assume that the superposition of cases and controls constitutes a partial realization of a stationary spatial point process.

Our description of the so-called *K-function* approach in the spatial epidemiological context follows that given in [7]. For an unlabeled stationary, isotropic point process of *events* the reduced second moment measure or the K-function is defined by

$$K(s) = \lambda^{-1}\mathrm{E}\ [\text{number of further events within distance } s \text{ from an arbitrary event}]$$

where λ is the *intensity,* or mean number of events per unit area. This definition of $K(s)$ requires an additional technical conditions which essentially preclude multiple coincident events.

For a *labeled* stationary, isotropic point process, in which the events are of qualitatively different types $j = 1, 2$ (here, cases and controls), we similarly define a set of K-functions

$$K_{ij}(s) = \lambda_j^{-1}\mathrm{E}\ [\text{number of type } j \text{ events within distance } s \text{ from an arbitrary type } i \text{ event}]$$

where λ_j is the intensity of type j events.

Under the random labeling hypothesis, the process of type j events constitutes a random thinning of the *unlabeled* point process defined as the superposition of type 1 and type 2 events. Also, it is clear from the above definitions that the K-functions are invariant under random thinning. It follows that under H_0^K,

$$K_{11}(s) = K_{22}(s) = K_{12}(s) = K_{21}(s)$$

for all s. Note that the above equality does not require any parametric assumptions about the underlying unlabeled process.

The above equality suggests that a useful way of investigating departures from H_0^K would be to assess the significance of differences amongst estimates of the three functions $K_{ij}(s)$. In particular, the difference $D(s) = K_{11}(s) - K_{22}(s)$ may be taken as a measure of the extra-clustering of the cases compared to the clustering of the controls. Thus, in the present context, significantly positive values of $D(s)$ would constitute evidence of spatial clustering of the disease in question.

For data $\mathbf{x}_i \in A$, $i = 1,...,n$ where $n = n_1 + n_2$ with the first n_1 events of type 1 and the remainder of type 2, unbiased estimators for the K_{ij} can be obtained as follows. Let $w(\mathbf{x}, s)$ be the reciprocal of the proportion of the circumference of the circle with centre $\mathbf{x}$ and radius s which lies within A. Let $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ be the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. Let $\delta_{ij}(s)$ be the indicator of the event $d_{ij} \leq s$. Put $w_{ij} = w(\mathbf{x}_i,d_{ij})$ for $j \neq i$ and $w_{ii} = 0$, then

$$\hat{K}_{11}(s) = |A|\{n_1(n_1 - 1)\}^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} w_{ij}\delta_{ij}(s),$$

$$\hat{K}_{22}(s) = |A|\{n_2(n_2 - 1)\}^{-1}\sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n} w_{ij}\delta_{ij}(s).$$

In order to yield unbiased estimators of $D(s)$ under $H_0{}^K$ the above expressions differ slightly from the usual definitions of K_{ij}. Note also that, for a convex region $A,$ the estimators are unbiased only for s less than the circumradius of $A,$ this restriction guarantees that the w_{ij} are all finite.

We are interested in the sampling distribution of the empirical function $\hat{D}(s) = \hat{K}_{11}(s) - \hat{K}_{22}(s)$. In [7] there is derived the mean and covariance structure of $\hat{K}_{ij}(s)$ under random labeling, from which it is a straightforward exercise to deduce the mean and variance of $\hat{D}(s)$.

In practice, the statistic $\hat{D}(s)$ is calculated for a range of values $s_1,...,s_m$. A plot of $\hat{D}(s)$ versus s (with tolerance limits under random labeling imposed) is useful for assessing at which distances departures from random labeling are seen. Care must be taken when such plots are interpreted since $\hat{D}(s_1)$ and $\hat{D}(s_2)$ for $s_1 \neq s_2$ are not independent. The range of distance to be examined is also, to some extent, arbitrary and likely to be important in the overall significance of D.

If a formal test of significance is required, we need somehow to combine the information from the m values $\hat{D}(s_k)$ into an appropriate test statistic, noted with D. Given a particular choice for D, we can implement an exact, albeit computationally intensive, Monte Carlo test. The Monte Carlo test consists of ranking the observed value D_1 of D amongst values $D_2,...,D_r$ obtained from $r - 1$ independent random permutations of the labels. If D_1 ranks kth largest, the exact p-value is k/r (see [8]). One sensible choice of test statistic is

$$D = \sum_{k=1}^{m} \hat{D}(s_k) / \sqrt{\operatorname{var}\{\hat{D}(s_k)\}}.$$

The approximate sampling distribution of D under H_0^K is normal, with $\mathrm{E}[D] = 0$ and

$$\operatorname{var}(D) = m + 2\sum_{j=2}^{m}\sum_{k=1}^{j-1} \operatorname{corr}\{\hat{D}(s_j), \hat{D}(s_k)\} \quad \text{(see [7])}.$$

Using the Monte Carlo test we verify the statistical significance of the null hypothesis $H_0^K : D = 0$ (no spatial clustering or inhibition) against the alternatives $H_A^K : D > 0$ (spatial clustering) or $H_{A'}^K : D < 0$ (spatial inhibition). In this paper, the P-value of the test is computed ranking $r = 19$ (i.e. a significance of 95%) Monte Carlo replications of D generated under the null hypothesis.,

2.2. The Flexibly Shaped Spatial Scan Statistic

Consider the situation where an entire study area A is divided into R regions (for example, county, enumeration districts, *et cetera*). The number of cases in the region i is denoted by the random variable N_i with observed value n_i, $i = 1,...,R$. Under the null hypothesis H_0^S of no clustering the N_i are independent Poisson variables such that

$$H_0^S : \mathrm{E}[N_i] = \xi_i, \ N_i \sim \mathrm{Po}(\xi_i), \ i = 1,...,R$$

where $\mathrm{Po}(e)$ denotes Poisson distribution with mean e and the ξ_i are the null expected number of cases in the region i. To specify the geographical position of each region, we will use the coordinates of the administrative population centroids.

Under this situation, the circular spatial scan statistic imposes a circular window **Z** on each centroid. For any of those centroids the radius of the circle varies from zero to a pre-set maximum distance d or a pre-set maximum number of regions C to be included in the cluster. If the window contains the centroid of a region, then that whole region is included in the window. In total, a very large number of different but overlapping circular windows are created, each with a different location and size, and each being a potential cluster. Let $\mathbf{Z}_{ik}$, $k = 1,..., C$ denote the window composed of the (k–1)-nearest neighbors to region i. Then all the windows to be scanned by the circular spatial scan statistic are included in the set

$$Z_1 = \{\mathbf{Z}_{ik} \,|1 \le i \le R,\ 1 \le k \le C\}.$$

The *flexibly shaped spatial scan statistic* ('flexscan' called hereafter) proposed in [9] imposes an irregularly shaped window Z on each region by connecting its adjacent regions. For any given region i the flexscan create the set of irregularly shaped windows with length k consisting of k connected regions including i and let k varying from 1 to the pre-set maximum C. In total, as in the circular spatial scan statistic, a very large number of different but overlapping arbitrarily shaped windows are created. Let $Z_{ik(j)}, j = 1,...,j_{ik}$ denote the j-th window which is a set of k regions connected starting from the region i, where j_{ik} is the number of j satisfying $\mathbf{Z}_{ik(j)} \subseteq \mathbf{Z}_{ik}$ for $k = 1,...,C$. Then all the windows to be scanned are included in the set

$$Z_2 = \{\mathbf{Z}_{ik(j)} \,|1 \le i \le R,\ 1 \le k \le C,\ 1 \le j \le j_{ik}\}.$$

In other words, for any given region i the circular spatial scan statistic consider C concentric circles, whereas the flexscan consider C concentric circles plus all the sets of connected regions (including the single region i) whose centroids are located within the C-th largest concentric circle. So, the size of Z_2 is far larger than that of Z_1 which is at most RC.

Under the alternative hypothesis, there is at least one window **Z** for which the underlying risk is higher inside the window when compared with outside. In other words, we are considering the following hypothesis:

$$H_0^S : \mathrm{E}[N(\mathbf{Z})] = \xi(\mathbf{Z}),\ \text{ for all } \mathbf{Z},$$
$$H_A^S : \mathrm{E}[N(\mathbf{Z})] > \xi(\mathbf{Z}),\ \text{ for some } \mathbf{Z}$$

where $N(\bullet)$ and $\xi(\bullet)$ denote the random number of cases and the null expected number of cases within the specified window, respectively. For each window it is possible to compute the likelihood to observe the observed number of cases within and outside the window, respectively. Under the Poisson assumption, the test statistic, which was constructed with the likelihood ratio test (see [10]), is given by

$$\sup_{\mathbf{Z}\in Z}\left\{\left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})}\right)^{n(\mathbf{Z})}\left(\frac{n(\mathbf{Z}^C)}{\xi(\mathbf{Z}^C)}\right)^{n(\mathbf{Z}^C)} I\left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})}>\frac{n(\mathbf{Z}^C)}{\xi(\mathbf{Z}^C)}\right)\right\}$$

where $\mathbf{Z}^C$ indicates all the regions outside the window $\mathbf{Z}$, $n()$ denotes the observed number of cases within the specified window and $I()$ is the indicator function. The window $\mathbf{Z}^*$ that attains the maximum likelihood is defined as the most likely cluster (MLC).

To find the distribution of the test statistic under the null hypothesis, the Monte Carlo hypothesis testing is required. In this paper, p-value of the test is based upon the null distribution of the statistic of the likelihood ratio test with a large number (we used 999) of the Monte Carlo replications of the data set generated under the null hypothesis. It should be noted that in the same manner as the circular spatial scan statistic the flexscan is also able to locate secondary clusters that do not overlap the most likely cluster but are still statistically significant.

3. Case Study: Childhood Leukemia in Nord-Pas-de-Calais

We used the above techniques in order to detect clusters of childhood (ages up to 15 years) acute leukemia cases in Nord Pas de Calais (NPC) region diagnosed in the 3 year period ending in 2003.

In this paper we consider the ‘canton’ division of the NPC area; there are, in total, 156 cantons. In the considered period, in 123 cantons, with a surface of 9,609 km^2, are distributed, 497 cases of acute childhood leukemia among a population of 573,500 children (data are provided by D.I.M. de C.H.R.U.-Lille). Hence, the mean intensity of the type 1 events (i.e. disease cases) is $\lambda_1 = 0.0517$ and the mean intensity of the type 2 events (i.e. healthy cases) is $\lambda_2 = 59.6836$. The resulting data are shown in Fig. 1.

In this study to specify the geographical position of each canton we have used the coordinates (obtained from Google Earth) of its main town (i.e. chef-lieu). In the NPC region we find out only 96 distinct chef-lieu for the 123 cantons with cases of acute childhood leukemia (because a town can be chef-lieu for more than one canton). Hence, for our purpose we consider only 96 sub-regions (some of them obtained by merging together the cantons with the same chef-lieu).

In order to apply the K-function method we generate for each of the 96 sub-regions:

– the spatial coordinates of all disease-cases of the sub-region;

– the spatial coordinates of a number of healthy-cases of the sub-region.

The total number of control-cases (i.e. healthy-cases) is 1,494; this drastic low number was chosen from computer-implementation consideration of the method. For each sub-region the number of control-cases is proportional to the healthy population.

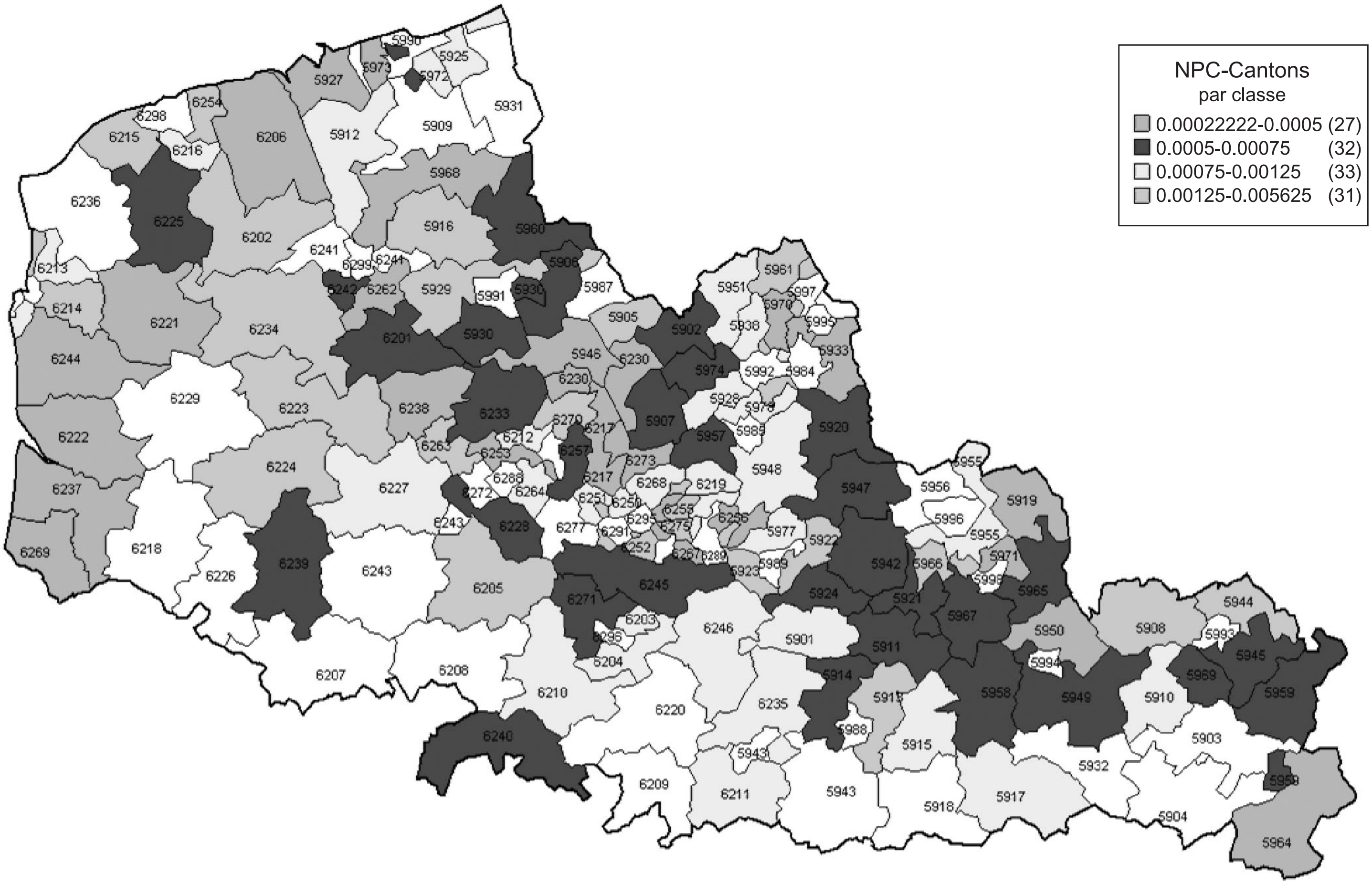

Fig. 1. The cantons of Nord Pas de Calais region illustrating the childhood leukemia diagnosed in the period 2001–2003

The spatial coordinates of an event, disease-case or healthy-case, are generated using a bivariate Gaussian distribution function with mean the geographical coordinates of the chef-lieu of the current sub-region and with variance-covariance matrix given by $\sigma_i^2\mathbf{I}$ (where **I** is the 2x2 identity matrix and σ_i is the radius/3 of the circle having the surface of the current sub-region *i*). The resulting data are given in Fig. 2.

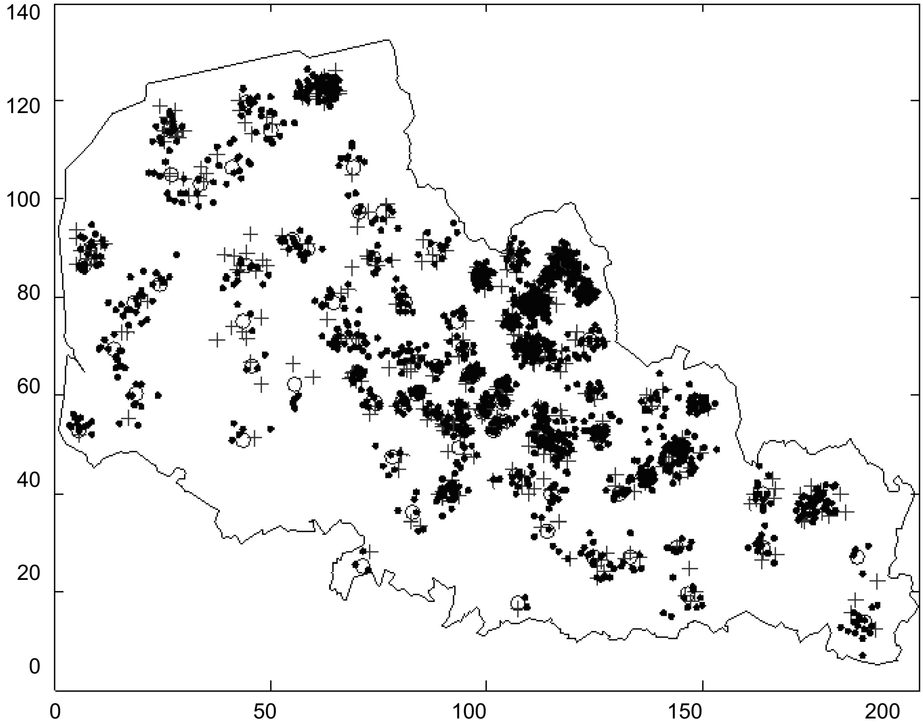

Fig. 2. Generated locations of 497 cases (+) and 1,494 controls (•) for childhood leukemia in Nord Pas de Calais for the period 2001–2003 (the 'o' are the locations of the chef-lieu of the 96 sub-regions). The units of the axes are in [km]

Table 1. The results of the Monte-Carlo tests of significance

The null hypothesis of no spatial inhibition is accepted	
the statistic 12.62233	the exact p-value 0.90
The null hypothesis of no spatial clustering is accepted	
the statistic 12.62233	the exact p-value 0.90

Figure 3 shows $\hat{D}(s)$ for the point labeled data of Fig. 2 evaluated at $s_k = 2,4,...,20$ together with approximate 95% tolerance limits $\pm 2\sqrt{\mathrm{var}\left[\hat{D}(s)\right]}$. The diagram suggests mild evidence of the spatial clustering.

The results of a formal Monte Carlo test of significance with $r = 19$ (i.e. a significance of 95%) random replications of D generated under the null hypothesis are listed in Table 1. Clearly, retrospective adjustment of m could give a more or less significant result, but we feel that this would enrich insignificantly the information conveyed by Fig. 3.

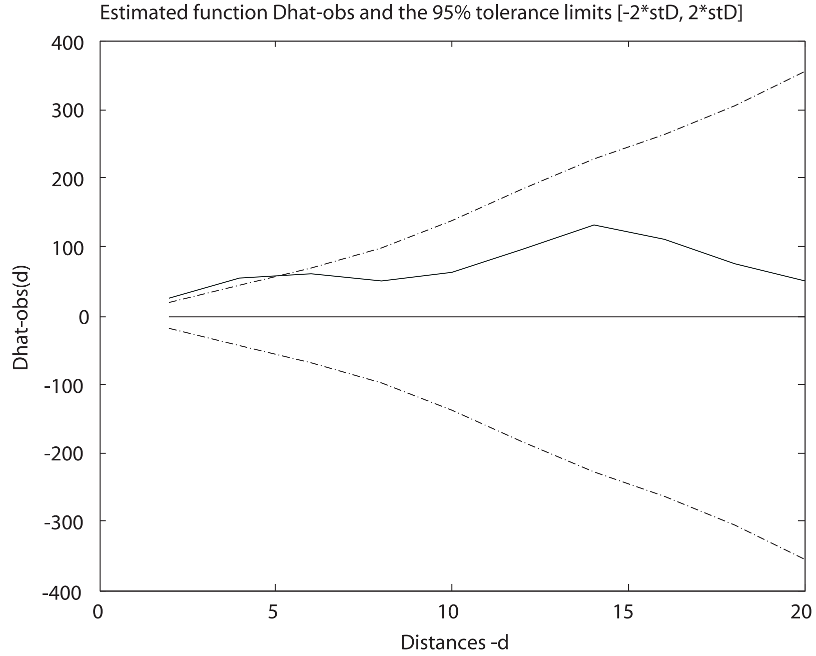

Fig. 3. K-function plot of $\hat{D}(s)$ and approximate 95% tolerance limits for D(s) = 0 (–.–). The units of the axes are in [Km]

The above numerical and graphical results are obtained in 16,800 sec. (≈ 4.66 h) using original subroutines written in Matlab 7.0 implemented on an IBM ThinkPad R40 equipped with an Intel Pentium M Processor 1.4GHz and 1Gb RAM.

The results obtained applying the flexscan and the Kulldorff's scan to the same data are synthesized comparatively in Table 2 and Fig. 4.

When the limit length of clusters is set to 20 then the flexscan method finds the <<ARDRES, CALAIS, FAUQUEMBERGUES, LUMBRES>> cluster to be the most likely cluster with a p-value of 0.019 and the <<TOURCOING>> and <<LIEVIEN>>

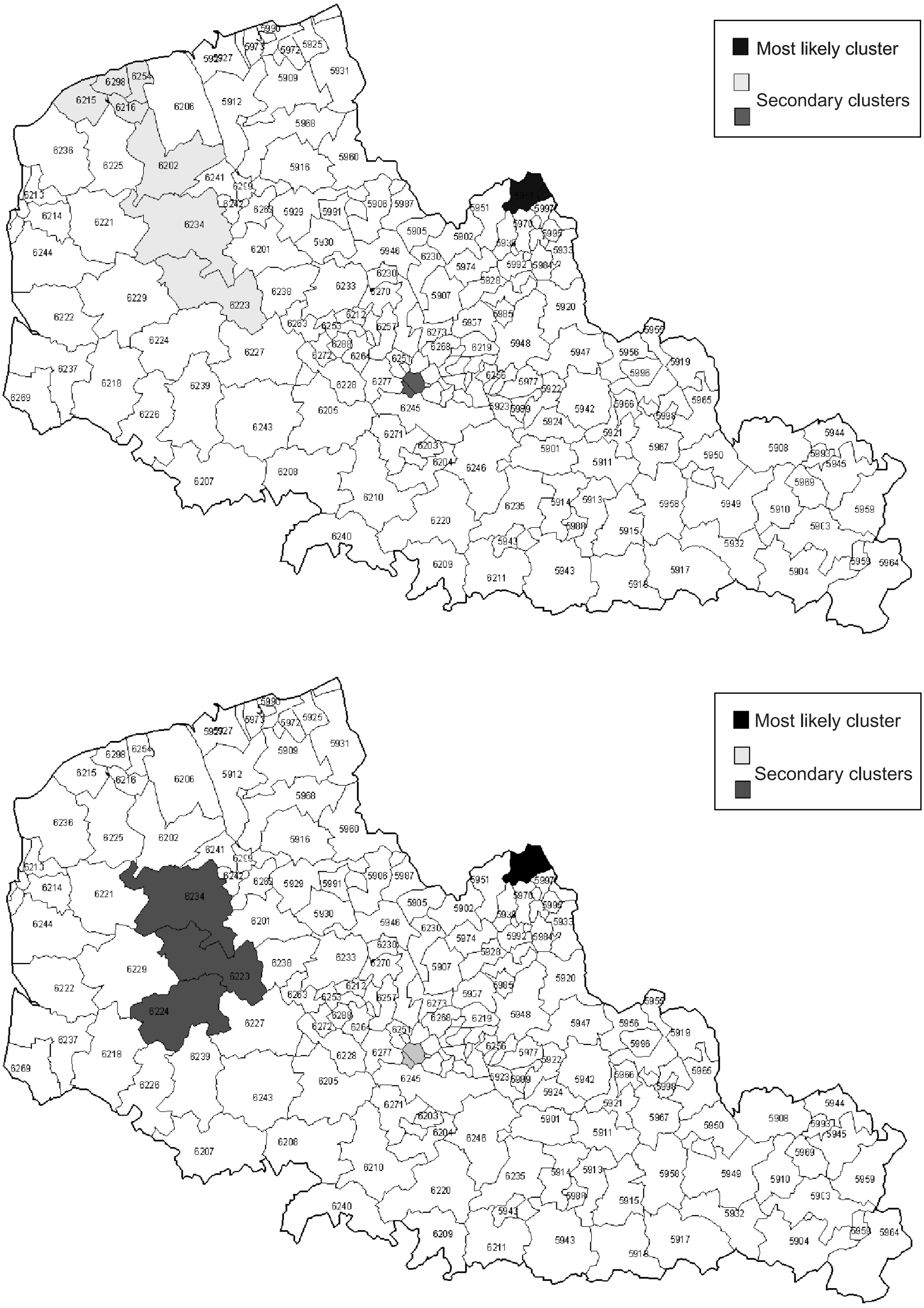

Fig. 4. The map of the 96 sub-regions the Nord-Pas de Calais region with the most likely leukemia clusters detected by flex-scan (the upper map) and Kulldorff scan (the lowert map)

Table 2. The most likely leukemia clusters detected by flex-scan and Kulldorff scan in the Nord-Pas de Calais region

Scanning method : Flex-scan	Scanning method : Kulldorff's method
Limit length of cluster: 15	Limit length of cluster: 15
MOST LIKELY CLUSTER	**MOST LIKELY CLUSTER**
TOURCOING	TOURCOING
Population ....: 14800.0	Population ....: 14800.0
Number of cases ......: 32 (12.82 expected)	Number of cases ......: 32 (12.82 expected)
Overall relative risk .: 2.49497	Overall relative risk .: 2.49497
Log likelihood ratio ..: 10.4674	Log likelihood ratio ..: 10.4674
P-value ....: 0.014	P-value ....: 0.002
SECONDARY CLUSTERS	**SECONDARY CLUSTERS**
ARDRES, CALAIS,	LIEVIN
FAUQUEMBERGUES, LUMBRES	
Population ....: 16500.0	Population ....: 3200.0
Number of cases ......: 34 (14.299 expected)	Number of cases ......: 13 (2.773 expected)
Overall relative risk .: 2.37778	Overall relative risk .: 4.68781
Log likelihood ratio .: 10.1564	Log likelihood ratio ..: 9.96426
P-value ....: 0.015	P-value ....: 0.002
LIEVIN	FAUQUEMBERGUES, FRUGES,
	LUMBRES
Population ....: 3200.0	Population ....: 7700.0
Number of cases ......: 13 (2.773 expected)	Number of cases ......: 17 (6.672 expected)
Overall relative risk .: 4.68781	Overall relative risk .: 2.54762
Log likelihood ratio .: 9.96426	Log likelihood ratio .: 5.68015
P-value ....: 0.016	P-value ....: 0.105

clusters to be secondary clusters with p-values 0.037 and 0.051, respectively. An other result of the limit length change is the dramatic increase of the total running time from 17s to 541s (the growing factor is 31.82).

In the case of Kulldorff method the change of the limit length of clusters from 15 to 20 doesn't change the ranking of the clusters nor the total running time.

All the results regarding the flexible spatial scan by data length are obtained using the FleXScan ver.1.1.2 software implemented on the same laptop as the above Matlab subroutines.

4. Conclusions

In this paper we described two methods for the detection of clustering. Certain points are worth stressing.

For point (case-control) data (i.e. the *K*-function method), matched cases and controls may be available but great care must be taken when point-based methods are utilized since, in general, the basic method for unmatched data will not be directly applicable and some adjustment of the procedure will be required. We view the *K*-function method as useful initial tool for the data exploration. However, the statistical properties make the interpretation difficult. A great care must be paid to presentation and interpretation of results to avoid unnecessary and unwarranted alarm among the local residents.

For the flexscan method it should be noted that the power estimate reflects the power to reject the null hypothesis for whatever reason and that the probability of both rejecting the null hypothesis and detecting the true cluster correctly is a different matter.

The conclusion is that the methods are complementary and work well for small to moderate number events and cluster sizes, respectively. For larger numbers of events and cluster sizes the methods are practically not feasible and more efficient algorithms are needed.

References

1. Marshall R.J.: A review of the statistical analysis of spatial patterns of disease. Journal of Royal Statistical Society 1991, Series A, 154, 421–441.
2. Lawson A., Bigger A., Böhning D., Lesaffre E., Viel J. F., Bertollini R. (Eds): Disease Mapping and Risk Assessment for Public Health. John Wiley & Sons, London 1999.
3. Elliot P., Wakefield J., Best N., Briggs D. (Eds.): Spatial Epidemiology, Methods and Applications, Oxford Univ. Press, 2000.
4. Waller L.A., Gotway C. A.: Applied Spatial Statistics for Public Health Data, John Wiley & Sons, New York 2004.
5. Ripley B. D.: Spatial Statistics. Wiley, New York 1981.
6. Cuzick J., Edwards R.: Spatial clustering for inhomogeneous populations (with Discussion). Journal of the Royal Statistical Society 1990, Series B, 52, 73–104.
7. Diggle P. J., Chetwynd A. G.: Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations. Biometrics 1991, 47, 1155–1163.
8. Besag J., Diggle P. J.: Simple Monte Carlo tests for spatial patterns. Applied Statistics 1977, 26, 327–333.
9. Tango T., Takahashi K.: A flexibly shaped spatial scan statistic for detecting cluster. Inert. J. of Health Geographics 2005, 4, 11, [Open Access], http://www.ij-healthgeographics .com/.
10. Kulldorff M.: A spatial scan statistic. Communications in Statistics, 1997, 26, 1481–1496.