CS-BIGS: 6(2) 1-9

http://www.csbigs.fr

Spatial cluster detection for socio-economic data

Lionel CUCALA
Institut Montpelliérain Alexander Grothendieck, France.
Christophe DEMATTEI
Centre Hospitalier Universitaire de Nîmes, France.

Abstract

This article focuses on spatial cluster detection methods, mainly the scan methods. We introduce the scan scatistics and the mathematical concepts they rely on and we discuss about the choice of the underlying model. Finally these methods are applied to two socio-economic data sets.

Keywords *: Cluster detection, Scan statistics, Spatial statistics.*

Introduction

Cluster detection has become a very fruitful research subject since the earlier work of Naus (1963): a thorough review of the proposed methods, which have been first applied to temporal data and then extended to spatial and spatio-temporal data, is given by Glaz et al. (2001).

Most of the spatial cluster detection methods have been set up to analyse epidemiological data, in order to identify areas with unusually high or low rates of disease outcome and estimate their significance (Kulldorff and Nagarwalla, 1995). Since then, these methods have been used in many other fields: astronomy, forestry, ecology, genetics, ... (Lawson and Denison, 2002). However, spatial cluster detection for socio-economic data is not very common: let us mention Minamisava et al. (2009) who worked on the spatial locations of murders, Exeter and Boyle (2007) who looked for areas exhibiting abnormal suicide rate, and Huang et al. (2009) who investigated transportation data. When analysing socio-economic data in geographical units, people often want to know whether the measure in one unit is close to the measures in the neighbouring units. Local indicators of spatial association (Anselin, 1995) are useful tools to do that but they only provide one indicator for every geographical unit and, contrary to spatial cluster detection methods, they do not investigate sets of neighbouring units.

Many spatial cluster detection methods can be found in the literature: some of them rely on nonparametric density estimation (Kelsall and Diggle, 1995), some are based on a specific point process model (Stoica et al., 2007). Dematteï et al. (2007) transform a spatial data

set into a unidimensional one before looking for clusters. In this article, we will focus on the most popular methods for local cluster detection named as spatial scan methods. Since the article by Cressie (1977), the scan statistic denotes the maximal concentration observed on a collection of potential clusters. Originally, the size of all the potential clusters had to be the same, so that the scan statistic was just the maximum number of events in a window of size d, d being fixed a priori. This major drawback vanished when Kulldorff (1997) introduced the scan statistic based on generalized likelihood-ratio (GLR) in a Poisson model, which allows to compare the concentration in windows having different sizes. In the same article, the Bernoulli model scan statistic is defined to analyse point processes with binary marks, such as case/control data: if the marks of the cases are 1 and those of the controls are 0, the goal is to identify the areas in which the marks are significantly higher, i.e. the areas where there are significantly more cases, taking into account the number of controls. Later on, Kulldorff et al. (2009) introduced the Gaussian model scan statistic which allows to analyse point processes with continuous marks.

In the first part of the paper, we give details about the Bernoulli and Gaussian model-based scan statistics and the way to evaluate their significance. Then we apply them to socio-economic data sets and we discuss the best choice according to the nature of the data. The paper is concluded with a discussion.

1. Likelihood-based scan statistics

Let $\{(x_i, s_i), i = 1, \cdots, n\}$ be a sample of spatial data, where $s_i \in D$ stands for the spatial location and x_i stands for the corresponding observation of a numeric variable X. The area $D \subset \mathbb{R}^d$ is the observation domain and the spatial locations are usually bidimensional ($d = 2$), sometimes tridimensional ($d = 3$). Our goal is to detect the spatial area $Z \subset D$ in which observations of X are significantly different (larger or smaller) than elsewhere. Let n_Z be the number of locations in area Z and $\sum_Z$ the sum over all these locations. The complementary set of Z is denoted by $\bar{Z}$.

The scan methods usually consist in maximizing a likelihood ratio induced by a parametric model in a collection of potential clusters. Thus the two questions to answer are: how to choose the potential clusters and which parametric model should be used?

Concerning the potential clusters, we will focus here on circular clusters, such as Kulldorff (1997). The set of potential clusters, denoted by $\mathcal{D}$, is the set of discs (or balls if $d = 3$) centered on a location and passing through another one:

$$\mathcal{D} = \{D_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n\}$$

where $D_{i,j}$ is the disc (or the ball) centred on s_i and passing through s_j. Since the disc may have null radius (if $i = j$), the number of potential clusters is n^2. This set of potential clusters is the most popular one, mainly because it allows a fast computation. However, people who are interested in detecting non-circular clusters (clusters along a river, for example) may focus on the wide family of elliptic windows with predetermined shape, angle and center introduced by Kulldorff et al. (2006). The major drawback of this alternative is the very large number of possible clusters to test. Alternatively, one can use arbitrarily-shaped spatial scan statistics that have been proposed by Patil and Taillie (2004), Duczmal and Assuncão (2004), Tango and Takahashi (2005) or Cucala et al. (2012): the potential clusters are completely data-based and their number is quite limited, but the computation of distances between all pairs of spatial locations is needed in order to find them. The ultimate solution would be to consider as possible clusters all the convex envelopes including any subset of the spatial locations. However, this is computationally infeasible when the number of locations is large.

Now, let us consider two parametric models associated to likelihood-based scan statistics.

1.1. *Bernoulli model*

This model was introduced by Nagarwalla (1996) for binary data: $x_i \in \{0,1\}$. The null hypothesis H_0, corresponding to the absence of cluster, is the following: "the x_i's are independent observations of a Bernoulli distribution with parameter p". To each potential cluster $Z \subset D$ is associated an alternative hypothesis, corresponding to the presence of a cluster in Z, $H_{1,Z}$: "the x_i's are independent observations of a Bernoulli distribution with parameter p_Z if $s_i \in Z$ and parameter $p_{\bar{Z}}$ otherwise". The likelihood of the sample under H_0 is

$$L_0(p) = p^{\sum_D x_i}(1-p)^{n-\sum_D x_i}$$

and the maximum is obtained for $p^* = \frac{\sum_D x_i}{n}$. The likelihood of the sample under $H_{1,Z}$ is

$$L_{1,Z}(p_Z, p_{\bar{Z}}) = p_Z^{\sum_Z x_i}(1-p_Z)^{n_Z-\sum_Z x_i} \times p_{\bar{Z}}^{\sum_{\bar{Z}} x_i}(1-p_{\bar{Z}})^{n_{\bar{Z}}-\sum_{\bar{Z}} x_i}$$

and the maximum is obtained for $p_Z^* = \frac{\sum_Z x_i}{n_Z}$ et $p_{\bar{Z}}^* = \frac{\sum_{\bar{Z}} x_i}{n_{\bar{Z}}}$. The likelihood ratio between both hypotheses is

$$\lambda(Z) = \frac{L_{1,Z}(p_Z^*, p_{\bar{Z}}^*)}{L_0(p^*)}$$

and the more this ratio, the more likely the presence of a cluster in Z. Thus, Kulldorff (1997) proposed to maximize this ratio over the set of potential circular clusters defined previously, denoted by $\mathcal{D}$. The scan statistic is

$$\lambda = \max_{Z \in \mathcal{D}} \lambda(Z).$$

Remark that if we look only for positive (resp. negative) clusters where X is significantly larger (resp. smaller) than elsewhere, we should focus on areas Z for which p_Z^* is larger (resp. smaller) than $p_{\bar{Z}}^*$.

1.2. *Gaussian model*

This model was introduced by Kulldorff et al. (2009) for continuous data: $x_i \in \mathbb{R}$. The null hypothesis H_0, corresponding to the absence of cluster, is the following: "the x_i's are independent observations of a Gaussian distribution with mean μ and variance σ^2". To each potential cluster $Z \subset D$ is associated an alternative hypothesis, corresponding to the presence of a cluster in Z, $H_{1,Z}$: "the x_i's are independent observations of a Gaussian distribution with common variance σ_Z^2 and mean μ_Z if $s_i \in Z$, $\mu_{\bar{Z}}$ otherwise". The likelihood of the sample under H_0 is

$$L_0(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\Big[-\frac{\sum_D (x_i-\mu)^2}{\sigma^2}\Big]$$

and the maximum is obtained for $\mu^* = \frac{\sum_D x_i}{n}$ and $\sigma^{2*} = \frac{\sum_D (x_i-\mu^*)^2}{n}$. The likelihood of the sample under $H_{1,Z}$ is

$$L_{1,Z}(\mu_Z, \mu_{\bar{Z}}, \sigma_Z^2) = (2\pi\sigma_Z^2)^{-n/2} \times \exp\Big[-\frac{\sum_Z (x_i-\mu_Z)^2 + \sum_{\bar{Z}} (x_i-\mu_{\bar{Z}})^2}{\sigma_Z^2}\Big]$$

and the maximum is obtained for $\mu_Z^* = \frac{\sum_Z x_i}{n_Z}$, $\mu_{\bar{Z}}^* = \frac{\sum_{\bar{Z}} x_i}{n_{\bar{Z}}}$ and $\sigma_Z^{2*} = \frac{\sum_Z (x_i-\mu_Z^*)^2 + \sum_{\bar{Z}} (x_i-\mu_{\bar{Z}}^*)^2}{n}$. The likelihood ratio between both hypotheses is

$$\lambda(Z) = \frac{L_{1,Z}(\mu_Z^*, \mu_{\bar{Z}}^*, \sigma_Z^{2*})}{L_0(\mu^*, \sigma^{2*})}.$$

Such as for Bernoulli model, this ratio is maximized over the set of potential circular clusters $\mathcal{D}$ and the scan statistic is still

$$\lambda = \max_{Z \in \mathcal{D}} \lambda(Z).$$

Once again, if we look only for positive (resp. negative) clusters where X is significantly larger (resp. smaller) than elsewhere, we should focus on areas Z for which μ_Z^* is larger (resp. smaller) than $\mu_{\bar{Z}}^*$.

1.3. *Estimating the significance*

Once the scan statistic is computed, we need to evaluate its significance. Unfortunately, the null distribution of λ is untractable due to the dependence between $\lambda(Z)$ and $\lambda(Z')$ if $n_{Z \cap Z'} \neq 0$. Another solution, chosen by Kulldorff (1997) or Kulldorff et al. (2009), would be to simulate random datasets under the null hypothesis. However, this solution is valid

only if the true distribution is really Bernoulli or Gaussian, assuring that the correct alpha level is maintained. Thus we decided to run a technique called random labelling: a simulated dataset is obtained by randomly associating the observations x_i to the spatial locations s_i. Let T denote the number of simulated datasets and $\lambda^{(1)}, \cdots, \lambda^{(T)}$ be the values of the scan statistic associated to these datasets. The p-value of the scan statistic λ, observed on the initial sample, is $\frac{\sum_{t=1}^{T} \mathbb{1}(\lambda^{(t)} > \lambda)}{T+1}$.

This randomization procedure is very interesting because, contrary to the simulation of data under the null hypothesis, it does not need any parameter estimation procedure so that its computation is much easier. Moreover, it is the only one for which the type I error remains equal to α whatever the underlying distribution of the data. However, if the labels are spatially autocorrelated, this may lead to overestimating the significance of the detected clusters. Note that the same problem arises with any likelihood-based scan statistic when the significance is estimated through Monte Carlo simulation. As mentioned by Haining (2003), restricted randomization procedures taking into account this spatial autocorrelation are usually applied to global clustering tests such as Ripley's K and derived methods. On the other hand, for local cluster detection tests such as scan statistics, this approach is much less frequent, except in a few articles including the ones by Loh and Zhu (2007) and Zhang et al. (2012).

Let us mention that the detection of clusters and inference with spatial scan statistics might be quite computer-intensive, specially when the number of spatial locations is large. However, the choices we made for the set of potential clusters and for the significance estimation are the ones minimizing the computation time. Indeed, finding all the circular potential clusters is straightforward once the distances between all pairs of spatial locations have been computed. Then, during the significance estimation process, this set of potential clusters remains the same since the spatial locations are not modified. Finally, a simple random permutation of the n first integers is needed to obtain a permuted sample and likelihood has to be maximized for each of the T permuted samples.

2. Applications

2.1. *Public housing*

We analysed a data set provided by the French statistics agency INSEE (Institut National de la Statistique et des Etudes Economiques) : for year 2009, the number of public housing accomodations and the total number of accomodations in each of the 94 French departments have been computed. From these binary data, we obtained continuous data by computing the ratio between the number of public housing accomodations and the total number of accomodations in each department. Figure 1 illustrates the spatial distribution of this ratio.

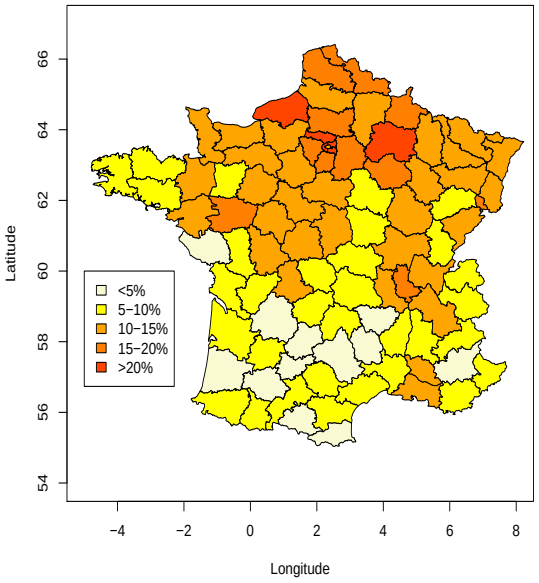

Figure 1: Ratio of public housing accomodations

This ratio seems to be larger in the Parisian area and in the North, smaller in the South-West. Since the ratio is a continuous data, we applied the scan method based on Gaussian

model and the results are given by Figure 2. Remark that the location associated to each department is the location of its capital city.

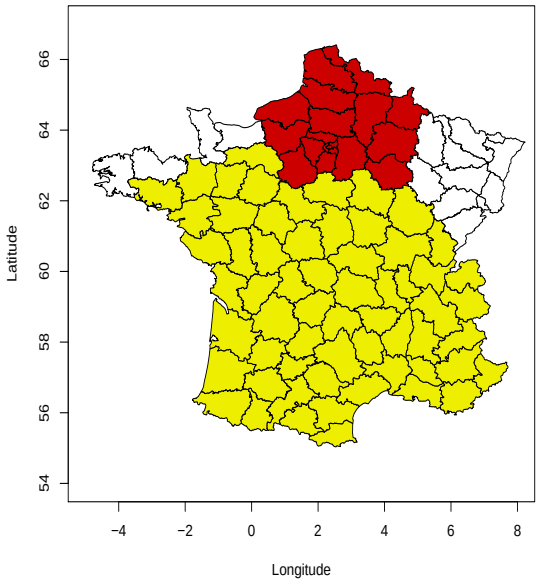

Figure 2: Ratio of public housing accomodations: clusters

The shaded area, including 19 departments in the Parisian area and the Northern part of France, is the most significant cluster. The logarithm of the scan statistic approximately equals 32.9 and the p-value estimation, based on $m = 1000$ simulations, equals 0.001. Of course this is a positive cluster: the ratio of public housing accomodations is 0.18 inside and only 0.09 outside. The lightened area, including 61 departments, is the most significant negative cluster. The logarithm of the negative scan statistic approximately equals 24.2 and the p-value estimation still equals 0.001. These results confirm the feeling from Figure 1: in France, public housing ratio is not uniformly distributed and is much larger in high-density regions such as the Parisian area and the North of France.

However, we may remark that, in the preceding analysis, the department with the largest number of accomodations (Paris, more than 1300000) has the same weight than the one with the smallest number of accomodations (Lozère, with less than 60000). We circumvented this issue by analysing the raw binary data (the number of accomodations, not the ratio) through the scan method based on Bernoulli model: for each accomodation, $x_i = 1$ if it is a public housing accomodation, otherwise $x_i = 0$. The results we obtained are exactly the same than the previous ones.

2.2. *Presidential election*

We applied the same scan methods to the results of the 2012 presidential election in France for François Hollande, head of the Socialist Party, in each department. Here again, the original binary data (number of votes for Hollande and total number of votes) can be transformed into continuous data (ratio of votes for Hollande). Figure 3 illustrates the spatial distribution of this ratio.

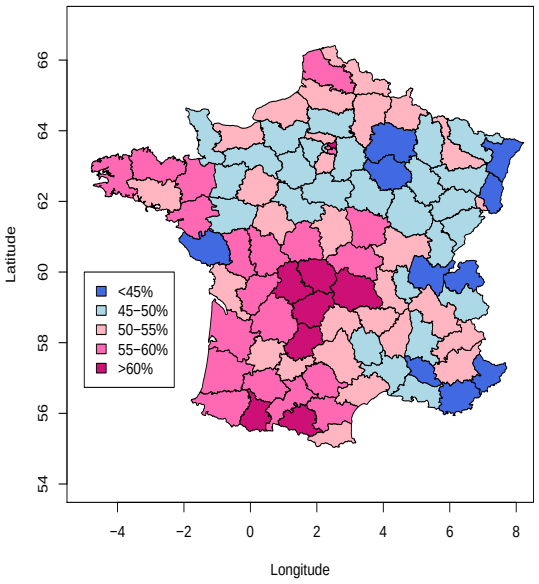

Figure 3: Presidential election results

The South-West of France is traditionally an area where more people vote for the Socialist Party. On the other side, the opponent of François Hollande, Nicolas Sarkozy, obtained his best results in the East of France. These correspond to the clusters exhibited by the scan

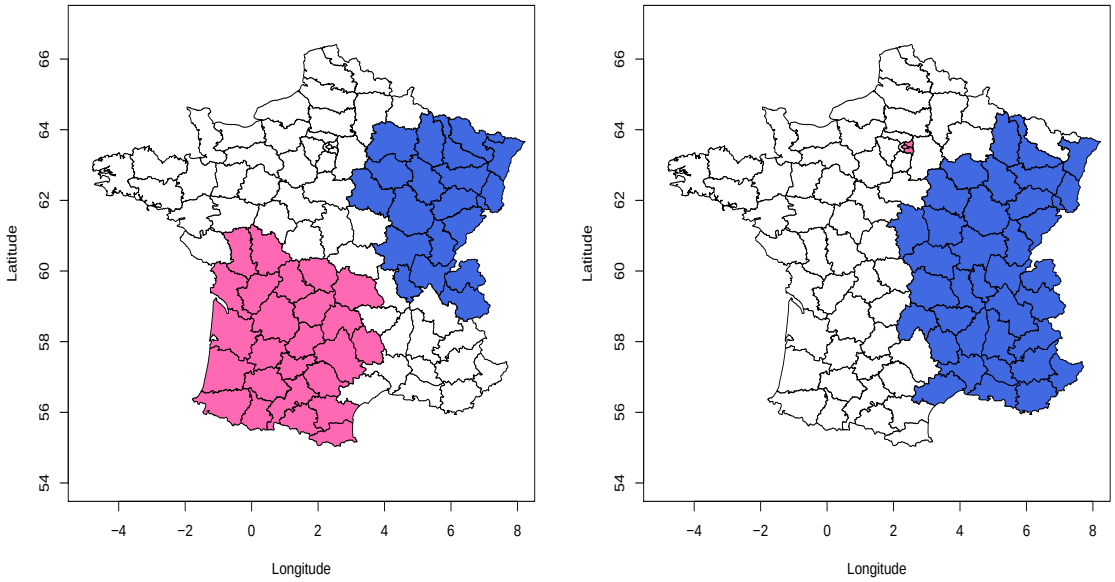

Figure 4: Presidential election results: clusters. Left: Gaussian model. Right: Bernoulli model.

method based on Gaussian model, plotted on the left part of Figure 4.

The most significant cluster contains 25 departments in the South-West in which 57% of the voters voted for Hollande. The most significant negative cluster contains 20 departments in the East in which 54% of the voters voted for Sarkozy. Both p-value estimations equal 0.001. However, the population density in the positive cluster is quite low so that we wonder whether taking into account the number of voters, using Bernoulli model on the original binary data, would change something. The results obtained through Bernoulli model are plotted on the right part of Figure 4.

The most significant cluster is now a negative one, including 37 departments in the East and in which Sarkozy obtained 5998531 votes among 11316095, that is around 53%. Compared to the negative cluster on the left part of Figure 4, this one includes neighbouring departments with high population density. The percentage of votes for Sarkozy has slightly decreased (53% instead of 54%) but this is balanced by the increase in the number of voters. On the other side, the most significant positive cluster is now a set of three departments in the Parisian area (Paris, Seine-Saint-Denis and Val-de-Marne) in which Hollande obtained 1247055 votes among 2139026, that is around 58,3%. Since this cluster only contains three departments, it was not that significant in the Gaussian model, but it is in the Bernoulli model because the number of population in it is very large.

2.3. *Implementation*

The analysis of both data sets has been done using personal C++ software which is available upon request. However, most people who perform scan analysis use the free SaTScan software (Kulldorff, M. and Information Management Services Inc, 2015), which allows to compute scan statistics based on any parametric model: Bernoulli and Gaussian, such as in our applications, but also Poisson, exponential, ordinal... The data may be aggregated in geographical units, as specified in the data sets we analysed, or there may be unique coordinates for each observation. The set of potential clusters can be either the set of circular clusters defined previously or a set of elliptic clusters with specific parameters. The SaTScan

software can be run as part of R environment using the R package rsatscan: this package provides functions for writing R data frames in SaTScan-readable formats, for setting SaTScan parameters, for running SaTScan in the OS, and for reading the files that SaTScan creates.

Alternatively, some likelihood-based scan statistics can be computed using other softwares. The Windows-based non-free ClusterSeer software (BioMedware, 2016) contains 24 different statistical methods for detecting and evaluating spatial, temporal and space-time clustering, including the spatial scan statistic based on Poisson model. The R package SpatialEpi (Chen et al., 2016) contains a function which performs the purely spatial scan statistic with either the Poisson or Bernoulli probability model. Let us also mention the R package graphscan (Loche et al., 2015) and the free software FlexScan (Takahashi et al., 2013) which compute a spatial scan statistic based on Poisson model but use a non-parametric rather than circular or elliptic definition of clusters.

Conclusion

Spatial cluster detection methods are useful to exhibit areas in which the distribution of a variable is significantly different than elsewhere. Even if these areas are sometimes obvious on graphic plots, the scan methods are necessary to obtain precise boundaries of the cluster and evaluate its significance. Moreover, as experienced with the election results application, the results may be quite different depending of which data is analysed: the binary raw data (the number of votes for Hollande and the total number of votes) or the continuous partial data (the ratio of votes for Hollande). To our knowledge, the scan methods are the only local spatial cluster detection methods who allow to analyse either continuous or binary spatial data.

Remark that, when the number of spatial locations is not that huge, these scan statistics can be computed quickly. As noted by Kulldorff et al. (2006), the main goal of a cluster detection technique is to generate an alarm so that the scientists can investigate more precisely the details of this excess of unusual values.

In this work we only focused on the most significant cluster but looking for secondary clusters is straightforward using the method proposed by Zhang et al. (2010): once a significant cluster is found, remove the data included in that cluster and restart the analysis.

Finally, we should underline that, if these scan methods take into account the population density, they could also be adjusted for any continuous covariate as proposed by Klassen et al. (2005), such as the age of an underlying population. This could be done by modeling a regression function of the marks depending on the adjusted covariates, and then analysing the corresponding residuals. This is for exemple the way López et al. (2015) analysed the housing prices in Madrid, Spain.

References

Anselin, L. (1995). Local indicators of spatial associationâĂŤlisa. *Geographical analysis*, 27(2):93–115.

BioMedware (2016). Clusterseer: Software for the detection and analysis of event clusters. www.biomedware.com.

Chen, C., Kim, A., Ross, M., and Wakelield, J. (2016). Spatialepi: Methods and data for spatial epidemiology. *R package version 1.2.2*.

Cressie, N. (1977). On some properties of the scan statistic on the circle and the line. *Journal of Applied Probability*, pages 272–283.

Cucala, L., Demattei, C., Lopes, P., and Ribeiro, A. (2012). A spatial scan statistic for case event data based on connected components. *Computational Statistics*, 28(1):357–369.

Dematteï, C., Molinari, N., and Daurès, J.-P. (2007). Arbitrarily shaped multiple spatial cluster detection for case event data. *Computational Statistics & Data Analysis*, 51(8):3931–3945.

Duczmal, L. and Assuncão, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45(2):269–286.

Exeter, D. J. and Boyle, P. J. (2007). Does young adult suicide cluster geographically in scotland? *Journal of epidemiology and community health*, 61(8):731–736.

Glaz, J., Naus, J., and Wallenstein, S. (2001). *Scan statistics*. Springer-Verlag, New York.

Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

Huang, L., Stinchcomb, D. G., Pickle, L., Dill, J., and Berrigan, D. (2009). Identifying clusters of active transportation using spatial scan statistics. *American journal of preventive medicine*, 37(2):157–166.

INSEE (Institut national de la statistique et des études économiques). www.insee.fr.

Kelsall, J. E. and Diggle, P. J. (1995). Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*, 14(21-22):2335–2342.

Klassen, A. C., Kulldorff, M., and Curriero, F. (2005). Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *International Journal of Health Geographics*, 4(1):1.

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

Kulldorff, M., Huang, L., and Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *International journal of health geographics*, 8(1):1.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in medicine*, 25(22):3929–3943.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in medicine*, 14(8):799–810.

Kulldorff, M. and Information Management Services Inc (2015). Satscan v9.4: Software for the spatial and space-time scan statistics. www.satscan.org.

Lawson, A. and Denison, D. (2002). *Spatial cluster modelling*. Chapman and Hall/CRC, London.

Loche, R., Giron, B., Abrial, D., Cucala, L., Charras-Garrido, M., and De-Goer, J. (2015). graphscan: Cluster detection with hypothesis free scan statistic. *R package version 1.1*.

Loh, J. M. and Zhu, Z. (2007). Accounting for spatial correlation in the scan statistic. *The Annals of Applied Statistics*, pages 560–584.

López, F., Chasco, C., and Le Gallo, J. (2015). Exploring scan methods to test spatial structure with an application to housing prices in madrid. *Papers in Regional Science*, 94(2):317–346.

Minamisava, R., Nouer, S. S., de Morais Neto, O. L., Melo, L. K., and Andrade, A. L. S. (2009). Spatial clusters of violent deaths in a newly urbanized region of brazil: highlighting the social disparities. *International journal of health geographics*, 8(1):1.

Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in Medicine*, 15(7-9):845–850.

Naus, J. (1963). *Clustering of random points in the line and plane*. PhD thesis, Ph.D. thesis, Rutgers University, New Brunswick, NJ.

Patil, G. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological statistics*, 11(2):183–197.

Stoica, R. S., Gay, E., and Kretzschmar, A. (2007). Cluster pattern detection in spatial data based on monte carlo inference. *Biometrical Journal*, 49(4):505–519.

Takahashi, K., Yokoyama, T., and Tango, T. (2013). Flexscan v3.1.2: Software for the flexible scan statistics. www.sites.google.com/site/flexscansoftware.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4(1):1.

Zhang, T., Zhang, Z., and Lin, G. (2012). Spatial scan statistics with overdispersion. *Statistics in medicine*, 31(8):762–774.

Zhang, Z., Assunção, R., and Kulldorff, M. (2010). Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 2010.

Correspondence: lionel.cucala@umontpellier.fr.