Spatial cluster detection using the number of connected components of a graph

Avner Bar-Hen, Michel Koskas, Nicolas Picard

Spatial cluster detection using the number of connected components of a graph

Avner Bar-Hen[a], Michel Koskas[b], Nicolas Picard[c]

[a] Université Paris Descartes, Laboratoire MAP5 (CNRS-UMR 8145), 45, rue des Saints-Pères, 75270 Paris, France
[b] AgroParisTech, 16 rue Claude Bernard, F-75231 Paris cedex 05, France
[c] CIRAD, BP 4035, Libreville, Gabon

Abstract

The aim of this work is to detect spatial clusters. We link Erdös graph and Poisson point process. We give the probability distribution function (pdf) of the number of connected component for an Erdös graph and obtain the pdf of the number of cluster for a Poisson process. Using this result, we directly obtain a test for complete spatial randomness and also obtain the clusters that violates the CSR hypothesis. Border effects are computed. We illustrate our results on a tropical forest example.

Keywords: Erdös graph, Point Process, cluster detection, 62M30, 05C80

1 Introduction

Point process modelling is a classical approach for spatial processes when locations of events are random [Cressie 1993, Diggle 1983, Ripley 1988, Stoyan et al. 1995, Stoyan and Stoyan 1994, Daley and Vere-Jones 1988]. The spatial pattern of a given population of trees in a forest for example can be viewed as the result of interactions between the biology of the population and other ecological

processes on the abiotic and biotic environments of the population. Then describing the spatial distribution of a species, as a print of these processes, is an important tool to understand its dynamic.

Exploration of a point process classically begins by descriptive statistics. These exploratory statistics are basic statistical summaries upon which further modeling studies are based. These analyses rely on tests based on comparisons of descriptive statistics to confidence bounds of their counterparts under complete spatial randomness (CSR) [Diggle 1983]. If CSR is rejected, tests will help to choose between aggregative or regular models. The shape of the descriptive statistics is generally used as a guide for parametric modeling of the point process. Many methods are based on various distance measurements between events or between sampled points and the nearest event ([Cressie 1993]). A common feature of these tests is to compare the empirical distribution to simulated distribution of the chosen statistic under H_0. More recently, ([Justel, Peña and Zamar1997, Cucala & Thomas-Agnan2006, Zimmerman1993]) proposed tests based on the empirical distribution in $\mathbb{R}^2$ but they rely on asymptotic properties. Speed of convergence on the asymptotics is a difficult task and from a practical point of view, simulated empirical percentiles should be preferred to the asymptotic distribution.

A second classical objective is to identify zone(s) in which events are most concentrated, usually named cluster. Spatial cluster detection affects several fields: medicine, cosmology with spatial clustering of galaxies, social sciences and criminology, agronomy and more. The question of whether events are clustered in space has received considerable attention in the literature. Among the cluster detection methods, the spatial scan statistic ([Kulldorff and Nagarwalla1995, Kulldorff1997]) 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. Many extensions were proposed ([Patil and Taillie2004, Duczmal and Assunção2004, Tango and Takahashi2005, Demattei C, Molinari N & Daurès2007]). Cluster significance is always obtained via Monte Carlo simulations.

Two main points can be noted: (i) at first CSR tests and cluster detection are done separately: The first aim of this article is to propose a unified approach for theses two tests; and (ii) CSR tests and cluster detection are generally based on Monte Carlo simulations: we derive exact distribution of the proposed statistics.

Given a cloud of points, we construct edge between two points if the distance between the two points is less than a given threshold r. The resulting graph G has k of connected components. In the next section we compute the probability that G is made of k connected components, for a given $k \geq 1$ under the hypothesis of a Poisson process for the points. One may expect that a regular process has too many connected components (may be isolate points) while aggregative processes are composed of a too low number of connected components. Relationships between random graphs and point processes were studied by some authors as [Penrose2003] from a probabilistic point of view. In this article we are not looking for asymptotic distributional properties but for finite distance statistics.

To illustrate the interest of considering the influence of these measurement errors, we selected one of the many tropical tree species of the French Guiana *terra firme* rainforest, mapped and surveyed on the plots of the Paracou Experimental Site [Gourlet-Fleury et al. 2004]. *Dicorynia guianensis* Amshoff, Caesalpiniaceae) is characterised by an aggregated spatial pattern [Loubry et al. 1993, Collinet 1997].

2 Connections between graph theory and spatial statistics

In the following, we denote $Y = (y_1, ..., y_n)$ a set of n points observed in a window W. Without loss of generality, we can assume $W = [0, 1]^2$.

The first process to obtain connected components is described above: points of Y are randomly chosen in a window, a radius r is chosen and given this radius, two points are connected if and only if they are at distances less than or equal to r.

A second process to build random graphs (and connected components) is the following: the vertices of a graph are given and a parameter p (a probability) is chosen. For all pairs of vertices u and v, one chooses to build the edge between u and v with the probability p. Once the edges are built, so are the connected components. Such a graph is called an Erdös graph of parameters p and n (see [Bollobás 2001] for example).

For a Poisson point process the two approaches are equivalent. In the first case, the points are random while in the second case, the edges are random. The next section is dedicated to a theoretical computation of the number of

connected components for a random graph.

2.1 Main results on graph

We consider an Erdös graph of n vertices. Let p the probability of connection for two vertices. In this section we consider that p is given. We denote, as usual, $q = 1 - p$. The number $p_{k,n}$ shall denote the probability that a graph build with the given probability p containing n vertices were made of exactly k connected components, where $1 \leq k \leq n$. The following Proposition permits to compute $p_{k,n}$ recursively from $p_{k',n'}$s with $k' < k$ and $n' < n$.

Proposition 1 *The numbers $p_{k,n}$ verify the following relation: $\forall k \geq 2$, $p_{k,n} = \frac{1}{k} \sum_{l=1}^{n-(k-1)} \binom{n}{l} p_{1,l} p_{k-1,n-l} q^{l(n-l)}$. Furthermore, $p_{1,n} = 1 - \sum_{k=2}^{n} p_{k,n}$.*

Proof: When a graph consisting of n vertices is constituted of k connected components, one can particularize one of these components, whose size lies between 1 and $n - (k - 1)$. There are k choices for this component and once its size is chosen, there are $\binom{n}{l}$ different ways to choose its elements. The probability that the remaining part is constituted of exactly $n-l$ vertices and $k - 1$ connected components is $p_{k-1,n-l}$ and none of the vertices of the particular connected component is connected to any vertex of the remaining part of the graph, which gives the term $q^{l(n-l)}$.

Remark 1 *This relation may be written $\forall k \geq 2$,*

$$p_{k,n} = \frac{1}{k} \sum_{\substack{l_1 \geq 1, l_2 \geq 1, \\ l_1 + l_2 = n}} \binom{n}{l_1} p_{1,l_1} p_{k-1,l_2} q^{l_1 l_2}.$$

This formula may be extended to relate $p_{k,n}$ to the size of each component, *i.e.* to write $p_{k,n}$ as a function of each $p_{1,l}$. It is the aim of the two following propositions.

Proposition 2 *For all $k \geq 2$, the numbers $p_{k,n}$ verify the following relation:*

$$p_{k,n} = \frac{1}{k!} \sum_{\substack{\forall 1 \leq i \leq k, l_i \geq 1, \\ l_1 + l_2 + \cdots + l_k = n}} \binom{n}{l_1, l_2, \ldots, l_k} p_{1,l_1} p_{1,l_2} \ldots p_{1,l_k} q^{\sum_{1 \leq a < b \leq k} l_a l_b}.$$

Proof: Let us prove this property by induction on k. It is true for $k = 2$.

Now let us suppose that it is true for $k-1$ (for a $k \geq 3$) and let us prove it is true for k.

Since $p_{k,n} = \frac{1}{k} \sum_{\substack{l_1 \geq 1, l'_2 \geq 1, \\ l_1 + l'_2 = n}} \binom{n}{l_1} p_{1,l_1} p_{k-1,l'_2} q^{l_1 l'_2}$ and

$$p_{k-1,l'_2} = \frac{1}{(k-1)!} \sum_{\substack{\forall 2 \leq i \leq k, l_i \geq 1, \\ l_2 + l_3 + \cdots + l_k = l'_2}} \binom{n - l_1}{l_2, l_3, \ldots, l_k} p_{1,l_2} p_{1,l_3} \ldots p_{1,l_k} q^{\sum_{2 \leq a < b \leq k} l_a l_b},$$

the following relation stands:

$$p_{k,n} = \frac{1}{k} \sum_{\substack{l_1 \geq 1, l'_2 \geq 1, \\ l_1 + l'_2 = n}} \binom{n}{l_1} p_{1,l_1} \frac{1}{(k-1)!} \sum_{\substack{\forall 2 \leq i \leq k, l_i \geq 1, \\ l_2 + l_3 + \cdots + l_k = l'_2}} \binom{l'_2}{l_2, l_3, \ldots, l_k}$$
$$p_{1,l_2} p_{1,l_3} \ldots p_{1,l_k} q^{\sum_{2 \leq a < b \leq k} l_a l_b} q^{l_1 (l_2 + \cdots + l_k)}$$

which is the wanted formula. We now obtain the formula regarding only p_{1,l_i}:

Proposition 3 *For all $n \geq 1$,*

$$p_{1,n} = 1 - \sum_{d=2}^{n} \frac{1}{d!} \sum_{\substack{l_1 + l_2 + \cdots + l_d = n \\ l_i \geq 1}} \binom{n}{l_1, l_2, \ldots, l_d} p_{1,l_1} p_{1,l_2} \ldots p_{1,l_d} q^{\sum_{1 \leq a < b \leq d} l_a l_b}.$$

To illustrate the formulas, we gather a few results for very little values of n in Table 1.

For $n = 3$ for instance, a graph has 3 connected components with a probability q^3 because it is the probability that the graph did not contain any edge. It has 2 connected components with a probability $p_{2,3} = 3q^2 p$ because such a graph must contain exactly one edge (there are three possibilities). Finally $p_{1,3} = 1 - (p_{2,3} + p_{3,3}) = p^2(2q + 1)$.

2.2 Connection between Poisson point process and graph.

For sake of simplicity we consider the unit square U. For computational reason (see later, border effects) we use L^∞ distance, but the results extend to other norms of $\mathbb{R}^2$.

n \ k	0	1	2	3	4	5
0	1	0	0	0	0	0
1	0	1	0	0	0	0
2	0	p	q	0	0	0
3	0	$p^2(2q+1)$	$3pq^2$	q^3	0	0
4	0	$p^3(6q^3+6q^2+3q+1)$	$p^2q^3(11q+4)$	$6pq^5$	q^6	0
5	0	$p^4(24q^6+36q^5+30q^4+20q^3+10q^2+4q+1)$	$5p^3q^4(10q^3+8q^2+3q+1)$	$5p^2q^7(7q+2)$	$10pq^9$	q^{10}

Table 1: The first values of $p_{k,n}$

Let neglect the border effect for the moment. For a Poisson process the probability to have two points closer than r is r^2 for $0 \leq r \leq 1$. This is a characteristic of Poisson process.

Therefore we can construct a Erdös graph with probability p by connecting all the points with distances less that $r = \sqrt{p}$.

The main idea is to notice that choosing an edge with probability p is equivalent to have two points of a Poisson process closer than $\sqrt{p}$. Thanks to Palm measure, this is also equivalent to the probability that a ball centred on a point of the observed process contained another point of the proces.

By comparing the empirical number of components with the theoretical distribution of components for a Poisson process we can test if the observed cluster are compatible with the Poisson hypothesis.

One difficult task with scan statistics is the border of the cluster which is directly determined by the shape of the scan. Since we only consider the points we do not have this problem with the proposed methodology. Another difficulty with scan statistics is test multiplicity. Since we only consider one empirical curve, we do not have multiple test.

Border effect occurs when the distance between a subject point (taken at random) and the centre C of the square W is greater than the distance from C to the border of W. Border effect can strongly affect the results for large p. Correcting border effects at distance r involves computing the intersection between W and the set of all points at distance less than r from C. Using the L^∞ distance, this is equivalent with computing the intersection of two squares, which simplifies computations. Taking border effects into account, the probability to have two randomly chosen points in U closer than r is less than r^2. This probability is equal to the mean surface of the intersection of

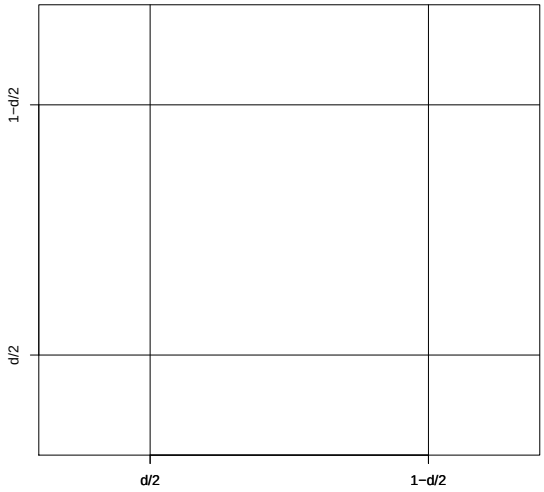

Figure 1: Definition of the corner, border and central part of the square for a given radius d

the unit square U and the square W_r with radius r centred on a randomly chosen point. The following proposition gives the border correction.

Proposition 4 *Let W_d be the square with radius d centred on the origin, and U the unit square of $\mathbb{R}^2$. The expectation of the area of $U \cap (W_d + x)$ where x is a random point with uniform distribution in U is $d^2(1 - d + \frac{3}{8}d^2)^2$*

Proof: Looking at Figure 1 we see that, for a given d three cases have to be considered: (i) the surface of the central part is $(1-d)^2$ while the surface of $U \cap W_d$ is d^2; (ii) the surface of a border part is $(1-d)d/2$ while the surface of $U \cap W_d$ is $d\int_0^{d/2}(d/2 + x)\mathrm{d}x = d\frac{3d}{4}$; (iii) the surface of a corner is $d^2/4$ while the surface of $U \cap W_d$ is $\left(\frac{3d}{4}\right)^2$. Easy computations finish the proof.

Hence, conditionnally to the observation of n points within U, generating an Erdös graph with parameters p and n is equivalent with connecting all the points with distances less that r in a point pattern generated by a Poisson process with intensity n, where r solves $r^2(1 - r + \frac{3}{8}r^2)^2 = p$.

3 Cluster detection and inference

From a practical point of view, we compute the matrix of distance between the n points. For a given distance r, we connect all points such that the distance is less than r and we compute the number of components. The theoretical distribution of the number of components of an Erdös graph with a probability of connection $\sqrt{(}r)$ is given in the various propositions. Construction of a test of H_0 : Complete spatial randomness (CSR) is therefore. When H_0 is rejected the clusters are directly identified.

4 Example

In this section we illustrate our method on the analysis of the spatial pattern of two tree species surveyed at the Paracou experimental site in French Guiana [Gourlet-Fleury et al. 2004]. This site is dedicated to ecological studies on the impact of various types of silvicultural treatments — timber and fuelwood logging, possibly followed by poison-girdling — on the functioning of the forest ecosystem. It is located in the coastal part of the region, approximately 15 km SSE of the town of Sinnamary and 50 km NW of the European space town of Kourou (5°18' N, 52°53' W). The climate is equatorial, with a well marked dry season (middle August to middle November) and a long lasting rainy season, often interrupted by a short drier period between March and April. Natural forest stands grow on shallow ferralitic soils. Within each plot, all trees with diameter at breast height (dbh) greater than 10 cm were located, botanically determined and monitored each year within a core area of 6.25 ha.

Each plot was square 250 m $\times$ 250 m with ropes placed at the edge of the plot with decametre and compass. The coordinates of a tree were then measured with respect to the nearest origin (of the system of ropes axis) with decametre and compass (to keep the orthogonality).

Dicorynia guianensis displayed punctual clusters of radius about 50 m, distant about 100 m from each other over the whole site [Dessard et al. 1999]. Various studies of this species [Loubry et al. 1993, Gourlet-Fleury et al. 2004] led to the conclusion that this clustered spatial pattern was the result of a limited dispersal distance, combined with a relatively high shade-tolerance at young stages and, possibly, an internal replacement dynamics inside the clusters due to a higher turnover inside than outside the clusters.

Figure 2 presents the studied population of *D. guianensis*, the theoretical mean and 95%confidence interval (red and green lines) and the empirical number of clusters. The x-axis represents the distance of interest and the y-axis represents the number of cluster. If the empirical curve is under the confidence interval, it means that the number of cluster is significantly lower than expected. This is a characterisation of clustered process. We zoomed the distance between 0 and 100 meters since after there is only one cluster of points (as expected).

The smallest distance between trees is 6 metres and the number of clusters is outside the confidence interval for all values between 6 and 50 metres. Under six metres, the number of clusters is equal to the number of points.

Figure 3 presents the detected clusters for distances between 10 and 90 metres. For a given distance, we connect all points closer than the distance. The number of connected components, i.e. clusters is given in Figure 2. One may notice that isolated points at the upper right and lower left corner can be part of cluster outside the window. The Figure gives the construction of the three clusters.

References

[Bollobás 2001] Bollobás, B. (2001). *Random Graphs*. Academic Press, London.

[Collinet 1997] Collinet, F. (1997) *Essai de regroupements des principales espèces structurantes d'une forêt dense humide d'après l'analyse de leur répartition spatiale (Forêt de Paracou - Guyane)*. Unpublished PhD thesis, Université Claude Bernard, Lyon I.

[Cressie 1993] Cressie, N. (1993). *Statistics for spatial data*. John Wiley, New York.

[Cucala & Thomas-Agnan2006] Cucala, L. & Thomas-Agnan, C. (2006). Spacings-based tests for spatial randomness and coordinate-invariant procedures. *Ann. I.S.U.P.*, **50:1-2** 31-45.

[Daley and Vere-Jones 1988] Daley, D. and Vere-Jones, D. (1988). *An introduction to the theory of point processes*. Springer-Verlag, New York.

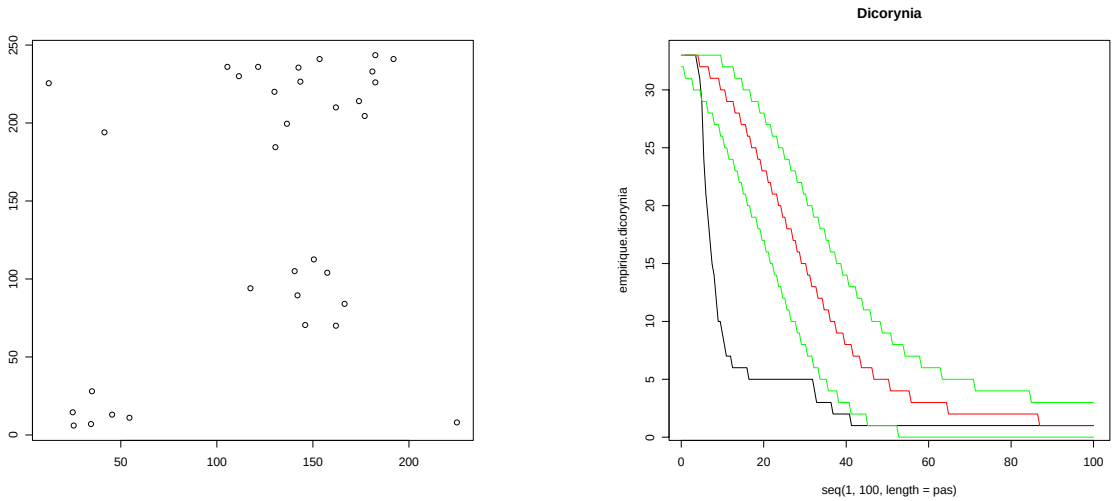

Figure 2: Left part: spatial repartition of *D. guianensis* population, right part: theoretical mean of the number of clusters for a Poisson process (red), the 95% confidence interval (green) and the empirical one (black). x-axis is distance between 0 and 100 meters; y-axis is the number of clusters

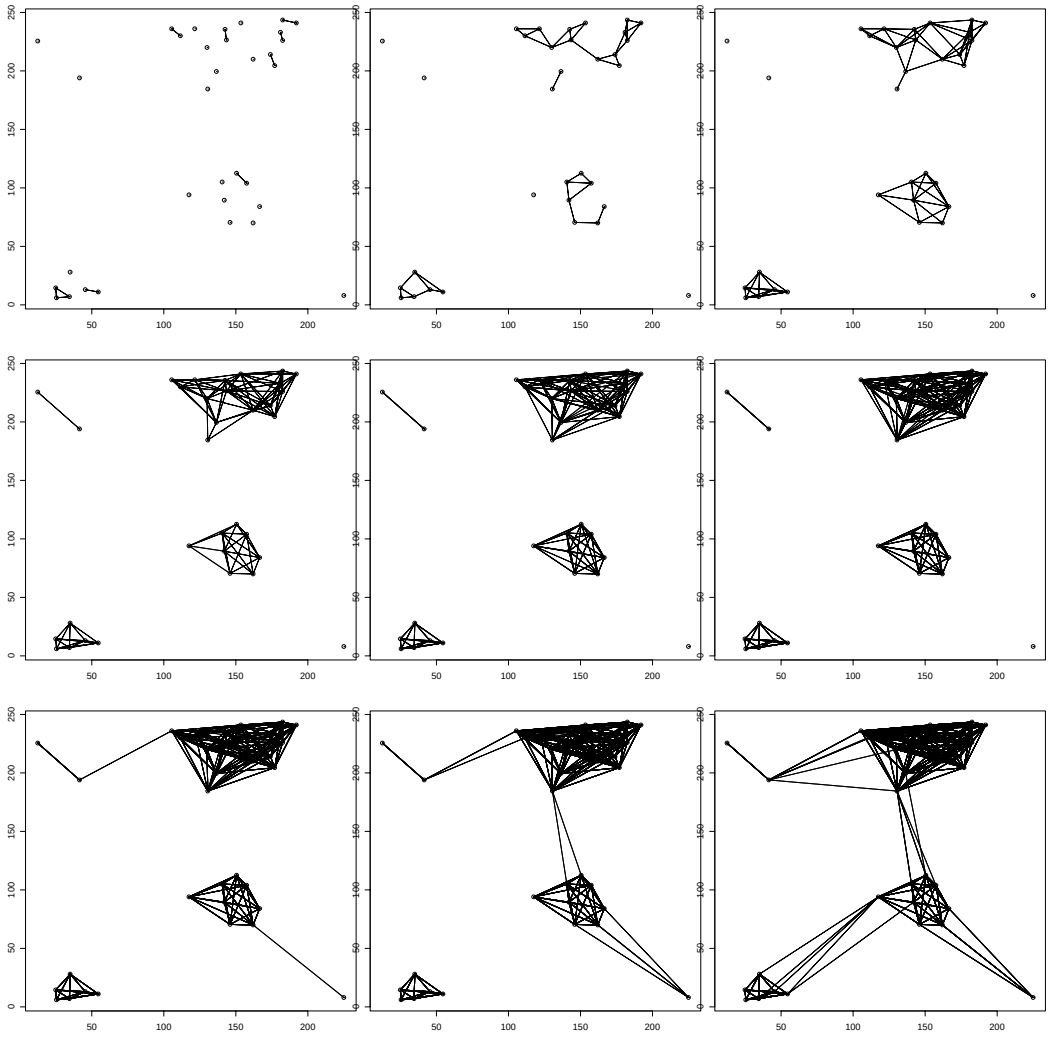

Figure 3: Connected components for distance less than 10, 20, 30, 40, 50, 70, 80 and 90 metres

[Dematteï C, Molinari N & Daurès2007] Dematteï C, Molinari N & Daurès JP, Arbitrarily shaped multiple spatial cluster detection for case event data, *Computational Statistics and Data Analysis*. doi:10.1016/j.csda.2006.03.011

[Dessard et al. 1999] Dessard, H., Picard, N., Pélissier, R., and Collinet-Vautier, F. (2004). *Spatial patterns of the most abundant tree species. In Ecology and Management of a Neotropical Rainforest. Lessons drawn from Paracou, a long-term experimental research site in French Guiana* (eds S. Gourlet-Fleury, J.M. Guehl & O. Laroussinie), pp. 177-190. Elsevier, Paris.

[Diggle 1983] Diggle, P. (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.

[Foresta 1994] De Foresta, H., Charles-Dominique, P., Erard, C. et Prévost, M.-F. (1984). Zoochorie et premiers stades de la régénération naturelle après coupe en forêt guyanaise. *Revue d'Ecologie : la Terre et la Vie* **3**9, 369-400.

[Gourlet-Fleury et al. 2004] Gourlet-Fleury, S., Ferry, B., Molino, J.-F. and Petronelli, P., S. L. (2004). Paracou experimental plots: key features. In S., G.-F., Guehl, J.-M. and Laroussinie, O., editors, *Paracou 15 years of inter-disciplinary research on the dynamics of tropical rainforest in French Guiana*, pages 17–34. Elsevier.

[Justel, Peña and Zamar1997] Justel, A., Peña D. and Zamar R. (1997). A multivariate Kolmogorov- Smirnov test of goodness of fit. *Statistics and Probability Letters*, **35**, 251-259.

[Duczmal and Assunção2004] Duczmal, L., Assunção , R., 2004. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Comput. Statist. Data Anal.* **45**, 269ñ286.

[Kulldorff1997] Kulldorff, M., 1997. A spatial scan statistic. *Comm. Statist. Theory Methods* **26**, 1481ñ1496.

[Kulldorff and Nagarwalla1995] Kulldorff, M., Nagarwalla, N., 1995. Spatial disease clusters: detection and inference. *Statist. Medicine* **14**, 799ñ810.

[Loubry et al. 1993] Loubry D. (1993). Les paradoxes de l'Angélique (Dicorynia guianensis Amshoff) : dissémination et parasitisme des graines avant dispersion chez un arbre anémochore de forêt guyanaise. Revue d'Ecologie (*Revue d'Ecologie : la Terre et la Vie*, **4**8:4, 353-363.

[Patil and Taillie2004] Patil, G.P., Taillie, C., 2004. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statist.* **11**, 183ñ197.

[Penrose2003] Penrose M. (2003). *Random geometric graphs*. Oxford University Press, Oxford

[Ripley 1988] Ripley, B. (1988). *Statistical inference for spatial processes*. Cambridge University Press, Cambridge.

[Ripley 1981] Ripley, B. D. (1981). *Spatial Statistics*. John Wiley, New York.

[Stoyan et al. 1995] Stoyan, D., Kendall, W. and Mecke, J. (1995). *Stochastic geometry and its applications*. John Wiley, Chichester.

[Stoyan and Stoyan 1994] Stoyan, D. and Stoyan, H. (1994). *Fractals, random shapes and point fields*. John Wiley, Chichester.

[Tango and Takahashi2005] T. Tango and K. Takahashi, A flexibly shaped spatial scan statistic for detecting clusters, *Internat. J. Health Geographics* **4** (2005), p. 11

[Zimmerman1993] Zimmerman, D.L. (1993). A Bivariate Cramer-Von Mises Type of Test for Spatial Randomness. *Applied Statistics*, **42**, 43-54.