

The R Package `HDSpatialScan` for the Detection of Clusters of Multivariate and Functional Data using Spatial Scan Statistics

by *Camille Frévent, Mohamed-Salem Ahmed, Julien Soula, Zaineb Smida, Lionel Cucala, Sophie Dabo-Niang and Michaël Genin*

Abstract This paper introduces the R package `HDSpatialScan`. This package allows users to easily apply spatial scan statistics to real-valued multivariate data or both univariate and multivariate functional data. It also permits plotting the detected clusters and to summarize them. In this article the methods are presented and the use of the package is illustrated through examples on environmental data provided in the package.

1 Introduction

Spatial cluster detection methods are useful tools for objective detection and localization of statistically significant aggregation of events indexed in space. Examples of the applications of these methods are numerous: in the field of epidemiology, these methods allow epidemiologists to detect spatial clusters of disease cases and to formulate etiological hypotheses; in the environmental sciences, researchers can be led to search for particularly polluted geographical areas, either by one pollutant in particular or by several pollutants simultaneously. In astronomy, researchers may want to identify star clusters from telescope image data.

Several cluster detection methods have been proposed in the literature. In particular, spatial scan statistics (originally proposed by [Kulldorff and Nagarwalla \(1995\)](#) and [Kulldorff \(1997\)](#) for Bernoulli and Poisson models) are powerful methods for detecting statistically significant spatial clusters, which can be defined by an aggregation of sites presenting an abnormal concentration (mean, etc) of an observed variable, with a variable scanning window and in the absence of pre-selection bias (objective detection of the cluster). Following on from Kulldorff's initial work, several researchers have adapted spatial scan statistics to other spatial data distributions: exponential ([Huang et al., 2007](#)), ordinal ([Jung et al., 2007](#)), normal ([Kulldorff et al., 2009](#)), Weibull ([Bhatt and Tiwari, 2014](#)), etc. Others use nonparametric approaches such as [Jung and Cho \(2015\)](#) and [Cucala \(2016\)](#) who respectively extend the Wilcoxon-Mann-Whitney test for spatial scan statistics and for temporal or spatial scan statistics. Note that in the case of spatial data the two approaches are equivalent by generalizing the method of [Jung and Cho \(2015\)](#) to detect either high or low clusters.

The applications of scan statistics are numerous. In the field of epidemiology, [Khan et al. \(2021\)](#) detected significant clusters of diabetes incidence in Florida between 2007 and 2010, which will help guide local health policies. [Marciano et al. \(2018\)](#) sought to detect spatial clusters of leprosy incidence in a hyperendemic Brazilian municipality between 2000 and 2005 and 2006 and 2010. The study showed a high percentage of contact between people which facilitates the transmission of the disease. [Genin et al. \(2020\)](#) detected high-risk clusters of Crohn's disease in France over the period 2007-2014. As the causes of this disease are still poorly understood, the detection of spatial clusters of Crohn's disease allows the researchers to make hypotheses on possible risk factors, such as high-social deprivation or high urbanization. In the context of environmental science, the detection of clusters of symptomatic exposure to pesticides in rural areas ([Sudakin et al., 2002](#)) would allow the monitoring and prevention of pesticide-related diseases. [Gao et al. \(2014\)](#) focused on the presence of iodine in drinking water in Shandong Province, China. The detection of spatial clusters of iodine presence in drinking water allows an improvement of the monitoring of drinking water quality in these geographical areas. Finally in the context of pollution data, [Wan et al. \(2020\)](#) and [Shi et al. \(2021\)](#) respectively detected clusters of high concentrations of $PM_{2.5}$ in America and China. Such results may allow local authorities to specifically monitor these areas and make decisions to reduce pollution.

When multiple variables are observed simultaneously at each spatial location, researchers may be interested in detecting spatial clusters with anomalous values of all measured variables. In this context, [Kulldorff et al. \(2007\)](#) proposed a multivariate spatial scan statistic using a combination of independent

univariate scan statistics. However it fails to take into account the potential correlations between the variables. A first spatial scan statistic for multivariate data taking into account the correlations was proposed by Cucala et al. (2017). Their method is based on a multivariate normal probability model and a likelihood ratio. Later, Cucala et al. (2019) proposed a nonparametric spatial scan statistic for multivariate data based on a multivariate Wilcoxon-Mann-Whitney test.

Technological developments in measurement tools and data storage capacity have yielded to the increasing use of sensors, cell phones and more generally connected devices that collect data continuously or almost continuously over time. This has led to the introduction of new analysis methods for functional data (Ramsay and Silverman, 2005), as well as the adaptation of classical statistical methods such as principal component analysis (Boente and Fraiman, 2000; Berrendero et al., 2011) or regression (Cuevas et al., 2002; Ferraty and Vieu, 2002; Chiou and Müller, 2007).

In the field of spatial scan statistics, Frévent et al. (2021a) and Smida et al. (2022) proposed new methods for univariate processes. However for example, in environmental surveillance, numerous variables are simultaneously measured, making a multivariate functional approach necessary to detect environmental black-spots. These can be defined as geographical areas characterized by elevated concentrations of multiple pollutants. Although Smida et al. (2022) only studied their approach in the univariate functional framework, they suggest that it could also be adapted for multivariate processes. Frévent et al. (2021b) studied this adaptation and also developed new efficient methods for multivariate functional spatial scan statistics.

In R several packages provide spatial scan statistics implementations. The best known is certainly the `rsatscan` (Kleinman, 2015) package which provides functions to interface R and the SaTScan software (Kulldorff, 2021), allowing the latter to be launched from R. It implements lots of univariate methods (ordinal, Bernoulli, Poisson, ...) but also the space-time spatial scan statistic (Kulldorff et al., 1998) and the multivariate extensions proposed by Kulldorff et al. (2007). The function `kulldorff` implemented in the R package `SpatialEpi` (Chen et al., 2018) also performs the spatial scan statistics based on the Poisson and the Bernoulli models. Other softwares were created to detect clusters such as ClusterSeer (Greiling et al., 2012; Durbeck et al., 2012) which performs spatial, temporal and space-time clustering, and TreeScan (Kulldorff, 2018) which implements the tree-based scan statistic (Kulldorff et al., 2003). We should also mention the R package `DCluster` (Gómez-Rubio et al., 2015) which implements the spatial scan statistics for Poisson or Bernoulli models. The R package `DClusterM` (Gómez-Rubio et al., 2019; Gomez-Rubio et al., 2020) also implements a cluster detection method. Briefly, it consists in considering a large number of generalized linear models by including potential cluster indicators one by one, and then to use a model selection procedure. The Shiny application `SpatialEpiApp` (Moraga, 2017b) and the R package `SpatialEpiApp` (Moraga, 2017a) allow the detection and visualization of clusters by using the scan statistics implemented in SaTScan. Finally the software `FlexScan` (Takahashi et al., 2010) and the R package `rflexscan` (Otani and Takahashi, 2021) implement the spatial scan statistic using a scan window with a non pre-defined shape, defined by Takahashi and Tango (2005). Other R packages also allow clusters detection such as `graphscan` (Loche et al., 2016) (the `cluster` function), `SPATCLUS` (Demattei et al., 2006) or `scanstatistics` (Allévius, 2018b,a) for spatial or space-time data. It should be noted that these last two packages are no longer available on the CRAN (The Comprehensive R Archive Network) repository. Although existing packages implement a large number of statistical spatial scan models, none of them propose multivariate scan models taking into account the potential correlations between variables or scan models for functional data. Thus, we have developed the R package `HDSpatialScan` for high-dimensional spatial scan statistics. The latter allows on the one hand the detection of spatial clusters in multivariate or functional data, and on the other hand, their display on a map and the description of their characteristics.

This paper is organized as follows: The following section presents the different models implemented in the R package `HDSpatialScan`. Then, the implementation of the methods is described and examples of use of the package are given. The last section concludes the paper.

2 Models

Let s_1, \dots, s_n be n different locations of an observation domain $S \subset \mathbb{R}^2$ and X_1, \dots, X_n be the observations of a variable X in s_1, \dots, s_n . Hereafter all observations are considered to be independent, which is a classical assumption in scan statistics. Three types of spatial data can be considered: either lattice data (the data are aggregated at the spatial level, e.g.: county), geostatistical data (the variable is defined on a continuous area and each individual measure corresponds to a unique fixed spatial location, e.g.: pollutant concentration measured by sensors over a region), or marked point data (each individual measure corresponds to a unique random spatial location, e.g.: height of the trees in a

forest, the location of the trees is random).

Spatial scan statistics aim at detecting spatial clusters and testing their statistical significance. Hence, one tests a null hypothesis \mathcal{H}_0 (the absence of a cluster) against a composite alternative hypothesis \mathcal{H}_1 (the presence of at least one cluster $w \subset S$ presenting abnormal values of X). For this purpose, a spatial scan statistic consists of two steps. The first one is a detection phase using a scanning window of variable size and shape. We will focus here on the approach of [Kulldorff and Nagarwalla \(1995\)](#) which use a circular scanning window of variable center and radius, however it should be noted that other shapes can be considered ([Kulldorff et al., 2006](#); [Cucala et al., 2013](#)). An approach often advised is to limit the maximum size to half of the studied region since otherwise it would be like detecting a “negative cluster” in the areas outside the clusters covering almost all the studied region ([Kulldorff and Nagarwalla, 1995](#)). Then the scanning window allows to define a set of potential clusters \mathcal{W} by

$$\mathcal{W} = \{w_{i,j} / 1 \leq |w_{i,j}| \leq \frac{n}{2}, 1 \leq i, j \leq n\}, \quad (1)$$

where $w_{i,j}$ is the disc centered on s_i that passes through s_j and $|w_{i,j}|$ corresponds to the number of sites in $w_{i,j}$. [Figure 1](#) illustrates the set of potential clusters defined with a circular scanning window with Equation 1 on a set of eight administrative areas in France.

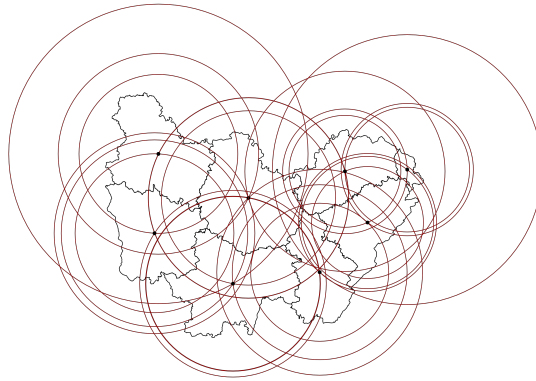


Figure 1: Set of potential clusters defined with a circular scanning window of variable center and radius with Equation 1 on a set of eight administrative areas in France. Each potential cluster is represented with a red circle.

Then the spatial scan statistic can be defined as the maximum of a concentration index over the set of potential clusters \mathcal{W} . The second step is the determination of the statistical significance of the spatial scan statistic. For this, since the distribution of the scan statistic is intractable under \mathcal{H}_0 due to the overlapping nature of \mathcal{W} , a common approach, which will be considered here, is to use a Monte-Carlo method (see Section [Computing the statistical significance of the MLC](#) for more details).

Spatial scan statistics for multivariate data

Here we consider the case where several continuous variables are simultaneously observed in each spatial location: $X = (X^{(1)}, \dots, X^{(p)})^\top$ is a p -dimensional variable ($p \geq 2$). In this context the objective is to identify multivariate spatial clusters that are aggregations of sites in which X takes higher or lower values (in terms of mean, median, etc.) than elsewhere. For example one could observe the average concentrations of several pollutants over a day: a vector can be associated with each site, each element of which corresponds to the average concentration of one pollutant. In this context a spatial cluster corresponds to a set of sites under or overexposed to multiple pollutants. Different approaches will be presented: a parametric method based on a Gaussian model and a nonparametric one.

[Figure 2](#) summarizes the different types of multivariate data with examples, and provides guidelines on the spatial scan statistics methods to be used for these data (and the argument to use in the scan function of the package). More precisely, we distinguish three types of spatial data: lattice data which are aggregated data for example at the scale of the regions of a country, geostatistical data which are defined on a continuous space (typically temperature, sunshine, or atmospheric pressure) although they are observed only at discrete sites, and marked point data for which the location is random (for example the distribution of trees in a forest) and we observe at each location the circum-

ference and height of the tree for example. To detect spatial clusters, in the case of Gaussian data we will prefer the Gaussian approach (MG) and otherwise we will use the nonparametric approach (MNP).


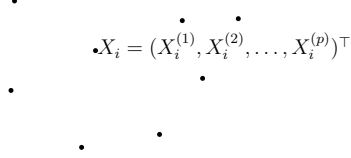
Data	$X_i = (X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(p)})^\top$ 			 $\bullet X_i = (X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(p)})^\top$
	Lattice data: <ul style="list-style-type: none"> • Unemployment rate and fraction of the population that has not graduated from high school 	Geostatistical data: <ul style="list-style-type: none"> • Temperature and air pressure 	Point pattern: <ul style="list-style-type: none"> • Circumference and height of trees 	
Question	Is there a statistically significant cluster of high unemployment rates and high fraction of the population with a low level of education?	Is there a statistically significant cluster of high temperatures and low air pressure?	Is there a statistically significant cluster of trees with larger circumferences and heights?	
Methods	Gaussian data: <ul style="list-style-type: none"> • Multivariate Gaussian spatial scan statistic • “MG” argument in the scan function of the package Non-Gaussian data: <ul style="list-style-type: none"> • Multivariate Nonparametric spatial scan statistic • “MNP” argument in the scan function of the package 			
Interpretation	There is a statistically significant cluster and by describing the mean or median of each variable, we can get an indication of which variables are dominant in the cluster, and which variables are higher or lower in that cluster.			

Figure 2: Summary of spatial scan statistics for multivariate data. The table indicates the question that can be asked for the detection of clusters. It then indicates the methods to be used according to the distribution of the data as well as the ways to interpret the detected clusters. Spatial scan statistics for multivariate data can be used to detect spatial clusters on any type of spatial data (lattice, geostatistical, point data) modeled by vectors. The detected clusters can be characterized by computing the mean or median of each variable inside and outside each cluster.

Cucala et al. (2017) proposed a parametric spatial scan statistic for multivariate data based on a multivariate normal model taking into account the correlations between the variables.

The null hypothesis \mathcal{H}_0 , corresponding to the absence of any cluster in the data, is the following:

$\forall i \in \llbracket 1; n \rrbracket, X_i \sim \mathcal{N}_p(\mu, \Sigma)$ and the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as: $\forall i \in \llbracket 1; n \rrbracket, X_i \sim \begin{cases} \mathcal{N}_p(\mu_w, \Sigma_{w,w^c}) & \text{if } s_i \in w \\ \mathcal{N}_p(\mu_{w^c}, \Sigma_{w,w^c}) & \text{otherwise} \end{cases}$.

Then we can compute the MLE estimates of $\mu, \mu_w, \mu_{w^c}, \Sigma$ and Σ_{w,w^c} : $\hat{\mu}, \hat{\mu}_w, \hat{\mu}_{w^c}, \hat{\Sigma}$ and $\hat{\Sigma}_{w,w^c}$, and we can show that the log-likelihood ratio between these two hypotheses is

$$\widehat{LLR}^w = -\frac{n}{2} \ln \left[\det \left(\sum_{s_i \in w} (X_i - \hat{\mu}_w)(X_i - \hat{\mu}_w)^\top + \sum_{s_i \in w^c} (X_i - \hat{\mu}_{w^c})(X_i - \hat{\mu}_{w^c})^\top \right) \right] + \frac{n}{2} \ln \left[\det \left(\sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^\top \right) \right],$$

where $\hat{\mu}_g = \frac{1}{|g|} \sum_{i, s_i \in g} X_i$ for $g \in \{w, w^c\}$ and $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$.

Finally the log-likelihood ratio is used as a concentration index and maximised over the set of potential clusters \mathcal{W} .

Thus we can show that the multivariate Gaussian (MG) scan statistic is

$$\lambda_{MG} = \min_{w \in \mathcal{W}} \det \left(\sum_{\substack{i \\ s_i \in w}} (X_i - \hat{\mu}_w)(X_i - \hat{\mu}_w)^\top + \sum_{\substack{i \\ s_i \in w^c}} (X_i - \hat{\mu}_{w^c})(X_i - \hat{\mu}_{w^c})^\top \right).$$

This test performs very well against Gaussian alternatives but faces problems when the data is not normal, which is often the case when dealing with environmental data exhibiting extreme values. For that reason [Cucala et al. \(2019\)](#) developed a nonparametric spatial scan statistic for multivariate data based on a multivariate extension of the Wilcoxon-Mann-Whitney test for multivariate data ([Oja and Randles, 2004](#)).

In this context the null hypothesis \mathcal{H}_0 can be rewritten as $\mathcal{H}_0 : X_1, \dots, X_n$ are identically distributed, whatever the associated location.

Let

$$\text{sgn} : \mathbb{R}^p \rightarrow \mathbb{R}^p$$

$$x \mapsto \begin{cases} \|x\|_2^{-1}x & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases},$$

then the multivariate ranks R_i are defined by $R_i = \frac{1}{n} \sum_{j=1}^n \text{sgn}(A_X(X_i - X_j))$ where the matrix A_X

makes the ranks such that $\frac{p}{n} \sum_{i=1}^n R_i R_i^\top = \frac{1}{n} \sum_{i=1}^n R_i^\top R_i I_p$. Note that this matrix can be easily computed using an iterative procedure. Then the multivariate extension of the Wilcoxon-Mann-Whitney statistic proposed by [Oja and Randles \(2004\)](#) is

$$U^2(w) = \frac{p}{c_X^2} \left[|w| \|\bar{R}_w\|_2^2 + |w^c| \|\bar{R}_{w^c}\|_2^2 \right], \text{ where } c_X^2 = \frac{1}{n} \sum_{i=1}^n R_i^\top R_i.$$

[Cucala et al. \(2019\)](#) used $U^2(w)$ as a concentration index to build the spatial scan statistic: the multivariate nonparametric (MNP) scan statistic is $\lambda_{MNP} = \max_{w \in \mathcal{W}} U^2(w)$.

It should be noted that in the case $p = 1$, these statistics are respectively equivalent to the ones introduced by [Kulldorff et al. \(2009\)](#) (which is equivalent to the scan statistic developed by [Cucala \(2014\)](#), UG), and [Jung and Cho \(2015\)](#) (UNP).

Spatial scan statistics for univariate functional data

Here we consider the case where a continuous variable is observed in each spatial location over time: $\{X(t), t \in \mathcal{T}\}$ is a real-valued stochastic process where \mathcal{T} is an interval of \mathbb{R} . In this context the objective is to identify functional spatial clusters that are aggregations of sites in which the curves are higher or lower than elsewhere. For example, one can observe the concentration of an air pollutant over time in different geographical areas. Then a cluster corresponds to an aggregation of sites in which the concentration of the air pollutant is higher or lower over the time than in the other spatial units. Several methods will be considered: a parametric method based on a functional ANOVA, a nonparametric approach using a Wilcoxon-Mann-Whitney test for high-dimensional data, a distribution-free approach based on a pointwise Student's t-test and finally a pointwise rank-based method. On Gaussian data, for non localized clusters in time all approaches show high power and high true positive rates. However the performances of the ANOVA-based method strongly decrease on non-normal data. For localized clusters in time (that are aggregations of sites that take higher or lower values for X only in a small interval of time (an interval of five days over a study period of one month for example)) the pointwise approaches should be favored.

Figure 3 summarizes the different types of univariate functional data with examples, and provides recommendations on the spatial scan statistics methods to be used for these data (and the argument to use in the scan function of the package). More precisely, we distinguish lattice functional data which are aggregated functional data for example at the scale of the administrative areas of a country (unemployment rate, percentage of the population over 65, etc), geostatistical functional data which are defined on a continuous space (typically temperature, sunshine, or atmospheric pressure over time) although they are observed only at discrete sites, and marked point data for which the location is random (for example the distribution of trees in a forest) and we observe at each location the circumference of the tree over time for example. To detect spatial clusters, as mentioned before, in the case of Gaussian data we will prefer the pointwise distribution-free functional approach (DFFSS) and otherwise we will use the pointwise rank-based approach (URBFSS).

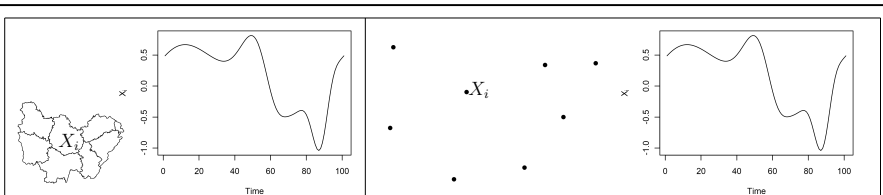
<p>Data</p>		
<p>Application example</p>	<p>Lattice data:</p> <ul style="list-style-type: none"> • Unemployment rate over time 	<p>Geostatistical data:</p> <ul style="list-style-type: none"> • Temperature over time <p>Point pattern:</p> <ul style="list-style-type: none"> • Circumference of trees over time
<p>Question</p>	<p>Is there a statistically significant cluster of high or low unemployment rate curves?</p>	<p>Is there a statistically significant cluster of high or low temperature curves?</p> <p>Is there a statistically significant cluster of trees with high or low circumference curves?</p>
<p>Methods</p>	<p>Gaussian data:</p> <ul style="list-style-type: none"> • Distribution-free functional spatial scan statistic • “DFSS” argument <p>Non-Gaussian data:</p> <ul style="list-style-type: none"> • Univariate rank-based functional spatial scan statistic • “URBFSS” argument 	
<p>Interpretation</p>	<p>There is a statistically significant cluster and by describing the mean or median curve of the variable, we can get an indication of the characteristics of the cluster.</p>	

Figure 3: Summary of spatial scan statistics for univariate functional data. The table indicates the question that can be asked for the detection of clusters on a set of curves. It then indicates the methods to be used according to the distribution of the data as well as the ways to interpret the detected clusters. Spatial scan statistics for univariate functional data can be used to detect spatial clusters on any type of spatial data (lattice, geostatistical, point data) observed over a period of time. The detected clusters can be characterized by computing the mean or median curve inside and outside each cluster.

The parametric spatial scan statistic for univariate functional data

Frévent et al. (2021a) supposed that the process X takes values in a semi-metric space, in particular in $\mathcal{L}^2(\mathcal{T}, \mathbb{R})$ and proposed a parametric spatial scan statistic for functional data, based on a functional ANOVA. Here the null hypothesis \mathcal{H}_0 can be rewritten: $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c} = \mu_S$, and the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as follows: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$, where μ_w, μ_{w^c} and μ_S stand for the mean functions in w , outside w and over S , respectively. Cuevas et al. (2004) and Górecki and Smaga (2015) proposed the following ANOVA test statistic:

$$F_n^{(w)} = \frac{|w| \|\bar{X}_w - \bar{X}\|_2^2 + |w^c| \|\bar{X}_{w^c} - \bar{X}\|_2^2}{\frac{1}{n-2} \left[\sum_{j,s_j \in w} \|X_j - \bar{X}_w\|_2^2 + \sum_{j,s_j \in w^c} \|X_j - \bar{X}_{w^c}\|_2^2 \right]}$$

where $\bar{X}_g(t) = \frac{1}{|g|} \sum_{i,s_i \in g} X_i(t)$ are empirical estimators of μ_g ($g \in \{w, w^c\}$), $\bar{X}(t) = \frac{1}{n} \sum_{i=1}^n X_i(t)$ is the empirical estimator of μ_S and $\|x\|_2^2 = \int_{\mathcal{T}} x^2(t) dt$.

Thus, Frévent et al. (2021a) proposed to use $F_n^{(w)}$ as a concentration index and the proposed parametric functional spatial scan statistic (PFSS) is $\Lambda_{PFSS} = \max_{w \in \mathcal{W}} F_n^{(w)}$.

This method gives high powers and F-measures on normal data but as in the multivariate framework the parametric method faces problems when the data is not normal. Smida et al. (2022) proposed a nonparametric spatial scan statistic for functional data based on a functional Wilcoxon-Mann-Whitney test (Chakraborty and Chaudhuri, 2014).

A nonparametric spatial scan statistic for functional data

Here X is a process of a smooth Banach space χ , with a Gâteaux differentiable norm $\|\cdot\|_\chi$. Let denote P_w and P_{w^c} the probability measures of X in w and in w^c respectively, then \mathcal{H}_0 corresponds to: $\mathcal{H}_0 : \forall w \in \mathcal{W}, P_w = P_{w^c}$ and the alternative hypothesis associated with a potential cluster w can be rewritten as $\mathcal{H}_1^{(w)} : P_w(X) = P_{w^c}(X - \Delta), \Delta \in \chi \setminus \{0\}$.

Chakraborty and Chaudhuri (2014) defined the sign function in the functional framework as

$$\forall h \in \chi, \text{Sgn}_X(h) = \begin{cases} \lim_{v \rightarrow 0^+} \frac{\|X + vh\|_\chi - \|X\|_\chi}{v} & \text{if } X \neq 0 \\ 0 & \text{if } X = 0 \end{cases}.$$

Then they proposed the following test statistic:

$$T_{WMW}(w) = \frac{1}{|w||w^c|} \sum_{i,s_i \in w} \sum_{j,s_j \in w^c} \text{Sgn}_{X_j - X_i}.$$

Under \mathcal{H}_0 , if $\frac{|w|}{n} \rightarrow \gamma \in [0;1]$ as $|w|, |w^c| \rightarrow \infty, \sqrt{\frac{|w||w^c|}{n}} T_{WMW}(w)$ converges weakly to a distribution that does not depend on $|w|$. Thus Smida et al. (2022) proposed to use $U(w) = \left\| \sqrt{\frac{|w||w^c|}{n}} T_{WMW}(w) \right\|$ as a concentration index: the nonparametric functional scan statistic (NPFSS) is $\Lambda_{\text{NPFSS}} = \max_{w \in \mathcal{W}} U(w)$. It should be noticed that although Smida et al. (2022) only studied the performances of the NPFSS in the univariate functional framework, their method is also applicable on multivariate functional data as shown by Frévent et al. (2021b).

A distribution-free spatial scan statistic for univariate functional data

Frévent et al. (2021a) also proposed to combine the distribution-free spatial scan statistic for univariate data proposed by Cucala (2014) and the max statistic of Lin et al. (2021). They supposed that for each time $t, \mathbb{V}[X_i(t)] = \sigma^2(t)$ for all $i \in \llbracket 1; n \rrbracket$. Then for each t , the concentration index proposed by Cucala (2014) to test $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w(t) = \mu_{w^c}(t) = \mu_S(t)$ was

$$I^{(w)}(t) = \frac{|\bar{X}_w(t) - \bar{X}_{w^c}(t)|}{\sqrt{\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)]}},$$

where $\hat{\mathbb{V}}[\bar{X}_w(t) - \bar{X}_{w^c}(t)] = \widehat{\sigma^2}(t) \left[\frac{1}{|w|} + \frac{1}{|w^c|} \right],$

$$\widehat{\sigma^2}(t) = \frac{1}{n-2} \left[\sum_{i,s_i \in w} (X_i(t) - \bar{X}_w(t))^2 + \sum_{i,s_i \in w^c} (X_i(t) - \bar{X}_{w^c}(t))^2 \right].$$

Then the idea is to globalize the information by maximizing the previous quantity over the time for each potential cluster w , as suggested by Lin et al. (2021):

$$I^{(w)} = \sup_{t \in \mathcal{T}} I^{(w)}(t).$$

For cluster detection, as for the PFSS, the null hypothesis \mathcal{H}_0 (the absence of cluster) is defined as follows: $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c} = \mu_S$. And the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as follows: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$.

Frévent et al. (2021a) considered $I^{(w)}$ as a concentration index and maximized it over the set of potential clusters \mathcal{W} yielding to the following distribution-free functional spatial scan statistic (DFSS): $\Lambda_{\text{DFSS}} = \max_{w \in \mathcal{W}} I^{(w)}$.

A new rank-based spatial scan statistic for univariate functional data

A pointwise approach based on ranks and on the nonparametric scan statistic for univariate data (Jung and Cho, 2015) can be proposed in the univariate functional framework by adapting the approach of Frévent et al. (2021b).

For a time t , Jung and Cho (2015) proposed to test $\mathcal{H}_0 : \forall w \in \mathcal{W}, F_{w,t} = F_{w^c,t}$ where $F_{w,t}$ and $F_{w^c,t}$ are the cumulative distribution functions of $X(t)$ in w and outside w , by using the Wilcoxon rank-sum test statistic. For a time t and a potential cluster w , the Wilcoxon rank-sum test statistic is $W(t)^{(w)} = \sum_{i: s_i \in w} R_i(t)$ where $R_i(t)$ is the rank of $X_i(t)$ in $\{X_1(t), \dots, X_n(t)\}$, using the average rank in the case of tied observations.

Then the standardized version of this statistic is

$$T(t)^{(w)} = \frac{W(t)^{(w)} - \mathbb{E}[W(t)^{(w)}]}{\sqrt{\mathbb{V}[W(t)^{(w)}]}}$$

where $\mathbb{E}[W(t)^{(w)}] = \frac{|w|(n+1)}{2}$ and $\mathbb{V}[W(t)^{(w)}] = \frac{|w||w^c|(n+1)}{12}$ are the expected value and the variance of $W(t)^{(w)}$ under \mathcal{H}_0 .

Jung and Cho (2015) proposed to minimize the p-value associated with $T(t)^{(w)}$ on the set of potential clusters \mathcal{W} . We propose to adapt their approach by simply using $|T(t)^{(w)}|$ as a pointwise statistic.

In the context of cluster detection, the null hypothesis is defined as $\mathcal{H}_0: \forall w \in \mathcal{W}, \forall t \in \mathcal{T}, F_{w,t} = F_{w^c,t}$. The alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w is $\mathcal{H}_1^{(w)}: \exists t \in \mathcal{T}, F_{w,t}(x) = F_{w^c,t}(x - \Delta_t), \Delta_t \neq 0$.

As before, we propose to globalize the information over the time with $T^{(w)} = \sup_{t \in \mathcal{T}} |T(t)^{(w)}|$ and to use this quantity as a concentration index, yielding to the following univariate rank-based functional spatial scan statistic (URBFSS): $\Lambda_{\text{URBFSS}} = \max_{w \in \mathcal{W}} T^{(w)}$.

Spatial scan statistics for multivariate functional data

Here we consider the case where several continuous variables are observed simultaneously in each spatial unit over time: $\{X(t), t \in \mathcal{T}\}$ is a p -dimensional vector-valued stochastic process ($p \geq 2$) where \mathcal{T} is an interval of \mathbb{R} . The objective is to detect multivariate functional spatial clusters that are aggregations of sites in which the curves are higher or lower than elsewhere. For example we can observe the concentration of several pollutants over time in different locations. Thus at each location we observe several processes (air pollutant concentrations) and these processes can be correlated. In this context a cluster is an aggregation of sites overexposed or underexposed to multiple pollutants over time. Several methods will be presented: a parametric method based on a functional MANOVA, a distribution-free approach based on a pointwise Hotelling T^2 -test and finally a pointwise rank-based method. On normal data, all approaches show high power and high true positive rates for non localized clusters in time. However the performances of the methods based on the MANOVA and the Hotelling T^2 -test decrease on non-normal data. For localized clusters in time the pointwise approaches should be favored, especially the pointwise rank-based method on non-Gaussian data. By localized clusters in time we mean aggregations of sites that take higher or lower values for X only in a small interval of time (an interval of five days over a study period of one month for example).

Figure 4 summarizes the different types of multivariate functional data with examples, and provides guidelines on the spatial scan statistics methods to be used for these data (and the argument to use in the scan function of the package). To be more precise, we can distinguish lattice functional data which are aggregated data for example at the scale of the regions of a country (unemployment rate and fraction of the population that has not graduated from high school, over time, for example), geostatistical functional data which are defined on a continuous space (temperature, sunshine, and atmospheric pressure over time) although they are observed only at discrete sites, and marked point data for which the location is random (for example the distribution of trees in a forest) and we observe at each location the circumference and height of the tree over time for example. As previously mentioned, to detect spatial clusters, in the case of Gaussian data we will prefer the multivariate distribution-free functional spatial scan statistic (MDFSS) and for non-Gaussian data we will use the pointwise rank-based approach (MRBFSS).

A parametric spatial scan statistic for multivariate functional data

Here, the process X is supposed to take values in a semi-metric space, in particular the Hilbert space $\mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$ of p -dimensional vector-valued square-integrable functions on \mathcal{T} , equipped with the inner

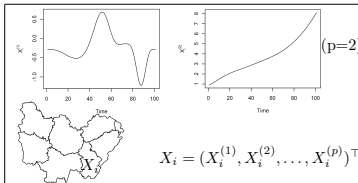
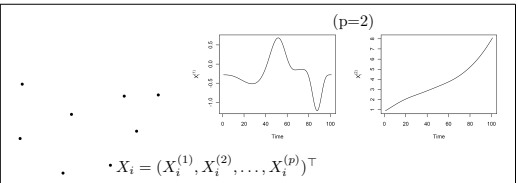
<p>Data</p>	 <p>$X_i = (X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(p)})^\top$</p>		 <p>$X_i = (X_i^{(1)}, X_i^{(2)}, \dots, X_i^{(p)})^\top$</p>	
<p>Application example</p>	<p>Lattice data:</p> <ul style="list-style-type: none"> • Unemployment rate and fraction of the population that has not graduated from high school over time 	<p>Geostatistical data:</p> <ul style="list-style-type: none"> • Temperature and air pressure over time 	<p>Point pattern:</p> <ul style="list-style-type: none"> • Circumference and height of trees over time 	
<p>Question</p>	<p>Is there a statistically significant cluster of high unemployment rate curves and high fraction of the population with a low level of education over time?</p>	<p>Is there a statistically significant cluster of high temperature and low air pressure curves?</p>	<p>Is there a statistically significant cluster of trees with high circumference and height curves?</p>	
<p>Methods</p>	<p>Gaussian data:</p> <ul style="list-style-type: none"> • Multivariate distribution-free functional spatial scan statistic • “MDFSS” argument in the scan function of the package <p>Non-Gaussian data:</p> <ul style="list-style-type: none"> • Multivariate rank-based functional spatial scan statistic • “MRBFSS” argument in the scan function of the package 			
<p>Interpretation</p>	<p>There is a statistically significant cluster and by describing the mean or median curve of each variable, we can get an indication of which variables are dominant in the cluster, and which variables present higher or lower curves in that cluster.</p>			

Figure 4: Summary of spatial scan statistics for multivariate functional data. The table indicates the question that can be asked for the detection of clusters of multivariate curves. It then indicates the methods to be used according to the distribution of the data as well as the ways to interpret the detected clusters. Spatial scan statistics for multivariate functional data can be used to detect spatial clusters on any type of spatial data (lattice, geostatistical, point data) composed of multivariate curves X_i (in the table example, $X_i = (X_i^{(1)}, X_i^{(2)})^\top$ is composed of two curves). Then, the detected clusters can be characterized by computing the mean or median curve of each variable inside and outside each cluster.

product $\langle X, Y \rangle = \int_{\mathcal{T}} X(t)^\top Y(t) dt.$

Frévent et al. (2021b) proposed a parametric scan statistic for multivariate functional data based on a functional MANOVA Lawley–Hotelling trace test (Górecki and Smaga, 2017).

In this context, the null hypothesis \mathcal{H}_0 is $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c} = \mu_S$, where μ_w, μ_{w^c} and μ_S stand for the mean functions in w , outside w and over S , respectively. And the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w is $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$. Górecki and Smaga (2017) presented the following adaptation of the Lawley-Hotelling trace test statistic:

$$LH^{(w)} = \text{Trace}(H_w E_w^{-1})$$

where $H_w = |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)] [\bar{X}_w(t) - \bar{X}(t)]^\top dt + |w^c| \int_{\mathcal{T}} [\bar{X}_{w^c}(t) - \bar{X}(t)] [\bar{X}_{w^c}(t) - \bar{X}(t)]^\top dt$
 and $E_w = \sum_{j, s_j \in w} \int_{\mathcal{T}} [X_j(t) - \bar{X}_w(t)] [X_j(t) - \bar{X}_w(t)]^\top dt + \sum_{j, s_j \in w^c} \int_{\mathcal{T}} [X_j(t) - \bar{X}_{w^c}(t)] [X_j(t) - \bar{X}_{w^c}(t)]^\top dt$

with $\bar{X}_g(t) = \frac{1}{|g|} \sum_{i, s_i \in g} X_i(t)$ the empirical estimators of $\mu_g(t)$ for $g \in \{w, w^c\}$ and $\bar{X}(t) = \frac{1}{n} \sum_{i=1}^n X_i(t)$ the empirical estimator of $\mu_S(t)$.

Then Frévent et al. (2021b) considered $LH^{(w)}$ as a concentration index and proposed the parametric multivariate functional spatial scan statistic (MPFSS): $\Lambda_{\text{MPFSS}} = \max_{w \in \mathcal{W}} LH^{(w)}$.

In fact Górecki and Smaga (2017) proposed four test statistics using the matrices H_w and E_w to compare the mean functions in w and w^c : (1) the Lawley–Hotelling trace test statistic $LH^{(w)} = \text{Trace}(H_w E_w^{-1})$, (2) the Pillai trace test statistic $P^{(w)} = \text{Trace}(H_w (H_w + E_w)^{-1})$, (3) the Roy’s largest root test statistic $R^{(w)} = \lambda_{\max}(H_w E_w^{-1})$ where $\lambda_{\max}(H_w E_w^{-1})$ is the maximum eigenvalue of $H_w E_w^{-1}$ and (4) the Wilks

$$\text{lambda test statistic } W^{(w)} = \frac{\det(E_w)}{\det(H_w + E_w)}.$$

Thus each of these quantities (or the opposite for the Wilks lambda test statistic) can be considered as a concentration index and maximized over \mathcal{W} which results in the following parametric multivariate functional spatial scan statistics:

$$\Lambda_{\text{LH}} = \max_{w \in \mathcal{W}} \text{LH}^{(w)}, \quad \Lambda_{\text{P}} = \max_{w \in \mathcal{W}} \text{P}^{(w)}, \quad \Lambda_{\text{R}} = \max_{w \in \mathcal{W}} \text{R}^{(w)}, \quad \Lambda_{\text{W}} = \min_{w \in \mathcal{W}} W^{(w)}.$$

These four approaches are implemented in the package **HDSpatialScan**.

A distribution-free spatial scan statistic for multivariate functional data

Frévent et al. (2021b) proposed a distribution-free spatial scan statistic for multivariate functional data which is the counterpart of the distribution-free spatial scan statistic for univariate functional data developed by Frévent et al. (2021a). They supposed that for each time t , $\mathbb{V}[X_i(t)] = \Sigma(t, t)$ for all $i \in \llbracket 1; n \rrbracket$, where Σ is a $p \times p$ covariance matrix function.

Thus, as previously, in the context of cluster detection, the null hypothesis \mathcal{H}_0 can be defined as follows: $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w = \mu_{w^c} = \mu_S$, where μ_w, μ_{w^c} and μ_S stand for the mean functions in w , outside w and over S , respectively. And the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w can be defined as follows: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$. Next, Qiu et al. (2021) proposed to compare the mean function μ_w in w with the mean function μ_{w^c} in w^c by using the following statistic:

$$T_{n,\max}^{(w)} = \sup_{t \in \mathcal{T}} T_n(t)^{(w)}$$

where $T_n(t)$ is a pointwise statistic defined by the Hotelling T^2 -test statistic

$$T_n(t)^{(w)} = \frac{|w||w^c|}{n} (\bar{X}_w(t) - \bar{X}_{w^c}(t))^\top \hat{\Sigma}(t, t)^{-1} (\bar{X}_w(t) - \bar{X}_{w^c}(t)).$$

$\bar{X}_w(t)$ and $\bar{X}_{w^c}(t)$ are the empirical estimators of the mean functions defined previously, and

$$\hat{\Sigma}(s, t) = \frac{1}{n-2} \left[\sum_{i, s_i \in w} (X_i(s) - \bar{X}_w(s)) (X_i(t) - \bar{X}_w(t))^\top + \sum_{i, s_i \in w^c} (X_i(s) - \bar{X}_{w^c}(s)) (X_i(t) - \bar{X}_{w^c}(t))^\top \right]$$

is the pooled sample covariance matrix function.

Then Frévent et al. (2021b) proposed to use $T_{n,\max}^{(w)}$ as a concentration index and to maximize it over the set of potential clusters \mathcal{W} : the multivariate distribution-free functional spatial scan statistic (MDFSS) is $\Lambda_{\text{MDFSS}} = \max_{w \in \mathcal{W}} T_{n,\max}^{(w)}$.

A rank-based spatial scan statistic for multivariate functional data

Finally Frévent et al. (2021b) also proposed to consider as a pointwise test statistic the multivariate extension of the Wilcoxon rank-sum test statistic developed by Oja and Randles (2004) and detailed in Section [Spatial scan statistics for multivariate data](#). They defined the pointwise multivariate ranks as

$$R_i(t) = \frac{1}{n} \sum_{j=1}^n \text{sgn}(A_X(t)(X_i(t) - X_j(t)))$$

where the pointwise transformation matrix $A_X(t)$ is so that

$$\frac{p}{n} \sum_{i=1}^n R_i(t) R_i(t)^\top = \frac{1}{n} \sum_{i=1}^n R_i(t)^\top R_i(t) I_p,$$

and the sgn function is the same as in Section [Spatial scan statistics for multivariate data](#).

Then for each time t , the pointwise multivariate extension of the Wilcoxon rank-sum test statistic is defined as $W(t)^{(w)} = \frac{pn}{\sum_{i=1}^n R_i(t)^\top R_i(t)} \left[|w| \|\bar{R}_w(t)\|_2^2 + |w^c| \|\bar{R}_{w^c}(t)\|_2^2 \right]$ where

$$\bar{R}_g(t) = \frac{1}{|g|} \sum_{i, s_i \in g} R_i(t) \quad (g \in \{w, w^c\}).$$

In the context of cluster detection, the null hypothesis is defined as $\mathcal{H}_0 : \forall w \in \mathcal{W}, \forall t \in \mathcal{T}, F_{w,t} = F_{w^c,t}$ where $F_{w,t}$ and $F_{w^c,t}$ correspond respectively to the cumulative distribution functions of $X(t)$ in w and outside w . The alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with a potential cluster w is $\mathcal{H}_1^{(w)}$:

$$\exists t \in \mathcal{T}, F_{w,t}(x) = F_{w^c,t}(x - \Delta_t), \Delta_t \neq 0.$$

Finally Frévent et al. (2021b) proposed to globalize the information over the time with the quantity $W^{(w)} = \sup_{t \in \mathcal{T}} W(t)^{(w)}$ and to use it as a concentration index to be maximized over the set of potential clusters \mathcal{W} . The multivariate rank-based functional spatial scan statistic (MRBFSS) is then $\Lambda_{\text{MRBFSS}} = \max_{w \in \mathcal{W}} W^{(w)}$.

Computing the statistical significance of the MLC

Once the most likely cluster (MLC) is detected, its statistical significance must be evaluated. The distribution of the scan statistic \mathcal{S} ($\mathcal{S} = \lambda_{\text{MG}}, \lambda_{\text{MNP}}, \Delta_{\text{PFSS}}, \Delta_{\text{NPFSS}}, \Delta_{\text{DFSS}}, \Delta_{\text{URBFSS}}, \Delta_{\text{MPFSS}}, \Delta_{\text{MDFSS}}$ or Λ_{MRBFSS}) is intractable under \mathcal{H}_0 due to the overlapping nature of \mathcal{W} . Then we choose to obtain a large set of simulated datasets by randomly permuting the observations X_i in the spatial locations. This technique was already used in spatial scan statistics (Kulldorff et al., 2009; Cucala et al., 2017).

Let M denote the number of random permutations of the original dataset and $\mathcal{S}^{(1)}, \dots, \mathcal{S}^{(M)}$ be the observed scan statistics on the simulated datasets. According to Dwass (1957) the p-value for \mathcal{S} observed in the real data is estimated by

$$\hat{p} = \frac{1 + \sum_{m=1}^M \mathbb{1}_{\mathcal{S}^{(m)} \geq \mathcal{S}}}{M + 1}. \quad (2)$$

Finally, the MLC is considered to be statistically significant if the associated \hat{p} is less than the type I error.

How to choose the method to apply to the data?

According to Cucala et al. (2019) the MNP method tends to present a better power and higher true positive rates for non-Gaussian data than the MG one. Although the false positive rates are often higher for this approach than the MG one, it remains moderate. The same conclusions are true for the UG (Kulldorff et al., 2009) and the UNP (Jung and Cho, 2015) which are their particular counterparts in the case of a single variable. In the functional framework, the approaches that present the best results are the DFFSS and the URBFSS in the univariate context and the MDFSS and the MRBFSS in the multivariate one (Frévent et al., 2021a,b). The URBFSS and the MRBFSS tend to show higher powers and higher true positive rates although they detect more false positives than the DFFSS and the MDFSS respectively. Table 1 summarizes the methods and their performances. The symbols \checkmark and \times indicate respectively a high and a low performance on the criterion. If there is no symbol it means that for this criterion the approach offers medium performances. The terminology "localized clusters in time" in the functional cases refers to aggregations of sites that take higher or lower values for the process only in a small interval of time (an interval of five days over a study period of one month for example). Table 2 gives an idea of the computation time of the different scanning methods proposed by the package. It should be noted that the computation time of the different spatial scan statistics methods is dependent on the size of the datasets, and in particular a function of the number of sites, the number of observation times and the number of variables considered.

3 Software

Computing the spatial scan statistic

The package **HDSpatialScan** provides a function `SpatialScan` to compute all the spatial scan statistics. The user chooses the method to apply by specifying the `method` argument: "MG" and "MNP" apply respectively the parametric and nonparametric spatial scan statistics approaches on multivariate data. Their univariate counterparts (when $p = 1$) can be computed with "UG" and "UNP" respectively. Then "PFSS", "DFSS" and "URBFSS" apply the parametric, the distribution-free and the new rank-based functional approaches on univariate functional data, and "MPFSS", "MDFSS", and "MRBFSS" are their multivariate counterparts. Finally "NPFSS" applies the nonparametric spatial scan statistic for functional data developed by Smida et al. (2022) on both univariate and multivariate functional data.

Table 1: Performance in terms of power, true positive rate and false positive rate of spatial scan statistics for multivariate data (MG and MNP), univariate functional data (PFSS, DFFSS, NPFSS and URBFS) and multivariate functional data (MPFSS, MDFSS, NPFSS and MRBFSS)

Method	Gaussian distribution			Non-Gaussian distribution		
	Power	True positive rate	False positive rate	Power	True positive rate	False positive rate
Univariate data						
UG ^a	✓	✓	✓	X	X	✓
UNP ^a	✓	✓		✓	✓	
Multivariate data						
MG ^b	✓	✓	✓	X	X	✓
MNP ^b	✓	✓		✓	✓	
Functional univariate data						
Non localized clusters in time						
PFSS ^c				X		✓
DFFSS ^d	✓	✓	✓	✓	✓	✓
NPFSS ^e		✓			✓	
URBFSS ^f	✓	✓		✓	✓	
Localized clusters in time						
PFSS ^c	X	X		X	X	
DFFSS ^d	✓	✓	✓	✓	✓	✓
NPFSS ^e	X			X		
URBFSS ^f	✓	✓	✓	✓	✓	✓
Functional multivariate data						
Non localized clusters in time						
MPFSS ^c				X	X	✓
MDFSS ^d	✓	✓	✓			✓
NPFSS ^e		✓		✓	✓	
MRBFSS ^f	✓	✓		✓	✓	
Localized clusters in time						
MPFSS ^c	X	X		X	X	✓
MDFSS ^d	✓	✓	✓			✓
NPFSS ^e	X			X		
MRBFSS ^f	✓	✓	✓	✓	✓	✓

^a The Univariate Gaussian (UG) and the Univariate Nonparametric (UNP) spatial scan statistics

^b The Multivariate Gaussian (MG) and the Multivariate Nonparametric (MNP) spatial scan statistics

^c The Parametric Functional (PFSS) and the Multivariate Parametric Functional (MPFSS) spatial scan statistics

^d The Distribution-free Functional (DFFSS) and the Multivariate Distribution-free Functional (MDFSS) spatial Scan Statistics

^e The Nonparametric Functional (NPFSS) spatial Scan Statistic

^f The Univariate Rank-based Functional (URBFSS) and the Multivariate Rank-based Functional (MRBFSS) spatial Scan Statistics

Table 2: Estimation of the computation time (over 100 repetitions) for the different scan statistics methods among 169 sites with *a priori* clusters comprising between 1 and 50% of the sites (default parameters) and 99 permutations for the estimation of the associated p-values (the default parameter of 999 permutations multiplies the computation time by about 10). Parallelization by running seven tasks in parallel was used (except for UG and UNP since these two methods are optimized to have a very low computation time without using CPUs) on two hexacores of type Intel(R) Xeon(R) CPU E5-2620 v2. For multivariate data (functional or not) 4 variables are considered and for functional data (univariate or multivariate) 56 observation times are considered.

Method	Computation time (in s)			
	Mean	Standard deviation	Minimum	Maximum
Univariate data				
UG ^a	4.31	0.18	4.01	4.77
UNP ^a	3.07	0.12	2.77	3.5
Multivariate data				
MG ^b	253.04	32.93	231.34	310.91
MNP ^b	14.6	1.44	13.48	17.77
Functional univariate data				
PFSS ^c	113.51	8.08	109.47	132.58
DFSS ^d	55.99	4.93	52.52	66.76
NPFSS ^e	12.91	1.23	11.74	15.6
URBFSS ^f	17.93	1.37	16.93	22.53
Functional multivariate data				
MPFSS ^c	72.52	9.21	65.23	100.17
MDFSS ^d	182.95	21.54	166.27	227.84
NPFSS ^e	22.95	1.7	21.77	28
MRBFSS ^f	244.04	21.83	234.05	305.56

^a The Univariate Gaussian (UG) and the Univariate Nonparametric (UNP) spatial scan statistics

^b The Multivariate Gaussian (MG) and the Multivariate Nonparametric (MNP) spatial scan statistics

^c The Parametric Functional (PFSS) and the Multivariate Parametric Functional (MPFSS) spatial scan statistics

^d The Distribution-free Functional (DFSS) and the Multivariate Distribution-free Functional (MDFSS) spatial Scan Statistics

^e The Nonparametric Functional (NPFSS) spatial Scan Statistic

^f The Univariate Rank-based Functional (URBFSS) and the Multivariate Rank-based Functional (MRBFSS) spatial Scan Statistics

Type of the data

Depending on the type of approach (univariate, multivariate, functional univariate or functional multivariate), the data must be formatted in a specific way. For univariate approaches, the data must be a vector in which each element corresponds to a site. If the data is individual and many individuals share the same site, the data can remain in an individual format with one element of the vector per individual. Then for real-valued multivariate methods or functional univariate methods, the data must be a matrix in which each row corresponds to a site (or an individual) and each column corresponds to a variable or an observation time in the functional framework. For multivariate functional methods the data must be a list in which each element is a matrix corresponding to a site (or an individual). In the matrices, the rows correspond to the variables and the columns to the observation times. Note that the observation times must be the same for each site or individual and they must be equally spaced for the methods "NPFSS", "PFSS" and "MPFSS". However if it is not the case of the raw data, they can be easily transformed by smoothing the data (Ramsay and Silverman, 2005), by using for example the R package `fda` (Ramsay et al., 2020).

Parameters of the scan function

The most important parameter is the method argument which has already been presented previously and allows to choose the spatial scan statistics to be applied. Note that you can choose one or more methods. Supplying "MPFSS" automatically computes the four strategies for the multivariate parametric functional spatial scan statistic (the Lawley-Hotelling trace (LH), the Roy's largest root (R), the Pillai's trace (P) and the Wilks' lambda (W)). If you only want the Lawley-Hotelling trace for example, you can simply supply "MPFSS-LH". Although the Lawley-Hotelling trace test is the most used statistic (Oja and Randles, 2004), it should be noted that all these methods usually provide very similar results. The other arguments are `data`, `sites_coord`, `system`, `mini`, `maxi`, `type_minimaxi`, `mini_post`, `maxi_post`, `type_minimaxi_post`, `sites_areas`, `MC`, `typeI`, `nbCPU`, `variable_names` and `times`. Note that `nbCPU` will be ignored for the methods "UG" and "UNP", `variable_names` is ignored for the univariate and univariate functional scan statistics and `times` is ignored for non-functional scan statistics.

The argument `data`, is the data vector, matrix or list on which the approaches must be applied. `MC` and `typeI` correspond respectively to the number of permutations of the data while computing the statistical significance of the clusters and the type I error i.e. a cluster is declared significant if its estimated p-value is below this threshold.

The arguments `sites_coord` and `system` are respectively a matrix of two columns corresponding to the coordinates of each site or individual, and to the system of coordinates ("Euclidean" or "WGS84"). The `sites_areas` argument is optional and corresponds to the areas of the sites (or the site of each individual if the data is individual).

The argument `nbCPU` permits to do parallelization and the arguments `mini`, `maxi`, `type_minimaxi`, `mini_post`, `maxi_post`, `type_minimaxi_post` are described further below.

`variable_names` is simply the names of the variables (in the same order as in the data) for multivariate or multivariate functional scan statistics and `times` corresponds to the times of observation, they must be numeric.

A priori filtering The clusters are computed automatically as circular clusters, so we need to define a minimum and a maximum size for these clusters. That is what we call "*a priori* filtering" and this allows to control the computation time. Three types of *a priori* filtering are possible through the argument `type_minimaxi`: "sites/indiv" (the filtering is applied on the number of sites or individuals in the potential clusters, it is the default value), "area" (it is applied on the area of the clusters and is available only if `sites_areas` is provided), or "radius" (the radius of the clusters).

The arguments `mini` and `maxi` are then respectively the minimum number of sites/individuals, or the minimal area or radius and the maximum number of sites/individuals, or the maximal area or radius. For the radius it is specified in km if `system` is "WGS84" or in the user units if `system` is "Euclidean". It should be noted that this filtering can bias the p-values obtained for the clusters. In order to perform a correct statistical inference, Kulldorff and Nagarwalla (1995) recommended to consider a maximum size of half the study region. Thus the default setting is to consider potential clusters comprising at least one site and at most 50% of the sites (Equation 1). If you want to select clusters according to size (number of sites or individuals), area or radius, it is better to select them *a posteriori* among the detected clusters and if you really want to decrease the computation time we recommend to increase the number of CPU (with the argument `nbCPU`). Changing the default settings can allow the user to investigate whether there appear to be clusters in a relatively quick first step, although the inference is biased, before applying the scan procedure with the default settings for the *a priori* filtering (50% of the studied region).

A posteriori filtering Sometimes after that the p-value of each potential cluster has been computing, the user may want to retrieve only the significant clusters that satisfy a certain size, area, or radius criteria. That is what we call *a posteriori* filtering. The corresponding arguments are `mini_post`, `maxi_post` and `type_minimaxi_post` and their definitions are the same as `mini`, `maxi` and `type_minimaxi`. If the user only wants to obtain clusters meeting size criteria, this *a posteriori* approach must be prioritized over the *a priori* approach which gives biased results and must therefore be used with great care.

Output of the scan function

The function `SpatialScan` returns a list of object of class `"ResScanOutput"` which is composed of many elements. The element `sites_clusters` is a list in which each element corresponds to a significant cluster and contains the index of the sites (or the individuals) included in this cluster. The clusters are listed in their order of detection. The secondary clusters are defined according to [Kulldorff \(1997\)](#): they correspond to potential clusters that also present large values for the concentration index. Their p-values are calculated as if they were the most likely cluster themselves which is a bit conservative since the secondary clusters are compared with the most likely cluster of the permutations ([Kulldorff, 1997](#)). Finally, only clusters that are significant at the `typeI` threshold and that do not overlap with a more likely cluster are returned, and `pval_clusters` corresponds to the associated p-values. The element `centres_clusters` corresponds to the coordinates of the centres of each detected cluster and `radius_clusters` is the radius of the clusters in km if `system` is `"WGS84"` or in the user units otherwise. `areas_clusters` corresponds to the areas of the clusters (in the same units as `sites_areas`). Finally the system of coordinates, the coordinates of the sites, the data and the name of the scan procedure are recalled respectively in the elements `system`, `sites_coord`, `data` and `method`.

Depending on the type of the method (univariate, multivariate, univariate functional or multivariate functional) the objects of class `"ResScanOutput"` are also of class `"ResScanOutputUni"`, `"ResScanOutputMulti"`, `"ResScanOutputUniFunc"` or `"ResScanOutputMultiFunc"`. The objects of class `"ResScanOutputMulti"` and `"ResScanOutputMultiFunc"` also include the element `variable_names`, and the objects of class `"ResScanOutputUniFunc"` and `"ResScanOutputMultiFunc"` include the element `time`.

Plot or summarize the results

It is possible to plot the detected clusters by using the classical `plot` function. Depending on the type parameter, the package `HDSpatialScan` provides three different types of plot.

The first one, `"map"`, allows the user to plot a map of the sites and draws the circles corresponding to the circular clusters. The second one, `"map2"`, plots the clusters in colors. For these two types of plot the argument `sproject` which is the spatial object corresponding to the sites, must be provided. If you do not have this object you can use the third type `"schema"` which simply draws a schema of the sites and the clusters, with the argument `system_conv` which allows to correctly project the coordinates. It must be entered as in the `PROJ` documentation ([PROJ contributors, 2021](#)).

One may also want to get some features of one's clusters.

The function `summary` allows to get a summary of the clusters, either the mean and the standard deviation of each of the variables (if many) if the argument `type_summ` is `"param"`, or the 25th percentiles, the medians and the 75th percentiles if the argument `type_summ` is `"nparam"`. This function also provides the p-values, the radius and the area if available (only if `sites_areas` is provided) for each cluster detected.

Other interesting functions are `plotCurves` that allows to display cluster curves (only in the functional case), and `plotSummary` which displays the average (if `type = "mean"`) or the median (if `type = "median"`) curves in the clusters, outside and the global mean or median curves in the functional case. For the multivariate non-functional framework it displays a spider chart of means or medians for each variable inside the cluster, outside, or in all the area. Note that all these functions take an argument only `.MLC` which allows to only plot or summarize the most likely cluster (by setting `only.MLC = TRUE`). Finally the `print` function shows the scan procedure used as well as the number of clusters detected and their p-value.

4 Illustrations

To show the simplicity of use of the package, we will apply the different approaches on the environmental data provided in the package. It should be noted that the codes presented in this section represent a total computation time of about one hour on a regular laptop, using 7 cores.

Air pollution in northern France

We considered data provided by the French national air quality forecasting platform PREV'AIR which is available in the package **HDSpatialScan**. This lattice data consists in the daily concentrations (from May 1, 2020 to June 25, 2020) in $\mu\text{g}\cdot\text{m}^{-3}$ of four pollutants for each of the 169 *cantons* (administrative subdivisions of France) of the *Nord-Pas-de-Calais* (a region in northern France) characterized by spatial polygons and located by their center of gravity s_1, \dots, s_{169} : nitrogen dioxide (NO_2), ozone (O_3) and fine particles PM_{10} and $\text{PM}_{2.5}$ corresponding respectively to particles whose diameter is less than $10\mu\text{m}$ and $2.5\mu\text{m}$. The package **HDSpatialScan** provides the full data: `fmulti_data` but also some reduced data for the univariate functional case which consists in considering only the NO_2 concentrations (`funi_data`), and for the multivariate non-functional framework (`multi_data`) which corresponds to the temporal mean concentrations of the four pollutants over the study period.

The first step is to load the data:

```
library(HDSpatialScan)
data("map_sites")
data("multi_data")
data("funi_data")
data("fmulti_data")
```

The second step is to visualize the pollutants daily concentration curves in each *canton* and the spatial distributions of the temporal mean concentrations for each pollutant over the studied time period (Figures 5 and 6). This step allows us to see if sites seem to aggregate and therefore if launching a cluster detection is relevant, and if a temporal variation of the concentrations is visible, in which case a functional method will be more relevant than a multivariate approach summarizing each curve by its mean.

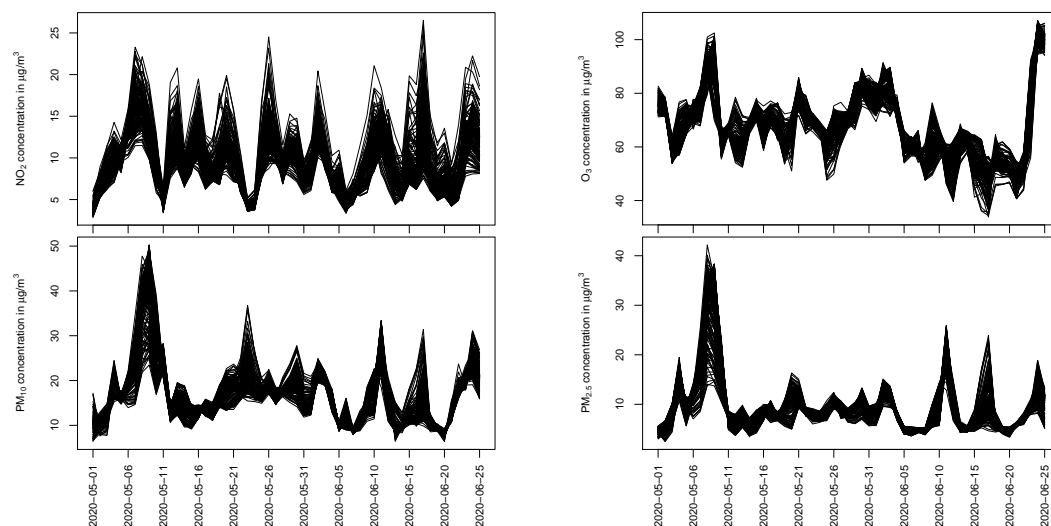


Figure 5: Daily concentration curves of NO_2 , O_3 , PM_{10} and $\text{PM}_{2.5}$ (from May 1, 2020 to June 25, 2020) in each of the 169 *cantons* of *Nord-Pas-de-Calais* (a region in northern France). A marked temporal variability in the concentrations of the four pollutants is observed over the study period.

The maps in Figure 6 show a spatial heterogeneity of the average concentration for each pollutant. Thus spatial scan statistics seem to be suitable to highlight the presence of *cantons*-level spatial clusters of pollutants concentrations. Moreover since the curves in Figure 5 show a marked temporal variability during the period from May 1, 2020 to June 25, 2020 a functional approach is more appropriate. However for sake of completeness we will also perform a multivariate spatial scan statistic approach anyway. Since small clusters of pollution are more relevant for interpretation because the sources of the pollutants are very localized, we will consider an *a posteriori* filtering of maximum radius equal to 10 km.

A multivariate spatial scan statistic

First we will investigate a multivariate spatial scan statistic. In this example the temporal component of the multivariate functional data was suppressed by averaging the components over the time and we looked for spatial clusters of the combination of the different air pollutants. This will pick up areas

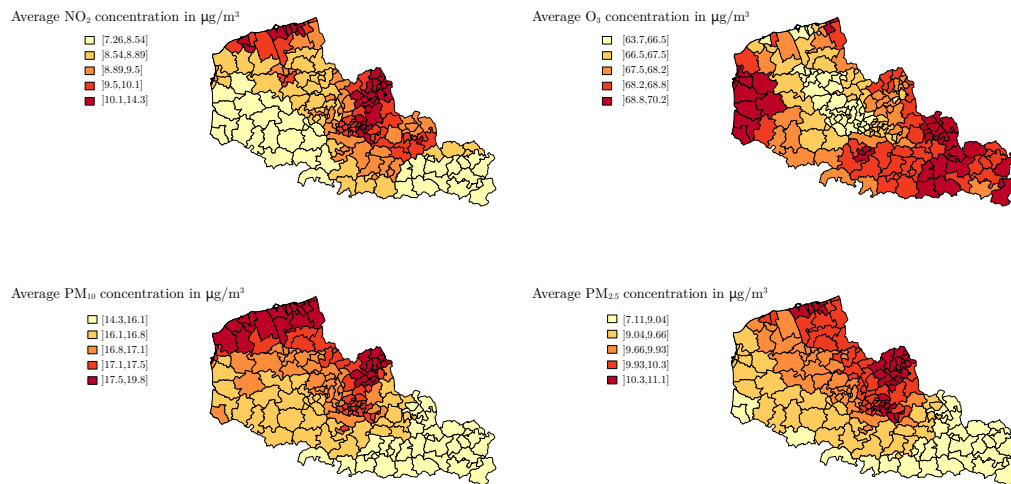


Figure 6: Spatial distributions of the average concentrations of NO₂, O₃, PM₁₀ and PM_{2.5} over the period from May 1, 2020 to June 25, 2020. A spatial heterogeneity of the average concentration (over the study period) is observed for the four pollutants. The spatial distributions of the average concentrations of PM₁₀ and PM_{2.5} are also observed to be similar.

of multiple exposure to pollutants or, on the contrary, areas with little pollution. We first checked the normality of each variable, which has been done by using a histogram and a qqplot. Since the distribution of the pollutants temporal mean concentrations is non-normal we decide to apply the MNP scan procedure. Here the system of coordinates is “WGS84”, it must be filled with the argument system. As explained in Section [Computing the spatial scan statistic](#), Kulldorff and Nagarwalla (1995) recommended to consider a maximum size of half the study region for the potential clusters so we use this *a priori* filtering with the parameters mini, maxi and type_minimaxi: the potential clusters are circular and they contain between 1 and 50% of the sites. Then as noticed in Section [Air pollution in northern France](#), we will apply an *a posteriori* filtering of maximum radius equal to 10 km (arguments mini_post, maxi_post and type_minimaxi_post). Here we only want to consider the significant clusters at the 5% threshold. Thus we leave the typeI parameter at its default value (0.05). However it should be noted that it is possible to obtain all the clusters (the MLC and the secondary clusters (Kulldorff, 1997)) by setting the typeI value at 1.

```
library(sp)
coords <- coordinates(map_sites)
res_mnp <- SpatialScan(method = "MNP", data = multi_data, sites_coord = coords,
+ system = "WGS84", mini = 1, maxi = nrow(coords)/2, type_minimaxi = "sites/indiv",
+ mini_post = 0, maxi_post = 10, type_minimaxi_post = "radius",
+ nbCPU = 7, MC = 99, variable_names = c("NO2", "O3", "PM10", "PM2.5"))$MNP
```

Once the scan procedure is completed, the plot function can be used. For brevity, we only focus on the MLC and for the sake of completeness we will show the use of the three possible visualizations of the clusters. Since we have a spatial object map_sites we can use the types “map” and “map2”. However for sake of completeness we also show the use of “schema” which allows to display the clusters otherwise (Figure 7). For the latter, since the system of the coordinates is “WGS84”, the plot function requires to complete the parameter system_conv which allows to correctly project the points. Here we choose the EPSG code 2154 corresponding to the Lambert 93 projection since the data is located in metropolitan France.

```
plot(x = res_mnp, type = "map", spobject = map_sites, only.MLC = TRUE)
plot(x = res_mnp, type = "map2", spobject = map_sites, only.MLC = TRUE)
plot(x = res_mnp, type = "schema", system_conv = "+init=epsg:2154", only.MLC = TRUE)
```

Finally users may want to get some summarized characteristics, such as the quantiles of the variables. This can be achieved by using the function summary with the argument type_summ equal to “nparam” (for the quantiles):

```
summary(res_mnp, type_summ = "nparam", only.MLC = TRUE)

## $basic_summary
##      Cluster 1
```

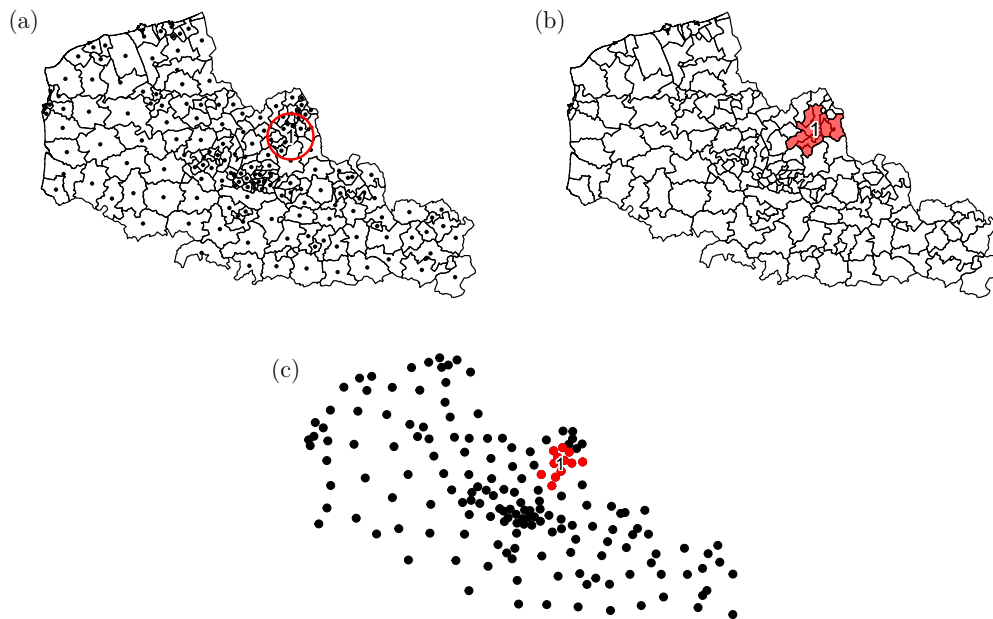


Figure 7: Visualization of the most likely cluster with the function plot with the types "map" (panel a), "map2" (panel b) and "schema" (panel c) for the MNP scan procedure computed with the function `SpatialScan` on the average concentrations of NO_2 , O_3 , PM_{10} and $\text{PM}_{2.5}$. The first two types of visualization require a spatial object corresponding to the spatial units (here the 169 *cantons* of the *Nord-Pas-de-Calais* region of northern France). The visualization option "schema" allows in the other case to draw a schema of the spatial units and the detected clusters. Here we focus the cluster visualization on the most likely cluster which is located in the urban area of Lille.

```
## p-value      0.001
## Radius      9.999
##
## $complete_summary
##           Overall Inside cluster 1 Outside cluster 1
## Number of sites 169.000           12.000           157.000
## Q25 NO2          8.673           11.327           8.635
## Median NO2       9.183           11.721           9.075
## Q75 NO2          9.848           12.382           9.692
## Q25 O3           66.778           67.527           66.721
## Median O3        67.895           67.609           67.961
## Q75 O3           68.564           67.922           68.658
## Q25 PM10         16.397           17.483           16.205
## Median PM10      16.970           17.877           16.933
## Q75 PM10         17.372           17.962           17.266
## Q25 PM2.5        9.132           10.584           9.113
## Median PM2.5     9.833           10.678           9.790
## Q75 PM2.5       10.213           10.919           10.107
```

The user can also use the function `plotSummary` to display the spider chart corresponding to the detected cluster (Figure 8).

```
plotSummary(res_mnp, type = "median", only.MLC = TRUE)
```

The MLC is located in the area of Lille. The summary and Figure 8 show that it is a cluster of overpollution (except for the pollutant O_3). This cluster is especially characterized by high concentrations of NO_2 and $\text{PM}_{2.5}$ which indicates pollution from road traffic and from the residential sector (auxiliary heating in particular). As the adverse health effects of air pollution (and their potential synergistic effect) are well established, such a result could inform local stakeholders about immediate interventions around the area of Lille to reduce the air pollution levels.

We have obtained some first results however the curves on Figure 5 present a marked temporal

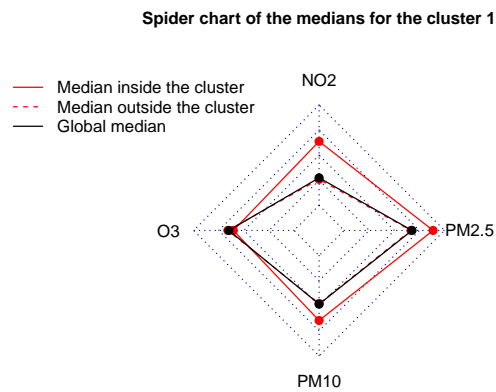


Figure 8: Spider chart obtained with the function `plotSummary` for the most likely cluster detected on the average concentrations of NO₂, O₃, PM₁₀ and PM_{2.5} by the MNP scan procedure in northern France. The most likely cluster is characterized by larger median concentrations of NO₂, PM₁₀ and PM_{2.5} than outside the cluster.

variability during the study period. Thus it could be interesting to apply functional spatial scan statistics.

A univariate functional spatial scan statistic

Here we only consider the pollutant NO₂. Applying a spatial scan statistic for univariate functional data will thus allow to highlight areas where the NO₂ concentration curves are abnormally high or, on the contrary abnormally low. We choose to use the URBFSF scan procedure since it often presents higher powers and true positive rates than the other univariate functional methods as its multivariate counterpart MRBFSS (Frévent et al., 2021b). As mentioned in Section [A multivariate spatial scan statistic](#) we decide to use the set of potential clusters *a priori* in the Equation 1 which corresponds to the recommended approach of Kulldorff and Nagarwalla (1995), and to the default values of the parameters `mini`, `maxi` and `type_minimaxi` in the scan functions. We also set a maximum radius equal to 10 km *a posteriori*.

```
res_urfss <- SpatialScan(method = "URBFSS", data = funi_data, sites_coord = coords,
+ system = "WGS84", mini = 1, maxi = nrow(coords)/2, type_minimaxi = "sites/indiv",
+ mini_post = 0, maxi_post = 10, type_minimaxi_post = "radius",
+ nbCPU = 7, MC = 99)$URBFSS
plot(res_urfss, type = "map2", sobject = map_sites, only.MLC = TRUE)
```

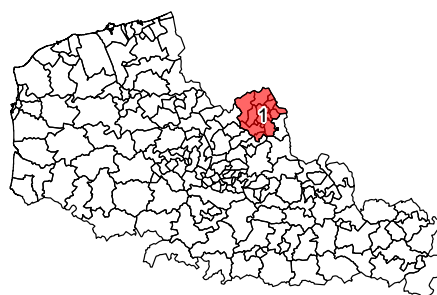


Figure 9: Visualization of the detected clusters with the function `plot` with `type = "map2"` for the URBFSF scan procedure computed using the concentrations of NO₂ in the 169 *cantons* of the *Nord-Pas-de-Calais* region of northern France over the period from May 1, 2020 to June 25, 2020. Here we focus the cluster visualization on the most likely cluster which is located in the urban area of Lille.

Again the MLC is located in the area of Lille (Figure 9).

For functional data another function is provided to give some characteristics of the clusters: we can visualize the curves in the cluster by adding the curve of the global median with the function

plotCurves. The function plotSummary allows to visualize the median curves inside and outside the cluster (Figure 10): this is a cluster of overexposure to NO₂, which indicates traffic-related air pollution. Since exposure to pollution impacts health negatively, these results can be used to intervene to reduce air pollution.

```
plotCurves(res_urbfss, add_median = TRUE, only.MLC = TRUE)
plotSummary(res_urbfss, type = "median", only.MLC = TRUE)
```

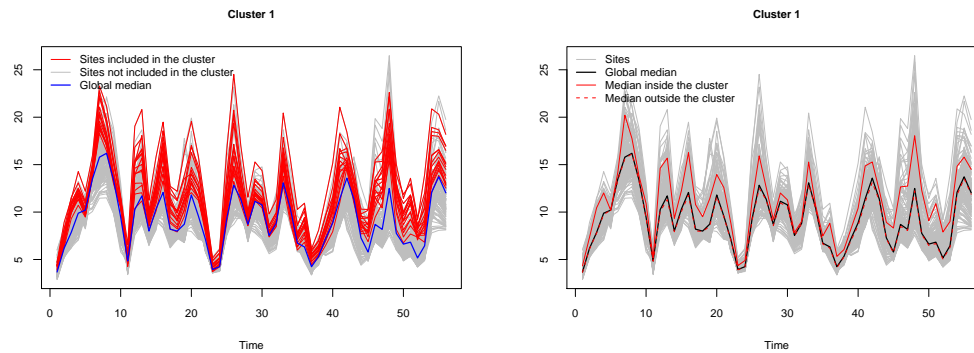


Figure 10: Characterization of the most likely cluster detected by the URBESS scan approach in the context of univariate functional data consisting of the NO₂ concentrations in northern France over the period from May 1, 2020 to June 25, 2020 with the functions plotCurves (left panel) and plotSummary (right panel). The most likely cluster is characterized by high concentration curves with a median concentration curve higher than outside the cluster.

A functional multivariate spatial scan statistic

Now we consider the four pollutants together. To detect spatial clusters of the combination of the four pollutants considering all available information on the time period, we apply a spatial scan statistic for multivariate functional data. It will identify geographical areas in which one or more of the pollutant concentration curves are abnormally high or abnormally low. For the same reason that we have previously chosen to apply the URBESS scan procedure, we use the MRBFSS in this context, with the same restrictions *a priori* and *a posteriori* as for the MNP and the URBESS scan approaches.

```
res_mrbfss <- SpatialScan(method = "MRBFSS", data = fmulti_data, sites_coord = coords,
+ system = "WGS84", mini = 1, maxi = nrow(coords)/2, type_minimaxi = "sites/indiv",
+ mini_post = 0, maxi_post = 10, type_minimaxi_post = "radius",
+ nbCPU = 7, MC = 99, variable_names = c("NO2", "O3", "PM10", "PM2.5"))$MRBFSS
plot(res_mrbfss, type = "map2", spobject = map_sites, only.MLC = TRUE)
```

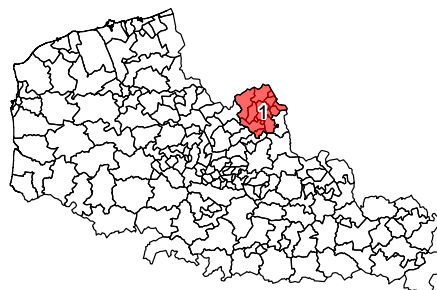


Figure 11: Visualization of the most likely cluster with the function plot with type = "map2" for the MRBFSS scan procedure computed using the concentrations of NO₂, O₃, PM₁₀ and PM_{2.5} in the 169 cantons of the Nord-Pas-de-Calais region of northern France over the period from May 1, 2020 to June 25, 2020. The most likely cluster is located in the urban area of Lille.

The detected cluster is exactly the same as before and is therefore located in the urban area of Lille (Figure 11).

Again we will display the curves in the cluster by adding the curve of the global median (Figure 12), as well as the median curves inside and outside the cluster which show that this is a cluster of high concentrations of NO_2 , PM_{10} and $\text{PM}_{2.5}$ (Figure 13). As mentioned in Section A *multivariate spatial scan statistic*, in environmental science it is well-known that NO_2 and $\text{PM}_{2.5}$ are more frequent in urban areas due to road traffic and population density so this is consistent with the cluster observed here. As the adverse health effects of air pollution and the combined effects of air pollutants are well established, this result could enable interventions by local authorities around the Lille area to reduce air pollution.

```
plotCurves(res_mrbfss, add_median = TRUE, only.MLC = TRUE)
plotSummary(res_mrbfss, type = "median", only.MLC = TRUE)
```

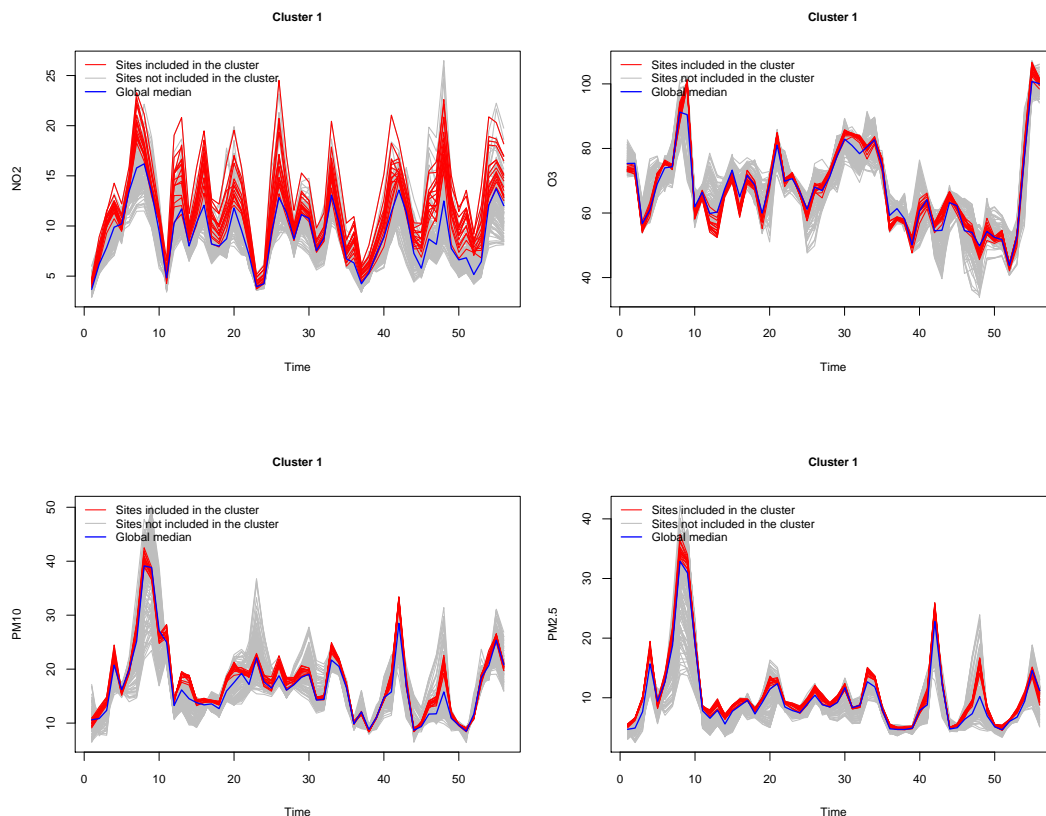


Figure 12: Characterization of the most likely cluster detected by the MRBFSS scan approach in the context of multivariate functional data consisting of the NO_2 , O_3 , PM_{10} and $\text{PM}_{2.5}$ concentrations in northern France over the period from May 1, 2020 to June 25, 2020 with the function `plotCurves`. The most likely cluster is characterized by high concentration curves of NO_2 , PM_{10} and $\text{PM}_{2.5}$.

5 Summary

In this article we presented the **HDSpatialScan** package. It makes it very easy to apply the existing scan statistics developed for multivariate data or functional data (univariate or multivariate), and the new rank-based scan statistic for univariate functional data presented in the Section *Models*. The potential clusters considered are of variable size and circular. In further updates of the package **HDSpatialScan** other shapes of scanning window such as elliptical or rectangular shapes will be implemented. Our package also allows to easily plot and summarize the detected clusters. Then examples of applications of the functions of the package have been shown. **HDSpatialScan** presents the advantage that all the scan procedures are applied using the same function `SpatialScan` and it uses the classical R functions `plot`, `print` and `summary` which makes it very quick to get started.

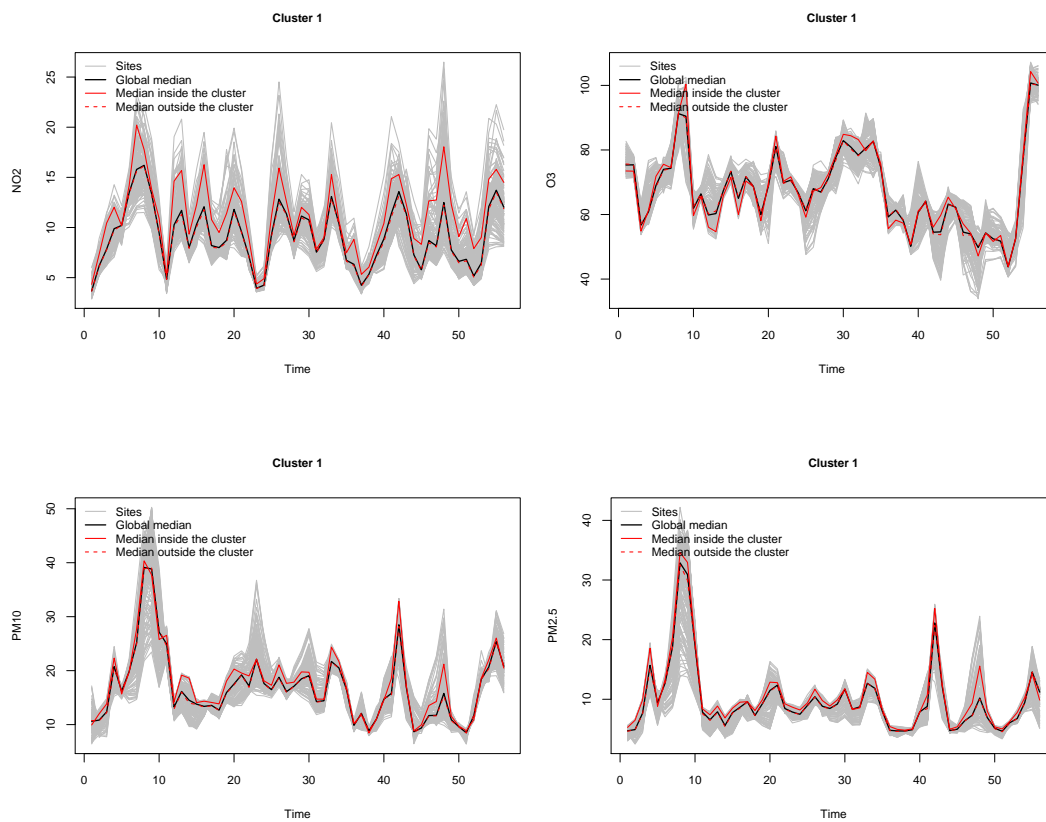


Figure 13: Characterization of the most likely cluster detected by the MRBFSS scan approach in the context of multivariate functional data consisting of the NO₂, O₃, PM₁₀ and PM_{2.5} concentrations in northern France over the period from May 1, 2020 to June 25, 2020 with the function `plotSummary`. The most likely cluster is characterized by median concentration curves of NO₂, PM₁₀ and PM_{2.5} higher than outside the cluster.

Bibliography

- B. Allévius. scanstatistics: Space-Time anomaly detection using scan statistics. *Journal of Open Source Software*, 3, 2018a. URL <https://doi.org/10.21105/joss.00515>. [p96]
- B. Allévius. *scanstatistics: Space-Time Anomaly Detection using Scan Statistics*, 2018b. URL <https://CRAN.R-project.org/package=scanstatistics>. [p96]
- J. Berrendero, A. Justel, and M. Svarc. Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55:2619–2634, 2011. URL <https://doi.org/10.1016/j.csda.2011.03.011>. [p96]
- V. Bhatt and N. Tiwari. A spatial scan statistic for survival data based on Weibull distribution. *Statistics in medicine*, 33(11):1867–1876, 2014. URL <https://doi.org/10.1002/sim.6075>. [p95]
- G. Boente and R. Fraiman. Kernel-based functional principal components. *Statistics & Probability Letters*, 48:335–345, 2000. URL [https://doi.org/10.1016/S0167-7152\(00\)00014-6](https://doi.org/10.1016/S0167-7152(00)00014-6). [p96]
- A. Chakraborty and P. Chaudhuri. A Wilcoxon-Mann-Whitney type test for infinite dimensional data. *Biometrika*, 102, 2014. URL <https://doi.org/10.1093/biomet/asu072>. [p100, 101]
- C. Chen, A. Y. Kim, M. Ross, and J. Wakefield. *SpatialEpi: Methods and Data for Spatial Epidemiology*, 2018. URL <https://CRAN.R-project.org/package=SpatialEpi>. [p96]
- J.-M. Chiou and H.-G. Müller. Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, 15:4849–4863, 06 2007. URL <https://doi.org/10.1016/j.csda.2006.07.042>. [p96]
- L. Cucala. A distribution-free spatial scan statistic for marked point processes. *Spatial Statistics*, 10: 117–125, 2014. URL <https://doi.org/10.1016/j.spasta.2014.03.004>. [p99, 101]
- L. Cucala. A Mann-Whitney scan statistic for continuous data. *Communications in Statistics-Theory and Methods*, 45(2):321–329, 2016. URL <https://doi.org/10.1080/03610926.2013.806667>. [p95]
- L. Cucala, C. Demattei, P. Lopes, and A. Ribeiro. A spatial scan statistic for case event data based on connected components. *Computational Statistics*, 28(1):357–369, 2013. URL <https://doi.org/10.1007/s00180-012-0304-6>. [p97]
- L. Cucala, M. Genin, C. Lanier, and F. Occelli. A multivariate Gaussian scan statistic for spatial data. *Spatial Statistics*, 21:66–74, 2017. URL <https://doi.org/10.1016/j.spasta.2017.06.001>. [p96, 98, 105]
- L. Cucala, M. Genin, F. Occelli, and J. Soula. A multivariate nonparametric scan statistic for spatial data. *Spatial statistics*, 29:1–14, 2019. URL <https://doi.org/10.1016/j.spasta.2018.10.002>. [p96, 99, 105]
- A. Cuevas, M. Febrero-Bande, and R. Fraiman. Linear functional regression: The case of fixed design and functional response. *The Canadian Journal of Statistics*, 30:285–300, 2002. URL <https://doi.org/10.2307/3315952>. [p96]
- A. Cuevas, M. Febrero-Bande, and R. Fraiman. An ANOVA test for functional data. *Computational Statistics & Data Analysis*, 47:111–122, 2004. URL <https://doi.org/10.1016/j.csda.2003.10.021>. [p100]
- C. Demattei, N. Molinari, and J.-P. Daurès. SPATCLUS: An R package for arbitrarily shaped multiple spatial cluster detection for case event data. *Computer Methods and Programs in Biomedicine*, 84(1): 42–49, 2006. ISSN 0169-2607. URL <https://doi.org/10.1016/j.cmpb.2006.07.008>. [p96]
- H. Durbeck, D. Greiling, L. Estberg, A. Long, G. Jacquez, Y. Pallicaris, and S. Hinton. *ClusterSeer: Software for the Detection and Analysis of Event Clusters, User Manual Book 2, Version 2.5*, 2012. [p96]
- M. Dwass. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187, 1957. URL <https://doi.org/10.1214/aoms/1177707045>. [p105]
- F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17:545–564, 2002. URL <https://doi.org/10.1007/s001800200126>. [p96]
- C. Frévent, M.-S. Ahmed, M. Marbac, and M. Genin. Detecting spatial clusters in functional data: New scan statistic approaches. *Spatial Statistics*, page 100550, 2021a. URL <https://doi.org/10.1016/j.spasta.2021.100550>. [p96, 100, 101, 104, 105]

- C. Frévent, M.-S. Ahmed, S. Dabo-Niang, and M. Genin. Investigating spatial scan statistics for multivariate functional data. arXiv:2103.14401v1, 2021b. [p96, 101, 103, 104, 105, 113]
- J. Gao, Z. Zhang, Y. Hu, J. Bian, W. Jiang, X. Wang, L. Sun, and Q. Jiang. Geographical distribution patterns of iodine in drinking-water and its associations with geological factors in Shandong province, China. *International Journal of Environmental Research and Public Health*, 11(5):5431–5444, 2014. URL <https://doi.org/10.3390/ijerph110505431>. [p95]
- M. Genin, M. Fumery, F. Occelli, G. Savoye, B. Pariente, L. Dauchet, J. Giovannelli, C. Vignal, M. Body-Malapel, H. Sarter, et al. Fine-scale geographical distribution and ecological risk factors for Crohn's disease in France (2007-2014). *Alimentary Pharmacology & Therapeutics*, 51(1):139–148, 2020. URL <https://doi.org/10.1111/apt.15512>. [p95]
- V. Gómez-Rubio, J. Ferrándiz-Ferragud, and A. López-Quílez. *DCluster: Functions for the Detection of Spatial Clusters of Diseases*, 2015. URL <https://CRAN.R-project.org/package=DCluster>. [p96]
- V. Gómez-Rubio, P. Moraga, J. Molitor, and B. Rowlingson. DClusterm: Model-based detection of disease clusters. *Journal of Statistical Software*, 90(14):1–26, 2019. URL <https://doi.org/10.18637/jss.v090.i14>. [p96]
- V. Gomez-Rubio, P. E. M. Serrano, and B. Rowlingson. *DClusterm: Model-Based Detection of Disease Clusters*, 2020. URL <https://CRAN.R-project.org/package=DClusterm>. [p96]
- T. Górecki and Ł. Smaga. A comparison of tests for the one-way ANOVA problem for functional data. *Computational Statistics*, 30:987–1010, 2015. URL <https://doi.org/10.1007/s00180-015-0555-0>. [p100]
- T. Górecki and Ł. Smaga. Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, 44(12):2172–2189, 2017. URL <https://doi.org/10.1080/02664763.2016.1247791>. [p103]
- D. Greiling, L. Estberg, A. Long, and G. Jacquez. *ClusterSeer: Software for the Detection and Analysis of Event Clusters, User Manual Book 1, Version 2.5*, 2012. [p96]
- L. Huang, M. Kulldorff, and D. Gregorio. A spatial scan statistic for survival data. *Biometrics*, 63(1): 109–118, 2007. URL <https://doi.org/10.1111/j.1541-0420.2006.00661.x>. [p95]
- I. Jung and H. J. Cho. A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics*, 14, 2015. URL <https://doi.org/10.1186/s12942-015-0024-6>. [p95, 99, 101, 102, 105]
- I. Jung, M. Kulldorff, and A. C. Klassen. A spatial scan statistic for ordinal data. *Statistics in Medicine*, 26(7):1594–1607, 2007. URL <https://doi.org/10.1002/sim.2607>. [p95]
- M. M. Khan, S. Roberson, K. Reid, M. Jordan, and A. Odoi. Geographic disparities and temporal changes of diabetes prevalence and diabetes self-management education program participation in Florida. *Plos one*, 16(7), 2021. URL <https://doi.org/10.1371/journal.pone.0254579>. [p95]
- K. Kleinman. *rsatscan: Tools, Classes, and Methods for Interfacing with SaTScan Stand-Alone Software*, 2015. URL <https://CRAN.R-project.org/package=rsatscan>. [p96]
- M. Kulldorff. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26:1481–1496, 1997. URL <https://doi.org/10.1080/03610929708831995>. [p95, 109, 111]
- M. Kulldorff. *TreeScan User Guide*, 2018. URL <https://www.treescan.org/>. [p96]
- M. Kulldorff. *SaTScan User Guide for Version 9.7*, 2021. URL <https://www.satscan.org/>. [p96]
- M. Kulldorff and N. Nagarwalla. Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995. URL <https://doi.org/10.1002/sim.4780140809>. [p95, 97, 108, 111, 113]
- M. Kulldorff, W. Athas, E. Feurer, B. Miller, and C. Key. Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos, New Mexico. *The American Journal of Public Health*, 88(9): 1377–1380, 1998. URL <https://doi.org/10.2105/AJPH.88.9.1377>. [p96]
- M. Kulldorff, Z. Fang, and S. J. Walsh. A tree-based scan statistic for database disease surveillance. *Biometrics*, 59(2):323–331, 2003. URL <https://doi.org/10.1111/1541-0420.00039>. [p96]
- M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal. An elliptic spatial scan statistic. *Statistics in Medicine*, 25:3929–3943, 2006. URL <https://doi.org/10.1002/sim.2490>. [p97]

- M. Kulldorff, F. Mostashari, L. Duczmal, W. K. Yih, K. Kleinman, and R. Platt. Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, 26:1824–1833, 2007. URL <https://doi.org/10.1002/sim.2818>. [p95, 96]
- M. Kulldorff, L. Huang, and K. Konty. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, 8(58), 2009. URL <https://doi.org/10.1186/1476-072X-8-58>. [p95, 99, 105]
- Z. Lin, M. E. Lopes, and H.-G. Müller. High-dimensional MANOVA via bootstrapping and its application to functional and sparse count data. *Journal of the American Statistical Association*, 2021. URL <https://doi.org/10.1080/01621459.2021.1920959>. [p101]
- R. Loche, B. Giron, D. Abrial, L. Cucala, M. CharrasGarrido, and J. De-Goer. *graphscan: Cluster Detection with Hypothesis Free Scan Statistic*, 2016. URL <https://CRAN.R-project.org/package=graphscan>. [p96]
- L. H. S. C. Marciano, A. de Faria Fernandes Belone, P. S. Rosa, N. M. B. Coelho, C. C. Ghidella, S. M. T. Nardi, W. C. Miranda, L. V. Barrozo, and J. C. Lastória. Epidemiological and geographical characterization of leprosy in a Brazilian hyperendemic municipality. *Cadernos de saude publica*, 34, 2018. URL <https://doi.org/10.1590/0102-311X00197216>. [p95]
- P. Moraga. *SpatialEpiApp: A Shiny Web Application for the Analysis of Spatial and Spatio-Temporal Disease Data*, 2017a. URL <https://CRAN.R-project.org/package=SpatialEpiApp>. [p96]
- P. Moraga. SpatialEpiApp: A Shiny Web Application for the Analysis of Spatial and Spatio-Temporal Disease Data. *Spatial and Spatio-temporal Epidemiology*, 23, 2017b. URL <https://doi.org/10.1016/j.sste.2017.08.001>. [p96]
- H. Oja and R. H. Randles. Multivariate nonparametric tests. *Statistical Science*, 18(4):598–605, 11 2004. URL <https://doi.org/10.1214/088342304000000558>. [p99, 104, 108]
- T. Otani and K. Takahashi. *rflexscan: The Flexible Spatial Scan Statistic*, 2021. URL <https://CRAN.R-project.org/package=rflexscan>. [p96]
- PROJ contributors. *PROJ coordinate transformation software library*. Open Source Geospatial Foundation, 2021. URL <https://proj.org/>. [p109]
- Z. Qiu, J. Chen, and J.-T. Zhang. Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis*, 157, 2021. URL <https://doi.org/10.1016/j.csda.2020.107160>. [p104]
- J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag, 2005. [p96, 108]
- J. O. Ramsay, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2020. URL <https://CRAN.R-project.org/package=fda>. [p108]
- G. Shi, J. Liu, and X. Zhong. Spatial and temporal variations of PM_{2.5} concentrations in Chinese cities during 2015-2019. *International Journal of Environmental Health Research*, pages 1–13, 2021. URL <https://doi.org/10.1080/09603123.2021.1987394>. [p95]
- Z. Smida, L. Cucala, A. Gannoun, and G. Durif. A Wilcoxon-Mann-Whitney spatial scan statistic for functional data. *Computational Statistics & Data Analysis*, 167:107378, 2022. URL <https://doi.org/10.1016/j.csda.2021.107378>. [p96, 100, 101, 105]
- D. L. Sudakin, Z. Horowitz, and S. Giffin. Regional variation in the incidence of symptomatic pesticide exposures: Applications of geographic information systems. *Journal of Toxicology: Clinical Toxicology*, 40(6):767–773, 2002. URL <https://doi.org/10.1081/CLT-120015837>. [p95]
- K. Takahashi and T. Tango. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4(11), 05 2005. URL <https://doi.org/10.1186/1476-072X-4-11>. [p96]
- K. Takahashi, T. Yokoyama, and T. Tango. *FleXScan v 3.1: Software for the Flexible Scan Statistic*, 2010. [p96]
- L. Wan, Y. Sun, I. Lee, W. Zhao, and F. Xia. Industrial pollution areas detection and location via satellite-based IIoT. *IEEE Transactions on Industrial Informatics*, 17(3):1785–1794, 2020. URL <https://doi.org/10.1109/TII.2020.2992658>. [p95]

Camille Frévent

*University of Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, INRIA-MODAL
F-59000 Lille France
camille.frevent@univ-lille.fr*

Mohamed-Salem Ahmed

*University of Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales
F-59000 Lille France*

Julien Soula

*University of Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales
F-59000 Lille France*

Lionel Cucala

*IMAG, Université de Montpellier, CNRS
Montpellier
France*

Zaineb Smida

*IMAG, University of Montpellier, CNRS
Montpellier
France*

Sophie Dabo-Niang

*Laboratoire Paul Painvelé UMR CNRS 8524, INRIA-MODAL, University of Lille
F-59000 Lille France*

Michaël Genin

*University of Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales
F-59000 Lille France*