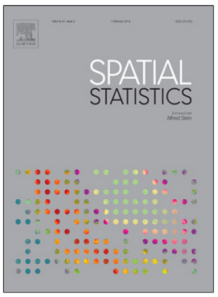
Accepted Manuscript



A multivariate nonparametric scan statistic for spatial data

Lionel Cucala, Michaël Genin, Florent Occelli, Julien Soula

PII: S2211-6753(18)30119-2
DOI: https://doi.org/10.1016/j.spasta.2018.10.002
Reference: SPASTA 330

To appear in: *Spatial Statistics*

Received date : 8 June 2018
Accepted date : 8 October 2018

Please cite this article as: Cucala L., et al., A multivariate nonparametric scan statistic for spatial data. *Spatial Statistics* (2018), https://doi.org/10.1016/j.spasta.2018.10.002

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

A Multivariate nonparametric scan statistic for spatial data

Lionel CUCALA[a], Michaël GENIN[b], Florent OCCELLI[c], Julien SOULA[b]

[a] *IMAG, Université de Montpellier, CNRS, Montpellier, France.*
[b] *Santé Publique: épidémiologie et qualité des soins EA 2694, Université de Lille, France.*
[c] *IMPact de l'Environnement Chimique sur la Santé humaine EA 4483, CHU Lille, Institut Pasteur de Lille, Université de Lille, France.*

Abstract

This paper introduces a nonparametric scan method for multivariate data indexed in space. Contrary to many other scan methods, it does not rely on a generalized likelihood ratio but is completely distribution-free as it is based on so-called multivariate ranks. This spatial scan test seems to be more reliable for analysing data that are not Gaussian, such as environmental measurements. We apply this method to a data set recording the levels of metallic pollutants for two areas in the North of France.

Keywords: Spatial statistics, Scan Statistics, Cluster detection.

1. Introduction

The original scan statistic was defined [22] to be the maximum number of events observed within a window with given shape and volume, known as the scanning window, as it moves over the observation domain in a continuous manner. The distribution of this statistic under the null hypothesis has been widely studied [12]. However, its main drawback is that the length or volume of the potential clusters must be fixed a priori. A second drawback is that, when dealing with spatial data, it is only suitable when the underlying population measure is uniform.

These issues were circumvented by Nagarwalla [23] and Kulldorff [16] using a concentration indicator based on likelihood ratios, which is able to

Email address: lionel.cucala@umontpellier.fr (Lionel CUCALA)

ACCEPTED MANUSCRIPT

compare all possible windows, whatever their sizes and their population measures. These innovations gave birth to a very large number of application papers in many different fields: astronomy, forestry, ecology, genetics, epidemiology, ... [20, 11, 1]. These likelihood-based methods were also adapted for random variables indexed in space: depending on the nature of the variable, the proposed scan statistic is a generalized likelihood-ratio issued from an adequate parametric model.

Sometimes, such as in environmental surveillance, numerous variables are collected on the same locations. A natural question arises: how to detect a spatial cluster in which the measurements of these variables are significantly different? The multivariate scan statistic proposed by Kulldorff et al. [18] combines the univariate scan statistics associated to each variable such as if they were independent. Very recently, Cucala et al. [7] introduced a multivariate Gaussian scan statistic which takes into account the covariances between different variables. The test based on this statistic performs very well against Gaussian alternatives but faces problems when the data are not Gaussian, which is often the case when dealing with environmental data exhibiting extreme values.

However, in the last few years, some alternatives to likelihood-ratio based scan statistics arised [5]. In the univariate setting, a nonparametric scan statistic, only relying on the ranks, has been introduced separately by Cucala [4] and Jung and Cho [14], based on the Wilcoxon-Mann-Whitney test. Since a multivariate extension of this test has been proposed by Oja and Randles [26], a scan procedure based on this multivariate nonparametric test should be investigated.

In this paper we develop a scan statistic for multivariate continuous data that is based on the multivariate ranks and is thus completely nonparametric. In Section 2, we explain how the statistic introduced by Oja and Randles [26] can be maximised on a finite set of potential clusters, giving birth to a spatial scan statistic. We also describe the permutation-based procedure that provides the statistical significance associated to this statistic. The behaviour of this method is investigated through a simulation procedure and the analysis of real environmental data in Section 3. Finally, we consider future development in the Conclusion.

ACCEPTED MANUSCRIPT

2. A nonparametric scan statistic for multivariate data

Consider p numerical variables, denoted by $X^1, \cdots, X^p$, which are measured in n different spatial locations $s_1, \cdots, s_n$ included in $D \subset \mathbb{R}^2$, the observation domain. All these measurements are recorded in a $n \times p$ matrix

$$X = (x_i^k), \quad 1 \leq i \leq n, \quad 1 \leq k \leq p.$$

The $1 \times p$ vector containing all the measures in s_i, and corresponding to the i^{th} row of matrix X, is denoted by X_i: following the terminology of point process theory, we will call X_i the mark associated to location s_i. Our goal is to detect the spatial area $Z \subset D$ in which the marks are significantly different than elsewhere.

A scan statistic is nothing but the maximum of a concentration index observed in a collection of potential clusters. It thus depends only on the set of potential clusters and on the concentration index that should be used.

There is abundant literature on the choice of variable-size potential clusters [20] but two main possibilities can be identified. On the one hand, one may focus on clusters having a constrained shape: circular [16], elliptic [17] or any other shape. On the other hand, you may set up a procedure only based on distances between spatial locations: for example, Duczmal and Assunção [9] investigate irregularly shaped windows via a simulated annealing strategy, while Dematteï et al. [8] introduce an ordering from one location to its closest neighbour.

In this article, for sake of simplicity, we consider the circular clusters introduced by Kulldorff [16]. The set of potential clusters, denoted by $\mathcal{D}$, is the set of discs 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 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.

Most of the time, the concentration index which is maximized on this set of potential clusters is nothing but a likelihood ratio. For example, as said in the Introduction, Cucala et al. [7] introduced a multivariate Gaussian-based scan statistic relying on the likelihood ratio between two hypotheses: the multivariate marks are supposed to be normally-distributed and independent but the null hypothesis considers equal mean vectors and covariance matrices whereas the alternative hypothesis considers equal covariance matrices but

ACCEPTED MANUSCRIPT

different mean vectors inside and outside the potential cluster. As stated by Cucala [5], likelihood ratios are not always the best tools to compare potential clusters. Thus, a nonparametric concentration index could be more efficient.

In the univariate setting, one of the most popular nonparametric methods to test whether the distributions of two samples of continuous observations are equal, or more precisely whether their medians are equal, is the Mann-Whitney test [21], also called Wilcoxon rank-sum test. It just relies on the comparison between the sums of the ranks related to the two samples. Recently, a multivariate extension of this test has been proposed by Oja and Randles [26]. The definition of multivariate ranks associated to marks $X_1, \cdots, X_n$ in $\mathbb{R}^p$ depends on the spatial sign function

$$S(x) = \begin{cases} ||x||^{-1}x & \text{if } x \neq 0, \\ 0 & \text{if } x = 0, \end{cases}$$

for all $x \in \mathbb{R}^p$, $||x||$ being the L_2 norm. The function value is just a direction (a point on the unit p sphere). The multivariate signs associated to $X_1, \cdots, X_n$ are

$$S_i = S(A_x X_i), \quad \text{for } i = 1, \cdots, n$$

where the matrix A_x, called Tyler's transformation, makes the covariance matrix of the multivariate signs equal to $\frac{1}{p} I_p$, i.e. the covariance matrix associated to the uniform distribution on the unit p sphere. Note that this matrix can be easily computed using an iterative procedure. The multivariate ranks can now be defined as

$$R_i = \frac{1}{n} \sum_{j=1}^{n} S_{i,j}$$

where $S_{i,j} = S\big(A_x(X_i - X_j)\big), \quad i, j = 1, \cdots, n$, are signs of transformed differences, still using the Tyler's transformation.

Now that the multivariate ranks are defined, we can go back to the cluster detection problem. From now on, the marks $X_1, \cdots, X_n$ are assumed to be independent: this is a very classical assumption when introducing scan statistics. The null hypothesis H_0, corresponding to the absence of any cluster in the data, is the following: "the random marks $X_1, \cdots, X_n$ are all identically distributed, whatever the associated locations". Note that, contrary to likelihood-based scan methods, we do not make any assumption concerning the distribution of the marks. Let $Z \in \mathcal{D}$ be any potential cluster and Z^c

ACCEPTED MANUSCRIPT

its complement. In order to test for a difference of distribution of the marks between Z and Z^c, Oja and Randles [26] propose to use the multivariate extension of the Wilcoxon-Mann-Whitney statistic

$$U^2(Z) = \frac{p}{c_x^2}[n_Z||\bar{R}_Z||^2 + n_{Z^c}||\bar{R}_{Z^c}||^2$$

where

$$\begin{cases} n_Z &= \sum_{i=1}^{n} \mathbb{1}(s_i \in Z) \\ \bar{R}_Z &= \frac{1}{n_Z} \sum_{i:s_i \in Z} R_i \\ c_x^2 &= \sum_{i=1}^{n} R_i^T R_i \end{cases}$$

Under H_0, the limiting distribution of $U^2(Z)$ is the chi-squared distribution with p degrees of freedom. Since this does not depend on n_Z, the number of observations in Z, we believe the concentration index $U^2(Z)$ is relevant to compare potential clusters having different population sizes.
This concentration index may now be maximised on the set of potential clusters previously defined. The multivariate nonparametric (MNP) scan statistic is

$$\lambda_{MNP} = \max_{Z \in \mathcal{D}} U^2(Z)$$

and the potential cluster for which this maximum is obtained,

$$\hat{C} = \arg\max_{Z \in \mathcal{D}} U^2(Z),$$

is called the most likely cluster. Note that, when $p = 1$, this scan statistic is exactly similar to the Mann-Whitney scan statistic introduced separately by Cucala [4] and Jung and Cho [14].

The most likely cluster $\hat{C}$ is the area in which the difference of marks is the most significant. However, it is important to evaluate this significance, i.e. to estimate the probability of such a difference under the null hypothesis. Computing the null distribution of a variable-window scan statistic is untractable since it is a maximum computed on intersecting areas, which

ACCEPTED MANUSCRIPT

implies multiple correlations. Therefore, the only way to estimate the significance is via random simulations. Contrary to Kulldorff et al. [19], we decided not to simulate new marks since the null hypothesis H_0 is distribution-free. Thus the only way to obtain simulated datasets satisfying H_0 is by randomly associating marks and locations: this method is called random labelling [3]. Note that the random labelling of the marks X_i's is exactly similar to the random labelling of their multivariate ranks R_i's: we decided to perform the permutation of the R_i's since it allows to compute the multivariate ranks only once and to reduce drastically the computation time. Once these datasets have been simulated, the associated multivariate nonparametric scan statistics $\lambda_{MNP}^{(1)}, \cdots, \lambda_{MNP}^{(T)}$ are computed and compared to λ_{MNP}, the multivariate nonparametric scan statistic computed on the original dataset. As stated by Dwass [10], the proportion of simulated statistics greater than the original statistic, i.e. the p-value, is a consistent estimator for the significance.

3. Applications

3.1. A simulation study

A simulation study was conducted to compare the multivariate Gaussian scan statistic (λ_{MG}) and the multivariate nonparametric scan statistic (λ_{MNP}). Artificial datasets were generated using the geographic locations of the $n = 94$ French *départements* (administrative areas). One should notice that each of the locations has been defined by the administrative center of each *département*. Let denote by C the simulated cluster defined as a set of *départements* according to three size configurations: 10, 15 and 20. As an appendix (Figure A.4), we provide maps of these clusters. We consider $p = 5$ variables distributed, in each location $i = 1, \ldots, 94$, according to three different multivariate distributions: Gaussian, Weibull and exponential. For each model, let consider μ_0 as the mean vector outside the cluster and the parameter β denoting the difference of means inside and outside the cluster. In what follows, we will refer to this parameter by using the expression *cluster intensity*. Let define, for each model, the correlation matrix $\mathrm{corr}(X)$ related to covariance matrix Σ as follows:

$$\mathrm{corr}(X) = \begin{pmatrix} 1 & \rho & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 & 0 \\ \rho & \rho & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho \\ 0 & 0 & 0 & \rho & 1 \end{pmatrix},$$

ACCEPTED MANUSCRIPT

where ρ is the parameter that controls the correlation between the X^j's.
For the Gaussian model, we considered $X_i \sim \mathcal{N}_p(\mu_i, \Sigma)$ where

$$\mu_i = \begin{cases} \left(\mu_0^j + \beta\right)_{1\leq j\leq 5}^T & \text{if } s_i \in C, \\ \\ (\mu_0)^T & \text{otherwise.} \end{cases}$$

For the Weibull model, we considered $X_i \sim \mathcal{W}_p(\lambda_i, k, \Sigma)$ where

$$\lambda_i = \begin{cases} \left(\frac{\mu_0^j+\beta}{\Gamma(1+\frac{1}{k_j})}\right)_{1\leq j\leq 5}^T & \text{if } s_i \in C, \\ \\ \left(\frac{\mu_0^j}{\Gamma(1+\frac{1}{k_j})}\right)_{1\leq j\leq 5}^T & \text{otherwise,} \end{cases}$$

and $k_j = 2$ corresponds to j^{th} element of the vector of shape parameters and $\Gamma(x)$ stands for the Gamma function at point x.
For the Exponential model, we considered $X_i \sim \mathcal{E}_p(\lambda_i, \Sigma)$ where

$$\lambda_i = \begin{cases} \left(\frac{1}{\mu_0^j+\beta}\right)_{1\leq j\leq 5}^T & \text{if } s_i \in C, \\ \\ \left(\frac{1}{\mu_0^j}\right)_{1\leq j\leq 5}^T & \text{otherwise.} \end{cases}$$

In this study, the values of μ_0 have been set to 10. For different values of the parameters, $S = 1000$ simulated data sets have been generated. The comparison of the two methods was performed using three distinct criteria: the power of the method, the true-positive rate (TP) and the false-positive rate (FP).
The power of the method was defined as the proportion of data sets highlighting a significant cluster, considering the type I error equal to 0.05 and $T = 999$ permuted samples. To calculate the TP, we considered for each simulated data set $s = 1, \ldots, S$, the number of *départements* included both in the most likely cluster $\hat{C}_s$ and in the simulated cluster C_s divided by the number of *départements* included in C_s. The TP was defined as the average of these proportions over all simulated data sets:

$$\text{TP} = \frac{1}{S}\sum_{s=1}^{S} \frac{\text{card}\left(\hat{C}_s \cap C_s\right)}{\text{card}(C_s)}.$$

ACCEPTED MANUSCRIPT

Similarly, to calculate the FP, we considered the number of *départements* included in $\hat{C}_s$ but not in C_s divided by the number of *départements* not included in C_s for each simulated data set s. The FP was defined as the average of these proportions over all simulated data sets:

$$\text{FP} = \frac{1}{S}\sum_{s=1}^{S}\frac{\text{card}\left(\hat{C}_s \cap C_s^c\right)}{\text{card}(C_s^c)}.$$

The results of the simulation study are presented in Figure 1 (see Appendix (Table B.2) for more detailed results).

For Gaussian and Weibull models, both methods tend to have equivalent powers when the size of the simulated cluster increases, regardless of the value of the parameters. For the exponential model, the nonparametric scan statistic shows a higher power, regardless of the size of the simulated cluster and the values of the parameters. The power difference between the two methods increases with the size of the simulated cluster. The true-positive rate is consistently higher for λ_{MNP} and the difference between the two methods increases with the skewness of the distribution. This implies that the parametric method tends to exhibit significant clusters smaller than the other one leading to a large number of false-negative *départements*. On the other hand, the false-positive rate is often higher for λ_{MNP}: the method tends to show significant clusters larger than the other one. However, for the non-parametric method, the false-positive rate does not exceed 11%.

3.2. An application to environmental data

We applied the scan statistic model to the same real environmental biomonitoring data as those presented in the study of Cucala et al. [7]. Briefly, thalli of the foliose lichen Xanthoria parietina were collected in respectively 128 and 59 point locations in the Lille European Metropole and the Dunkerque agglomeration for the analysis of 14 trace elements (TE) concentrations: aluminium (Al), antimony (Sb), arsenic (As), cadmium (Cd), cobalt (Co), chrome (Cr), copper (Cu), lead (Pb), manganese (Mn), mercury (Hg), nickel (Ni), titanium (Ti), zinc (Zn), and vanadium (V). Most of these TE are highly correlated to the others, but some are slightly less correlated to the others. All TE show heavy-tailed distributions characterized by very high extreme values. Plots of these statistical distributions can be found in the Appendix (Figures C.5 and C.6). We also calculated the Mean Impregna-

ACCEPTED MANUSCRIPT

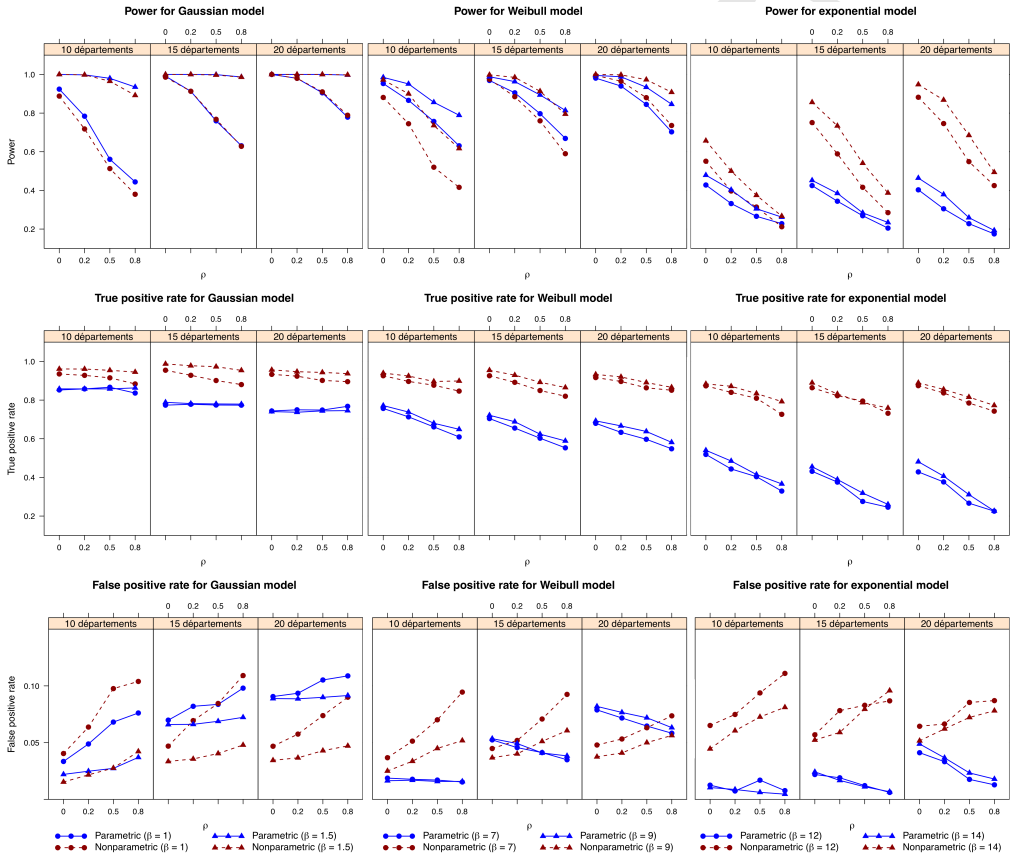


Figure 1: Simulation study - Comparison of the multivariate Gaussian scan statistics (λ_{MG}) and the multivariate nonparametric scan statistic (λ_{MNP}) through three different multivariate models (Gaussian, Weibull, Exponential) and according to 3 different size of the simulated cluster : 10, 15 and 20 *départements*. For each combination of parameters, power, true and false-positive rates are presented.

ACCEPTED MANUSCRIPT

tion Ratio (MIR) for each point location. This multimetallic index allows to consider the general pollution status. See Occelli et al. [24] for more details.

In order to identify areas exhibiting an excessive level of pollutants, the parametric (λ_{MG}) and the non-parametric (λ_{MNP}) multivariate scan statistics were applied to each dataset, considering $T = 999$ permutations. Both methods detected a significant cluster on the Lille European Metropole ($p = 0.001$ for λ_{MG} and $p = 0.002$ for λ_{MNP}; Figure 2). Only one location, situated on the North (the grey circle), was included in the λ_{MG} cluster. The median [interquartile range : IQR] of each TE for locations inside/outside the cluster are presented in Table 1. The location identified inside the cluster presents the highest contamination for 12 of the 14 TE, and is amongst the highest values for the two others. The latter is considered as an extreme data which is between 4 and 48 times higher than the median observed outside the cluster. MIR values are 16.35 inside and 1.32 [1.01-1.92] outside the cluster. The λ_{MNP} cluster includes 42 locations centered on the city of Lille (the red circle), and does not contain the one detected with λ_{MG}. MIR values are 1.74 [1.28-2.46] inside and 1.18 [0.91-1.66] outside the cluster, and TE concentrations are always higher inside than outside.

For the Dunkerque agglomeration, each method detected a significant cluster ($p = 0.001$ for both; Figure 3). The λ_{MG} cluster includes only one location (the grey circle), which presents the highest contamination for 7 of the 14 TE, and is amongst the highest values for 4 others. Concentrations are amongst the lowest for Ti, medium for As, and no data were available for V, due to the laboratory analysis. The latter is also considered as an extreme data which is between 1.5 and 128 times higher than the median observed outside the cluster. MIR values are 49.26 inside and 2.16 [1.42-4.20] outside. The λ_{MNP} cluster includes 3 locations (the red circle), which comprise the one detected with λ_{MG}. MIR values are 10.16 [7.51-29.71] inside and 2.14 [1.39-4.04] outside the cluster, and TE concentrations are higher inside than outside, expected for Ti. All these results are detailed in Table 1.

4. Discussion

The method introduced in this paper is an automatic tool for investigating multivariate data in a spatial context without choosing any distribution and setting up any parameter. Such as any scan method, it can be used by non-statisticians to highlight spatial areas in which things are different, leading to further investigation. The nonparametric multivariate scan statistic

ACCEPTED MANUSCRIPT

Table 1: Comparison of the multivariate gaussian scan statistic (λ_{MG}) and the multivariate nonparametric scan statistic (λ_{MNP}) on environmental data (Trace elements) considering the Lille European Metropole and the Dunkerque agglomeration areas. For each area and method, the trace elements are described (median [first quartile ; third quartile]) inside and outside the significant detected cluster.

	Lille European Metropole			
	λ_{MG} ($p = 0.001$)		λ_{MNP} ($p = 0.002$)	
Trace elements	Out ($n = 127$)	In ($n = 1$)	Out ($n = 84$)	In ($n = 44$)
Aluminium	690 [496;1162]	3057	648 [458;1119]	761 [576;1260]
Chrome	3.21 [2.26;5.17]	45.1	3.00 [2.06;5.03]	3.68 [2.76;6.54]
Copper	12.6 [8.83;19.2]	259	10.9 [7.69;17.4]	16.9 [11.8;21.8]
Arsenic	0.78 [0.61;1.21]	5.10	0.77 [0.56;1.18]	0.85 [0.67;1.41]
Mercury	0.11 [0.09;0.13]	0.81	0.10 [0.09;0.11]	0.12 [0.10;0.16]
Cadmium	0.46 [0.30;0.83]	10.6	0.35 [0.25;0.72]	0.80 [0.46;1.34]
Manganese	44.0 [33.0;55.9]	309	39.5 [30.8;52.5]	50.2 [39.0;58.0]
Cobalt	0.52 [0.36;0.76]	9.62	0.48 [0.31;0.74]	0.57 [0.45;0.84]
Antimony	1.09 [0.66;2.00]	14.0	0.87 [0.55;1.42]	1.54 [1.08;2.62]
Nickel	2.33 [1.58;3.21]	72.9	2.15 [1.44;2.98]	2.52 [1.97;3.71]
Vanadium	2.79 [2.00;4.29]	15.0	2.62 [1.90;4.29]	3.00 [2.14;4.41]
Lead	18.0 [9.89;33.0]	873	13.0 [8.00;24.5]	34.5 [18.0;46.5]
Titanium	13.0 [10.0;19.5]	76.0	12.0 [9.00;19.2]	14.5 [12.0;20.0]
Zinc	87.1 [59.5;124]	1583	72.7 [52.3;111]	113 [83.5;168]
	Dunkerque agglomeration			
	λ_{MG} ($p = 0.001$)		λ_{MNP} ($p = 0.001$)	
Trace elements	Out ($n = 58$)	In ($n = 1$)	Out ($n = 56$)	In ($n = 3$)
Aluminium	1126 [806;1553]	3179	1105 [788;1523]	3179 [2379;3336]
Chrome	10.0 [5.78;18.8]	1172	9.60 [5.65;17.9]	194 [112;683]
Copper	11.5 [8.72;26.6]	95.5	11.1 [8.57;21.9]	31.6 [29.1;63.5]
Arsenic	1.58 [0.94;2.57]	2.33	1.58 [0.92;2.55]	2.33 [2.00;2.91]
Mercury	0.16 [0.12;0.21]	1.02	0.16 [0.12;0.21]	0.21 [0.20;0.62]
Cadmium	0.58 [0.40;0.98]	30.1	0.57 [0.40;0.95]	4.79 [3.45;17.4]
Manganese	273 [142;584]	1730	260 [140;568]	859 [700;1295]
Cobalt	0.80 [0.50;1.47]	13.1	0.77 [0.48;1.39]	3.58 [2.89;8.34]
Antimony	0.94 [0.52;1.87]	3.49	0.91 [0.52;1.61]	3.57 [3.53;3.79]
Nickel	6.17 [4.38;9.75]	795	6.15 [4.28;9.37]	107 [71.7;451]
Vanadium	8.52 [5.04;14.0]	N.A.	7.77 [4.90;14.7]	12.0 [6.00;12.3]
Lead	17.0 [9.01;37.8]	372	16.0 [9.00;35.5]	53.0 [46.0;212]
Titanium	21.5 [9.50;40.0]	0.18	21.5 [10.5;41.2]	0.24 [0.21;19.1]
Zinc	97.9 [71.6;182]	2408	96.7 [71.3;160]	512 [379;1460]

ACCEPTED MANUSCRIPT

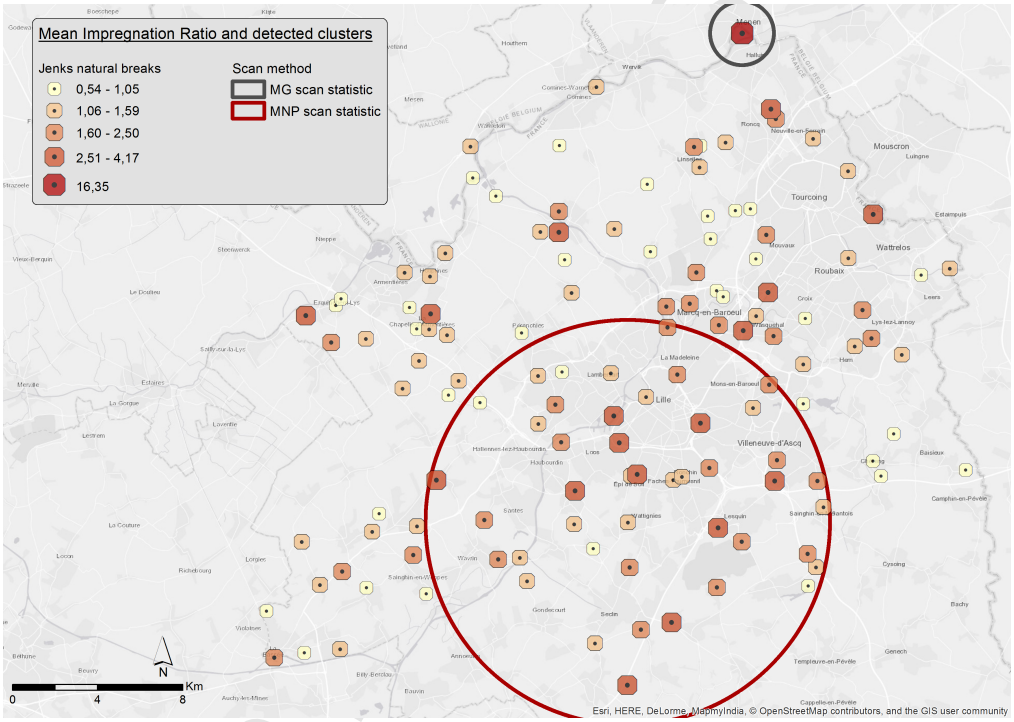


Figure 2: Application study - Mean Impregnation Ratios by Jenks classification and detected clusters for the multivariate Gaussian scan statistics (MG) and the multivariate nonparametric scan statistic (MNP), European Lille Metropole.

ACCEPTED MANUSCRIPT

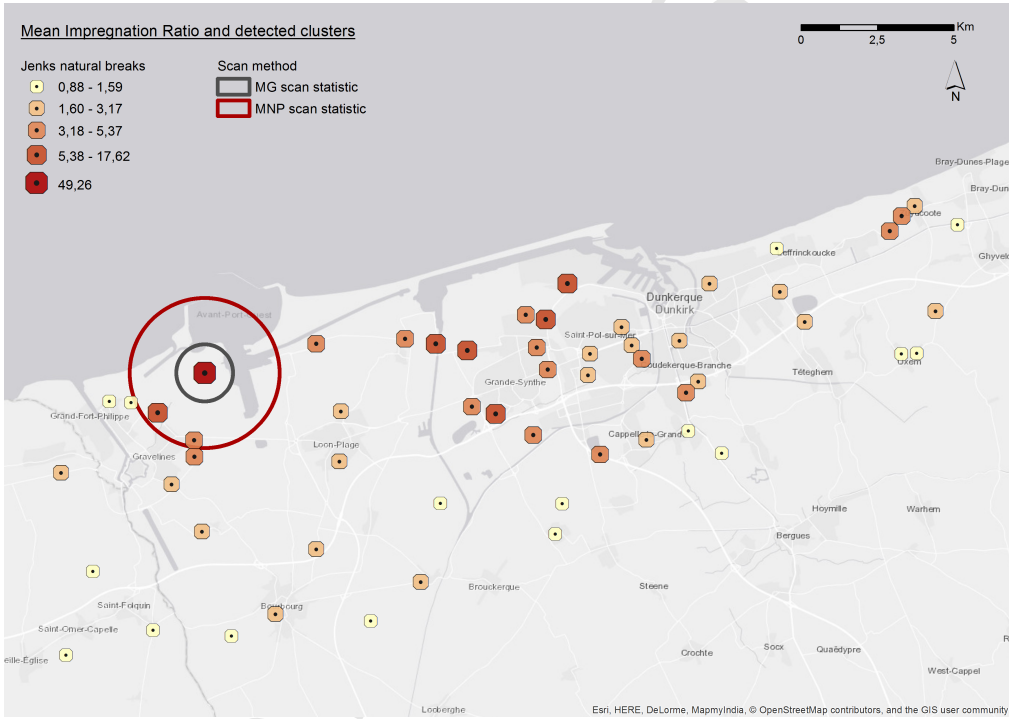


Figure 3: Application study - Mean Impregnation Ratios by Jenks classification and detected clusters for the multivariate Gaussian scan statistics (MG) and the multivariate nonparametric scan statistic (MNP), Dunkerque agglomeration.

ACCEPTED MANUSCRIPT

mainly addresses a major environmental-health issue: detecting areas where populations are multi-exposed to environmental pollutions.

We used simulation (artificial dataset) and real (environmental dataset) applications to compare our nonparametric method to the multivariate Gaussian scan statistic. As expected, the simulation study shows that both methods tend to have equivalent powers when the distribution of variables has a low asymmetry, regardless of the correlation between variables and the size of the cluster. Conversely, for heavy-tailed distribution, the non-parametric method shows a higher power than the parametric one whatever the correlation between variables and the size of the cluster. The non-parametric method also shows a higher true-positive rate, which implies that the parametric method tends to identify smaller clusters characterized by the presence of extreme values. This characteristic has also been observed on the environmental dataset. In our two examples, extreme high values for most of the 14 pollutants considered were observed for one location, compared to the others, leading the parametric method to the identification of a cluster containing this unique location. Our non-parametric method is less sensitive to these extreme individuals and thus can be considered as more robust.

As said previously, we only focused on circular potential clusters, which is the simplest procedure. However, the detection of clusters with variable shape should be considered when analysing the environmental dataset, as TE contamination could be strongly influenced by road traffic, which has not a circular pattern. A natural extension would be to consider also elliptic clusters, as proposed by Kulldorff et al. [17].

We should notice that we computed the nonparametric scan statistic associated to our environmental dataset even if there was missing data. The solution we adopted is the following: if x_i^k is unobserved, the k^{th} component of the vector of differences $X_i - X_j$ is set to 0, whatever the value of j.

Finally, we may wonder how to deal when observed variables are not all continuous. For example, the observations could be count data for different diseases in different administrative cells. Using the nonparametric multivariate scan statistic allows to take into account the correlations between diseases, but not the different populations in the administrative cells. Using the likelihood-based multivariate scan statistic introduced by Kulldorff et al. [18] leads to the exact opposite. A solution might be a transformation of the count data using the underlying population, before using the multivariate nonparametric scan statistic.

ACCEPTED MANUSCRIPT

Appendix A. Maps of the simulated clusters

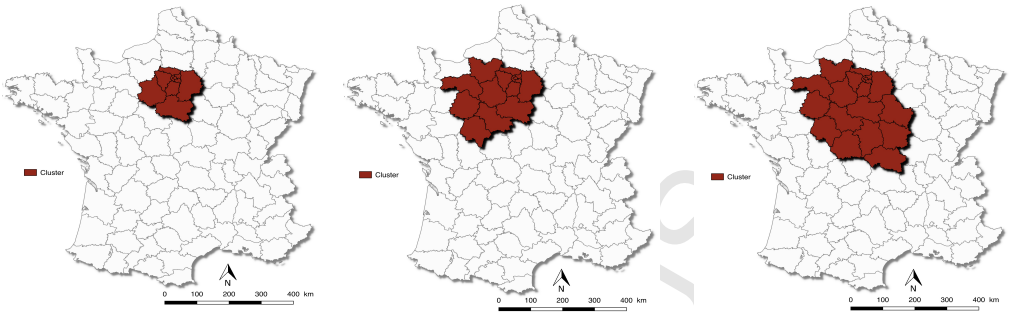


Figure A.4: Simulated clusters (10, 15 and 20 "départements")

Appendix B. Detailed results for the simulation study

Appendix C. Application: distributions of the observed variables

References

[1] Castra, L. , Genin, M., Escutnaire, J., Baert, V., Agostinucci, J.M., Revaux, F., Ursat, C., Tazarourte, K., Adnet, F. and Hubert, H. (2018). Socioeconomic status and incidence of cardiac arrest: a spatial approach to social and territorial disparities. *European Journal of Emergency Medicine* doi: 10.1097/MEJ.0000000000000534

[2] Cressie, N. (1977). On some properties of the scan statistic on the circle and the line. *Journal of Applied Probability* **14,** 272–283.

[3] Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes. *Spatial Statistics* **10,** 117–125.

[4] Cucala, L. (2016). A Mann-Whitney scan statistic for marked point processes. *Communications in Statistics. Theory and Methods* **45,** 321–329.

[5] Cucala, L. (2018). Variable window scan statistics: alternatives to generalized likelihood ratio tests. In *Handbook of scan statistics* (ed. J. Glaz and M. Koutras). Springer.

ACCEPTED MANUSCRIPT

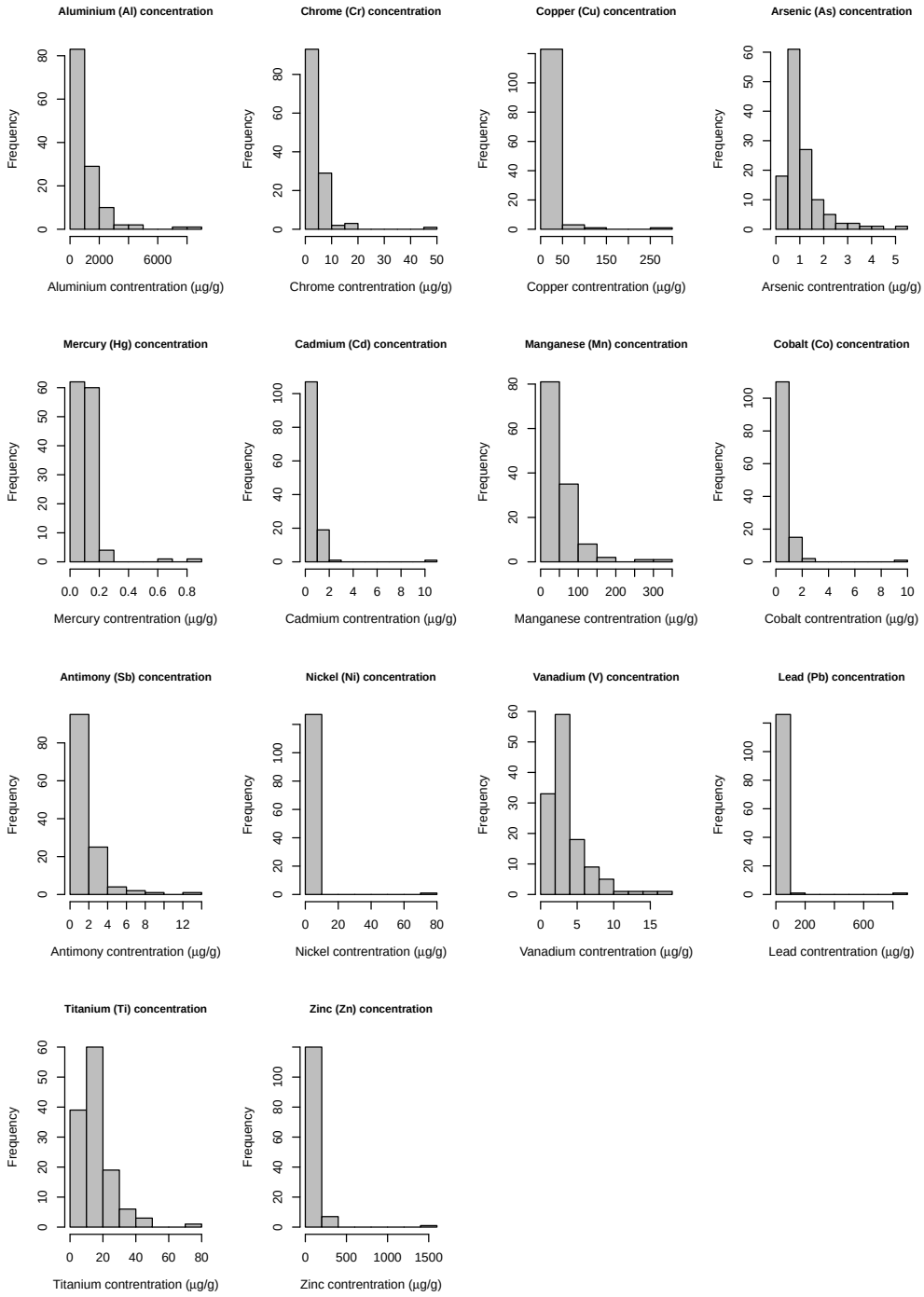


Figure C.5: Application study - Distributions of metal pollutants, European Lille Metropole.

ACCEPTED MANUSCRIPT

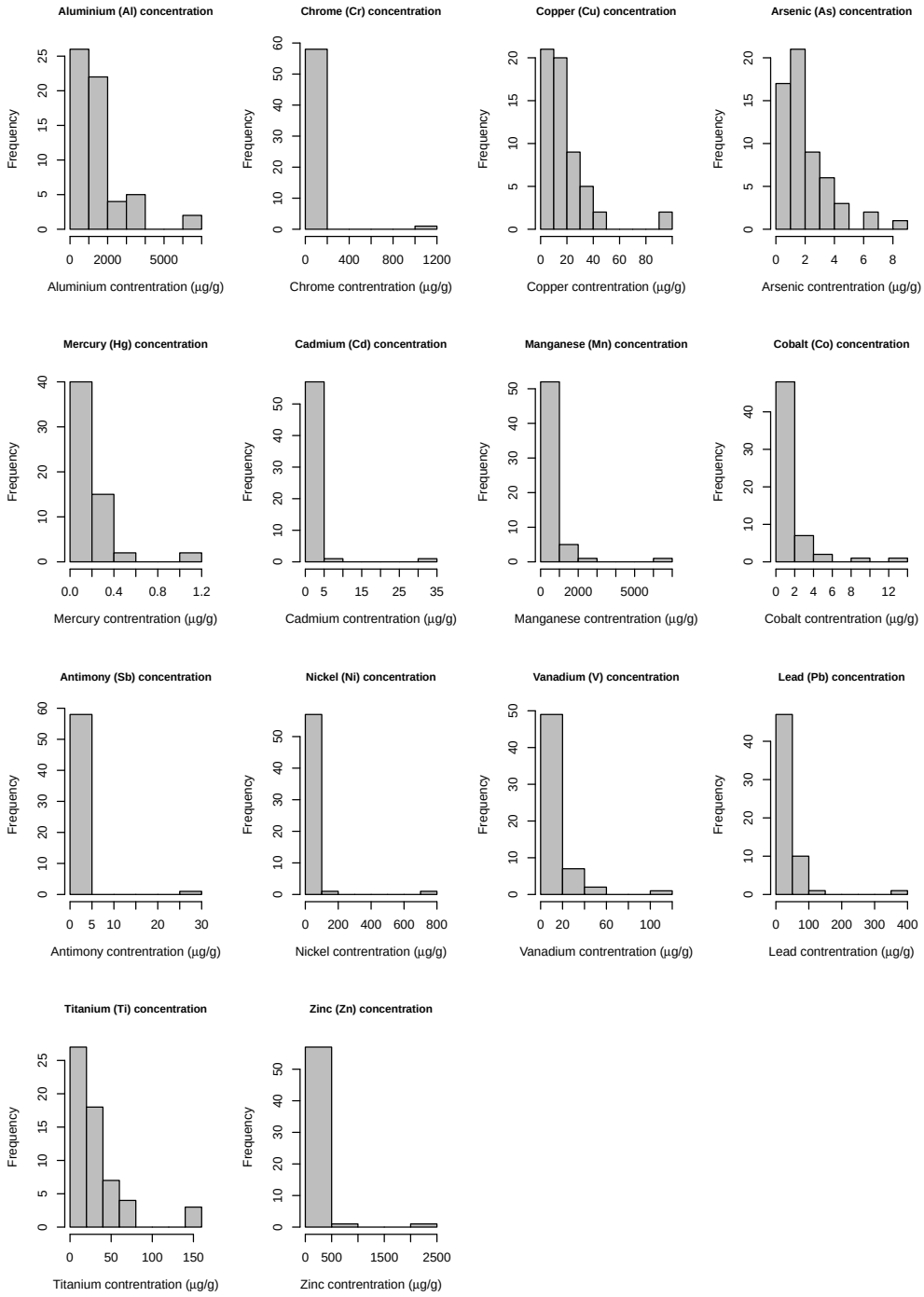


Figure C.6: Application study - Distributions of metal pollutants, Dunkerque agglomeration.

ACCEPTED MANUSCRIPT

[6] Cucala, L., Dematteï, C., Lopes, P. and Ribeiro, A. (2012). Spatial scan statistics for case event data based on connected components. *Computational Statistics* **28,** 357–369.

[7] Cucala, L., Genin, M., Lanier, C. and Occelli, F. (2017). A multivariate Gaussian scan statistic for spatial data. *Spatial Statistics* **21,** 66–74.

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

[9] Duczmal, L. and Assuncão, R. (2004). A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* **45,** 269-286.

[10] Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* **28,** 181–187.

[11] Genin, M., Duhamel, A., Preda, C., Fumery, M., Savoye, G., Peyrin-Biroulet, L., Salleron, J., Lerebours, E., Vasseur, F., Cortot, A., Colombel, J.F. and Gower-Rousseau, C. (2013). Space-time clusters of Crohn's disease in northern France. *Journal of Public Health* **21:6** 497–504

[12] Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer-Verlag, New York.

[13] Glaz, J., Pozdnyakov, V. and Wallenstein, S. (2009). *Scan Statistics. Methods and Applications*, Birkhauser, Basel.

[14] Jung, I. and Cho, H. (2015). A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics* **14:** 30.

[15] Klassen, A., 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.

[16] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics. Theory and Methods* **26,** 1481–1496.

ACCEPTED MANUSCRIPT

[17] Kulldorff, M., Huang, L., Pickle, L. and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine* **25,** 3929–3943.

[18] Kulldorff, M., Mostashari, F., Duczmal, L., Yih, K., Kleinman, K. and Platt, R. (2007). Multivariate spatial scan statistics for disease surveillance. *Statistics in Medicine* **26,** 1824–1833.

[19] 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:** 58.

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

[21] Mann, H. and Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18,** 50–60.

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

[23] Nagarwalla, N. (1996). A scan statistic with a variable window. *Statistics in Medicine* **15,** 845–850.

[24] Occelli, F., Bavdek, R., Deram, A., Hellequin, A.-P., Cuny, M.-A., Zwarterook, I. and Cuny, D. (2016). Using lichen biomonitoring to assess environmental justice at a neighbourhood level in an industrial area of Northern France. *Ecological Indicators* **60,** 781-788.

[25] Occelli, F., Cuny, M.-A., Devred, I., Deram, A., Quarré, S. and Cuny, D. (2014). Étude de l'imprégnation de l'environnement de trois bassins de vie de la région Nord-Pas-de-Calais par les éléments traces métalliques. Vers une nouvelle utilisation des données de biosurveillance lichénique. *Pollution Atmosphérique* **220,** 2268-3798.

[26] Oja, H. and Randles, R. (2004). Multivariate nonparametric tests. *Statistical Science* **19,** 598–605.

[27] Zhang, Z., Kulldorff, M. and Assuncão, R. (2010). Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, Article ID 642379.

ACCEPTED MANUSCRIPT

Table B.2: Simulation study - Comparison of the multivariate Gaussian scan statistics (λ_{MG}) and the multivariate nonparametric scan statistic (λ_{MNP}) through three different multivariate models (Gaussian, Weibull, Exponential). The bold values are the best results obtained in each procedure.

			Gaussian model				**Weibull model**				**Exponential model**		
Nb. Dep.	ρ	β		λ_{MG}	λ_{MNP}	β		λ_{MG}	λ_{MNP}	β		λ_{MG}	λ_{MNP}
10	0.0	1.0	Power	**0.924**	0.888	7	Power	**0.953**	0.881	12	Power	0.428	**0.551**
			%TP	0.853	**0.935**	7	%TP	0.757	**0.926**	12	%TP	0.519	**0.874**
			%FP	**0.033**	0.040	7	%FP	**0.019**	0.037	12	%FP	**0.013**	0.065
	0.2		Power	**0.784**	0.718	7	Power	**0.866**	0.745	12	Power	0.332	**0.397**
			%TP	0.858	**0.928**	7	%TP	0.713	**0.897**	12	%TP	0.444	**0.840**
			%FP	**0.049**	0.064	7	%FP	**0.018**	0.051	12	%FP	**0.007**	0.075
	0.5		Power	**0.561**	0.513	7	Power	**0.757**	0.520	12	Power	0.266	**0.314**
			%TP	0.867	**0.915**	7	%TP	0.661	**0.877**	12	%TP	0.404	**0.810**
			%FP	**0.068**	0.098	7	%FP	**0.017**	0.070	12	%FP	**0.017**	0.094
	0.8		Power	**0.444**	0.380	7	Power	**0.631**	0.416	12	Power	**0.228**	0.212
			%TP	0.837	**0.884**	7	%TP	0.610	**0.847**	12	%TP	0.329	**0.726**
			%FP	**0.076**	0.104	7	%FP	**0.015**	0.095	12	%FP	**0.008**	0.111
	0.0	1.5	Power	1.000	1.000	9	Power	**0.985**	0.972	14	Power	0.479	**0.657**
			%TP	0.858	**0.961**	9	%TP	0.772	**0.940**	14	%TP	0.540	**0.884**
			%FP	0.022	**0.015**	9	%FP	**0.016**	0.025	14	%FP	**0.010**	0.044
	0.2		Power	0.998	0.998	9	Power	**0.951**	0.900	14	Power	0.403	**0.500**
			%TP	0.858	**0.961**	9	%TP	0.738	**0.925**	14	%TP	0.485	**0.872**
			%FP	0.025	**0.022**	9	%FP	**0.017**	0.034	14	%FP	**0.009**	0.060
	0.5		Power	**0.980**	0.965	9	Power	**0.856**	0.736	14	Power	0.305	**0.375**
			%TP	0.858	**0.954**	9	%TP	0.680	**0.896**	14	%TP	0.415	**0.834**
			%FP	**0.027**	0.028	9	%FP	**0.016**	0.045	14	%FP	**0.006**	0.073
	0.8		Power	**0.935**	0.892	9	Power	**0.789**	0.617	14	Power	0.262	**0.267**
			%TP	0.863	**0.945**	9	%TP	0.649	**0.899**	14	%TP	0.366	**0.792**
			%FP	**0.037**	0.042	9	%FP	**0.016**	0.052	14	%FP	**0.005**	0.081
15	0.0	1.0	Power	0.991	**0.986**	7	Power	0.969	**0.975**	12	Power	0.425	**0.751**
			%TP	0.774	**0.955**	7	%TP	0.705	**0.926**	12	%TP	0.431	**0.864**
			%FP	0.070	**0.047**	7	%FP	0.052	**0.045**	12	%FP	**0.022**	0.057
	0.2		Power	0.913	0.913	7	Power	**0.905**	0.885	12	Power	0.344	**0.589**
			%TP	0.778	**0.928**	7	%TP	0.655	**0.892**	12	%TP	0.376	**0.823**
			%FP	0.082	**0.069**	7	%FP	**0.046**	0.052	12	%FP	**0.019**	0.078
	0.5		Power	0.760	**0.768**	7	Power	**0.797**	0.760	12	Power	0.269	**0.416**
			%TP	0.775	**0.901**	7	%TP	0.603	**0.850**	12	%TP	0.276	**0.795**
			%FP	**0.084**	0.085	7	%FP	**0.041**	0.071	12	%FP	**0.012**	0.083
	0.8		Power	**0.631**	0.628	7	Power	**0.669**	0.590	12	Power	0.205	**0.285**
			%TP	0.774	**0.880**	7	%TP	0.553	**0.820**	12	%TP	0.246	**0.732**
			%FP	**0.098**	0.109	7	%FP	**0.035**	0.092	12	%FP	**0.006**	0.087
	0.0	1.5	Power	1.000	1.000	9	Power	0.988	0.998	14	Power	0.452	**0.856**
			%TP	0.788	**0.987**	9	%TP	0.722	**0.954**	14	%TP	0.455	**0.890**
			%FP	0.066	**0.033**	9	%FP	0.053	**0.037**	14	%FP	**0.024**	0.052
	0.2		Power	1.000	1.000	9	Power	0.963	**0.985**	14	Power	0.385	**0.734**
			%TP	0.782	**0.978**	9	%TP	0.688	**0.929**	14	%TP	0.389	**0.832**
			%FP	0.066	**0.036**	9	%FP	0.049	**0.040**	14	%FP	**0.017**	0.059
	0.5		Power	**0.999**	0.997	9	Power	0.894	**0.913**	14	Power	0.284	**0.541**
			%TP	0.780	**0.973**	9	%TP	0.624	**0.892**	14	%TP	0.319	**0.787**
			%FP	0.069	**0.040**	9	%FP	**0.041**	0.051	14	%FP	**0.011**	0.080
	0.8		Power	**0.987**	0.986	9	Power	**0.814**	0.795	14	Power	0.234	**0.387**
			%TP	0.779	**0.954**	9	%TP	0.589	**0.865**	14	%TP	0.260	**0.759**
			%FP	0.072	**0.048**	9	%FP	**0.038**	0.061	14	%FP	**0.007**	0.096
20	0.0	1.0	Power	1.000	1.000	7	Power	0.981	**0.995**	12	Power	0.403	**0.882**
			%TP	0.744	**0.933**	7	%TP	0.680	**0.917**	12	%TP	0.428	**0.875**
			%FP	0.091	**0.047**	7	%FP	0.079	**0.048**	12	%FP	**0.041**	0.064
	0.2		Power	0.980	0.980	7	Power	0.940	**0.965**	12	Power	0.305	**0.746**
			%TP	0.750	**0.924**	7	%TP	0.633	**0.896**	12	%TP	0.377	**0.837**
			%FP	0.093	**0.058**	7	%FP	0.072	**0.053**	12	%FP	**0.033**	0.066
	0.5		Power	0.905	**0.910**	7	Power	0.845	**0.880**	12	Power	0.228	**0.549**
			%TP	0.749	**0.902**	7	%TP	0.597	**0.864**	12	%TP	0.267	**0.785**
			%FP	0.105	**0.074**	7	%FP	0.065	**0.063**	12	%FP	**0.018**	0.085
	0.8		Power	0.779	**0.789**	7	Power	0.703	**0.736**	12	Power	0.175	**0.425**
			%TP	0.768	**0.895**	7	%TP	0.548	**0.851**	12	%TP	0.225	**0.743**
			%FP	0.109	**0.090**	7	%FP	**0.058**	0.074	12	%FP	**0.013**	0.087
	0.0	1.5	Power	1.000	1.000	9	Power	0.995	**1.000**	14	Power	0.464	**0.948**
			%TP	0.741	**0.956**	9	%TP	0.693	**0.933**	14	%TP	0.481	**0.889**
			%FP	0.089	**0.034**	9	%FP	0.082	**0.037**	14	%FP	**0.049**	0.052
	0.2		Power	1.000	1.000	9	Power	0.988	**0.998**	14	Power	0.379	**0.868**
			%TP	0.737	**0.947**	9	%TP	0.667	**0.920**	14	%TP	0.407	**0.856**
			%FP	0.089	**0.037**	9	%FP	0.077	**0.041**	14	%FP	**0.037**	0.062
	0.5		Power	1.000	1.000	9	Power	0.934	**0.973**	14	Power	0.259	**0.685**
			%TP	0.744	**0.943**	9	%TP	0.638	**0.891**	14	%TP	0.311	**0.816**
			%FP	0.090	**0.043**	9	%FP	0.072	**0.050**	14	%FP	**0.023**	0.072
	0.8		Power	0.997	**0.998**	9	Power	0.846	**0.908**	14	Power	0.193	**0.494**
			%TP	0.746	**0.937**	9	%TP	0.582	**0.866**	14	%TP	0.226	**0.773**
			%FP	0.091	**0.047**	9	%FP	0.063	**0.056**	14	%FP	**0.018**	0.078