Detecting spatial clusters in functional data: New scan statistic approaches

Camille Frevent, Mohamed-Salem Ahmed, M. Marbac, Michaël Genin

Detecting spatial clusters in functional data: new scan statistic approaches

Camille Frévent[1], Mohamed-Salem Ahmed[1] , Matthieu Marbac[2] and Michaël Genin[1]

[1]Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France.

[2]Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France.

Abstract

We have developed two scan statistics for detecting clusters of functional data indexed in space. The first method is based on an adaptation of a functional analysis of variance and the second one is based on a distribution-free spatial scan statistic for univariate data. In a simulation study, the distribution-free method always performed better than a nonparametric functional scan statistic, and the adaptation of the anova also performed better for data with a normal or a quasi-normal distribution. Our methods can detect smaller spatial clusters than the nonparametric method. Lastly, we used our scan statistics for functional data to search for spatial clusters of abnormal unemployment rates in France over the period 1998-2013 (divided into quarters).

Keywords: Cluster detection, functional data, spatial scan statistics

1 Introduction

Spatial cluster detection has been studied for many years. The goal is usually to develop new tools capable of detecting the aggregation of spatial sites that behave "differently" from other sites. There are three main categories of cluster detection tests: (i) global clustering tests, designed to determine whether some spatial observations tend to aggregate (but without locating the corresponding cluster), (ii) focused cluster detection tests, which screen for the presence of one or more clusters around a fixed focus (but which require prior knowledge, since a focus determined on the data generates pre-selection bias), and (iii) non-focused cluster detection tests, designed to detect statistically significant spatial clusters in the absence of any a priori information about the latter's location. Spatial scan statistics are well-known members of the third category and have been applied in many fields of applied research, such as epidemiology (Kulldorff, 1999; Luquero et al., 2011; Genin et al., 2020), environmental sciences (Chong et al., 2013; Duncan et al., 2016), and geology (Gao et al., 2014).

Scan statistics were first introduced by Naus for the one-dimensional case Naus (1965b) and then the two-dimensional case Naus (1965a). In fact, spatial scan statistics can be viewed as the extension of two-dimensional scan statistics to spatial data. This approach was initially suggested

by Kulldorff & Nagarwalla (1995) and Kulldorff (1997) for Bernouilli and Poisson models, respectively. The researchers used a method based on the likelihood ratio and Monte-Carlo testing to detect significant clusters of various sizes and shapes. Following on from Kulldorff's initial work, several other researchers have adapted spatial scan statistics to other spatial data distributions, such as ordinal (Jung et al., 2007), normal (Kulldorff et al., 2009), exponential (Huang et al., 2007), and Weibull models (Bhatt & Tiwari, 2014). Kulldorff et al. (2007) developed a parametric scan statistic for application to multivariate spatial data. However, this method is based on the combination of independent univariate scan statistics and fails to take account of the correlations between variables. This obstacle was circumvented by Cucala et al. (2017), who developed a spatial scan statistic based on a likelihood ratio and a multivariate normal probability model that takes account of the correlations between variables.

Thanks to progress in sensing and data storage capacity, data are increasingly being measured continuously over time. This led to the introduction of functional data analysis (FDA) by Ramsay & Silverman (2005a). A considerable amount of work has gone into the representation, exploration and modelling of functional data and the adaptation of classical statistical methods to the functional framework (e.g. principal component analysis, Boente & Fraiman, 2000; Berrendero et al., 2011) and regression (Cuevas et al., 2002; Ferraty & Vieu, 2002; Chiou & Müller, 2007). Nonparametric methods have also been developed for FDA (for a review, see Ferraty & Vieu, 2006). FDA is thus an active research topic with potential applications in many fields.

In the field of spatial scan statistics, the application of a "naive" univariate approach (based on the time-averaged data in each spatial location over the studied period) would lead to a huge loss of information. A multivariate method that considers each observation time point as a variable would be face with high dimensionality and high correlation issues. Recently, Smida et al. (2020) developed a nonparametric approach based on a functional Wilcoxon-Mann-Whitney test. To the best of our knowledge, a parametric scan statistic for functional data has not previously been suggested. We reasoned that since an analysis of variance (ANOVA) for functional data has been described by Cuevas et al. (2004) it should be possible to develop a scan procedure based on this test. Then in the last few years, statistical tests for high-dimensional data were developed by resuming the information after the computation of the statistics for each component of the data (Lin et al., 2021). Hence we proposed a scan statistic based on the combination of this approach and the distribution-free scan statistic approach proposed by Cucala (2014).

Here, we describe a parametric spatial scan statistic for functional data based on a functional ANOVA and another one based on the distribution-free scan statistic. The methodology is presented in Section 2. In Section 3, the new approaches' performances in a simulation study are described and compared with that of the Smida et al. (2020) method. Our methods' application to a real dataset is presented in Section 4. Lastly, the results of this work are discussed in Section 5.

2

2 Methodology

2.1 General principle

Let $s_1, \ldots, s_n$ be n non-overlapping locations of an observation domain $S \subset \mathbb{R}^2$ and $X_1, \ldots, X_n$ be the observations of X in $s_1, \ldots, s_n$. Hereafter, all observations are considered to be independent; this is a classical assumption in scan statistics. Spatial scan statistics are designed to detect spatial clusters and to test 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).

When $X \in \mathbb{R}$, parametric univariate scan statistics can detect aggregates of locations in which the mean of X is higher or lower than the mean of X outside. To this end, they often use likelihood ratios (Kulldorff et al., 2009). In such a case, a cluster w can be defined by $\mathbb{E}[X_i \mid s_i \in w] = \mathbb{E}[X_i \mid s_i \notin w] + \Delta$ where $\Delta \neq 0$.

In the same way, in the multivariate case ($X \in \mathbb{R}^d$, $d \geq 2$), parametric spatial scan statistics also use likelihood ratios (Kulldorff et al., 2007; Cucala et al., 2017) to detect aggregates of locations in which the values of the random vector X are abnormally high or low. Hence, a cluster w in this framework can be characterized by $\mathbb{E}[X_i \mid s_i \in w] = \mathbb{E}[X_i \mid s_i \notin w] + \Delta$ where $\Delta = (\Delta_1, \ldots, \Delta_d)^\top \neq 0$, and for all $i \in [\![1;d]\!]$, $\Delta_i \geq 0$ or for all $i \in [\![1;d]\!]$, $\Delta_i \leq 0$.

In the functional framework, $\{X(t),\ t \in \mathcal{T}\}$ is a real-valued stochastic process where $\mathcal{T}$ is an interval of $\mathbb{R}$. For functional data, and in line with Dai & Genton's work (Dai & Genton, 2019) on magnitude outlyingness and shape outlyingness, two types of clusters can be distinguished: *magnitude clusters* and *shape clusters*. Thus, by analogy with the univariate and multivariate definitions, a *magnitude cluster w* can be defined as:

$$\forall t \in \mathcal{T},\ \mathbb{E}[X_i(t) \mid s_i \in w] = \mathbb{E}[X_i(t) \mid s_i \notin w] + \Delta(t), \tag{1}$$

where Δ is of constant sign, and non-zero over at least one sub-interval of $\mathcal{T}$.
In the same way, a *shape cluster w* can be defined as:

$$\forall t \in \mathcal{T},\ \mathbb{E}[X_i(t) \mid s_i \in w] = \mathbb{E}[X_i(t) \mid s_i \notin w] + \Delta(t) \tag{2}$$

where Δ is not constant almost everywhere.
Cressie (1977) defined a spatial scan statistic as the maximum of a concentration index over a set of potential clusters $\mathcal{W}$. The spatial scan statistic thus depends on the choice of the concentration index and of $\mathcal{W}$. In the following and without loss of generality, $\mathcal{W}$ is composed of circular clusters (as introduced by Kulldorff (1997)), with $|w|$ denoting the number of sites in w:

$$\mathcal{W} = \{w_{i,j}\ /\ 1 \leq |w_{i,j}| \leq \frac{n}{2},\ 1 \leq i,j \leq n\}, \tag{3}$$

where $w_{i,j}$ is the disc centred on s_i that passes through s_j: a cluster cannot cover more than 50% of the studied region. It should be noted that researchers have suggested many other shapes, such as

elliptical clusters (Kulldorff et al., 2006), rectangular clusters (Chen & Glaz, 2009) or graph-based clusters (Cucala et al., 2013).

From now on, we focused to the case of functional univariate data. In subsection 2.2 we proposed a parametric scan statistic and a distribution-free one in subsection 2.3.

2.2 A parametric spatial scan statistic for univariate functional data

As in the univariate and multivariate frameworks, the parametric concentration index defined here will compare the mean functions between the potential clusters and the outside.

In this subsection, the process X is supposed to take values in the space $L^2(\mathcal{T}, \mathbb{R})$ of real valued, square-integrable functions on $\mathcal{T}$.

Cuevas et al. (2004) and Górecki & Smaga (2015) adapted the classical ANOVA F-statistic for L^2 processes. Without loss of generality, and considering two independent samples of trajectories drawn from two L^2 processes X_{g_1} and X_{g_2} in two groups g_1 and g_2, the test compares the mean functions μ_{g_1} and μ_{g_2} where $\mu_{g_i}(t) = \mathbb{E}[X_{g_i}(t)]$, $i = 1, 2$.

Thus, in the context of cluster detection, the null hypothesis $\mathcal{H}_0$ (the absence of cluster) 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. 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, the functional ANOVA can be used to compare the mean function in w with the mean function in w^c by using the following statistic:

$$F_n^{(w)} = \frac{|w|\,||\bar{X}_w - \bar{X}||_2^2 + |w^c|\,||\bar{X}_{w^c} - \bar{X}||_2^2}{\frac{1}{n-2}\left[\sum_{j, s_j \in w} ||X_j - \bar{X}_w||_2^2 + \sum_{j, s_j \in w^c} ||X_j - \bar{X}_{w^c}||_2^2\right]}, \tag{4}$$

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 $||x||_2^2 = \int_{\mathcal{T}} x^2(t)\,\mathrm{d}t$.

Thus, $F_n^{(w)}$ is considered to be a concentration index and is maximized over the set of potential clusters $\mathcal{W}$, yielding to the following definition of the parametric functional spatial scan statistic (PFSS):

$$\Lambda_{\text{PFSS}} = \max_{w \in \mathcal{W}} F_n^{(w)}. \tag{5}$$

The potential cluster for which this maximum is obtained (namely, the most likely cluster (MLC)) is therefore

$$\text{MLC} = \arg\max_{w \in \mathcal{W}} F_n^{(w)}. \tag{6}$$

2.3 A distribution-free spatial scan statistic for univariate functional data

Here, we propose to combine the distribution-free spatial scan statistic for univariate data proposed by Cucala (2014) and the max statistic of Lin et al. (2021). Briefly, the latter proposed a new approach to the problem of the functional ANOVA by maximizing a statistic over the time.

In line with the work of Cucala (2014), we suppose that for each time t, $\mathbb{V}[X_i(t)] = \sigma^2(t)$ for all $i \in [\![1;n]\!]$. Then for each t, the concentration index proposed by Cucala (2014) to test $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w(t) = \mu_{w^c}(t) = \mu_S(t)$ is

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

where $\bar{X}_g(t) = \dfrac{1}{|g|} \sum_{i, s_i \in g} X_i(t)$ are empirical estimators of μ_g ($g \in \{w, w^c\}$) and

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

Here $\widehat{\sigma^2}(t)$ needs to be computed since it depends on t:

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

is the pooled sample variance at time t.

So now 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: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$.

$I^{(w)}$ can be considered as a concentration index and maximized over the set of potential clusters $\mathcal{W}$ yielding to the following distribution-free functional spatial scan statistic (DFFSS):

$$\Lambda_{\text{DFFSS}} = \max_{w \in \mathcal{W}} I^{(w)}. \tag{7}$$

The potential cluster for which this maximum is obtained (namely, the most likely cluster (MLC)) is therefore

$$\text{MLC} = \arg\max_{w \in \mathcal{W}} I^{(w)}. \tag{8}$$

2.4 Computing the significance of the MLC

Once the MLC has been detected, its statistical significance must be evaluated. However, due to the overlapping nature of $\mathcal{W}$, the distribution of the scan statistic Λ (Λ_{PFSS} or Λ_{DFFSS}) does not have closed form under $\mathcal{H}_0$. Hence, we created a large set of random datasets by randomly

permuting the observations X_i in the spatial locations. This technique is called "random labelling" and was already used in spatial scan statistics (Kulldorff et al., 2009; Cucala et al., 2017, 2019). Let M denote the number of random permutations of the original dataset and $\Lambda^{(1)}, \ldots, \Lambda^{(M)}$ be the scan statistics observed in the simulated datasets. According to Dwass (1957) the p-value for Λ with real data is estimated by

$$\hat{p} = \frac{1 + \sum_{m=1}^{M} \mathbb{1}_{\Lambda^{(m)} \geq \Lambda}}{M + 1}. \tag{9}$$

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

3   The simulation study

In a simulation study, we compared (i) the parametric functional spatial scan statistic (PFSS) Λ_{PFSS}, (ii) the distribution-free functional spatial scan statistic (DFFSS) Λ_{DFFSS} and (iii) the nonparametric functional spatial scan statistic (NPFSS) Λ_{NPFSS} developed by Smida et al. (2020). As stated by Smida et al. (2020), calculation of the NPFSS is very time-consuming because $|w|(n - |w|)$ terms have to be summed for each potential cluster. Although Smida et al. (2020) suggested a computational improvement, the number of calculations was still high. Here, we found a way of improving the NPFSS computation that significantly reduced the computing time. Details of this optimization and examples of computation time are provided in the Appendix (Section B).

3.1   Design of the simulation study

Artificial datasets were generated by using the geographic locations of the 94 French *départements* (county-type administrative areas; Figure 4). The location of each *département* was defined by its administrative capital. A true cluster w (composed of eight *départements* in the Paris region; the red area, see Figure 4 in Supplementary materials) was defined and simulated for each artificial dataset.

3.1.1   Generation of the artificial datasets

The X_i are generated using the following model (see Qiu et al., 2021, for more details) and measured at 101 equally spaced times on $[0; 1]$:

$$\text{for each } i \in [\![1; 94]\!], \ X_i(t) = \sin\left[2\pi t^2\right]^5 + \Delta(t)\mathbb{1}_{s_i \in w} + \varepsilon_i(t), \ t \in [0; 1]$$

where $\varepsilon_i(t) = \sum_{k=1}^{7} \sqrt{1.5 \times 0.2^k}(v_{i,1,k} - v_{i,2,k})\Psi_k(t)$ and

$$\Psi_k(t) = \begin{cases} 1 & \text{if } k = 1 \\ \sqrt{2}\sin\left[k\pi t\right] & \text{if } k \text{ even} \\ \sqrt{2}\cos\left[(k-1)\pi t\right] & \text{if } k \text{ odd and } k > 1 \end{cases}$$

6

Three distributions for the $v_{i,j,k}, j = 1, 2$ were considered: (i) $v_{i,j,k} \sim \mathcal{N}(0,1)$, (ii) $v_{i,j,k} \sim t(4)/\sqrt{2}$ and (iii) $v_{i,j,k} \sim (\chi^2(4) - 4)/(2\sqrt{2})$.

Three types of clusters were simulated with an intensity controlled by a parameter $\alpha > 0$. The three chosen shifts Δ vary over time and are positive and non-zero (except possibly when $t = 0$ or $t = 1$): $\Delta_1(t) = \alpha t$, $\Delta_2(t) = \alpha t(1 - t)$ and $\Delta_3(t) = \alpha \exp\left[-100(t - 0.5)^2\right]/3$. Thus, these clusters correspond to both magnitude clusters and shape clusters. Different values of the true cluster intensity were considered for each Δ. Since Δ_2 and Δ_3 takes lower values over time than Δ_1 for the same value of α, it needs greater α values. Thus, $\alpha \in \{0; 0.75; 1.5; 2.25; 3\}$ for Δ_1, $\alpha \in \{0; 2; 4; 6; 8\}$ for Δ_2 and $\alpha \in \{0; 2.5; 5; 7.5; 10\}$ for Δ_3. It is noteworthy that $\alpha = 0$ was also tested in order to evaluate the maintenance of the nominal type I error. An example of the data for the Gaussian distribution for the $v_{i,j,k}$ is given in the Appendix (Figure 5).

3.1.2 Comparison of the methods

For each type of process, each type of Δ, and each value of α, we simulated 1000 artificial datasets. The p-value associated with each MLC was estimated by generating 999 random permutations of the data. A type I error of 5% was considered for the rejection of $\mathcal{H}_0$. The three methods' respective performances were compared with regard to four criteria: the power, the true positive rate, the false positive rate, and the F-measure.

The power was defined as the proportion of simulations leading to the rejection of $\mathcal{H}_0$, according to the type I error. For the simulated datasets yielding to the rejection of $\mathcal{H}_0$, the true positive rate is the mean proportion of sites correctly detected among the sites in w, the false positive rate is the mean proportion of sites in w^c that were included in the detected cluster, and the F-measure corresponds to the average harmonic mean of the proportion of sites in w within the detected cluster (positive predictive value) and the true positive rate.

3.1.3 The results of the simulation study

The results of the simulation are shown in Figure 1. When $\alpha = 0$ the different methods seem to maintain the correct type I error (0.05), regardless of the type of process and the Δ (see power curves in Figure 1).

The DFFSS almost always shows higher powers than the two other methods for all the scenarios, especially in the case of the local shift Δ_3. While the NPFSS and the PFSS presents similar powers when the $v_{i,j,k}$ are gaussian for Δ_1 and Δ_2 but when the data were not normal or the shift is local (Δ_3) the NPFSS shows better powers than the PFSS.

Generally, as expected since it is parametric, the PFSS performed less well when the data were not distributed normally. Even if it is presented as "distribution-free", the DFFSS is based on a Student test statistic which works better for normal data: we can see that its performances also slightly decrease when the $v_{i,j,k}$ follow a Student distribution. The DFFSS and the NPFSS show higher true positive rates than the PFSS, for the shift Δ_2 the NPFSS also presents slightly higher true positive rates than the DFFSS. However the true positive rates of the DFFSS are much higher than the NPFSS for the local shift Δ_3. In terms of false positives, the DFFSS always presents the
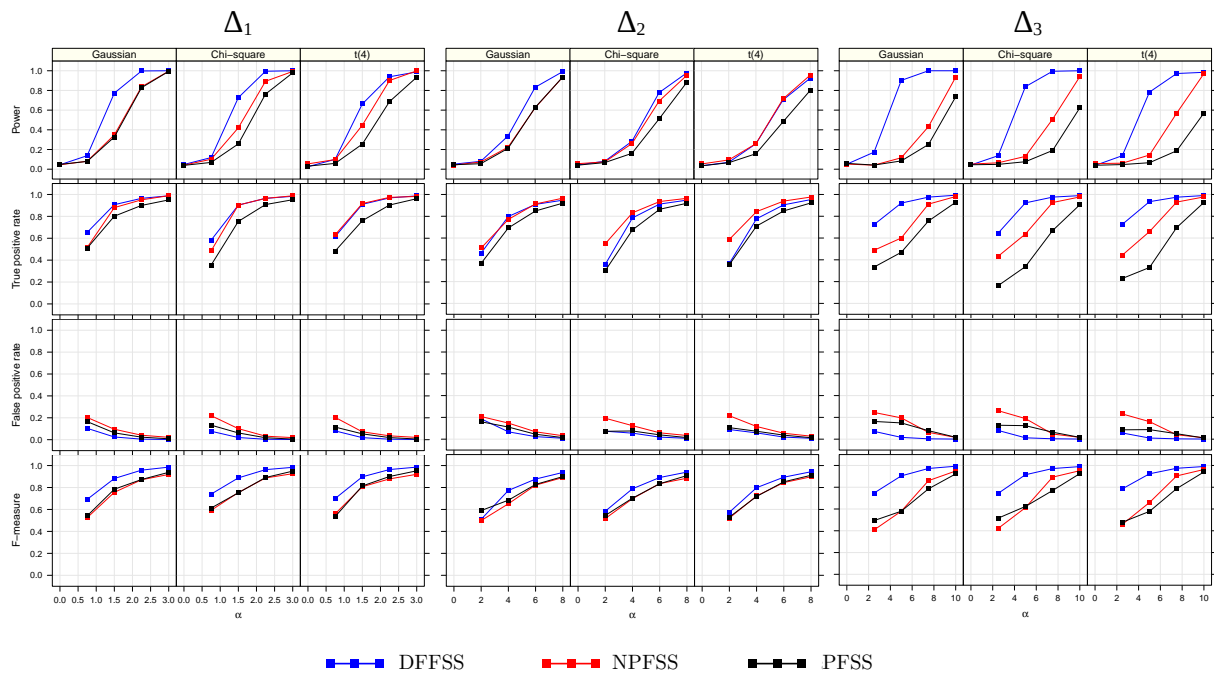
Figure 1: The simulation study: comparison of the NPFSS, DFFSS and PFSS methods for three types of shift Δ: $\Delta_1(t) = \alpha t$ (left panel), $\Delta_2(t) = \alpha t(1 - t)$ (middle panel) and $\Delta_3(t) = \alpha \exp\left[-100(t - 0.5)^2\right]/3$ (right panel). For each method, the power curves, the true-positive and false-positive rates, and the F-measure values for detection of the true cluster as the MLC are shown. α is the parameter that controls the cluster intensity.

lowest false positive rates. The NPFSS presents the highest false positive rates, yielding to low F-measures. In contrast the DFFSS presents the highest F-measures. The PFSS also shows quite low F-measures which is due to its low true positive rates. However its F-measures are often higher than for the NPFSS.

4 Application to real data

4.1 Unemployment rates in France

We considered data on unemployment rates in France; these were provided by the *Institut National de la Statistique et des Etudes Economiques* (Paris, France) for each quarter between 1998 and 2013 (64 values) and for each of the 94 French *départements* (Figure 2, left panel). It can be seen that the unemployment rates varied markedly over time. Furthermore, the spatial distribution of the mean unemployment rate between 1998 and 2013 (Figure 2, right panel) was heterogeneous. High

unemployment rates tended to aggregate in northern France and south-eastern France. These two observations suggested that functional spatial scan statistics could be used to detect *départements*-level spatial clusters of unemployment.
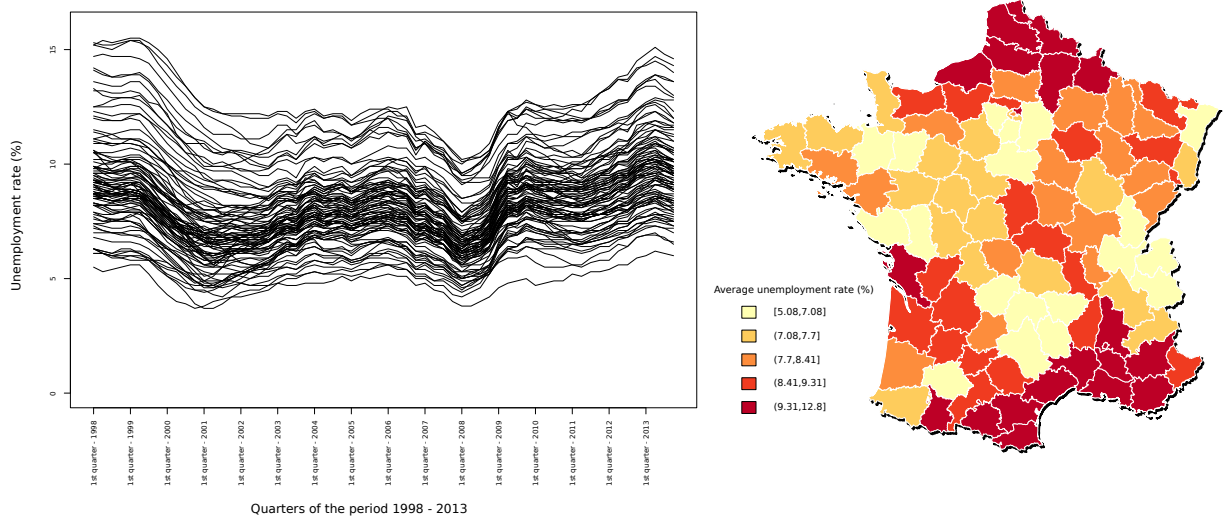

Figure 2: Unemployment rate curves from 1998 to 2013 in each of the 94 French *départements* (left panel), and the spatial distribution of the mean unemployment rates over the period (right panel).

4.2   Spatial clusters detection

In order to detect spatial clusters of low or elevated unemployment rates, we used the PFSS, the DFFSS and the NPFSS to detect the MLC and secondary clusters that had a high value of the concentration index ($F_n^{(w)}$ for PFSS, $I^{(w)}$ for DFFSS, and $U(w)$ for NPFSS; see Smida et al. (2020) for details) and that did not cover the MLC (Kulldorff, 1997). For both methods, we considered a circular variable scanning window whose maximum size was set so that a cluster never contained more than half the total number of *départements*. The statistical significance of the MLC and the secondary clusters was evaluated in 999 Monte-Carlo permutations. A spatial cluster was considered to be statistically significant when the associated p-value was below 0.05.

4.3   Results

The DFFSS, the PFSS and the NPFSS methods each detected two significant spatial clusters (Figure 3 and Table 1). The PFSS and the DFFSS identified the same two significant spatial clusters of elevated unemployment rates: the MLC (7 *départements*, $\hat{p}_{\text{PFSS}} = 0.011$, $\hat{p}_{\text{DFFSS}} = 0.005$) was located in south-eastern France and the secondary cluster (9 *départements*, $\hat{p}_{\text{PFSS}} = 0.039$,

9

$\hat{p}_{\text{DFFSS}} = 0.026$) was located in northern France. These clusters were homogeneous because they contained only curves that were above the national average unemployment rate (except for two *départements* in the secondary cluster at the beginning of the period studied).

The NPFSS identified a significant large (40-*départements*) MLC ($\hat{p}_{\text{NPFSS}} = 0.013$) with low unemployment rates in the centre of France. The unemployment rate curves were heterogeneous, and the MLC included some *départements* with an unemployment rate above the national average (Figure 3). The secondary cluster (9 *départements*, $\hat{p}_{\text{NPFSS}} = 0.039$) detected by the NPFSS was exactly the same as that detected by the PFSS and the DFFSS.

Table 1: Statistically significant spatial clusters of higher or lower unemployment rates detected using the NPFSS, the PFSS and the DFFSS

	Cluster	# *départements*	p-value
NPFSS	1	40	0.013
	2	9	0.039
PFSS	1	7	0.011
	2	9	0.039
DFFSS	1	7	0.005
	2	9	0.026

5   Discussion

Here, we developed a PFSS and a DFFSS for detecting clusters of functional data indexed in space. They are respectively based on the adaptation of the ANOVA F-statistic initially developed by Cuevas et al. (2004) and the "distribution-free" scan statistic of Cucala (2014) as a concentration index for spatial scan statistics. For functional spatial data, our methods appear to be more relevant than univariate spatial scan statistics because the latter data average data over a time period and thus lose a large amount of information. Furthermore, our methods appear to be more relevant than multivariate spatial scan statistics because it deals with two key problems: high dimensionality, and strong correlation between time measurements.

In a simulation study, we compared our new methods with the NPFSS developed by Smida et al. (2020). The results of the simulation study showed that all the methods maintain the nominal type I error. In case of normally distributed data, the PFSS performed better than the NPFSS: the two methods had similar powers (except for Δ_3) but the NPFSS tended to detect larger clusters. The PFSS has the advantage of detecting smaller clusters, which might be more relevant in many applications. Moreover, the F-measure value was almost always higher for the PFSS than for the NPFSS, which thus increase the level of confidence in the detected clusters. For all distributions the DFFSS appears to perform better than the NPFSS and the PFSS. Moreover it should be noted than it also presents a better power to detect spatial clusters that appear in a short period of time (Δ_3 case in the present simulation study).

10

We came to the same conclusions by studying the application to real data. The PFSS, the DFFSS and the NPFSS were used to detect clusters of unemployment rates in the 94 French *départements*. The NPFSS detected an MLC composed of 40 *départements* in the centre of France. Although this was a "low unemployment cluster", it contained many *départements* with high unemployment rate curves and so was not relevant – especially since the cluster covered approximately half of France. In comparison, the MLC for the PFSS and the DFFSS was the same and was only composed of 7 *départements* in south-eastern France, and all of these had high unemployment rates. All approaches detected the same secondary cluster of high unemployment rates in northern France.

It should be noted that the DFFSS and the PFSS applied here dealt solely with round-shaped clusters. In the application to unemployment data, an elongated cluster of *départements* of high unemployment rates was found in southern France – suggesting that other cluster shapes should be considered. In fact, the DFFSS and the PFSS can easily be adapted to other cluster shapes, such as elliptic clusters (Kulldorff et al., 2006) and graph-based clusters (Cucala et al., 2013).

It is also noteworthy that in the simulation study and the study of real data, the observation times were identical at each spatial location. Although this is an ideal situation for functional data, it does not apply to many real datasets. However, our new methods can be still applied to the case in which observation times are not identical at each spatial location – notably by using well-known methods (e.g. projection into a B-spline basis) to smooth the raw data (for details, see Ramsay & Silverman (2005b)).

The PFSS is derived from the ANOVA test for functional data introduced by Cuevas et al. (2004). More recently, Srivastava et al. (2013) developed another two-sample test statistic for functional data. However, the latter appears to be very time consuming and we did not consider it for the present spatial scan statistic. In fact, the concentration index is calculated first for each potential MLC and then again when computing the latter's statistical significance (i.e. the detection step is performed for each permutation of the data). Thus, we considered Cuevas' approach because it yields a reasonable computing time for the spatial scan statistic.

Although the DFFSS and the PFSS were designed to handle univariate functional data, the multivariate functional case should also be investigated. By way of an example, one could consider the data from sensors located in different geographical locations (e.g. sensors simultaneously measuring several air pollutants) over time. In such a case, the detection of significant spatial clusters might help experts to identify environmental black spots. The PFSS could be extended to the multivariate functional framework by adapting the multivariate ANOVA for functional data suggested by Górecki & Smaga (2017). It should be noted that the functional Wilcoxon-Mann-Whitney test developed by Chakraborty & Chaudhuri (2014) can also be adapted for use with multivariate functional data by taking account of a suitable scalar product for calculation of the sign function.

References

Berrendero, J., Justel, A., & Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, *55*, 2619–2634. doi:`10.1016/j.csda.2011.03.011`.

Bhatt, V., & Tiwari, N. (2014). A spatial scan statistic for survival data based on weibull distribution. *Statistics in medicine*, *33*, 1867–1876.

Boente, G., & Fraiman, R. (2000). Kernel-based functional principal components. *Statistics & Probability Letters*, *48*, 335–345. doi:`10.1016/S0167-7152(00)00014-6`.

Chakraborty, A., & Chaudhuri, P. (2014). A wilcoxon-mann-whitney type test for infinite dimensional data. *Biometrika*, *102*. doi:`10.1093/biomet/asu072`.

Chen, J., & Glaz, J. (2009). Scan statistics. chapter Approximations for Two-Dimensional Variable Window Scan Statistics. (pp. 109 – 128). Birkhäuser Boston. doi:`https://doi.org/10.1007/978-0-8176-4749-05`.

Chiou, J.-M., & Müller, H.-G. (2007). Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, *15*, 4849–4863. doi:`10.1016/j.csda.2006.07.042`.

Chong, S., Nelson, M., Byun, R., Harris, L., Eastwood, J., & Jalaludin, B. (2013). Geospatial analyses to identify clusters of adverse antenatal factors for targeted interventions. *International journal of health geographics*, *12*. doi:`10.1186/1476-072X-12-46`.

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

Cucala, L. (2014). A distribution-free spatial scan statistic for marked point processes. *Spatial Statistics*, *10*, 117–125. doi:`http://dx.doi.org/10.1016/j.spasta.2014.03.004`.

Cucala, L., Demattei, C., Lopes, P., & Ribeiro, A. (2013). A spatial scan statistic for case event data based on connected components. *Computational Statistics*, *28*, 357–369. doi:`https://doi.org/10.1007/s00180-012-0304-6`.

Cucala, L., Genin, M., Lanier, C., & Occelli, F. (2017). A multivariate gaussian scan statistic for spatial data. *Spatial Statistics*, *21*, 66–74. doi:`10.1016/j.spasta.2017.06.001`.

Cucala, L., Genin, M., Occelli, F., & Soula, J. (2019). A multivariate nonparametric scan statistic for spatial data. *Spatial statistics*, *29*, 1–14.

Cuevas, A., Febrero-Bande, M., & Fraiman, R. (2002). Linear functional regression: The case of fixed design and functional response. *Canadian Journal of Statistics*, *30*, 285–300. doi:`10.2307/3315952`.

Cuevas, A., Febrero-Bande, M., & Fraiman, R. (2004). An anova test for functional data. *Computational Statistics & Data Analysis*, *47*, 111–122. doi:`10.1016/j.csda.2003.10.021`.

Dai, W., & Genton, M. G. (2019). Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, *131*, 50–65. doi:`10.1016/j.csda.2018.03.017`.

Duncan, D., Rienti, M., Kulldorff, M., Aldstadt, J., Castro, M., Frounfelker, R., Williams, J., Sorensen, G., Johnson, R., Hemenway, D., & Williams, D. (2016). Local spatial clustering in youths' use of tobacco, alcohol, and marijuana in boston. *The American journal of drug and alcohol abuse*, *42*, 412–421. doi:`10.3109/00952990.2016.1151522`.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, *28*, 181–187. doi:`10.1214/aoms/1177707045`.

Ferraty, F., & Vieu, P. (2002). Functional nonparametric model and application to sprectrometric data. *Computational Statistics*, *17*, 545–564. doi:`10.1007/s001800200126`.

Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer.

Gao, J., Zhang, Z., Hu, Y., Bian, J., Jiang, W., Wang, X., Sun, L., & Jiang, Q. (2014). 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*, 5431–5444. doi:`10.3390/ijerph110505431`.

Genin, M., Fumery, M., Occelli, F., Savoye, G., Pariente, B., Dauchet, L., Giovannelli, J., Vignal, C., Body-Malapel, M., Sarter, H. et al. (2020). Fine-scale geographical distribution and ecological risk factors for crohn's disease in france (2007-2014). *Alimentary Pharmacology & Therapeutics*, *51*, 139–148.

Górecki, T., & Smaga, Ł. (2015). A comparison of tests for the one-way anova problem for functional data. *Computational Statistics*, *30*, 987–1010. doi:`10.1007/s00180-015-0555-0`.

Górecki, T., & Smaga, Ł. (2017). Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, *44*, 2172–2189. doi:`10.1080/02664763.2016.1247791`.

Huang, L., Kulldorff, M., & Gregorio, D. (2007). A spatial scan statistic for survival data. *Biometrics*, *63*, 109–118.

Jung, I., Kulldorff, M., & Klassen, A. C. (2007). A spatial scan statistic for ordinal data. *Statistics in medicine*, *26*, 1594–1607.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics - Theory and Methods*, *26*, 1481–1496. doi:`10.1080/03610929708831995`.

Kulldorff, M. (1999). Spatial scan statistics: Models, calculations, and applications. (pp. 303–322). doi:`10.1007/978-1-4612-1578-314`.

Kulldorff, M., Huang, L., & Konty, K. (2009). A scan statistic for continuous data based on the normal probability model. *Int J Health Geogr*, *8*. doi:`https://doi.org/10.1186/1476-072X-8-58`.

Kulldorff, M., Huang, L., Pickle, L., & Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in medicine*, *25*, 3929–3943. doi:`10.1002/sim.2490`.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W., Kleinman, K., & Platt, R. (2007). Multivariate scan statistics for disease surveillance. *Statistics in medicine*, *26*, 1824–1833. doi:`10.1002/sim.2818`.

Kulldorff, M., & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, *14*, 799–810. doi:`10.1002/sim.4780140809`.

Lin, Z., Lopes, M. E., & Müller, H.-G. (2021). High-dimensional manova via bootstrapping and its application to functional and sparse count data, .

Luquero, F., Banga, C., Remartínez, D., Urrutia, P. P. P., Baron, E., & Grais, R. (2011). Cholera epidemic in guinea-bissau (2008) : The importance of "place". *PloS one*, *6*. doi:`https://doi.org/10.1371/journal.pone.0019005`.

Naus, J. I. (1965a). Clustering of random points in two dimensions. *Biometrika*, *52*, 263–267. doi:`10.2307/2333829`.

Naus, J. I. (1965b). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, *60*, 532–538. doi:`10.2307/2282688`.

Qiu, Z., Chen, J., & Zhang, J.-T. (2021). Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis*, *157*. doi:`https://doi.org/10.1016/j.csda.2020.107160`.

Ramsay, J., & Silverman, B. (2005a). *Functional Data Analysis*. Springer Series in Statistics. Springer.

Ramsay, J., & Silverman, B. (2005b). Principal components analysis for functional data. *Functional data analysis*, (pp. 147–172).

Smida, Z., Cucala, L., & Gannoun, A. (2020). A nonparametric spatial scan statistic for functional data. *hal-02908496*, .

Srivastava, M. S., Katayama, S., & Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, *114*, 349–358. doi:`https://doi.org/10.1016/j.jmva.2012.08.014`.

A Supplementary materials

Figure 5 shows an example of the data generated when the $v_{i,j,k}$ are Gaussian.

B   Reduction of the computation time for the NPFSS

Smida et al. (2020) defined a nonparametric scan statistic for independent observations of a process in a separable Banach space χ, $\mathcal{S}_{\mathcal{W}}$ by

$$\mathcal{S}_{\mathcal{W}} = \max_{w \in \mathcal{W}} \left\| \frac{1}{\sqrt{|w| \, |w^{\mathsf{c}}| \, (|w| + |w^{\mathsf{c}}|)}} \sum_{i,s_i \in w} \sum_{j,s_j \in w^{\mathsf{c}}} \mathrm{sign}(X_j - X_i) \right\|$$

where the sign function is the Gâteaux-derivative of the norm on χ.

Calculation of the sign function in the case of square-integrable functions

Let $X \in L^2(\mathcal{T}, \mathbb{R}) n \{0\}$. In this section, we shall compute the $\mathrm{sign}(X)$ which is the Gâteaux-derivative of $||X||_2 = \sqrt{\langle X, X \rangle}$ where $\langle X, Y \rangle = \int_{\mathcal{T}} X(t) Y(t) \, \mathrm{d}t$.

Let $h \in L^2(\mathcal{T}, \mathbb{R})$,

$$
\begin{aligned}
\lim_{v \to 0} \frac{||X + hv||_2 - ||X||_2}{v} &= \lim_{v \to 0} \frac{||X + hv||_2^2 - ||X||_2^2}{v(||X + hv||_2 + ||X||_2)} \\
&= \lim_{v \to 0} \frac{\langle X + hv, X + hv \rangle - \langle X, X \rangle}{v(||X + hv||_2 + ||X||_2)} \\
&= \lim_{v \to 0} \frac{2v\langle h, X \rangle + v^2 ||h||_2^2}{v(||X + hv||_2 + ||X||_2)} \\
&= \lim_{v \to 0} \frac{2\langle h, X \rangle + v ||h||_2^2}{||X + hv||_2 + ||X||_2} \\
&= \frac{\langle h, X \rangle}{||X||_2}.
\end{aligned}
$$

Then $\mathrm{sign}(X)(h) = \dfrac{\langle h, X \rangle}{||X||_2}$ if $X \in L^2(\mathcal{T}, \mathbb{R}) n \{0\}$.

Reduction of the computation time

Let X be a stochastic process taking values in $L^2(\mathcal{T}, \mathbb{R})$, and $X_1, \ldots, X_n$ be the realisations of X in the locations $s_1, \ldots, s_n$.

Since the X_i are in $L^2(\mathcal{T}, \mathbb{R})$, $\mathrm{sign}(X_i) = \begin{cases} \dfrac{X_i}{||X_i||_2} & \text{if } X_i \neq 0 \\ 0 & \text{if } X_i = 0 \end{cases}$.

Then:

$$\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w^c}}^{n}\text{sign}(X_j-X_i)=\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{j=1}^{n}\text{sign}(X_j-X_i)-\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w}}^{n}\text{sign}(X_j-X_i)$$

$$=\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{j=1}^{n}\text{sign}(X_j-X_i)-\sum_{\substack{i=2\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w}}^{i-1}\text{sign}(X_j-X_i)-\sum_{\substack{i=1\\s_i\in w}}^{n-1}\sum_{\substack{j=i+1\\s_j\in w}}^{n}\text{sign}(X_j-X_i)$$

$$=\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{j=1}^{n}\text{sign}(X_j-X_i)-\sum_{\substack{i=2\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w}}^{i-1}\text{sign}(X_j-X_i)+\sum_{\substack{j=1\\s_j\in w}}^{n-1}\sum_{\substack{i=j+1\\s_i\in w}}^{n}\text{sign}(X_j-X_i)$$

$$=\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{j=1}^{n}\text{sign}(X_j-X_i)-\sum_{\substack{i=2\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w}}^{i-1}\text{sign}(X_j-X_i)+\sum_{\substack{i=2\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w}}^{i-1}\text{sign}(X_j-X_i)$$

$$=\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{j=1}^{n}\text{sign}(X_j-X_i)$$

Thus the matrix of signs (in which each line i corresponds to $\sum_{j=1}^{n}\text{sign}(X_j-X_i)$) can be computed. Lastly, to obtain $\sum_{\substack{i=1\\s_i\in w}}^{n}\sum_{\substack{j=1\\s_j\in w^c}}^{n}\text{sign}(X_j-X_i)$, one sums the rows of the matrix that correspond to the sites in w.

In order to evaluate the order of magnitude of the reduction in computing time, let consider the following example: 1000 datasets of the previous simulation were simulated. The improved method also took into account the computation time of the matrix of signs. The computation times for the two methods are shown in Table 2. In this example, our method divided the computing time by a factor of approximately 100.

Table 2: Comparison of the computation times (in seconds) in a simulation before and after the improvement with the matrix of signs. The results are for 1000 simulated datasets observed at 101 equally spaced times of $[0;1]$.

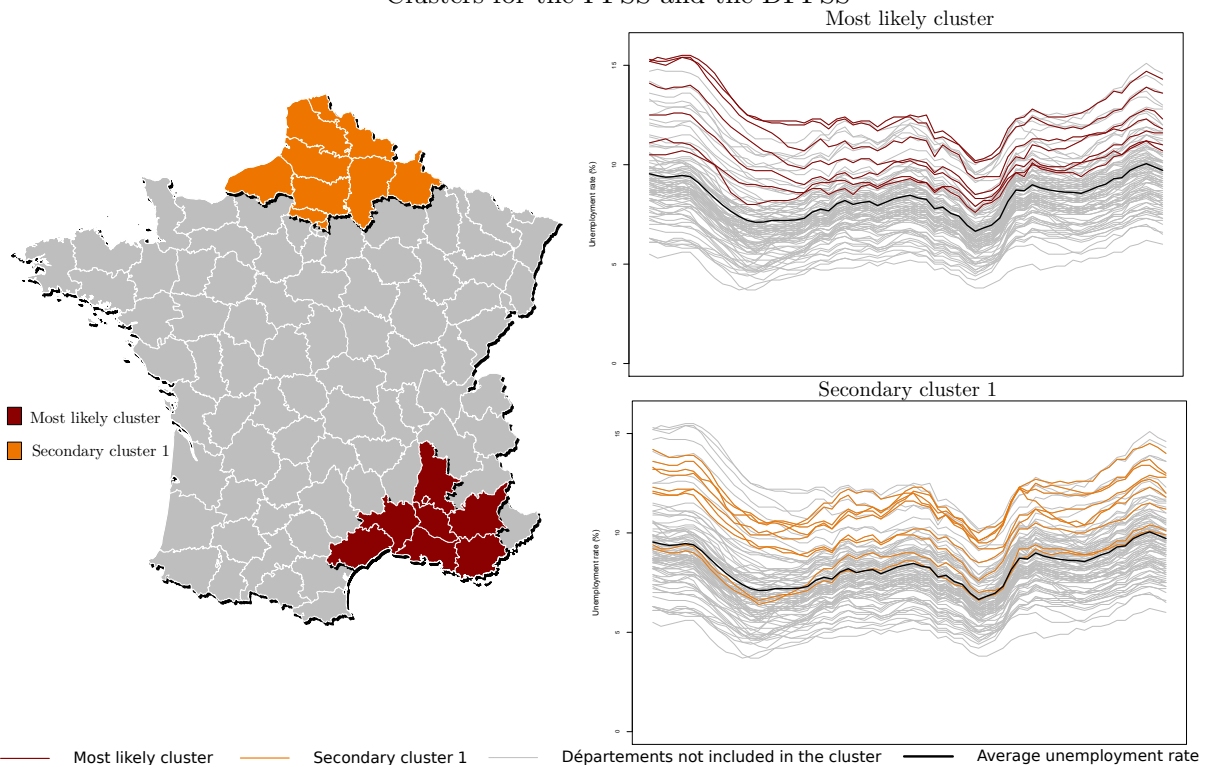| Method	Range	Mean (sd)
Basic method	[78 ; 120]	101 (8)
Improved method	[0.25 ; 0.87]	0.34 (0.06)
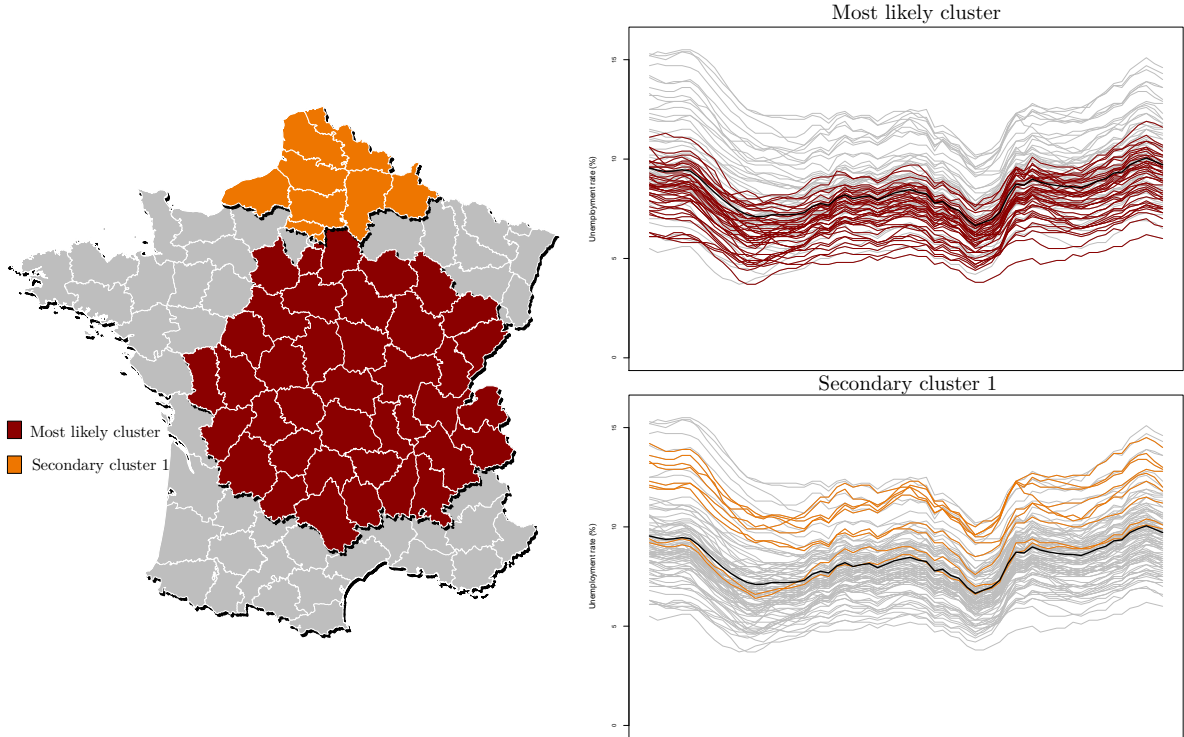
Figure 3: Significant spatial clusters of high or low unemployment rates detected by the NPFSS (top panel) and the PFSS and the DFFSS (bottom panel). For each method, the MLC is shown in red and the secondary cluster is shown in orange. The unemployment rate curves (from 1998 to 2013) for the *département* in the MLC and in the secondary cluster are presented as correspondingly coloured lines in the graphs. The black curve is the national average unemployment rate.
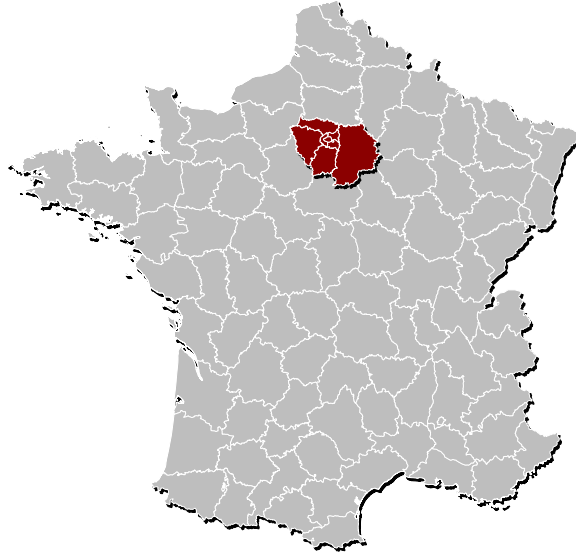
Figure 4: The 94 French *départements* and the true cluster (in red) simulated for each artificial dataset.
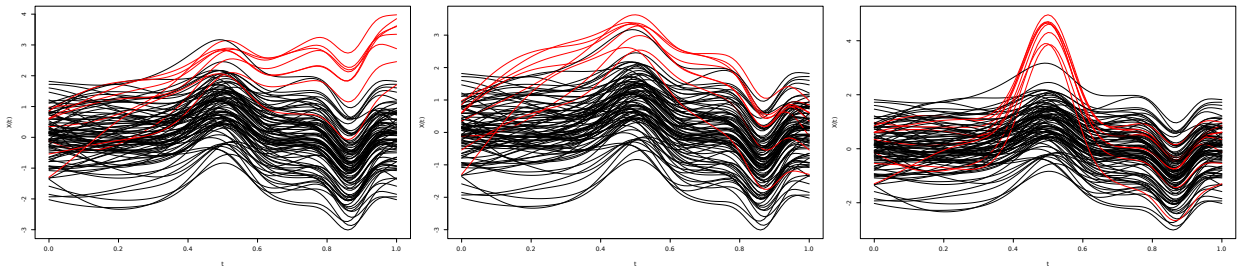

Figure 5: The simulation study: an example of the data generated for the Gaussian process, with $\Delta(t) = \Delta_1(t) = 3t$ (left panel), $\Delta(t) = \Delta_2(t) = 8t(1 - t)$ (middle panel) and $\Delta(t) = \Delta_3(t) = 10 \exp\left[-100(t - 0.5)^2\right]/3$ (right panel). The red curves correspond to the observations in the cluster.