Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2016, Article ID 7591680, 8 pages
http://dx.doi.org/10.1155/2016/7591680

Research Article

Scan Statistics for Detecting High-Variance Clusters

Lionel Cucala

Institut Montpelliérain Alexander Grothendieck, 34000 Montpellier, France

Correspondence should be addressed to Lionel Cucala; lionel.cucala@univ-montp2.fr

Received 28 September 2015; Revised 8 December 2015; Accepted 15 December 2015

Academic Editor: Dejian Lai

Scan statistics are mostly used to detect spatial areas or time intervals in which the mean level of a given variable is more important. Sometimes, when this variable is continuous, there is an interest in looking for clusters in which its variability is more important. In this paper, two scan statistics are introduced for identifying clusters of values exhibiting higher variance. Like many classical scan statistics, the first one relies on a generalized likelihood ratio test whereas the second one is based on ratios of empirical variances. These methods are useful to identify spatial areas or time intervals in which the variability of a given variable is more important. In an application of the new methods, I look for geographical clusters of high-variability income in France and then for residuals exhibiting higher variance in a linear regression context.

1. Introduction

Cluster detection has become a very fruitful research subject since the earlier work of Naus [1] looking for an unusual clustering of random points on the line. Thereafter, it was extended to identify time intervals or spatial areas where the observations of a given random variable are different than elsewhere. These methods are nowadays very popular in disease surveillance for the detection of disease clusters, and they are also used in many other fields, such as forestry, astronomy, and criminology. A thorough review is given by Glaz et al. [2].

Most of the cluster detection methods are designed for count data, that is, point processes made of the random coordinates of n events observed in S, a bounded subset of $\mathbb{R}^d$: the goal is to identify, if they exist, the areas in which the concentration of events is abnormally high. According to the article by Cressie [3], 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 [4] introduced the scan statistic based on generalized likelihood ratio in either Poisson model or Bernoulli model. This scan statistic is defined to analyse point processes with binary marks, such as case/control data. Later on, Kulldorff et al. [5] introduced the Gaussian model scan statistic which allows analysing point processes with continuous marks.

This scan method is useful to detect high-mean clusters, that is, areas where the marks are significantly higher than elsewhere. However, when analysing income inequalities, for example, it might be useful to look for high-variance clusters: in these areas, inequalities are more obvious and this could generate more violence or more crimes. Thus, existing scan methods need to be modified in order to detect such clusters.

In this paper I introduce two scan statistics for detecting high-variance clusters. Section 2 describes the scan statistics and their computational aspects in the framework of marked point processes. Their performances are compared through a simulation study in Section 3. In Section 4, they are applied to two real data sets: on the first one, which describes the spatial distribution of incomes in France, the high-variance scan statistics are directly applied; on the second one, which illustrates the link between expenditure on public schools and per capita income, a linear regression model is fitted and its residuals are analysed through high-variance scan statistics. The paper is concluded with a discussion.

2. Two High-Variance Scan Statistics

Let $\{(x_i, s_i),\ i = 1, \ldots, n\}$ denote the realization of a marked point process, where $s_i \in S$ is the location of the event and

$x_i \in \mathbb{R}$ its associated mark. The area $S \subset \mathbb{R}^d$ is the observation domain and the locations can be either one-dimensional ($d = 1$) or spatial ($d = 2$ or $d = 3$). Our goal is to detect the area $Z \subset S$ in which the marks exhibit a significantly higher variance than elsewhere.

From now on, let us consider that the locations are spatial. The setting of one-dimensional locations will be handled in Section 4.2. Most of the spatial cluster detection methods rely on likelihood ratio between two hypotheses depending on a potential cluster: the scan statistic is nothing but the maximum of all these likelihood ratios. Thus the two questions to answer are how to choose the potential clusters and which likelihood ratio should be used?

Concerning the potential clusters, I will focus on circular clusters, such as Kulldorff [4]. The set of potential clusters, denoted by $\mathscr{D}$, is the set of discs (if $d = 2$) or balls (if $d = 3$) centered on a location and passing through another one:

$$\mathscr{D} = \left\{D_{i,j},\ 1 \le i \le n,\ 1 \le j \le n\right\}, \tag{1}$$

where $D_{i,j}$ is the disc (or the ball) centered on s_i and passing through s_j: the number of potential clusters is n^2.

2.1. A Generalized Likelihood Ratio Scan Statistic. As said in the Introduction, Kulldorff et al. [5] introduced a Gaussian-based scan statistic to detect clusters exhibiting higher means than elsewhere. Their concentration index is based on the likelihood ratio between two hypotheses. In order to detect high-variance clusters instead of high-mean clusters, I decided to modify their alternative hypothesis.

Let $X_1, \ldots, X_n$ denote the random variables associated with the marks, which are assumed to be independent. Under the null hypothesis H_0, corresponding to the absence of any cluster, all the marks come from the same Gaussian distribution:

$$X_1, \ldots, X_n \sim \mathcal{N}\left(\mu^*, \sigma^{*2}\right), \tag{2}$$

where (μ^*, σ^{*2}) is the maximum-likelihood (ML) estimate associated with this distribution; that is,

$$\left(\mu^*, \sigma^{*2}\right) = \underset{(\mu,\sigma^2)}{\arg\max} \prod_{i=1}^{n} f_{(\mu,\sigma^2)}\left(x_i\right), \tag{3}$$

$f_{(\mu,\sigma^2)}(\cdot)$ being the density function associated with the Gaussian distribution with mean μ and variance σ^2.

Let the spatial area $Z \subset S$ be a potential cluster and $\overline{Z} = S \setminus Z$ its complement. Under the alternative hypothesis $H_{1,Z}$, corresponding to a high-variance cluster in Z, the marks come from a mixture Gaussian distribution with equal means but different variances:

$$\begin{aligned} \left(X_i : s_i \in Z\right) &\sim \mathcal{N}\left(\mu^*_{Z,\overline{Z}}, \sigma^{*2}_{Z}\right), \\ \left(X_i : s_i \in \overline{Z}\right) &\sim \mathcal{N}\left(\mu^*_{Z,\overline{Z}}, \sigma^{*2}_{\overline{Z}}\right), \end{aligned} \tag{4}$$

where $(\mu^*_{Z,\overline{Z}}, \sigma^{*2}_{Z}, \sigma^{*2}_{\overline{Z}})$ is the ML estimate associated with this mixture distribution; that is,

$$\begin{aligned} &\left(\mu^*_{Z,\overline{Z}}, \sigma^{*2}_{Z}, \sigma^{*2}_{\overline{Z}}\right) \\ &\quad = \underset{(\mu_{Z,\overline{Z}},\sigma^2_Z,\sigma^2_{\overline{Z}})}{\arg\max} \prod_{i:s_i\in Z} f_{(\mu_{Z,\overline{Z}},\sigma^2_Z)}\left(x_i\right) \prod_{i:s_i\in\overline{Z}} f_{(\mu_{Z,\overline{Z}},\sigma^2_{\overline{Z}})}\left(x_i\right). \end{aligned} \tag{5}$$

Both of these ML estimates have closed forms. Let

$$\begin{aligned} n_Z &= \sum_{i=1}^{n} \mathbb{1} \quad (s_i \in Z), \\ \overline{X}_Z &= \frac{1}{n_Z} \sum_{i=1}^{n} X_i \mathbb{1} \quad (s_i \in Z), \\ \overline{X^2}_Z &= \frac{1}{n_Z} \sum_{i=1}^{n} X_i^2 \mathbb{1} \quad (s_i \in Z) \end{aligned} \tag{6}$$

denote, respectively, the number and the mean of marks and the mean of squares of marks in Z. The ML estimate under H_0 is

$$\left(\mu^*, \sigma^{*2}\right) = \left(\overline{X}_S, \overline{X^2}_S - \left(\overline{X}_S\right)^2\right), \tag{7}$$

the vector containing the mean and the (biased) empirical variance of the marks in the whole domain. As shown in the Appendix, the ML estimate under $H_{1,Z}$ satisfies

$$\begin{aligned} \sigma^{*2}_{Z} &= \overline{X^2}_Z - 2\mu^*_{Z,\overline{Z}}\overline{X}_Z + \left(\mu^*_{Z,\overline{Z}}\right)^2, \\ \sigma^{*2}_{\overline{Z}} &= \overline{X^2}_{\overline{Z}} - 2\mu^*_{Z,\overline{Z}}\overline{X}_{\overline{Z}} + \left(\mu^*_{Z,\overline{Z}}\right)^2 \end{aligned} \tag{8}$$

and $\mu^*_{Z,\overline{Z}}$ is the root of a cubic function which can be obtained using, for example, Cardano's method [6]. The log-likelihood under the null hypothesis is

$$\begin{aligned} \mathrm{LL}_0 &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^{*2}\right) \\ &\quad - \frac{1}{2\sigma^{*2}} \sum_{i=1}^{n} \left(X_i - \mu^*\right)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^{*2}\right) - \frac{n}{2} \end{aligned} \tag{9}$$

and the one under the alternative hypothesis is

$$\begin{aligned} \mathrm{LL}_{1,Z} &= -\frac{n_Z}{2}\log(2\pi) - \frac{n_Z}{2}\log\left(\sigma^{*2}_{Z}\right) \\ &\quad - \frac{1}{2\sigma^{*2}_{Z}} \sum_{s_i\in Z} \left(X_i - \mu^*_{Z,\overline{Z}}\right)^2 - \frac{n_{\overline{Z}}}{2}\log(2\pi) \\ &\quad - \frac{n_{\overline{Z}}}{2}\log\left(\sigma^{*2}_{\overline{Z}}\right) - \frac{1}{2\sigma^{*2}_{\overline{Z}}} \sum_{s_i\in\overline{Z}} \left(X_i - \mu^*_{Z,\overline{Z}}\right)^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n_Z}{2}\log\left(\sigma^{*2}_{Z}\right) - \frac{n_{\overline{Z}}}{2}\log\left(\sigma^{*2}_{\overline{Z}}\right) \\ &\quad - \frac{n_Z}{2} - \frac{n_{\overline{Z}}}{2}, \end{aligned} \tag{10}$$

so that the log-likelihood ratio between both hypotheses reduces to

$$\begin{aligned} \mathrm{LLR}_Z &= \mathrm{LL}_{1,Z} - \mathrm{LL}_0 \\ &= \frac{1}{2}\left[n\log\left(\sigma^{*2}\right) - n_Z\log\left(\sigma_Z^{*2}\right) - n_{\overline{Z}}\log\left(\sigma_{\overline{Z}}^{*2}\right)\right]. \end{aligned} \tag{11}$$

The next step is only the maximization of this log-likelihood ratio on the set of potential clusters previously defined. The generalized likelihood ratio (GLR) high-variance scan statistic is

$$\lambda = \max_{Z\in\mathscr{D}} \mathrm{LLR}_Z \tag{12}$$

and the potential cluster for which this maximum is obtained,

$$\widehat{C} = \arg\max_{Z\in\mathscr{D}} \mathrm{LLR}_Z, \tag{13}$$

is called the most likely cluster.

Once the scan statistic is computed, I need to evaluate its significance. Unfortunately, the null distribution of λ is untractable due to the dependence between the log-likelihood ratios: indeed, the log-likelihood ratios LLR_Z and $\mathrm{LLR}_{Z'}$ associated with two potential clusters Z and Z' are independent only if $n_{Z\cap Z'} = 0$. Another solution, chosen by Kulldorff [4], would be to simulate random data sets under the null hypothesis. However, by doing the randomization this way, the correct alpha level will not be maintained if the marks do not truly come from a normal distribution. Thus I decided to run a technique used by Kulldorff et al. [5] called random labelling: a simulated data set is obtained by randomly associating the marks with the spatial locations. Let T denote the number of simulated data sets and let $\lambda^{(1)}, \ldots, \lambda^{(T)}$ be the observations of the scan statistic associated with these data sets: these simulated scan statistics must be compared to λ, the scan statistic observed on the real data set. According to Dwass [7], the Monte Carlo based p value of the scan statistic λ is $R/(T+1)$, where R is the rank of λ in the $(T+1)$-sample $(\lambda^{(1)}, \ldots, \lambda^{(T)}, \lambda)$. Note that this p value is unbiased in the sense that under the null hypothesis, the probability of observing a p value less than or equal to p is exactly p. According to the classical test theory, the most likely cluster $\widehat{C}$ is said to be significant if the associated p value is less than type I error α.

2.2. A Generalized Variance-Ratio Scan Statistic. Even if the likelihood-based scan statistics are broadly used, Cucala [8] proposed, in a different context, an alternative scan statistic showing better results than the one based on likelihood ratio: the power was slightly higher, such as the ability to detect small clusters. Here I propose a method to build a high-variance scan statistic not relying on likelihood ratio.

A classical test for equality of variances between the marks in Z and in $\overline{Z}$, usually known as F-test, relies on the ratio of (unbiased) empirical variances

$$R_Z = \frac{S_Z^{*2}}{S_{\overline{Z}}^{*2}}, \tag{14}$$

where

$$S_Z^{*2} = \frac{1}{n_Z - 1}\sum_{i=1}^{n}\left(X_i - \overline{X}_Z\right)^2 \mathbb{1} \quad (s_i \in Z). \tag{15}$$

Note that, since the X_i's are assumed to be independent, the empirical variances S_Z^{*2} and $S_{\overline{Z}}^{*2}$ are also independent. Moreover, under Gaussian assumption, the equality of variances of marks in Z and $\overline{Z}$ ensures that R_Z follows, by definition, a Fisher distribution $F(n_Z - 1, n_{\overline{Z}} - 1)$, as mentioned by Saporta [9]. Let me introduce the high-variance index for potential cluster Z

$$I_Z = F_{n_Z-1, n_{\overline{Z}}-1}\left(R_Z\right), \tag{16}$$

where $F_{n_Z-1, n_{\overline{Z}}-1}(\cdot)$ denotes the cumulative distribution function associated with the Fisher distribution $F(n_Z - 1, n_{\overline{Z}} - 1)$. If the variances are equal, the distribution of I_Z is uniform $(0, 1)$ and does not depend anymore on n_Z: this method is called the probability integral transform [10]. Moreover, a value of I_Z close to 1 shows that the variance in Z is significantly larger than in $\overline{Z}$. Therefore, the generalized variance-ratio (GVR) scan statistic is

$$\widetilde{\lambda} = \max_{Z\in\mathscr{D}} I_Z. \tag{17}$$

The way I evaluate the significance of $\widetilde{\lambda}$ is completely similar to the method for λ.

3. A Simulation Study

I decided to run a simulation study in order to compare the results of the tests based on the scan statistics previously introduced, λ and $\widetilde{\lambda}$. I generate artificial data sets according to three different models. Whatever the model, the geographical locations are the locations of the 94 French departments. Note that the location associated with each department is the location of its capital city. The true cluster, called C, is a set of eight departments around Paris called Ile-de-France: in this area, the variance of the marks is larger than elsewhere. In the first model, denoted by homogeneous Gaussian model, the associated marks are independent and follow a Gaussian distribution with equal means and different variances:

$$X_i \sim \begin{cases} \mathscr{N}(0, r) & \text{if } s_i \in C, \\ \mathscr{N}(0, 1) & \text{otherwise.} \end{cases} \tag{18}$$

In the second model, denoted by Exponential model, the associated marks are independent and are derived from exponentially distributed random variables with different rate parameters, $X_i = Y_i - \mathbb{E}(Y_i)$, where

$$Y_i \sim \begin{cases} \mathscr{E}\left(r^{-1/2}\right) & \text{if } s_i \in C, \\ \mathscr{E}(1) & \text{otherwise,} \end{cases} \tag{19}$$

so that the means are all equal to 0. In the third model, denoted by heterogeneous Gaussian model, the associated

Table 1: Homogeneous Gaussian model.

Cluster intensity r		Results of the following λ	$\widetilde{\lambda}$
1.5	Power	0.080	**0.118**
	TP	0.567	**0.741**
	FP	3.702	**3.590**
2.0	Power	0.265	**0.322**
	TP	1.723	**2.009**
	FP	3.998	**3.400**
2.5	Power	0.528	**0.553**
	TP	3.492	**3.500**
	FP	2.844	**2.118**

Table 2: Exponential model.

Cluster intensity r		Results of the following λ	$\widetilde{\lambda}$
1.5	Power	0.056	**0.057**
	TP	0.191	**0.352**
	FP	**1.441**	2.384
2.0	Power	0.070	**0.071**
	TP	0.349	**0.472**
	FP	**1.806**	2.195
2.5	Power	0.074	**0.093**
	TP	0.332	**0.659**
	FP	**1.538**	2.987

marks are independent and follow a Gaussian distribution with different means and different variances:

$$X_i \sim \begin{cases} \mathcal{N}(l, r) & \text{if } s_i \in C, \\ \mathcal{N}(0, 1) & \text{otherwise.} \end{cases} \tag{20}$$

Note that, for every model, the parameter r is the ratio between the variance inside the cluster and the variance outside the cluster: I should call it the cluster intensity. For the third model, the parameter l, which denotes the difference of means inside the cluster and outside of the cluster, is called the mean lag.

For each model and each value of the cluster intensity r, I generated 1000 simulated data sets. The high-variance scan statistics λ and $\widetilde{\lambda}$ have been computed, using the set of circular potential clusters described in Section 2, and their p values are estimated based on $T = 999$ permutations. In both methods, the most likely cluster $\widehat{C}$ is said to be significant if the associated p value is less than $\alpha = 5\%$.

When applied to such data sets exhibiting a true cluster, scan methods are useful when they manage simultaneously to identify that there is a significant cluster and to recover as precisely as possible the true cluster. Therefore, in order to compare the different scan methods, I computed three different criterions. The first one is the power of the method, that is, the proportion of data sets exhibiting a significant cluster. The second one is the mean number of true positive (TP) departments, that is, the departments included both in the most likely cluster $\widehat{C}$ and in the true cluster C. The third one is the mean number of false positive (FP) departments, that is, the departments included in the most likely cluster $\widehat{C}$ but not in the true cluster C. Note that the sum of TP and FP is the mean number of departments in the most likely cluster $\widehat{C}$. Table 1 gives the results obtained with the homogeneous Gaussian model, Table 2 the results obtained with the Exponential model, and Table 3 the results obtained with the heterogeneous Gaussian model.

The bold values are the best results obtained in each procedure. As expected, the power of the methods increases with the cluster intensity. Under the homogeneous Gaussian model, the GVR method is always more powerful than the GLR one, whatever the cluster intensity is. These results are very similar to the ones obtained by Cucala [8] stating that, for detecting clusters in count data, GLR methods are not always the most powerful ones. However, although these power results are quite satisfactory for the homogeneous Gaussian model, they are quite poor when the marks are exponentially distributed: even when the variance is 2.5 times larger in the cluster, both methods hardly detect it. The results in Table 3 also reveal that, even if those scan statistics are designed for detecting high-variance clusters, both are sensitive to a difference of means. Indeed, the cluster intensity being equal, the powers of both methods increases when there is a mean lag. Finally, looking at the FP results, I also conclude that the GLR method tends to exhibit larger significant clusters than the GVR one, which leads to a larger number of departments exhibited by mistake.

Table 3: Heterogeneous Gaussian model.

Cluster intensity r	Mean lag l		Results of the following λ	$\widetilde{\lambda}$
1.0	1.0	Power	**0.082**	0.075
		TP	**0.612**	0.504
		FP	5.499	**3.480**
1.5	1.0	Power	**0.168**	0.135
		TP	**1.196**	0.960
		FP	4.548	**4.236**
2.0	1.0	Power	**0.390**	0.345
		TP	**2.597**	2.317
		FP	4.822	**4.712**
2.5	1.0	Power	**0.603**	0.571
		TP	**3.970**	3.800
		FP	**2.618**	3.354

4. Applications

4.1. An Application to Economic Data. For many years now, the spatial analysis of income inequality is an interesting scheme for many economists; for example, Shelnutt and Yao [11] investigated the income inequalities in the different counties of Arkansas and the interaction with economic growth, whereas Atems [12] analysed the Gini index of the incomes in the 3109 American counties. Among others, Deutsch et al.

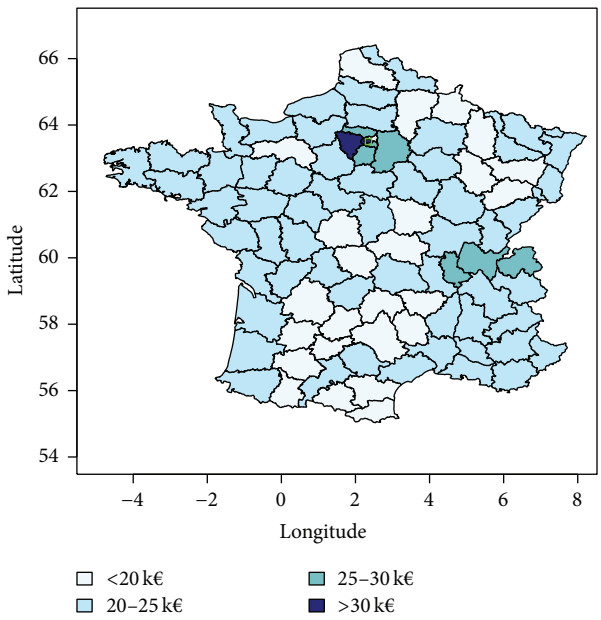

Figure 1: Median income.

[13] compared the spatial distribution of crime and economic inequalities in the US and found a correlation. These studies confirm that knowing precisely the geographic areas in which inequalities are more significant is of great interest. However, this search is usually done in a nonmathematical way. In this subsection, I show how the high-variance scan methods introduced later on can be useful for such a purpose. I apply the scan methods to an economic data set provided by a French institute, l'Institut National de la Statistique et des Etudes Economiques [14]. For each of the 94 departments in France, the median income for year 2009 has been computed and Figure 1 illustrates the results.

This figure clearly exhibits a cluster of wealthy departments in the North of France, next to Paris (whose location is given by the light square), whose significance has been assessed by Cucala [15] using the classical Gaussian-based scan statistic. However, we may wonder whether there is any high-variance cluster, enlightening a region where inequalities between departments are significant. To answer this question, I applied both high-variance scan statistics to this data set, once again using the set of circular potential clusters described in Section 2. Their associated p values are also estimated based on $T = 9999$ permuted samples. The most likely clusters are given by Figure 2. Note that, for convenience, the clusters I highlighted on this figure are not the circular areas exhibited by the scan methods, but the set of departments whose capital cities are included in those circular areas.

The most likely cluster exhibited by GLR scan statistic λ is exactly the area around Paris called Ile-de-France. This is not surprising since it includes 7 wealthy departments but also the department called Seine-Saint-Denis which is one of the French departments in which the incomes are much lower. This cluster is quite significant since its p value equals 0.0001: none of the 9999 likelihood-based scan statistics computed from the permuted samples was larger than the likelihood-based scan statistic computed from the observed sample. The cluster exhibited by GVR scan statistic $\widetilde{\lambda}$ is a bit less significant: its p value equals 0.0075. It contains 7 departments in Ile-de-France but also 6 neighboring departments in which the incomes are much lower.

4.2. An Application to Residuals Analysis. Even if the high-variance scan statistics have been introduced in a spatial setting, they can also be adapted to data with one-dimensional locations, like high-mean scan statistics. This includes data indexed by time but also any kind of data indexed along an axis. In order to illustrate this, I apply the high-variance scan methods to a set of residuals in a regression framework.

The data set I use results from an investigation of public school spending in the United States: the sample consists of 51 observations of per capita expenditure on public schools and per capita income for each state and the District of Columbia in 1979 [16]. As recommended by Greene [17], a linear regression model is fitted to explain the expenditure: the explanatory variables are the income and the squared income. A classical way to check the homoscedasticity assumption, that is, that the errors variance is constant across all the observations, is the analysis of the plot of the studentized residuals versus one of the explanatory variables, or alternatively the fitted values. In Figure 3, the studentized residuals are plotted versus the income. Note that, since the other explanatory variable of the model is the squared income, the plot of the residuals versus this explanatory variable or versus the fitted values would have been quite similar since the ordering of the residuals would have been exactly the same.

The variance of the residuals seems to be constant, except for larger values of the income. A few methods have been introduced in order to check that the residuals have equal variance: the Breusch-Pagan [18] test and the White test [19] are powerful to do so but they do not mention which of the residuals exhibit higher variance. The use of high-variance scan methods is of great interest to obtain such information.

Let $x_1, \ldots, x_n$ denote the studentized residuals, ordered such that x_i corresponds to the ith observation of the income in ascending order. For any one-dimensional cluster detection method, the potential clusters are all the sets of observations $\{x_i, \ldots, x_j\}$, $1 \leq i \leq j \leq n$. Here, since the variance of the marks needs to be estimated inside and outside of the potential cluster, intervals containing one observation ($j = i + 1$) or $n - 1$ observations ($j = i + n - 2$) are excluded. I computed both high-variance scan statistics on these data. The most likely cluster given by λ corresponds to the four last residuals, from solid line in Figure 3. The most likely cluster given by $\widehat{\lambda}$ only contains the two last residuals, from dotted line in Figure 3. This matches the conclusion of the simulation study: on this specific data set, the most likely cluster exhibited by the likelihood-based scan statistic is smaller than the one exhibited by the variance-ratio scan statistic. Based on $T = 9999$ permuted samples, their p values are, respectively, 0.0535 and 0.0426: this clearly indicates that the regression model is not valid when the income overcomes a certain level. Note that the Breusch-Pagan test and the White test both gave the same conclusion.

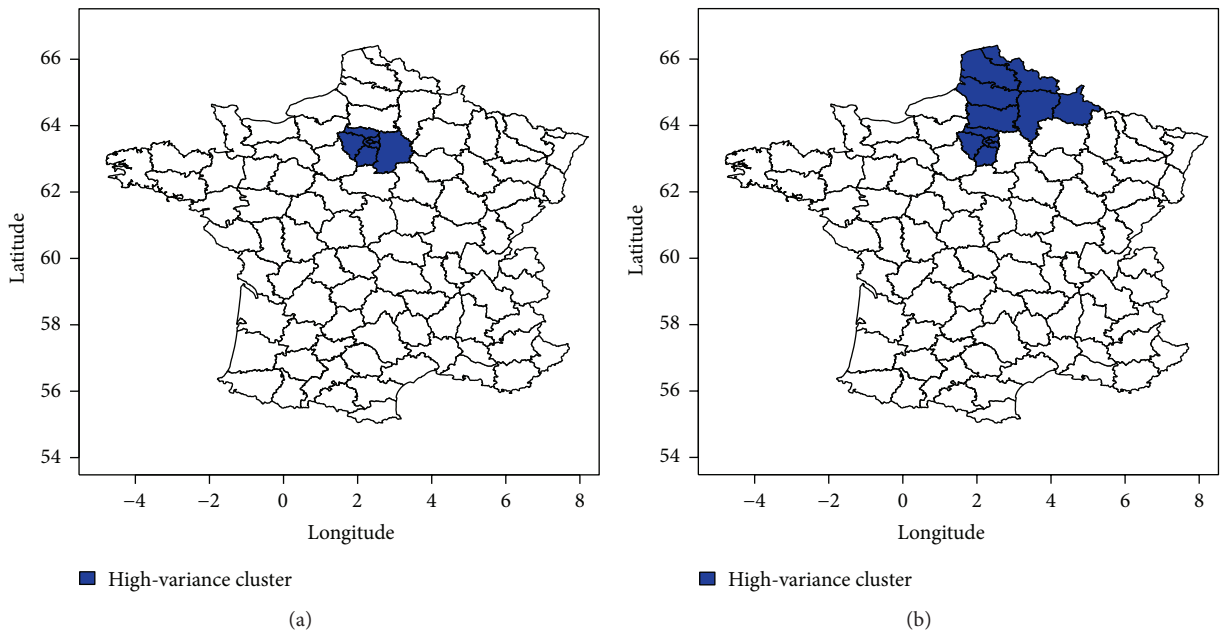

Figure 2: (a) Most likely cluster given by λ. (b) Most likely cluster given by $\widetilde{\lambda}$.

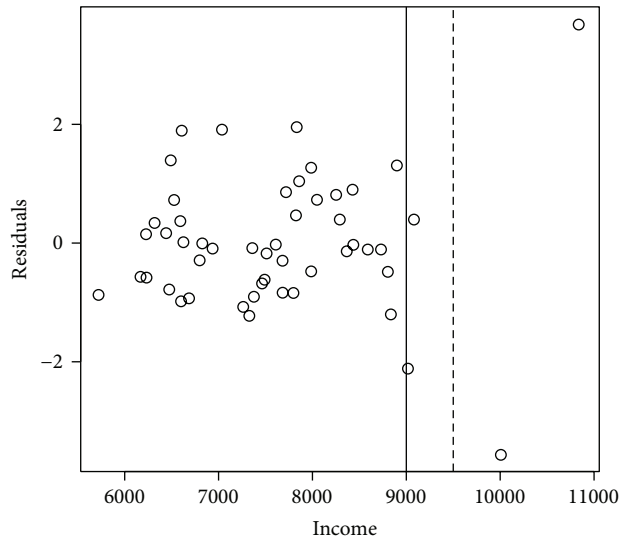

Figure 3: Residuals versus income.

5. Discussion

The scan statistics introduced in this paper allow one to detect high-variance clusters in marked point processes without any parameter to set up. According to the simulation study, the GVR scan statistic is slightly more powerful than the GLR one against any homogeneous clustering alternative: this result confirms that GLR scan statistics are not always the most efficient, even when the distribution of the marks is known. It also seems that, such as in the regression residuals analysis, the GLR scan statistic hardly detects clusters containing very few observations. Note that the GLR scan methods compare likelihood ratios which are, under the null hypothesis, not exactly equally distributed but only asymptotically when n tends to infinity, as stated by Wilks [20]. On the other hand, the GVR scan method compares the I_Z's which are equally distributed under the null hypothesis. Therefore, I should recommend using the GVR scan statistic instead of the GLR one.

In Section 4, I gave two examples of real data analysis through high-variance scan statistics: the search for inequalities hotspots on economic data and a homoscedasticity checking procedure on regression residuals. In the first example, the scan statistics clearly indicate a significant high-variance cluster, which is not very surprising, but also indicate precisely where social programs should be focused on in order to reduce the inequalities. Note also that a finer analysis could be done using the median income in each city instead of each department. In the second example, a new possibility is given by the high-variance scan methods to analyse regression residuals: they lead to the rejection of homoscedasticity hypothesis and focus on the individuals for which the linear regression model seems to be nonvalid. Let me mention that, among others, another application possibility would be the analysis of meteorological time series such as annual rainfall records in a specific station. Indeed, a high-variance scan test could determine whether the amplitude of rainfall increased in the last decades, because of global warming, in certain regions, leading to a higher frequency of very dry and very wet years.

The GVR scan statistic I introduce is based on the most famous test for equality of variances, the F-test. However, many other tests are available in the same purpose, as mentioned by Gartside [21]. For example, a scan statistic derived from the squared rank test introduced by Conover [22] could be set up, similar to the Mann-Whitney scan statistic for high-mean clusters defined by Cucala [23]. It

might be more suitable to marks whose distribution is not Gaussian.

The randomization procedure used to estimate the significance of both high-variance scan statistics is the most basic one, that is, random labelling. I focused on this procedure because it is the only one for which type I error remains equal to α whatever the underlying distribution of the data is. Since the GVR scan statistic itself is also distribution-free, this choice sounds natural. 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 other GLR based scan statistics when the significance is estimated through Monte Carlo simulation. As mentioned by Haining [24], 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 [25] and Zhang et al. [26]. Since the methods they introduced are designed for high-mean cluster detection, it could be interesting to adapt them for high-variance cluster detection: this might be the subject of a future work.

In the last few years, there have been more and more papers dealing with spatiotemporal cluster detection, such as Viel et al. [27]. It should be noted that the high-variance scan statistics can also be applied to spatiotemporal data using the spatiotemporal distance introduced by Demattei and Cucala [28].

In this work I only focused on the most likely cluster but the detection of secondary clusters is straightforward using the method proposed by Zhang et al. [29]: once a significant cluster is found, remove the data included in that cluster and restart the analysis.

Finally, I should emphasize that the high-variance scan tests could be adjusted for any continuous covariate adjustment as proposed by Klassen et al. [30], 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.

Appendix

Computing the ML Estimate under the Alternative Hypothesis

The ML estimate under $H_{1,Z}$ is the value for $(\mu_{Z,\overline{Z}}, \sigma_Z^2, \sigma_{\overline{Z}}^2)$ maximizing the log-likelihood

$$\begin{aligned} &-\frac{n_Z}{2}\log\left(2\pi\sigma_Z^2\right) - \frac{1}{2\sigma_Z^2}\sum_{i:s_i\in Z}\left(x_i-\mu_{Z,\overline{Z}}\right)^2 \\ &-\frac{n_{\overline{Z}}}{2}\log\left(2\pi\sigma_{\overline{Z}}^2\right) - \frac{1}{2\sigma_{\overline{Z}}^2}\sum_{i:s_i\in\overline{Z}}\left(x_i-\mu_{Z,\overline{Z}}\right)^2. \end{aligned} \tag{A.1}$$

When $\mu_{Z,\overline{Z}}$ is fixed, the value for σ_Z^2 maximizing

$$-\frac{n_Z}{2}\log\left(2\pi\sigma_Z^2\right) - \frac{1}{2\sigma_Z^2}\sum_{i:s_i\in Z}\left(x_i-\mu_{Z,\overline{Z}}\right)^2 \tag{A.2}$$

is

$$\begin{aligned} \sigma_Z^{*2} &= \frac{1}{n_Z}\sum_{i:s_i\in Z}\left(x_i-\mu_{Z,\overline{Z}}\right)^2 \\ &= \overline{X^2}_Z - 2\mu_{Z,\overline{Z}}\overline{X}_Z + \left(\mu_{Z,\overline{Z}}\right)^2. \end{aligned} \tag{A.3}$$

The same result holds for $\sigma_{\overline{Z}}^{*2}$. Thus, inserting these expressions in the log-likelihood function, $\mu_{Z,\overline{Z}}^*$ is the value for μ maximizing the function

$$\begin{aligned} g(\mu) &= -\frac{n_Z}{2}\log(2\pi) - \frac{n_Z}{2}\log\left(\overline{X^2}_Z - 2\mu\overline{X}_Z + \mu^2\right) \\ &\quad - \frac{n_Z}{2} \\ &= -\frac{n_{\overline{Z}}}{2}\log(2\pi) - \frac{n_{\overline{Z}}}{2}\log\left(\overline{X^2}_{\overline{Z}} - 2\mu\overline{X}_{\overline{Z}} + \mu^2\right) \\ &\quad - \frac{n_{\overline{Z}}}{2}. \end{aligned} \tag{A.4}$$

The derivative of this function is

$$\begin{aligned} &g'(\mu) \\ &= -\frac{n_Z}{2}\frac{-2\overline{X}_Z + 2\mu^2}{\overline{X^2}_Z - 2\mu\overline{X}_Z + \mu^2} \\ &\quad - \frac{n_{\overline{Z}}}{2}\frac{-2\overline{X}_{\overline{Z}} + 2\mu^2}{\overline{X^2}_{\overline{Z}} - 2\mu\overline{X}_{\overline{Z}} + \mu^2} \\ &= \frac{n}{2}\frac{h(\mu)}{\left(\overline{X^2}_Z - 2\mu\overline{X}_Z + \mu^2\right)\left(\overline{X^2}_{\overline{Z}} - 2\mu\overline{X}_{\overline{Z}} + \mu^2\right)}, \end{aligned} \tag{A.5}$$

where

$$\begin{aligned} h(\mu) &= \mu^3 + \frac{-n_Z\overline{X}_Z - 2n_Z\overline{X}_{\overline{Z}} - n_{\overline{Z}}\overline{X}_{\overline{Z}} - 2n_{\overline{Z}}\overline{X}_Z}{n}\mu^2 \\ &\quad + \frac{n_Z\overline{X^2}_{\overline{Z}} + 2n\overline{X}_Z\overline{X}_{\overline{Z}} + n_{\overline{Z}}\overline{X^2}_Z}{n}\mu \\ &\quad + \frac{-n_Z\overline{X}_Z\overline{X^2}_{\overline{Z}} - n_{\overline{Z}}\overline{X}_{\overline{Z}}\overline{X^2}_Z}{n}. \end{aligned} \tag{A.6}$$

Following Cardano's method [6], the three roots of this cubic function $h(\cdot)$ can be computed. Therefore, $\mu_{Z,\overline{Z}}^*$ is either the only real root (if the two others are complex) or the one of the three roots in which the function $g(\cdot)$ is maximized (if the three roots are real).

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

References

[1] J. Naus, *Clustering of random points in the line and plane [Ph.D. thesis]*, Rutgers University, New Brunswick, NJ, USA, 1963.

[2] J. Glaz, J. Naus, and S. Wallenstein, *Scan Statistics*, Springer, New York, NY, USA, 2001.

[3] N. Cressie, "On some properties of the scan statistic on the circle and the line," *Journal of Applied Probability*, vol. 14, no. 2, pp. 272–283, 1977.

[4] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics. Theory and Methods*, vol. 26, no. 6, pp. 1481–1496, 1997.

[5] M. Kulldorff, L. Huang, and K. Konty, "A scan statistic for continuous data based on the normal probability model," *International Journal of Health Geographics*, vol. 8, article 58, 2009.

[6] N. Jacobson, *Basic Algebra*, Dover, 2009.

[7] M. Dwass, "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, vol. 28, pp. 181–187, 1957.

[8] L. Cucala, "A hypothesis-free multiple scan statistic with variable window," *Biometrical Journal*, vol. 50, no. 2, pp. 299–310, 2008.

[9] G. Saporta, *Probabilités, Analyse des Données et Statistique*, Technip, Paris, France, 2011.

[10] Y. Dodge, *The Oxford Dictionary of Statistical Terms*, Oxford University Press, 2003.

[11] J. Shelnutt and V. Yao, "A spatial analysis of income inequality in Arkansas at the county level: evidence from tax and commuting data," *Regional Economic Development*, vol. 1, pp. 52–65, 2005.

[12] B. Atems, "A note on the differential regional effects of income inequality: empirical evidence using U.S. county-level data," *Journal of Regional Science*, vol. 53, no. 4, pp. 656–671, 2013.

[13] J. Deutsch, U. Spiegel, and J. Templeman, "Crime and income inequality: an economic approach," *Atlantic Economic Journal*, vol. 20, no. 4, pp. 46–54, 1992.

[14] Institut National de la Statistique et des Études Économiques, http://www.insee.fr/.

[15] L. Cucala, "A distribution-free spatial scan statistic for marked point processes," *Spatial Statistics*, vol. 10, pp. 117–125, 2014.

[16] US Department of Commerce, *Statistical Abstract of the United States*, US Government Printing Office, Washington, DC, USA, 1979.

[17] W. Greene, *Econometric Analysis*, Macmillan Publishing Company, New York, NY, USA, 2nd edition, 1993.

[18] T. S. Breusch and A. R. Pagan, "A simple test for heteroscedasticity and random coefficient variation," *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979.

[19] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, vol. 48, no. 4, pp. 817–838, 1980.

[20] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[21] P. S. Gartside, "A study of methods for comparing several variances," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 342–346, 1972.

[22] W. Conover, *Practical Nonparametric Statistics*, Wiley, New York, NY, USA, 1980.

[23] L. Cucala, "A Mann-Whitney scan statistic for continuous data," *Communications in Statistics. Theory and Methods*, In press.

[24] R. Haining, *Spatial Data Analysis: Theory and Practice*, Cambridge University Press, 2003.

[25] J. M. Loh and Z. Zhu, "Accounting for spatial correlation in the scan statistic," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 560–584, 2007.

[26] T. Zhang, Z. Zhang, and G. Lin, "Spatial scan statistics with overdispersion," *Statistics in Medicine*, vol. 31, no. 8, pp. 762–774, 2012.

[27] J.-F. Viel, N. Floret, and F. Mauny, "Spatial and space-time scan statistics to detect low rate clusters of sex ratio," *Environmental and Ecological Statistics*, vol. 12, no. 3, pp. 289–299, 2005.

[28] C. Demattei and L. Cucala, "Multiple spatio-temporal cluster detection for case event data: an ordering-based approach," *Communications in Statistics. Theory and Methods*, vol. 40, no. 2, pp. 358–372, 2011.

[29] Z. Zhang, M. Kulldorff, and R. Assunção, "Spatial scan statistics adjusted for multiple clusters," *Journal of Probability and Statistics*, vol. 2010, Article ID 642379, 11 pages, 2010.

[30] A. C. Klassen, M. Kulldorff, and F. Curriero, "Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors," *International Journal of Health Geographics*, vol. 4, article 1, 2005.