Université de Lille
Ecole Doctorale Biologie Santé de Lille

THÈSE DE DOCTORAT
Discipline : Mathématiques Appliquées

présentée par
Camille FRÉVENT

Contribution to spatial statistics for high-dimensional and survival data

---

Contribution à la statistique spatiale pour données en grande dimension et données de survie

---

dirigée par Michaël Genin et Sophie Dabo-Niang

Soutenue le 2 décembre 2022 devant le jury composé de :

Mᵐᵉ Liliane Bel	AgroParisTech	Présidente et examinatrice
Mʳ Lionel Cucala	Université de Montpellier	Membre invité
Mᵐᵉ Sophie Dabo-Niang	Université de Lille	Directrice de thèse
Mᵐᵉ Edith Gabriel	INRAE	Rapportrice
Mʳ Michaël Genin	Université de Lille	Directeur de thèse
Mʳ Mohamed Lemdani	Université de Lille	Examinateur
Mʳ Jorge Mateu	Universitat Jaume I (Espagne)	Rapporteur
Mʳ Cristian Preda	Université de Lille	Examinateur
Mᵐᵉ Christine Thomas-Agnan	Toulouse School of Economics	Examinatrice

Résumé français

Dans ce mémoire de thèse nous nous intéressons aux méthodes d'apprentissage statistique pour données spatiales en grande dimension et données de survie. L'objectif est de développer des méthodes de détection de clusters non supervisées avec des statistiques de scan spatiales, à la fois dans le cadre de l'analyse de données fonctionnelles, mais aussi pour l'analyse de données de survie.

Nous considérons tout d'abord des données fonctionnelles univariées ou multivariées mesurées spatialement dans une région géographique. Nous proposons des statistiques de scan paramétriques et non paramétriques dans ce contexte. Ces approches fonctionnelles univariées et multivariées évitent la perte d'information respectivement d'une méthode univariée ou multivariée appliquée sur des observations moyennes au cours de la période d'étude. Nous étudions également les performances de ces approches sur des études de simulation, avant de les appliquer sur des données réelles économiques et environnementales.

Nous nous intéressons également à la détection de clusters spatiaux de temps de survie. Bien qu'il existe déjà dans la littérature des approches de statistiques de scan spatiale dans ce cadre, celles-ci ne permettent pas de prendre en compte une éventuelle corrélation entre les temps de survie des individus d'une même unité spatiale. De plus, la nature spatiale des données implique une potentielle dépendance spatiale entre les unités spatiales, qui doit être prise en compte. L'originalité de l'approche que nous proposons est le développement d'une nouvelle statistique de scan spatiale basée sur un modèle de Cox à fragilité spatiale, permettant à la fois la prise en compte de la corrélation entre les temps de survie des individus d'une même unité spatiale, et une éventuelle dépendance spatiale entre les unités spatiales. Nous avons comparé les performances de cette nouvelle approche avec les méthodes existantes et nous les avons appliquées sur des données réelles de temps de survie des personnes âgées atteintes d'insuffisance rénale chronique terminale dans le nord de la France.

Enfin, nous proposons un certain nombre de perspectives à notre travail, à la fois avec des prolongements directs à cette thèse dans le cadre des statistiques de scan spatiales pour données en grande dimension et données de survie, mais également avec des perspectives dans un cadre plus large d'analyse spatiale non supervisée (clustering spatial pour données en grande dimension modélisées par des tenseurs), et d'apprentissage spatial supervisé (régression).

Mots-clés : Analyse non supervisée, Analyse supervisée, Détection de clusters, Données de survie, Données en grande dimension, Statistique spatiale

English abstract

In this thesis, we are interested in statistical spatial learning for high-dimensional and survival data. The objective is to develop unsupervised cluster detection methods by means of spatial scan statistics in the contexts of functional data analysis in one hand and survival data analysis in the other hand.

In the first two chapters, we consider univariate and multivariate functional data measured spatially in a geographical area. We propose both parametric and nonparametric spatial scan statistics in this framework. These univariate and multivariate functional approaches avoid the loss of information respectively of a univariate method or a multivariate method applied on the average of the observations during the study period. We study the new methods' performances in simulation studies before applying them on economic and environmental real data.

We are also interested in spatial cluster detection of survival data. Although there exist already spatial scan statistics approaches in this framework in the literature, these do not take into account a potential correlation of survival times between individuals of the same spatial unit. Moreover, the spatial nature of the data implies a potential spatial dependence between the spatial units, which should be taken into account. The originality of our proposed method is to introduce a spatial scan statistic based on a Cox model with a spatial frailty, allowing to take into account both the potential correlation between the survival times of the individuals of the same spatial unit and the potential spatial dependence between the spatial units. We compare the performances of this new approach with the existing methods and apply them on real data corresponding to survival times of elderly people with end-stage kidney failure in northern France.

Finally, we propose a number of perspectives to our work, both in a direct extension of this thesis in the framework of spatial scan statistics for high-dimensional and survival data, but also perspectives in a broader context of unsupervised spatial analysis (spatial clustering for high-dimensional data (tensors)), and supervised spatial learning (regression).

Keywords: Clusters detection, High-dimensional data, Spatial statistic, Supervised analysis, Survival data, Unsupervised analysis

Contents

List of Figures

List of Tables

Remerciements

Tout d'abord je souhaiterais remercier Sophie Dabo-Niang et Michaël Genin. Merci de m'avoir fait découvrir le monde de la recherche, de m'avoir fait confiance et d'avoir accepté d'encadrer cette thèse. Merci pour votre gentillesse, votre bienveillance et votre disponibilité malgré vos emplois du temps respectifs. Merci pour vos nombreuses relectures et vos conseils qui m'ont permis de progresser et d'améliorer la qualité de mes travaux. J'espère que nous aurons l'opportunité de continuer à collaborer ces prochaines années.

Je remercie sincèrement Edith Gabriel et Jorge Mateu d'avoir accepté de rapporter ce travail de thèse. Je remercie également Liliane Bel, Lionel Cucala, Mohamed Lemdani, Cristian Preda, et Christine Thomas-Agnan d'avoir accepté de faire partie de mon jury de thèse.

Je souhaiterais également remercier l'ensemble des membres de l'équipe METRICS et en particulier du CERIM pour leur accueil et leur bonne humeur. Merci à Renaud et Julien pour leur aide précieuse en informatique et leur disponibilité. Merci également à Mélanie pour toutes les démarches administratives lors de mes déplacements. Je remercie mes collègues de bureau : Anaïs, Laurine, Mohamed, Paul pour nos échanges et leur bonne humeur. Mohamed, travailler avec toi a été très enrichissant. J'espère que nous pourrons continuer à travailler ensemble dans le futur.
Je remercie également le professeur Jean-Baptiste Beuscart pour m'avoir permis de réaliser cette thèse.

Merci également aux membres de l'INRIA-MODAL pour leur accueil chaleureux, et en particulier Axel, Eglantine, Etienne, Ernesto, Filippo, Issam, et Rim pour nos discussions le midi.

Je remercie mes proches, qui me soutiennent depuis toujours. Merci de m'avoir encouragée, tout cela n'aurait pas été possible sans vous.

Chapter 1

General introduction

Contents

1   Presentation of the thesis

Spatial analysis started with cartography in Antiquity (Bailly et al., 1995). However, although in many application fields data naturally present a spatial component, spatial statistics was only popularized later, at the end of the 50's with the work of Danie Krige on spatial prediction and its mathematical formalization by Georges Matheron (Matheron, 1962, 1963). This has since led to the development of regression (Anselin, 2009; Ward and Gleditsch, 2018), interpolation (Cressie, 1988; Belkhiri et al., 2020), or clustering (Bourgault et al., 1992; D'Urso and Vitale, 2020) methods for spatial data, and to their applications in many domains such as biology (Khanal et al., 2010), environment (Scoggins et al., 2004; Kuechly et al., 2012) or econometrics (Anselin, 1988) to name a few.

In the field of public health, spatial analysis goes back further. Actually, it started at the end of the 18th century, in particular with the publication of the book of Lind (1768) and then the study of the doctor John Snow, who is considered as the initiator of spatial epidemiology. During the cholera epidemic that ravaged London in 1854 (500 deaths between August 31 and September 10 (Waller and Gotway, 2004)), the dominant theory was that the epidemic was spread through the air by inhalation. Skeptical of this idea, the doctor John Snow thought that the contamination was done by ingestion of a contaminated element, in this case water. He then mapped cholera cases in the Soho district of London and found a cluster of cases around a public water pump. This discovery led to the closing of the water pump and the decline of the epidemic. In 1883, the microbiologist Robert Koch discovered that cholera was caused by the bacterium *Vibrio cholerae*. The latter living in water, the hypothesis emitted by John Snow is confirmed.

Actually in public health, three types of spatial analysis can be distinguished depending on the objective of the study: disease mapping which aims at studying the spatial distribution of health events through cartography, cluster detection whose purpose is to determine if there

exist geographical areas with an over-incidence of health events, and the characterization of the relationship between the spatial distribution of health events and risk factors (e.g. through ecological regressions). These ecological studies make it possible to explore hypotheses at minimal cost before conducting more confirmatory studies, such as epidemiological approaches at the individual level.

This thesis focuses on spatial cluster detection methods. Spatial clusters are defined as aggregations of spatial units presenting an unusual concentration of a measure (disease incidence, pollutant concentration, etc.) compared to the rest of the studied area. In the field of public health, these methods make it possible to (i) objectify the detection of spatial clusters, (ii) evaluate their statistical significance, and (iii) provide robust information to decision-makers in order to adapt public health policies territorially. In particular spatial scan statistics are well-known cluster detection methods. Initially popularized by Kulldorff and Nagarwalla (1995) and Kulldorff (1997) for Bernoulli and Poisson models, they have been extended to other data distributions such as zero-inflated (Cançado et al., 2014; de Lima et al., 2015), ordinal (Jung et al., 2007), Gaussian (Kulldorff et al., 2009), and multivariate Gaussian (Cucala et al., 2017) distributions. A few authors have also proposed nonparametric spatial scan statistics for real univariate (Jung and Cho, 2015; Cucala, 2016) and multivariate (Cucala et al., 2019) data.

In the last few decades, technological advances have led to the emergence of functional data (Ramsay and Ramsey, 2002). This led to the adaptation of spatial analysis methods to the spatial functional framework. In the context of spatial scan statistics, summarizing the information of univariate functional data by the mean of the curves and then applying a spatial scan statistic for univariate real data leads to a huge loss of information. Moreover, considering each measurement time as a variable and applying a spatial scan statistic for multivariate data would be faced with high correlations and high dimensionality issues. Thus, Smida et al. (2022) proposed a first nonparametric spatial scan statistic for univariate functional data. However to our knowledge there is no parametric spatial scan statistic for univariate functional data.

In the context of multivariate functional data, a possible approach would be to summarize the curves by their mean and then apply a spatial scan statistic for multivariate data. However, this approach would result in a huge loss of information. Although the approach of Smida et al. (2022) seems to be adaptable to the framework of multivariate functional data, this has never been considered in this context and there exists no spatial scan statistic for multivariate functional data.

The spatial scan statistics described above generally consider data aggregated at the scale of spatial units (e.g. an incidence rate is measured within each spatial unit). However, in the context of spatial survival data, the observations are individual although for reasons of anonymity, their location is only available at an aggregated level such as an administrative one. Thus, within the same spatial unit, several individuals are observed. In this framework, few spatial scan statistics have been proposed. Huang et al. (2007) and Bhatt and Tiwari (2014) respectively proposed parametric spatial scan statistics based on an exponential and a Weibull model. Although these methods are widely used in practice to detect spatial clusters of time-to-event data (Gregorio et al., 2007; Henry et al., 2009; Wan et al., 2012), they require a distribution assumption. A first semi-parametric method using a Cox model has been proposed by Cook et al. (2007). However all these existing methods assume the independence of the observations. However, the survival times of individuals within the same spatial unit may be correlated, e.g., because of unobserved confounding factors at the level of the spatial units. Furthermore, the spatial units may be spatially dependent. In this

context, several authors have shown that not taking spatial correlation into account in spatial scan statistics leads to an increase in the type I error (Loh and Zhu, 2007; Ahmed et al., 2021b).

In this thesis we are interested in developing new spatial scan statistics methods for high-dimensional data and survival data. More precisely, we focus our developments on (i) univariate and multivariate functional data, and (ii) taking into account both the correlation between the survival times of individuals within the same spatial unit, and the spatial dependence between spatial units, in the context of spatial survival data. The rest of this manuscript is divided into six chapters, organized as follows:

Chapter 2. This chapter presents the fundamental concepts for a good understanding of this manuscript. First, it develops the notions related to functional data and survival data. Secondly, it presents the principal notions related to spatial analysis, in particular the categories of spatial data, spatial survival analysis, as well as the clustering and spatial cluster detection methods. Particular attention is given to spatial scan statistics.

Chapter 3. We first propose several definitions of a spatial cluster in the context of univariate functional data. In particular, we distinguish the notion of shape clusters and magnitude clusters. We then propose two new spatial scan statistics in this context, respectively based on an ANOVA test statistic for functional data and on a pointwise Student's t-statistic. In a simulation study we evaluate the performance of our methods and compare them with the nonparametric approach proposed by Smida et al. (2022). Finally, we apply them on unemployment rate data in France in order to detect clusters of abnormally high or low unemployment rates. This work has been realized in collaboration with Mohamed-Salem Ahmed, Matthieu Marbac and Michaël Genin. It was enhanced by an accepted publication in the journal *Spatial Statistics*.

Chapter 4. In this chapter we adapt the definitions of a spatial cluster to the framework of multivariate functional data. We then propose three spatial scan statistics for multivariate functional data. These are based on a MANOVA test statistic for multivariate functional data, a pointwise Hotelling's T^2-test statistic, and a pointwise Wilcoxon-Mann-Whitney test statistic for multivariate data. We also propose an adaptation of the approach of Smida et al. (2022) to the multivariate functional framework in order to evaluate the performance of our three methods and compare them with the one of Smida et al. (2022) in a simulation study. Then we apply these methods on air quality data in northern France. More precisely, it consists in the daily concentration of the pollutants NO_2, O_3, PM_{10} and $PM_{2.5}$. This chapter is the result of a collaborative work with Mohamed-Salem Ahmed, Sophie Dabo-Niang and Michaël Genin. It is currently under minor revision in the *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Chapter 5. This chapter proposes a new method of spatial scan statistics for survival data. This type of data presents the particularity to be composed of individual observations located (for anonymity reasons) at the level of spatial units. The Bayesian approach that we propose is based on a Cox model with spatial frailty. It presents the advantage of taking into account the possible correlation between the survival times of individuals within the same spatial unit, as well as the possible spatial dependence between spatial units. It also allows to adjust the detection of clusters easily on confounding factors. We perform several simulation studies to study both the impact of the correlation between the survival times of individuals within the same spatial unit on the approaches of the literature, and the performance of our approach in the absence and presence of censoring. Finally, we use the method on epidemiological data to detect atypical survival time clusters in elderly people with end-stage kidney failure in

northern France.

This work resulted in an article submitted in *Statistics in Medicine* and available on *arXiv*.

Chapter 6. In this chapter we present the R package **HDSpatialScan**. It is available on the Comprehensive R Archive Network (CRAN) and allows easy use by practitioners of the spatial scan statistics for functional data presented in this thesis. Note that it allows not only the detection of clusters but also the display of their location as well as the visualization of the data in the detected clusters (compared with the data outside the clusters). This package also implements other spatial scan statistics for which no implementation was yet available. The work presented in this chapter was done in collaboration with Mohamed-Salem Ahmed, Julien Soula, Zaineb Smida, Lionel Cucala, Sophie Dabo-Niang and Michaël Genin. It has resulted in an article accepted in *The R Journal*.

Chapter 7. This chapter concludes the manuscript with a conclusion and a number of perspectives. These are based both on direct extensions of spatial scan statistics for functional data and survival data, but also in a broader framework of spatial analysis. We distinguish between new developments in unsupervised spatial analysis, namely spatial clustering of high dimensional data modeled by tensors, and supervised learning methods in the context of spatial functional data (regression).

2 Written and oral communications

2.1 Articles

- A Bayesian shared-frailty spatial scan statistic model for time-to-event data (in collaboration with M.S. Ahmed, S. Dabo-Niang and M. Genin).
  Pre-print submitted in Statistics in Medicine, November 2022 and available on arXiv: https://doi.org/10.48550/arXiv.2209.00279

- The R package **HDSpatialScan** for the detection of clusters of multivariate and functional data using spatial scan statistics (in collaboration with M.S. Ahmed, J. Soula, Z. Smida, L. Cucala, S. Dabo-Niang and M. Genin). *The R Journal* **14/3** (September 2022).

- Investigating spatial scan statistics for multivariate functional data (in collaboration with M.S. Ahmed, S. Dabo-Niang and M. Genin).
  In minor revision for the Journal of the Royal Statistical Society: Series C (August 2022)

- Detecting spatial clusters on functional data: new scan statistic approaches (in collaboration with M.S. Ahmed, M. Marbac and M. Genin). *Spatial Statistics* **46** (December 2021)

2.2 Software

- R package **HDSpatialScan** (in collaboration with M.S. Ahmed, J. Soula, Z. Smida, L. Cucala, S. Dabo-Niang and M. Genin).
  Avaible on the CRAN: https://CRAN.R-project.org/package=HDSpatialScan

2.3   Oral communications

2.3.1   Invited communications in international congresses

- A Bayesian shared-frailty spatial scan statistic model for time-to-event data. 15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022), King's College, London, December 2022

- Investigating spatial scan statistics for multivariate functional data. 14th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2021), King's College, London, December 2021

2.3.2   Communications in national congresses

- Investigating spatial scan statistics for multivariate functional data. Congrès des Jeunes Chercheuses et Chercheurs en Mathématiques et leurs Applications (CJC-MA 2022), Calais, France, September 2022

- Investigating spatial scan statistics for multivariate functional data. 53th Journées de la Statistique (JdS'2022), Lyon, France, June 2022

- Detecting spatial clusters in functional data: new scan statistic approaches. Forum des Jeunes Mathématiciens-nes 2021, Besançon, France, December 2021

- Detecting spatial clusters in functional data: new scan statistic approaches. 52th Journées de la Statistique (JdS'2021), Nice, France, June 2021

2.3.3   Seminars

- *Statistiques de scan spatiales pour données fonctionnelles univariées et multivariées*: Presentation of my PhD work in the internal seminar of PhD students of the probability and statistics team of Lille, France, May 2022

- Introduction of the R package **HDSpatialScan** at ISSeP (Institut Scientifique de Service Public) in Liège, Belgium, April 2022

- Presentation of my PhD work in the internal seminar of the METRICS team, Lille, France, December 2021

2.4   Organization of congresses

- Co-organization (with Sophie Dabo-Niang) of the session "Recent advances in statistics for health" in the 15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2022), King's College, London, December 2022

Chapter 2

Fundamental concepts

Contents

1   High dimensional and survival data

1.1   Functional data

In the last decades, advances in measurement and data storage capacities have led to the emergence of data that are measured in a quasi-continuous manner over time. This new type of data, called "functional data", can be found in many application fields such as medicine (e.g., electroencephalography measures, growth and weight curves, etc. (Sørensen et al., 2013; Shangguan et al., 2020)), economy (e.g., stock prices (Ramsay and Ramsey, 2002)), environmental sciences (e.g., measurement of the concentration of a pollutant every hour for a month (Cardot et al., 2007; Bouveyron et al., 2022)), etc.

This led to the popularization of functional data analysis (FDA) by Ramsay and Silverman (2005b). Since then, a considerable amount of work has been done to adapt classical statistical methods to the functional framework. Especially, this has led to new regression (Cuevas et al., 2002; Chiou et al., 2003; Ferraty et al., 2011; Chiou et al., 2016), principal component analysis (Viviani et al., 2005; Bali and Boente, 2014) and clustering (Jacques and Preda, 2014a; Delaigle et al., 2019; Schmutz et al., 2020; Golovkine et al., 2022) methods for univariate and multivariate functional data. The interested reader can refer to Ferraty and Vieu (2006); Hsing and Eubank (2015) for reviews on functional data analysis.

Functional data analysis considers a p-dimensional stochastic process

$$\{X(t) = (X^{(1)}(t), \dots, X^{(p)}(t))^\top, \ t \in \mathcal{T}\}$$

with $\mathcal{T}$ an interval of $\mathbb{R}$. Note that the special case $p = 1$ corresponds to the univariate functional case whereas $p \geq 2$ refers to multivariate functional data.

If $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$ the space of p-dimensional vector-valued square-integrable functions on $\mathcal{T}$, equipped with the inner product $\langle X, Y \rangle = \int_{\mathcal{T}} X(t)^\top Y(t) \, \mathrm{d}t$, then the mean function of X is defined as

$$\mu_X(t) = \mathbb{E}[X(t)] \in \mathbb{R}^p$$

and the covariance function of X is defined as

$$\mathrm{Cov}_X(s, t) = \mathbb{E}[(X(t) - \mu_X(t))(X(s) - \mu_X(s))^\top] \in \mathbb{R}^{p \times p}.$$

Let $X_1, \ldots, X_n$ be an independent sample of observations of X. Then μ_X and Cov_X can be estimated by

$$\hat{\mu}_X(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t)$$

and

$$\widehat{\mathrm{Cov}}_X(s, t) = \frac{1}{n-1} \sum_{i=1}^{n} [X_i(t) - \mu_X(t)][X_i(s) - \mu_X(s)]^\top$$

the empirical mean and covariance functions respectively.

In practice, the X_i $(1 \leq i \leq n)$ are measured (possibly with an error) only at discrete times of $\mathcal{T}$. Let $\{X_{i,1}^{(j)}, \ldots, X_{i,m_{i,j}}^{(j)}\}$ denotes the observed measurements of the jth variable of X_i at $m_{i,j}$ time points of $\mathcal{T}$. $X_i^{(j)}$ can be reconstructed from the longitudinal observations $\{X_{i,1}^{(j)}, \ldots, X_{i,m_{i,j}}^{(j)}\}$ in a parametric or a nonparametric way. This is the subject of the following two sections.

1.1.1 Parametric reconstruction

A classical approach to reconstruct the X_i is to assume that the $X_i^{(j)}$ $(1 \leq j \leq p)$ can be written in a basis of functions $\{f_k^{(j)}(t), t \in \mathcal{T}, 1 \leq k \leq K_j\}$ (where $K_j \leq \min\{m_{1,j}, \ldots, m_{n,j}\}$):

$$X_i^{(j)}(t) = \sum_{k=1}^{K_j} a_{i,k}^{(j)} f_k^{(j)}(t), \qquad j = 1, \ldots, p, \quad \text{and} \quad i = 1, \ldots, n.$$

This is equivalent to writing X_i as

$$X_i(t) = f(t) a_i$$

where

$$a_i = (a_{i,1}^{(1)}, \ldots, a_{i,K_1}^{(1)}, a_{i,1}^{(2)}, \ldots, a_{i,K_2}^{(2)} \ldots, a_{i,1}^{(p)}, \ldots, a_{i,K_p}^{(p)})^\top$$

is computed using an interpolation method (if the observations are assumed to be errorless) or, with an ordinary or penalized least squares method (if some observations may be erroneous), which smooths the data, and

$$f(t) = \begin{pmatrix} f_1^{(1)}(t) & \cdots & f_{K_1}^{(1)}(t) & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & f_1^{(2)}(t) & \cdots & f_{K_2}^{(2)}(t) & 0 & \cdots & 0 \\ & & & & \cdots & & & & \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & f_1^{(p)}(t) & \cdots & f_{K_p}^{(p)}(t) \end{pmatrix}.$$

Depending on the nature of the data, various choices for the f_k are possible.

In the case of periodic data, a Fourier basis is appropriate:

$$f_k^{(j)}(t) = \begin{cases} 1 & \text{if } k = 1 \\ \cos(\frac{k}{2} wt) & \text{if } k \text{ even} \\ \sin\left(\frac{k-1}{2} wt\right) & \text{if } k \text{ odd}, k \geq 3 \end{cases}$$

is suitable for data of period $\frac{2\pi}{w}$.

For non-periodic data, possible choices are polynomial basis or splines basis. A polynomial basis $(f_k^{(j)}(t) = (t - t_0^{(j)})^k)$ presents two major drawbacks: a poor fit on the boundaries and it does not allow a good fit if the behavior differs according to time t. A splines basis allows to adjust different behaviors with time: $\mathcal{T}$ is split in bins and piecewise polynomials are adjusted on each bin. Splines bases and especially cubic B-splines bases are very popular choices in practice for functional data analysis.

An example of smoothing with a Fourier basis and a cubic B-splines basis is presented in Figure 2.1. Figure 2.1 shows that the spline basis achieves to capture the slight drop in $t = 300$ contrary to the Fourier basis which imposes a periodicity on the data. Note that other functional bases are possible such as wavelets for example. The reader may refer to Ramsay and Silverman (2005a) for a more detailed review on smoothing functional data.

(a)
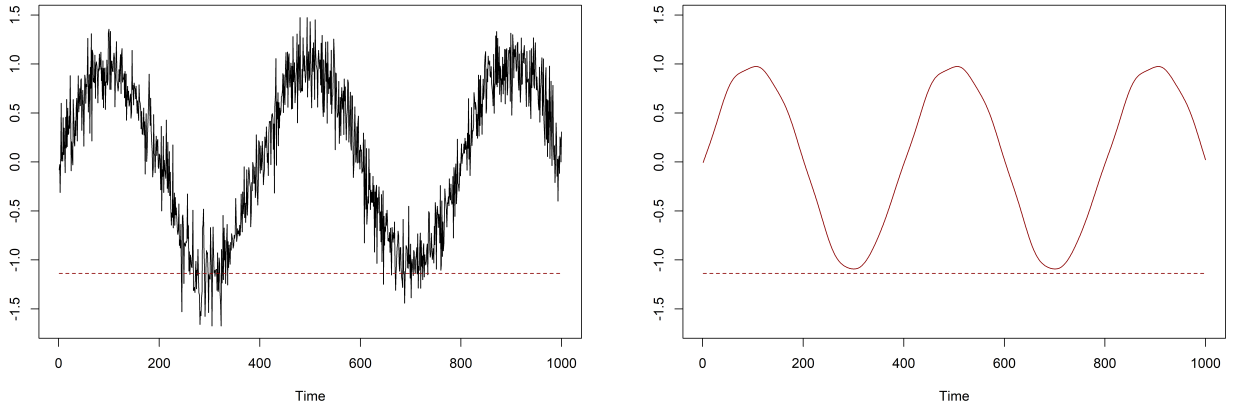
(b)
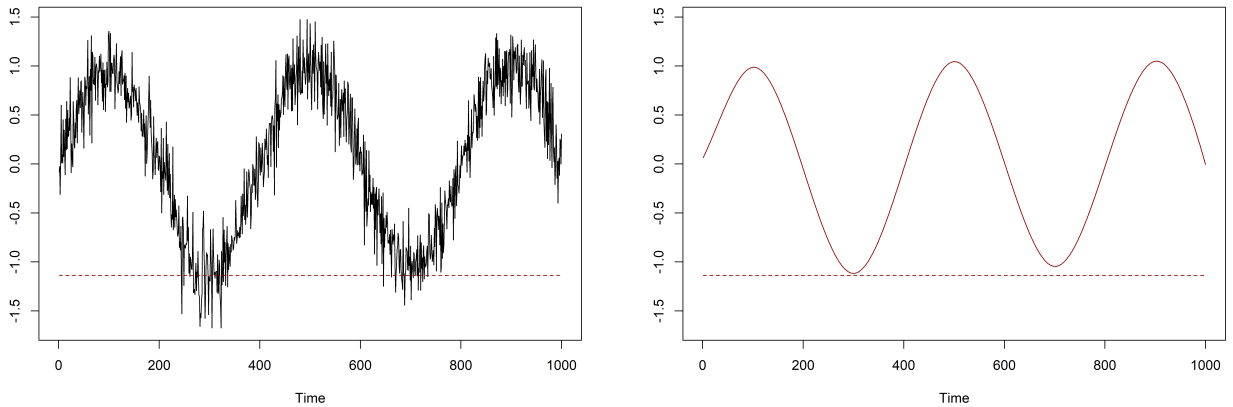
Figure 2.1: Example of smoothing with a Fourier basis (panel a) and a cubic B-splines basis (panel b)

1.1.2 Nonparametric reconstruction

Parametric approaches to reconstruct the $X_i^{(j)}$ require the specification of a basis of functions. However the $X_i^{(j)}$ can also be reconstructed in a nonparametric way, without using any basis.

Common approaches are methods using local weighting, in particular kernel methods.

The idea of local weighting around a fixed time point t_0 is to define a weight w for each observation time t such that w is greater the closer t and t_0 are. By using a kernel approach, w can be defined as

$$w(t) = \frac{1}{h} K\left(\frac{t - t_0}{h}\right)$$

where K is a kernel such that K is a non-negative real-valued integrable function and $\int_{-\infty}^{+\infty} K(u)\,\mathrm{d}u = 1$, and $h > 0$ is the bandwidth parameter.

Then to estimate the $X_i^{(j)}$, one can use the Nadaraya-Watson estimator (Nadaraya, 1964):

$$\hat{X}_i^{(j)}(t) = \frac{\dfrac{1}{h} \sum_{r=1}^{m_{i,j}} X_{i,r}^{(j)} K\left(\dfrac{t - t_r}{h}\right)}{\dfrac{1}{h} \sum_{r=1}^{m_{i,j}} K\left(\dfrac{t - t_r}{h}\right)}$$

where t_r is the rth observation time of $X_i^{(j)}$. Note that the choice of the bandwidth parameter h must be done carefully since it has a great influence on the estimates obtained (Figure 2.2). This choice can be made by cross-validation for example.

1.2   Survival data

1.2.1   Time-to-event data, truncation, censoring

The term survival time refers to the time until the occurrence of an event. This can be death, but also, for example, the occurrence of a disease, a relapse from remission or a recovery. This time is also referred to as time-to-event data. In the following we note $Y \in \mathbb{R}^+$ the random variable corresponding to the survival time. By noting f its density function, the survival function of Y is

$$S(y) = \mathbb{P}(Y > y).$$

In survival analysis the instantaneous hazard rate λ is also often used. It is defined by

$$\lambda(y) = \frac{f(y)}{S(y)} = \lim_{h \to 0} \mathbb{P}(Y \in [y, y + h[\mid Y \geq y).$$

Thus λ represents the rate of event at time y, given that the event did not occur before y.

Remark that $S(y) = \exp\left[-\Lambda(y)\right]$ where $\Lambda(y) = \int_0^y \lambda(t)\,\mathrm{d}t$ is the cumulative risk.

Although Y is the variable of interest it is not always observed in practice. Indeed, when a study is time-bound, individuals may have experienced the event of interest before the beginning of the study or after the study. In this case the data can be truncated or censored. The examples of this section are inspired from Kleinbaum et al. (2012).

1.2.1.1   Censoring

Several types of censoring can be distinguished: right censoring, left censoring or interval censoring. Let T be the observed variable and C the censoring time variable. An important assumption about censoring is that it is non informative and the censoring times are independent of the event times.

$$h = 3$$
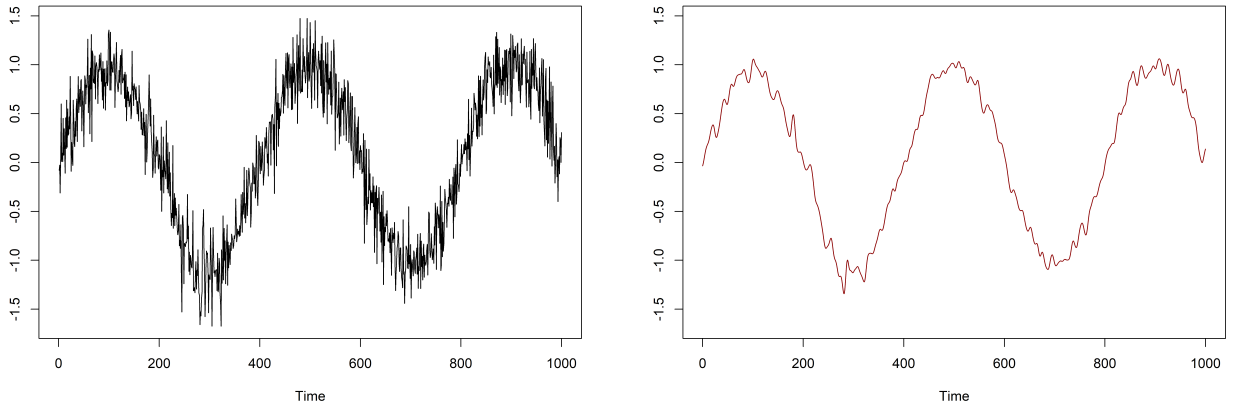
$$h = 15$$
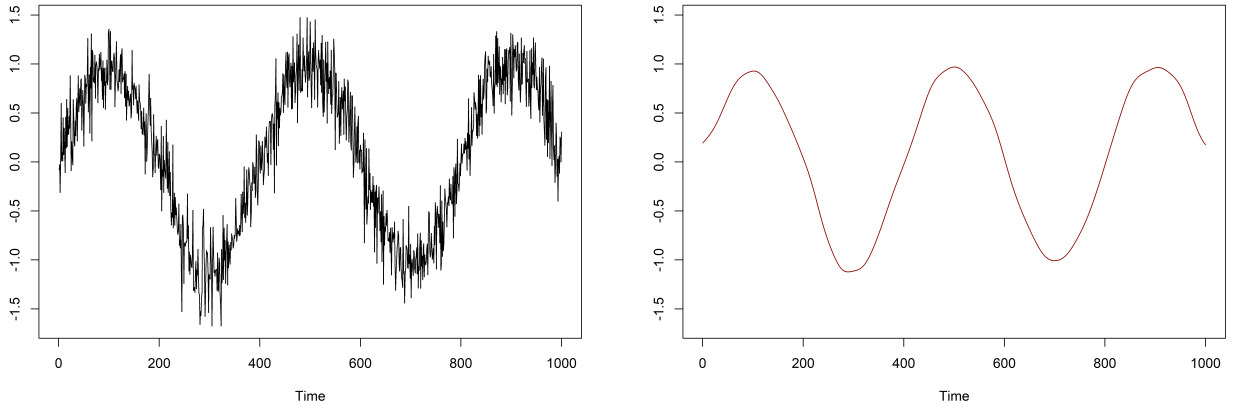
$$h = 75$$
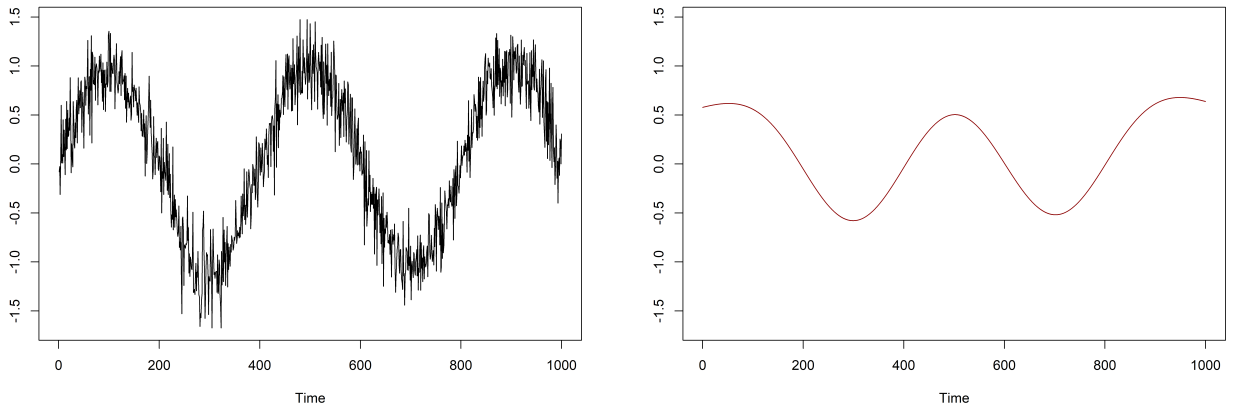
Figure 2.2: Influence of the bandwidth parameter h in nonparametric smoothing for functional data. Here the kernel used is the Gaussian kernel: $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

Right-censoring occurs when an individual has not experience the event before the end of the study. Then the only information available is that he survived at least to the end of the study. In this case C corresponds to the time until the end of the study, Y is the true survival time and $T = \min(Y, C)$. This type of censoring may also occur, for example, because the individual is lost to follow-up (see example (a) on Figure 2.3 representing the recovery time after starting treatment).

Left-censoring occurs when the true time-to-event Y is less than the observed time T: $T = \max(Y, C)$. For example Y is the time until a person becomes HIV positive. However one does not know a patient is positive until he or she tests positive. Thus you only know that the true time to HIV positivity is less than the time to test positive (example (b) on Figure 2.3).

Finally, interval-censoring occurs when you only know that $Y \in [C_1, C_2]$. For example individuals are observed until they are tested HIV positive. At time C_1 an individual is tested negative but he is tested positive at time C_2 then you only know that he became positive to HIV between C_1 and C_2 (example (c) on Figure 2.3).
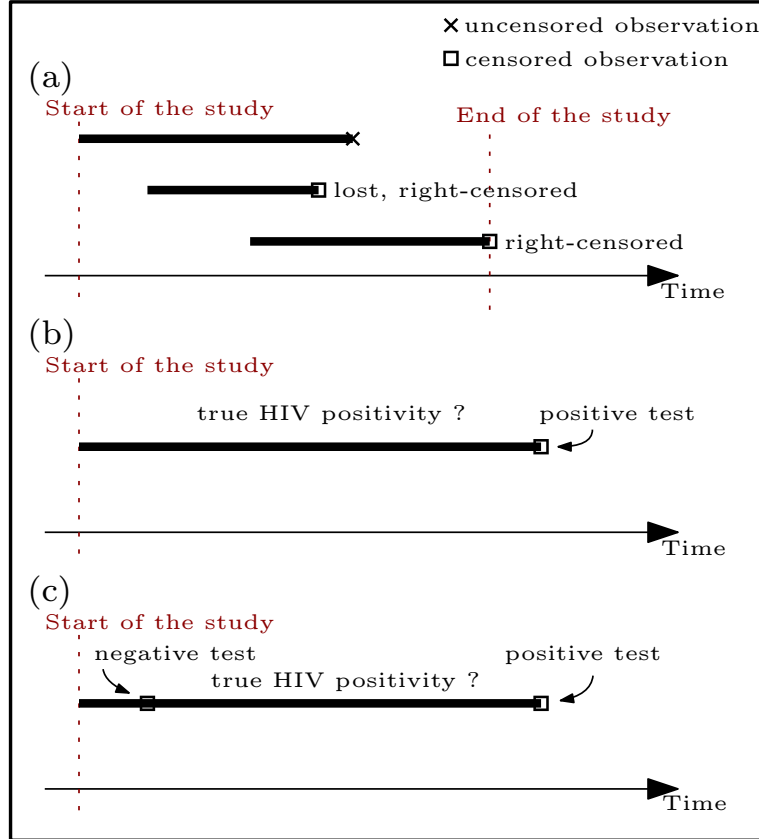
Figure 2.3: Example of right-censoring (a), left-censoring (b) and interval-censoring (c)

1.2.1.2 Truncation

Truncation occurs when only the individuals for which the time-to-event Y belongs to a time interval $[T_1, T_2]$ are observed. In particular, right-truncation occurs when only the individuals for which $Y \leq T_2$ are observed and left-truncation, on the contrary occurs when only the individuals for which $Y \geq T_1$ are observed.

In the following we only consider right censoring. Let (T, δ) be the observed time-to-event and the censoring indicator taking the value 1 if the event is observed (non censored) and 0 otherwise. In this context the contribution of the observation (T_i, δ_i) to the likelihood is

$$f(T_i)^{\delta_i} S(T_i)^{1-\delta_i}.$$

The following sections present the different approaches to model the survival function S, parametrically (Section 1.2.2), nonparametrically (Section 1.2.3) or semi-parametrically (Section 1.2.4).

1.2.2 Parametric modeling

In order to model the distribution of Y parametrically, two probability distributions are commonly used: the exponential model and the Weibull model.

1.2.2.1 Exponential model

This model assumes that $\lambda(y) = \lambda$, $\lambda > 0$. Then $S(y) = \exp\left[-\lambda y\right]\mathbb{1}_{y \geq 0}$ and $f(y) = \lambda \exp\left[-\lambda y\right]\mathbb{1}_{y \geq 0}$ (Figure 2.4). Thus the likelihood for the observed data $(T_1, \delta_1), \dots, (T_n, \delta_n)$ is

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} \exp\left[-\lambda T_i\right].$$

In this model, the instantaneous hazard rate is assumed to be constant which is quite simplistic in practice. Thus the Weibull model appears to be a good alternative to the exponential one.
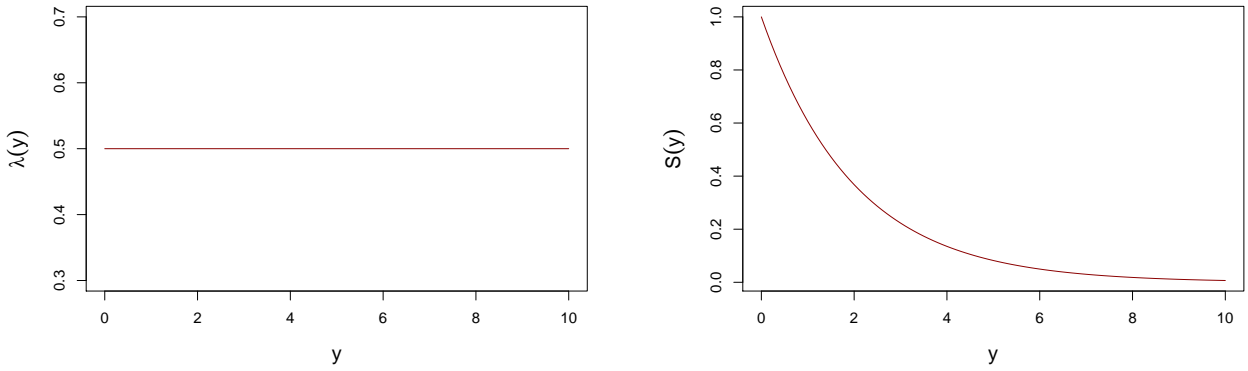
Figure 2.4: Example for the instantaneous hazard rate λ and the survival function S for the exponential model ($\lambda = 0.5$)

1.2.2.2 Weibull model

The Weibull model supposes that $\lambda(y) = k\lambda(\lambda y)^{k-1}$, $\lambda, k > 0$, then $S(y) = \exp\left[-(\lambda y)^k\right]\mathbb{1}_{y \geq 0}$ and $f(y) = k\lambda(\lambda y)^{k-1} \exp\left[-(\lambda y)^k\right]\mathbb{1}_{y \geq 0}$. Thus if $k > 1$, the instantaneous risk λ increases with time, while it decreases with time if $k < 1$ (Figure 2.5). Note that if $k = 1$ then it corresponds to the exponential model.

1.2.2.3 Adjustment on covariates

It is possible to adjust these parametric models relatively easily on covariates $X^{(1)}, \dots, X^{(p)}$. To do so, it is possible to make the parameters depend on the covariates. For example in the exponential model, $\lambda(y) = \lambda(X^{(1)}, \dots, X^{(p)}) = \exp\left[\beta_1 X^{(1)} + \dots + \beta_p X^{(p)}\right]$ (where $X^{(1)} = 1$) and the likelihood is then

$$\mathcal{L}(\{\beta_1, \dots, \beta_p\}) = \prod_{i=1}^{n} \exp\left[\beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)}\right]^{\delta_i} \exp\left[-\exp\left(\beta_1 X_i^{(1)} + \dots + \beta_p X_i^{(p)}\right)T_i\right].$$

1.2.3 Nonparametric modeling

The distribution of Y can also be modeled nonparametrically. The two most commonly used estimators in this context are the Kaplan-Meier estimator of the survival function S and
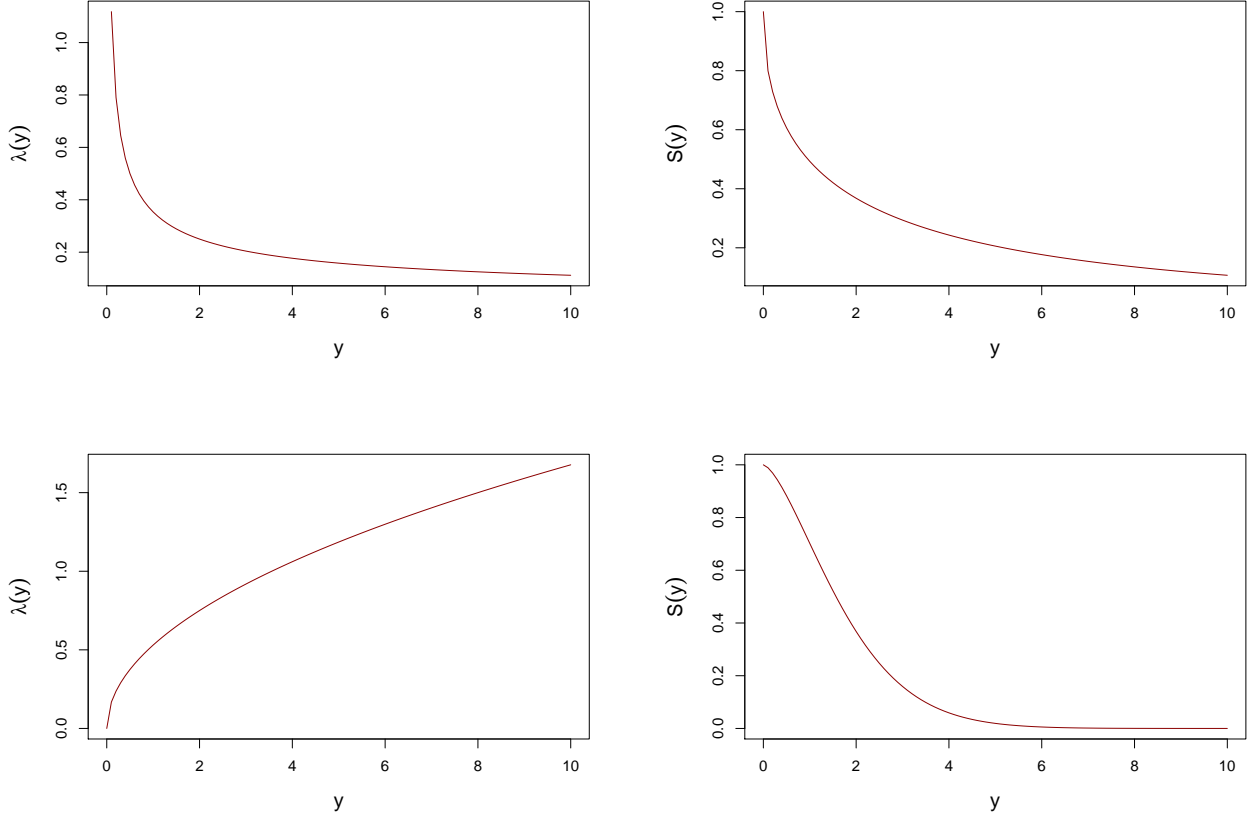
Figure 2.5: Example for the instantaneous hazard rate λ and the survival function S for the Weibull model with $k = 0.5$ (top panel) and with $k = 1.5$ (bottom panel) ($\lambda = 0.5$)

the Nelson-Aalen estimator of the cumulative risk Λ. The advantage of these nonparametric approaches is that they are more flexible than parametric approaches since they do not assume any distributional hypothesis. However, incorporating covariates in these approaches is not easy.

1.2.3.1 Kaplan-Meier estimator

The Kaplan-Meier estimator comes from the following. Considering two times y, y' such that $0 \leq y' \leq y$, then

$$S(y) = \mathbb{P}(Y > y) = \mathbb{P}(Y > y, Y \geq y') = \mathbb{P}(Y > y | Y > y')\mathbb{P}(Y > y').$$

Thus by noting $t_1, t_2, \ldots, t_D$ the D ordered distinct observed times among $T_1, \ldots, T_n$, where $T_i = \min(Y_i, C_i)$, and with the convention $t_0 = 0$,

$$S(t_d) = \mathbb{P}(Y > t_d) = \prod_{d'=1}^{d} \mathbb{P}(Y > t_{d'} | Y > t_{d'-1}).$$

Moreover $\mathbb{P}(Y > t_{d'} | Y > t_{d'-1}) = 1 - \mathbb{P}(Y \leq t_{d'} | Y > t_{d'-1}) = 1 - \mathbb{P}(Y \in]t_{d'-1}, t_{d'}] \mid Y > t_{d'-1})$.

$\mathbb{P}(Y \in]t_{d'-1}, t_{d'}] \mid Y > t_{d'-1})$ can be estimated by the number of events in $]t_{d'-1}, t_{d'}]$ (that is the number of events at $t_{d'}$: $\sum_{i=1}^{n} \delta_i \mathbb{1}_{T_i = t_{d'}}$) divided by the number of individuals that have not yet experienced the event just after time $t_{d'-1}$ (that is the number of individuals "at risk" as time $t_{d'}$): $\sum_{i=1}^{n} \mathbb{1}_{T_i > t_{d'-1}} = \sum_{i=1}^{n} \mathbb{1}_{T_i \geq t_{d'}}$.
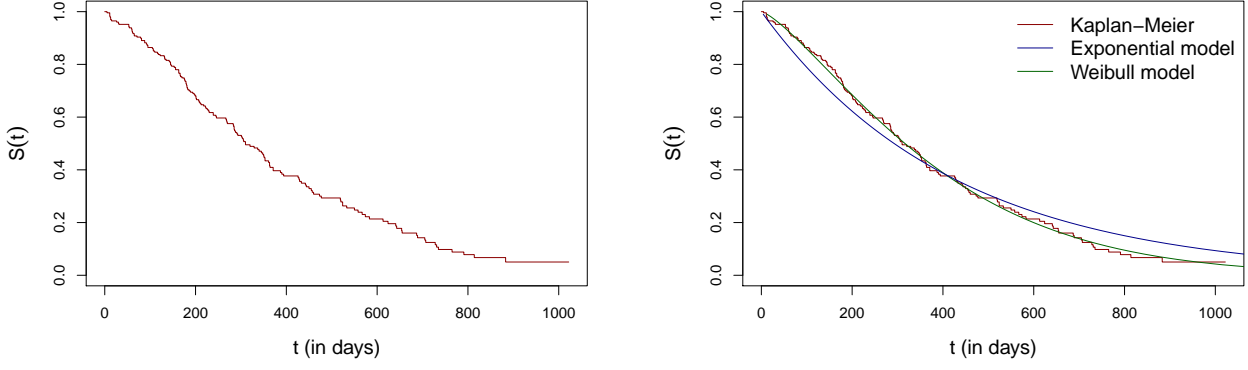
Figure 2.6: Examples of the Kaplan-Meier estimator (left panel), along with exponential and Weibull models (right panel) for the estimation of S on the *lung* data set from the R package *survival* corresponding to the survival time of 228 patients with advanced lung cancer from the North Central Cancer Treatment Group.

Thus

$$\hat{S}(t_d) = \prod_{d'=1}^{d} \left[1 - \frac{\displaystyle\sum_{i=1}^{n} \delta_i \mathbb{1}_{T_i=t_{d'}}}{\displaystyle\sum_{i=1}^{n} \mathbb{1}_{T_i \geq t_{d'}}} \right]$$

is the Kaplan-Meier estimator of S at each unique observed time $t_1, \ldots, t_D$: $\hat{S}$ is a right-continuous decreasing step function contrary to the parametric methods that result in a smooth estimation of S (Figure 2.6).

1.2.3.2   Nelson-Aalen estimator

The Nelson-Aalen estimator allows to estimate the cumulative hazard Λ.
Its formula comes from the fact that the instantaneous hazard rate λ can be estimated in each unique observed time t_d as

$$\hat{\lambda}(t_d) = \frac{\displaystyle\sum_{i=1}^{n} \delta_i \mathbb{1}_{T_i=t_d}}{\displaystyle\sum_{i=1}^{n} \mathbb{1}_{T_i \geq t_d}},$$

that is the number of events at t_d divided by the number of individuals "at risk" at time t_d.
Thus the Nelson-Aalen estimator of Λ at time t_d is

$$\hat{\Lambda}(t_d) = \sum_{d'=1}^{d} \hat{\lambda}(t_{d'}) = \sum_{d'=1}^{d} \frac{\displaystyle\sum_{i=1}^{n} \delta_i \mathbb{1}_{T_i=t_{d'}}}{\displaystyle\sum_{i=1}^{n} \mathbb{1}_{T_i \geq t_{d'}}}.$$

Remark that $\hat{\Lambda}$ is an increasing right-continuous step function.

1.2.4   Semi-parametric modeling: the Cox model

Let $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)})^\top$ be the vector of p covariates associated with the survival. Then the Cox model assumes that the instantaneous hazard rate at time $y > 0$ can be written as

$$\lambda(y|\boldsymbol{X}) = \lambda_0(y) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}\right].$$

This is a semi-parametric model since it assumes a parametric form concerning the effect of the covariates on the instantaneous hazard rate but makes no assumption on the form of λ_0.

This is also a proportional hazard model since for two individuals i and j, the ratio of their instantaneous hazard rates is supposed to be independent of time:

$$\frac{\lambda(y|\boldsymbol{X}_i)}{\lambda(y|\boldsymbol{X}_j)} = \frac{\exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}_i\right]}{\exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}_j\right]}.$$

This model is widely used in practice as it combines both the flexibility of nonparametric methods (no form is assumed on λ_0) and the possibility to easily take into account covariates (as in parametric models).

With this model we get

$$S(y|\boldsymbol{X}) = \exp\left[-\exp\left(\boldsymbol{\beta}^\top \boldsymbol{X}\right) \int_0^y \lambda_0(t)\, \mathrm{d}t\right] = S_0(y)^{\exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}\right]}$$

where $S_0(y) = \exp\left[-\int_0^y \lambda_0(t)\, \mathrm{d}t\right]$ and

$$f(y|\boldsymbol{X}) = \lambda(y|\boldsymbol{X}) S(y|\boldsymbol{X}) = \lambda_0(y) \exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}\right] S_0(y)^{\exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}\right]}.$$

Thus the likelihood can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^n f(T_i|\boldsymbol{X}_i)^{\delta_i} S(T_i|\boldsymbol{X}_i)^{1-\delta_i}$$

$$= \prod_{i=1}^n [\lambda_0(T_i) \exp\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right)]^{\delta_i} S_0(T_i)^{\exp\left[\boldsymbol{\beta}^\top \boldsymbol{X}_i\right]}.$$

Since λ_0 is unknown, the idea is to estimate $\boldsymbol{\beta}$ by considering the maximization of a partial likelihood expressed as (Cox, 1972)

$$\prod_{i=1}^n \left[\frac{\exp\left(\boldsymbol{\beta}^\top \boldsymbol{X}_i\right)}{\displaystyle\sum_{j, T_j \geq T_i} \exp\left(\boldsymbol{\beta}^\top \boldsymbol{X}_j\right)}\right]^{\delta_i}.$$

1.2.5   Frailty models

Sometimes factors, such as environmental factors, associated with the survival are not observed. In order to take into account this unknown heterogeneity in the data, frailty models have been developed. Here we focused on shared frailty models.

These shared frailties allow to take into account patient-specific heterogeneity when dealing with recurrent events (the event studied is repeated for an individual), or heterogeneity specific to groups of patients (individuals from the same family, the same municipality, the same hospital for example). In each case the Cox frailty model can be written as

- $\lambda_{i,r}(y|\{\boldsymbol{X}_{i,r}, \phi_i\}) = \lambda_0(y)\phi_i \exp[\boldsymbol{\beta}^\top \boldsymbol{X}_{i,r}]$ for the rth event of individual i. The frailty associated with the patient i is then ϕ_i with $\mathbb{E}[\phi_i] = 1$;

- $\lambda_{i,k}(y|\{\boldsymbol{X}_i, \phi_k\}) = \lambda_0(y)\phi_k \exp[\boldsymbol{\beta}^\top \boldsymbol{X}_i]$ for the individual i in group k. Thus the frailty associated with group k is ϕ_k with $\mathbb{E}[\phi_k] = 1$.

Note that these models can be rewritten respectively as

$$\lambda_{i,r}(y|\{\boldsymbol{X}_{i,r}, \varphi_i\}) = \lambda_0(y) \exp[\boldsymbol{\beta}^\top \boldsymbol{X}_{i,r} + \varphi_i]$$ for the rth event of individual i and
$$\lambda_{i,k}(y|\{\boldsymbol{X}_i, \varphi_k\}) = \lambda_0(y) \exp[\boldsymbol{\beta}^\top \boldsymbol{X}_i + \varphi_k]$$ for the individual i in group k,

and that although the term "frailty" refers to the ϕ_k (or ϕ_i), the φ_k (or φ_i) are also often referred to by the term "frailty".

Estimation of the frailty model
In order to estimate the frailty models a commonly used method is the EM algorithm (Klein, 1992). However, Therneau and Grambsch (2000) noted that this algorithm is slow and that a penalized partial likelihood method could be used instead. These methods require to assume a distribution of the frailties. The most used are the gamma, inverse Gaussian (Hougaard, 1986) and log-normal distributions. Note that for the latter we do not assume $\mathbb{E}[\phi] = 1$ but $\mathbb{E}[\varphi] = 0$ (McGilchrist and Aisbett, 1991).

2   Spatial data analysis

In many areas, data naturally exhibit a spatial component (Hung, 2016). For example, demographic data, such as the number of births, are measured at the level of administrative geographic units. In meteorology, temperature data is measured by sensors distributed over a geographical area. In addition, the expansion and availability of geo-located data has led to the development of new spatial analysis methods. For example, spatial regression methods (Chi and Zhu, 2008; Anselin, 2009; Ward and Gleditsch, 2018) have been developed to take into account the spatial dependence of the variable to be predicted. Moreover, when one considers data such as soil pollution, it is not feasible to measure the pollution at any point in space: in practice, the pollution is only measured at specific measurement points. Therefore, interpolation methods (such as kriging or co-kriging (Cressie, 1988; Holdaway, 1996; Legleiter and Kyriakidis, 2008; Belkhiri et al., 2020)) have been developed in order to estimate the pollution at any point. Finally, spatial clustering methods (Bourgault et al., 1992; Fouedjio, 2016c; D'Urso and Vitale, 2020) have been developed to take into account both the proximity of observations and the proximity of their spatial location.
Spatial methods are actually widely used in public health. In particular, they allow to study the spatial distribution of events and to determine the factors conditioning them. In particular, three types of analysis can be distinguished: (i) disease mapping which studies the spatial distribution of health events through cartographic representations, (ii) cluster detection, i.e., the detection of geographical areas with an over- or under-incidence of these events, and (iii) characterization of the relationships between spatial variations of health events and risk factors. In this last category one can mention ecological regressions that allow to model the link between the events and ecological variables (measured at the scale of geographical units), and the characterization of the spatial clusters detected in (ii) by ecological variables. These ecological studies present the advantage to allow the exploration of etiological hypotheses at low cost before confirmatory studies, such as epidemiological approaches at the individual level (cohort or case-control studies, etc.).

In Section 2.1 we present the different categories of spatial data. Then in Section 2.2 we focus on the spatial analysis of survival data. Finally in Section 2.3 we present the methods of spatial clustering and spatial clusters detection, with a focus on spatial scan statistics in Section 2.3.2.2.

2.1 Spatial data

Here we present the general concept of spatial data. In spatial data analysis, observations are considered to be realizations of a random spatial process $X = \{X_s, s \in S\}$ indexed by a spatial set $S \subset \mathbb{R}^d$ and taking values in a state space E. In most applications, S is considered to be two-dimensional ($S \subset \mathbb{R}^2$) or three-dimensional ($S \subset \mathbb{R}^3$, e.g., pollutant concentrations in a soil may be measured as a function of its "surface" location and depth). Furthermore, E can take its values in a continuous quantitative space ($E \subset \mathbb{R}^{+p}, p \geq 1$, e.g., measurement of the concentration of pollutants), a discrete quantitative space ($E \subset \mathbb{N}^p, p \geq 1$, e.g., counting of cases of diseases), or a qualitative space (e.g., presence or absence of cases of diseases: $E \subset \{0, 1\}^p, p \geq 1$). In practice, only n spatial units are observed. In the following we denote $s_1, \ldots, s_n$ these spatial units and $X_1, \ldots, X_n$ the realizations of X in $s_1, \ldots, s_n$ (see Figure 2.7).
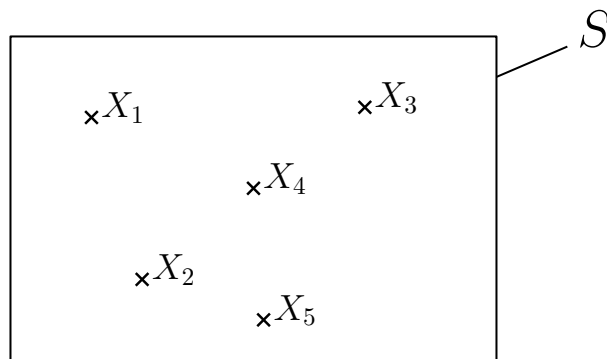
Figure 2.7: Example of a process X in five locations of an observation domain S

The nature of the spaces S and E determines the type of data. In particular we will distinguish geostatistical, lattice and point data (Gaetan and Guyon, 2010).

2.1.1 Geostatistical data

When $s_1, \ldots, s_n$ are fixed and S is a continuous subspace of $\mathbb{R}^d, d \geq 1$ then the data is said to be geostatistical. The repartition of $s_1, \ldots, s_n$ in S can be regular or not.
These data are often found in the fields of environmental sciences, meteorology (temperature or sunshine data for example) or oceanography. For example, we consider data from an environmental bio-monitoring campaign in the Lille Metropolis. Briefly, a particular type of lichen (Xanthoria parietina) was collected at 159 locations to analyze the concentrations of 14 trace elements. We consider here the concentrations of cadmium. These concentrations can be measured everywhere on the territory (S is continuous). However, in practice they are only measured at a finite set of 159 locations determined by environmentalists. Figure 2.8 (left panel) shows the observed locations $s_1, \ldots, s_{159}$ as well as the measured concentrations. Then the data can be spatially interpolated (via a kriging method for example) to obtain an estimate of the cadmium concentration over the entire study area (Figure 2.8, right panel). The interested refer may refer to Matheron (1963); Chiles and Delfiner (1999); Wackernagel (2003) for more details on geostatistical data and their modeling.
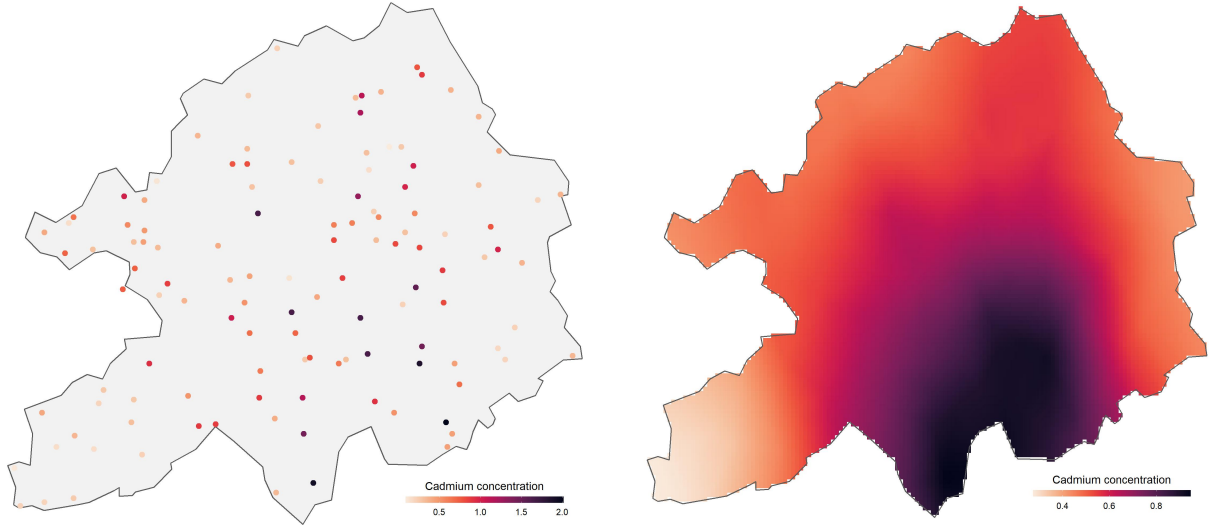
Figure 2.8: Observed cadmium concentrations from a lichen biomonitoring campaign in the territory of the European Metropolis of Lille (left panel) and spatially interpolated cadmium concentrations (*via* a kriging method) (right panel).

2.1.2 Lattice data

When S is a fixed discrete subset of $\mathbb{R}^d, d \geq 1$ and $s_1, \ldots, s_n$ correspond to geographical spatial units connected by a neighborhood graph of order n, then the data is said to be lattice. The neighborhood graph $\mathcal{G}$ associated with S is characterized by the set of its vertices S and a set of edges $\mathcal{A}$}. Then one can define a spatial weight matrix $V_n = (v_{i,j,n})_{1 \leq i,j \leq n}$ such that

$$
v_{i,j,n} = \begin{cases} v_{i,j,n} > 0 & \text{if } (i,j) \in \mathcal{A} \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}.
$$

In practice the weights $v_{i,j,n}$ can be defined in several manners among which (i) the inverse distance weights ($v_{i,j,n} = ||s_i - s_j||^{-\alpha}, \ \alpha > 0$), (ii) the weights based on the k-nearest neighbors ($v_{i,j,n} = 1$ if s_j belongs to the set of k-nearest neighbors of s_i, 0 otherwise) and (iii) the weights based on the notion of adjacency (i.e., $v_{i,j,n} = 1$ if s_i and s_j share a common boundary, 0 otherwise).

This type of data corresponds to data aggregated at the level of spatial units and located by their center (centroid or administrative center, for example).

This is the most common type of data in the field of epidemiology, especially for anonymity reasons. More precisely, cases of a disease can be located precisely by the residence of the individuals concerned. However, for anonymity reasons, the data are aggregated at the level of administrative spatial units such as the municipality or the department, for which only the number of cases of the disease under study is then available.

Figure 2.9 shows an example of lattice data: the choropleth map represents the average premature mortality rate in each department of metropolitan France (except Corsica) during the period 1998 to 2013. The premature mortality rate is aggregated at the level of a department (more specifically at its centroid s_i) and the connectivity between the departments can be represented by a neighborhood graph. The refer may refer to Ripley (1981); Cressie (1993) for more details on lattice data.

Distribution of premature mortality rates
France | 1998-2013

Premature mortality rates
(/1000 inhabitants)
■ [2.44,3.04]
■ [2.22,2.44]
■ [2.08,2.22]
■ [1.9,2.08]
■ [1.59,1.9]
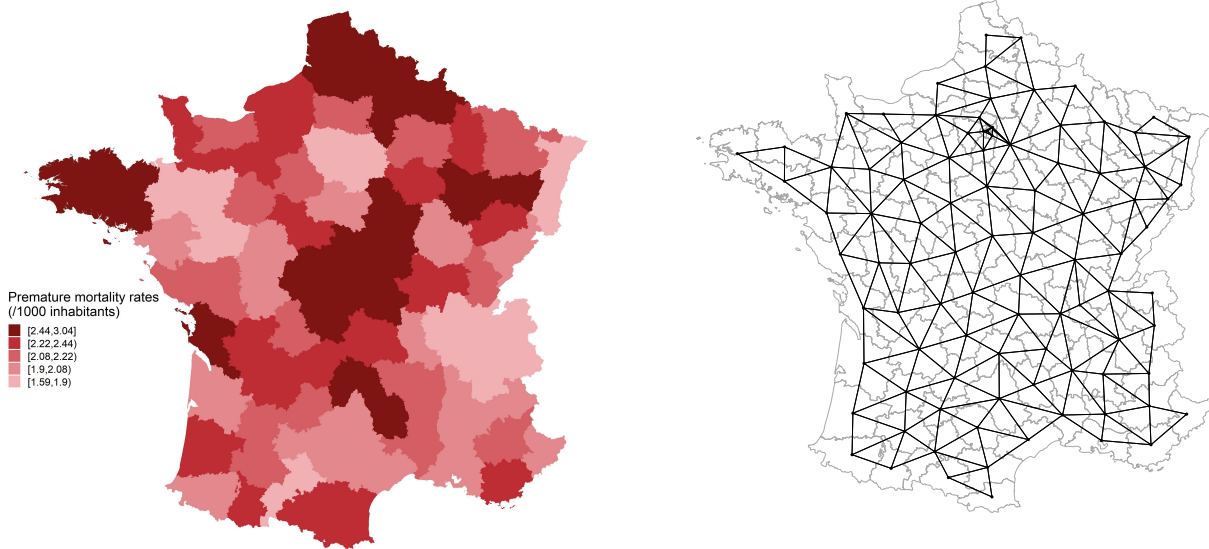
Figure 2.9: Mean premature mortality rate (for 1000 inhabitants) in France during the period 1998 to 2013 (*Source: Insee*) (left panel) and neighborhood graph of the French *departments* (right panel)

2.1.3 Spatial point data

When the number n of observation sites and their locations are random, the process is said to be a spatial point process. Here the realizations of X correspond to the locations of the sites: $x_1, \ldots, x_n$. E is then the set of locally finite configurations on S. The distribution of $x_1, \ldots, x_n$ can be completely random, aggregated or rather regular as shown in Figure 2.10 (the data comes from the R package *spatstat*).

Moreover X is said to be a marked point process if a value (a mark) is observed at each x_i.

Then Figure 2.10b shows a spatial point process (unmarked) with rather regular locations, whereas Figures 2.10a and 2.10d show spatial marked point processes with qualitative (the type of ants) and quantitative (the diameter of the pine trees) marks respectively. In Figure 2.10c the locations tend to be aggregated whereas the locations of the ants (Figure 2.10a) seem to be completely random. Note that a more comprehensive review of spatial point processes is provided by Cressie (1993) and Gaetan and Guyon (2010).

In this thesis we focus on the particular case $S \subset \mathbb{R}^2$ and on two categories of lattice spatial data, namely spatial functional data and spatial survival data. The spatial analysis of the latter requires defining the notion of spatial frailty which is explained in the next section.
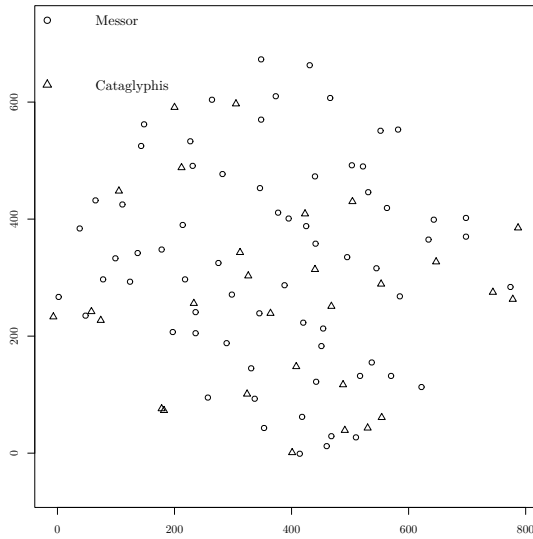
(a) Locations of two species of ants in Northern Greece
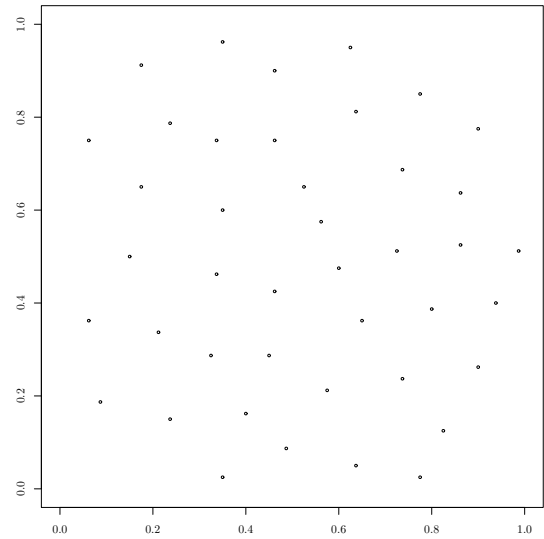
(b) Location of 42 cell centers of a histological section
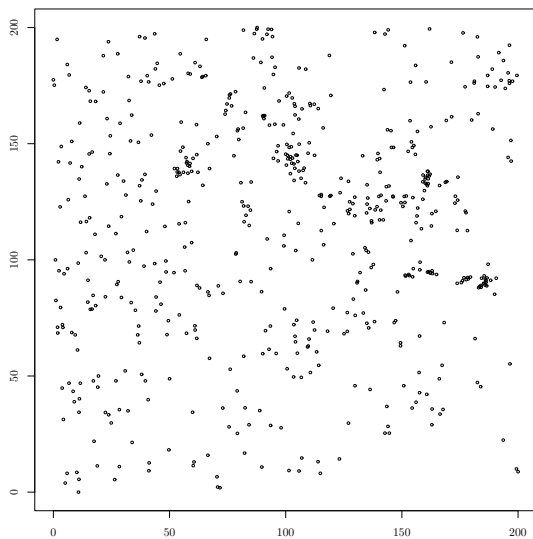
(c) Location of 584 pine trees in South Georgia
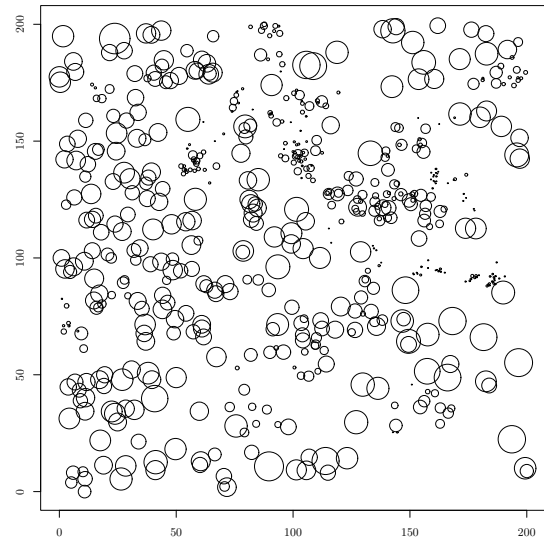
(d) Location and diameter of the 584 pine trees

Figure 2.10: Examples of spatial point data

2.2   Spatial analysis of survival data

Section 1.2 introduced the concept of survival data. When these data are spatially localized we refer to them as spatial survival data. For example, in spatial epidemiology, it is possible to observe the survival time of individuals following a renal transplant. These individuals can be localized by their place of residence. However, it should be noted that for reasons of anonymity, the exact location of individuals is rarely known in practice: their location is only available at a larger scale such as an administrative area (Figure 2.11). Furthermore, it should be noted that the survival times for individuals in the same spatial unit may be correlated: e.g., these individuals benefit from similar access to health care. In what follows, we will refer to this type of correlation as intra-spatial unit correlation. Moreover, the spatial units can be spatially dependent: e.g., access to health care can be facilitated around large cities. Hence, these

potential dependencies must be taken into account when dealing with spatial survival data.
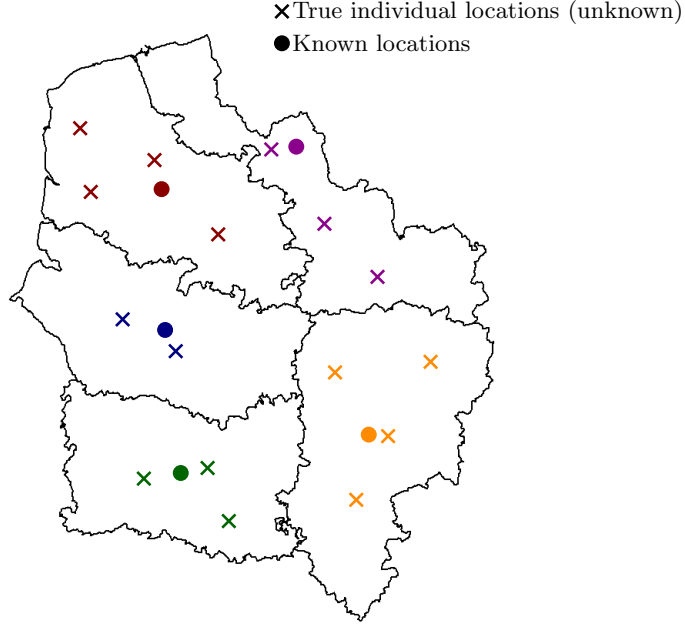
Figure 2.11: Example of survival data locations: In the case of survival data, the observation are individual. However for privacy reasons the researcher does not always get the precise location of the individuals.

In order to take the potential intra-spatial unit correlation into account, the literature proposes to consider shared frailty models (Clayton, 1978; Liang et al., 1995; Balan and Putter, 2020; Mahanta et al., 2021). As explained in Section 1.2.5, this makes it possible to take into account unobserved factors at the level of spatial units. However, these models do not take into account a possible spatial dependence between the spatial units, which may be due to unobserved spatially correlated socio-economic or environmental factors such as air pollution. This can lead to incorrect inference of the association between covariates and survival times (Li and Ryan, 2002). Thus, spatial frailty models were introduced to take into account both a potential intra-spatial unit correlation and a potential spatial dependence between the spatial units. They have been used in many application studies (Hennerfeind et al., 2003; Aswi et al., 2020; Daniel et al., 2021) and the following sections give examples of spatial models for the spatial frailties as well as estimation methods.

2.2.1   Spatial frailty distributions

Let $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K)^\top$ be the vector of the K shared frailties associated with K regions $s_1, \ldots, s_K$. Then the spatial dependence can be introduced in the frailties by several spatial models. For example one can consider geostatistical models such as an exponential or a Matérn structure (Banerjee et al., 2003):

- Exponential model: $\boldsymbol{\varphi} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{k,l} = \sigma^2 \exp\left[-C||s_k - s_l||_2\right]$, $C > 0$;

- Matérn model: $\boldsymbol{\varphi} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{k,l} = \sigma^2 \dfrac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\dfrac{||s_k - s_l||_2}{\rho}\right)^\nu \mathcal{K}_\nu\left(\sqrt{2\nu}\dfrac{||s_k - s_l||_2}{\rho}\right)$, $\mathcal{K}_\nu$ is the modified Bessel function of the second kind, and $\rho, \nu > 0$.

One can also consider lattice models such as

- A Leroux conditional autoregressive (CAR) model (Leroux et al., 2000): $\boldsymbol{\varphi} \sim \mathcal{N}(0, \Sigma)$ where $\Sigma = \sigma^2[\rho R + (1 - \rho)\mathrm{I}_K]^{-1}$, R is the square matrix with elements
$R_{k,l} = \begin{cases} \sum_{j=1}^{K} v_{k,j} & \text{if } k = l \\ -v_{k,l} & \text{otherwise} \end{cases}$, with $v_{k,l} = 1$ if s_k and s_l are adjacent and 0 otherwise, and $\rho \in [0, 1]$;

- An intrinsic CAR model (Besag et al., 1991) which is equivalent to the CAR model with $\rho = 1$.

2.2.2   Estimation

Some frequentist approaches have been proposed in the literature to estimate these models (Li and Ryan, 2002; Li and Lin, 2006; Hanson and Zhou, 2014). However, these approaches are few in number and because of the complexity of the likelihood function in these models, a frequentist estimation is relatively complex, and time-consuming (Motarjem et al., 2019). Moreover, to our knowledge, no implementation of these methods is available.

In contrast, much more literature is available concerning Bayesian approaches and many are implemented in R packages (Taylor and Rowlingson, 2017; Zhou et al., 2017). This has led to a large number of applications using this estimation method (Su et al., 2020; Thamrin et al., 2021). Actually among the Bayesian estimation methods, two approaches can be distinguished: the Markov chain Monte Carlo (MCMC) method (Hennerfeind et al., 2003; Banerjee et al., 2003; Aswi et al., 2020), and the integrated nested Laplace approximation (INLA) (Bivand and Gómez-Rubio, 2021). For more information on these methods the reader is invited to refer to Karandikar (2006) and Rue et al. (2009).

2.3   Spatial clustering and spatial clusters detection

Spatial clustering methods (Bourgault et al., 1992; D'Urso and Vitale, 2020) aim at partitioning a geographical area into a set of K optimal groups by taking into account both the proximity of observations and their geographical locations. The number of clusters K can be fixed in advance but it is usually optimized on the data. This type of approach is particularly useful in public health. Indeed, partitioning geographical area into more specific zones according to their socio-economic status or behavioral factors such as alcohol or tobacco consumption allows to adapt public health policies or prevention campaigns locally (Carvalho et al., 2009) and therefore to increase their efficiency. Another objective may be to detect atypical geographical areas, characterized by statistically higher alcohol consumption, or environmental black spots (i.e. geographical areas characterized by abnormally high pollution) in order to act specifically in these areas, for preventive health purposes or to reduce pollution. In this case the relevant methods are cluster detection approaches. In particular, spatial scan statistics are well-known cluster detection methods allowing to localize clusters and to determine their statistical significance.

Note that clustering and clusters detection methods are different although both consider the notion of "clusters". Indeed, spatial clustering methods consist in partitioning a geographical area in an optimal way and in this context a "cluster" corresponds to a group of observations presenting similar behaviors, whereas spatial cluster detection methods are statistical tests that seek to determine whether there exist spatial aggregations of outliers, called "spatial clusters".

Clustering and cluster detection methods are the subject of the next two sections. First, we present the different approaches of clustering in a non-spatial framework (Section 2.3.1.1) in order to then present their adaptations in the spatial framework (Section 2.3.1.2). Then, Section 2.3.2.1 presents the categories of tests in spatial analysis before focusing on cluster detection tests and more specifically spatial scan statistics in Section 2.3.2.2.

2.3.1 Spatial clustering

2.3.1.1 Non-spatial model-based clustering

The purpose of clustering methods is to group observations into clusters such that observations within a cluster share similar characteristics and the characteristics of each cluster are distinct from each other. These methods can be divided into several categories: in particular we can distinguish hierarchical methods (using similarity measures), partitional approaches (considering the optimization of an objective function) and model-based methods. Here we focus on model-based approaches. However the interested reader can find details about the other methods in Everitt et al. (2011); Madhulatha (2012); Saxena et al. (2017).

Model-based approaches assume that the observations $X_1, \ldots, X_n$ are realizations of a mixture model

$$f(x|\Theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k)$$

with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0, \forall k \in [\![1, K]\!]$. f is the density of X, f_k is the density of X in the group k, characterized by a vector of parameters θ_k, and π_k is the probability of group k. We note $\Theta = (\theta_1^\top, \ldots, \theta_K^\top, \pi_1, \ldots, \pi_K)$.

A classical model is to assume that the f_k are Gaussian densities parameterized by θ_k. θ_k is then estimated by an Expectation-Maximization (EM) algorithm.

EM algorithm

By noting $Z_1, \ldots, Z_n$ the cluster memberships (non observed) corresponding to $X_1, \ldots, X_n$, the log-likelihood of the complete data (X, Z) is

$$
\begin{aligned}
\ell(\Theta; X_1, \ldots, X_n, Z_1, \ldots, Z_n) &= \sum_{i=1}^{n} \log f((X_i, Z_i)|\Theta) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{Z_i=k}[\log f(X_i|Z_i = k, \Theta) + \log \mathbb{P}(Z_i = k|\Theta)] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{Z_i=k}[\log f_k(X_i|\theta_k) + \log \pi_k].
\end{aligned}
$$

The EM algorithm is an iterative algorithm presented in Algorithm 1. By noting $\Theta^{(t)}$ the value obtained for Θ at iteration t, then the idea is to maximize at the iteration $t + 1$

$$\mathbb{E}_{Z|X_1,\ldots,X_n,\Theta^{(t)}}[\ell(\Theta; X_1, \ldots, X_n, Z_1, \ldots, Z_n)]$$

that is to maximize

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}(Z_i = k|X_i, \Theta^{(t)})[\log f_k(X_i|\theta_k) + \log \pi_k]$$

with

$$\mathbb{P}(Z_i = k|X_i, \Theta^{(t)}) = \frac{f(X_i|Z_i = k, \Theta^{(t)})\mathbb{P}(Z_i = k|\Theta^{(t)})}{\displaystyle\sum_{j=1}^{K} f(X_i|Z_i = j, \Theta^{(t)})\mathbb{P}(Z_i = j|\Theta^{(t)})}$$

$$= \frac{\pi_k^{(t)} f_k(X_i|\theta_k^{(t)})}{\sum\limits_{j=1}^{K} \pi_j^{(t)} f_j(X_i|\theta_j^{(t)})}$$

from the Bayes theorem.

The expectation and maximization steps are repeated until convergence (or until a maximal number of iterations is reached). The interested reader can find more details about these approaches in Bouveyron and Brunet-Saumard (2014); Melnykov et al. (2015) for example.

Algorithm 1: The EM algorithm for model-based clustering

Input: $X_1, \ldots, X_n, K$, maximum number of iterations T

Initialization: initial values $\Theta^{(1)} = (\theta_1^{(1)\top}, \ldots, \theta_K^{(1)\top}, \pi_1^{(1)}, \ldots, \pi_K^{(1)})$, $t \leftarrow 1$;

while $t \leq T$ **do**

  E-step:

  Compute $\tau_{i,k}^{(t+1)} = \mathbb{P}(Z_i = k|X_i, \Theta^{(t)}) = \dfrac{\pi_k^{(t)} f_k(X_i|\theta_k^{(t)})}{\sum\limits_{j=1}^{K} \pi_j^{(t)} f_j(X_i|\theta_j^{(t)})}$ $\forall i, k$;

  M-step:

  Compute $\pi_k^{(t+1)} = \dfrac{1}{n} \sum\limits_{i=1}^{n} \tau_{i,k}^{(t+1)}$;

  Compute $\Theta^{(t+1)}$ by maximizing $\sum\limits_{i=1}^{n} \sum\limits_{k=1}^{K} \tau_{i,k}^{(t+1)} \log f_k(X_i|\theta_k)$;

  $t \leftarrow t + 1$;

  Stop if the algorithm has converged ;

Output: $\hat{\Theta} = (\hat{\theta}_1^{\top}, \ldots, \hat{\theta}_K^{\top}, \hat{\pi}_1, \ldots, \hat{\pi}_K)$ and $\hat{Z}_1, \ldots, \hat{Z}_n$ with $\hat{Z}_i = \underset{k \in [\![1,K]\!]}{\arg \max} \hat{\tau}_{i,k}$

2.3.1.2   Spatial clustering approaches

When the data present a spatial component, when one seeks to identify groups of observations, one generally wishes that the clusters obtained are not scattered in the study area. Moreover, spatial data often present, by nature, a spatial dependence such that nearby observations present similar characteristics. In this context, several works have focused on spatial data clustering. Spatial clustering methods can be separated into four categories (Fouedjio, 2017).

The first one considers non-spatial clustering approaches and includes spatial information as an attribute on which to perform clustering. However the clusters obtained are very scattered since the method does not distinguish between spatial and non-spatial attributes (Fouedjio, 2016c).

The second category consists of methods that modify classical similarity or dissimilarity measures to include the spatial information. For example, Oliver and Webster (1989) and Bourgault et al. (1992) proposed to weight an existing dissimilarity measure by a variogram. However Romary et al. (2015) and Fouedjio (2016c) noted that the clusters obtained with these approaches are still rather scattered.

Thus, Fouedjio (2016c) proposed a first solution by estimating the spatial dependency structure with a nonparametric kernel estimator to derive a new dissimilarity measure. Then the idea is to apply an agglomerative hierarchical clustering on the obtained dissimilarity matrix. A

similar approach was proposed by Fouedjio (2016a). Fouedjio (2016b) and Fouedjio (2017) also considered a nonparametric kernel estimator to build a similarity matrix before applying a spectral clustering method (Von Luxburg, 2007; Filippone et al., 2008). Finally D'Urso and Vitale (2020) proposed a modified version of the method of Fouedjio (2016c) by combining an agglomerative hierarchical clustering and a new dissimilarity measure which is robust to the possible presence of outliers in the data.

Methods of **the third category** use non spatial clustering approaches while imposing a spatial constraint. Among others, we can highlight the works of Pawitan and Huang (2003) and Romary et al. (2015) who proposed agglomerative hierarchical clustering approaches by adding the constraint that at each step only neighbor clusters can be merged. The method therefore requires a neighborhood structure. Depending on this, the results may vary. A structure that works quite well in practice is the one built with the Delaunay triangulation.

Finally **the last category** corresponds to model-based approaches. For example, Allard and Guillot (2000) assumes that the geographical area S is a partition of K unknown clusters $S_1, \ldots, S_K$. Let $X_1, \ldots, X_n$ be n real univariate observations in n spatial locations $s_1, \ldots, s_n \in S$ of a random function X such that

$$\mathbb{E}[X(s)] = \mu_k \quad \forall s \in S_k$$

$$Cov(X(s), X(s')) = \sigma_k^2 \rho_k(s' - s) \ \forall s, s' \in S_k$$

$$Cov(X(s), X(s')) = 0 \text{ if } s \text{ and } s' \text{ do not belong to the same cluster.}$$

ρ_k is the correlation function and σ_k^2 is the variance. Then by noting R_k the correlation matrix of the n_k observations of S_k, $\theta_k = (\mu_k, \sigma_k^2, R_k)$ and $\Theta = (\theta_1, \ldots, \theta_K)$, the likelihood associated with the Gaussian model is

$$\mathcal{L}(\Theta; X_1, \ldots X_n) = \frac{1}{(2\pi)^{n/2}} \prod_{k=1}^{K} \frac{1}{\sigma_k^{n_k} |R_k|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{X}_k - \mu_k \mathbb{1})^\top (\boldsymbol{X}_k - \mu_k \mathbb{1}) \sigma_k^{-2} R_k^{-1}\right]$$

where $\boldsymbol{X}_k$ is the vector of the n_k observations in S_k and $\mathbb{1}$ is the column vector made up only of 1.

The idea is then to use an adaptation of the EM algorithm presented in Algorithm 1 for the spatial clustering, assuming that the data are realizations of a Gaussian mixture

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

with $\forall k \in [\![1, K]\!], \pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

Spatial clustering approaches for functional data

In domains in which data naturally involve a spatial component, the emergence of functional data (Ramsay and Silverman, 2005b) has led to the introduction of spatial functional data (Delicado et al., 2010); these are characterized by the observation of one or more curves in each spatial location. Consequently, the combination of spatial clustering methods with FDA has prompted the development of new spatial clustering methods for functional data. For instance, several clustering methods for spatial functional data have combined the "dynamic clusters method" (Diday, 1971) with a trace-semivariogram (Romano et al., 2010, 2011). Furthermore, Giraldo et al. (2012) developed a hierarchical clustering method. Like Jacques and Preda (2012) and Jacques and Preda (2014b), Vandewalle et al. (2022) developed a model-based clustering method in the specific context of spatial functional data. Note that other approaches have also been suggested (Jiang and Serban, 2012; Romano et al., 2017).

2.3.2   Cluster detection

In the context of cluster detection, Knox (1989) defined a cluster as "a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance". In many application fields, researchers are interested in determining whether or not spatial clusters exist. For example, in the field of public health, researchers are interested in detecting geographic areas with a higher risk of disease than elsewhere. The identification of such geographical areas allows, in the case of a disease whose etiology remains elusive, to conduct local investigations in order to make hypotheses on the potential causes of the disease. Another example of application is the detection of geographical areas in which individuals consume more alcohol or tobacco than elsewhere. This can then be used, for example, to conduct more targeted prevention campaigns. Finally, in the context of environmental sciences, it is essential to be able to detect environmental black spots, that is, abnormally polluted geographical areas, in order to carry out localized actions in these areas to reduce the level of pollution.

These motivations require the development of statistical methods to determine whether spatial clusters exist, in an objective way, in particular by avoiding any pre-selection bias. The latter is defined by the use of the observed data to determine both the location of a cluster and its statistical significance. As an example, consider the mean PM_{10} concentrations in northern France in 2015 (Figure 2.12). It appears that an area (hatched) in the Lille metropolis presents higher PM_{10} concentrations than elsewhere. Then, an idea might be to simply perform a classical statistical test to determine whether the average PM_{10} concentration in this area is statistically significantly higher than outside the area. However, this "cluster" has been determined in a totally subjective way from the observed data, which leads to a pre-selection bias and to an overestimation of the statistical significance of this cluster.
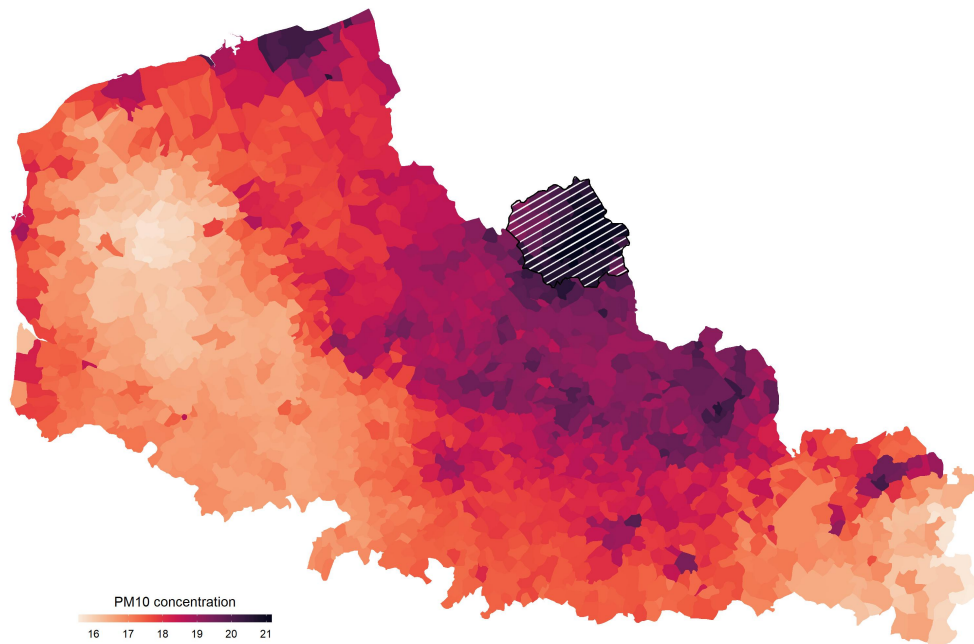
Figure 2.12: Mean PM_{10} concentrations in northern France in 2015 and illustration of the pre-selection bias for cluster detection

2.3.2.1 Cluster detection methods

In the literature, the methods are divided into three categories: global tests, focused tests and detection tests.

Global tests. This category includes tests that determine whether there is a global tendency to spatial aggregation. Thus, they identify whether nearby spatial units tend to present similar values of the measured variable, thus testing for the presence of spatial autocorrelation. In particular, we can mention the methods of Moran (1950); Diggle and Chetwynd (1991); Besag and Newell (1991). It should be noted that the methods of this category do not allow to detect the precise location of spatial clusters nor to test their statistical significance. The interested reader may consider Getis (2010) for a more detailed review of these methods.

Focused tests. These tests use an *a priori* defined source point to test the aggregation of events around this source point. For example, these approaches can be used in the case where, following a chemical leak from a factory, researchers wish to study the cases of cancer around this factory (Figure 2.13).

Note that since the knowledge of the source point is defined *a priori* and not from the observed
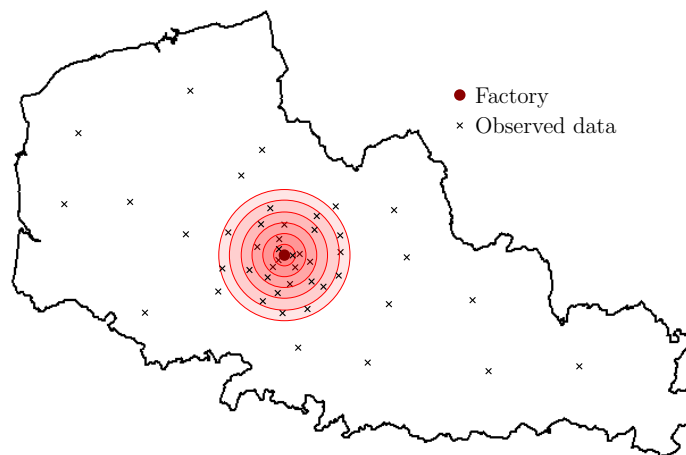
Figure 2.13: Example of a focused test: following a chemical leak from a factory, researchers wish to study the cases of cancer around this factory

data, these approaches do not suffer from pre-selection bias. In this category one can mention the methods of Stone (1988) and Bithell (1995).

For example, Michelozzi et al. (2002) used Stone's test to analyse leukemia cases near a radio station in Rome and found an over-incidence of leukemia deaths in men in the area within two kilometers of the station, as well as a decreased risk of developing the disease further away. Similar results were obtained by Anderson and Henderson (1986) and Maskarinec et al. (1994).

Detection tests. These approaches allow to objectively detect spatial clusters without any pre-selection bias and to test their statistical significance. In the field of health spatial analysis, these methods are mostly used due to the fact that epidemiologists rarely know the location of spatial clusters. In particular, the local indicators of spatial association (LISA) proposed by Anselin (1995); Bouayad Agha and De Bellefon (2018) including the local Moran's I, belong

to this category. Openshaw et al. (1987) proposed an approach named "Geographical Analysis Machine". It consists in performing a test for each cluster w of a set of potential clusters $\mathcal{W}$, which results in multiple testing and thus in the increasing of the type I error. In contrast, the spatial scan statistics approach originally proposed by Kulldorff and Nagarwalla (1995) is recognized to be particularly powerful (Kulldorff et al., 2003b; Goujon-Bellec et al., 2011) and does not suffer from multiple testing problems. These methods have been widely applied in epidemiology (Kulldorff, 1999; Luquero et al., 2011; Marciano et al., 2018; Genin et al., 2020; Khan et al., 2021) but also in other research fields such as environmental science (Sudakin et al., 2002; Gao et al., 2014; Wan et al., 2020; Shi et al., 2021), criminology (Minamisava et al., 2009) or astronomy (Bidin et al., 2010). This thesis focuses on these methods, and Section 2.3.2.2 therefore presents a review of the literature on spatial scan statistics.

2.3.2.2   Spatial scan statistics

General principle

First scan statistics were proposed by Naus (1965) in the uni-dimensional discrete framework. These are random variables used as test statistics considering a null hypothesis $\mathcal{H}_0$ of absence of cluster against a composite alternative hypothesis $\mathcal{H}_1$ (presence of at least one cluster). More formally, by noting $\{X_i\}_{i \geq 1}$, the observed variables, a scan statistic aims at testing $\mathcal{H}_0$: the X_i are independent and identically distributed variables according to a probability distribution $\mathcal{F}$ with parameter p: $\forall i, X_i \sim \mathcal{F}(p)$, against the alternative hypothesis $\mathcal{H}_1$: the X_i are independent and there is a cluster w such that $\forall i \in w, X_i \sim \mathcal{F}(p_w), \forall i \in w^c, X_i \sim \mathcal{F}(p_{w^c}), p_w > p_{w^c}$.

In the following we consider that the observations $X_1, \ldots, X_n$ take values in $\mathbb{N}$. For example, X_i corresponds to a number of events at a certain trial i. Then for a window of fixed size $T \in \mathbb{N}^*$, we denote

$$N_i = \sum_{i'=i}^{i+T-1} X_{i'}$$

the total number of events observed in the interval of trials $[i, i+T-1]$ (see Figure 2.14), and Naus (1965) defines the discrete one-dimensional scan statistic associated with the window size T as

$$\Lambda(T) = \max_{1 \leq i \leq n-T+1} N_i$$

which corresponds to the maximum number of events obtained within T contiguous trials. The example presented in Figure 2.14 leads to $\Lambda(4) = 7$.

Then by noting $\mathcal{W}(T) = \{w_i = [i, i+T-1], 1 \leq i \leq n-T+1\}$, we have

$$\Lambda(T) = \max_{w_i \in \mathcal{W}(T)} N^{(w_i)},$$

where $N^{(w_i)} = \sum_{i' \in w_i} X_{i'}$. $\mathcal{W}(T)$ is the "set of potential clusters of fixed size T" and $N^{(w_i)}$ is called "concentration index associated with the potential cluster w_i". In this context, $\mathcal{H}_0$ can be rewritten as $\mathcal{H}_0 : \forall i \in [1, n], X_i \sim \mathcal{F}(p)$ and $\mathcal{H}_1$ can be rewritten as $\mathcal{H}_1 : \exists w \in \mathcal{W}(T)$ such that $\forall i \in w, X_i \sim \mathcal{F}(p_w), \forall i \in w^c, X_i \sim \mathcal{F}(p_{w^c}), p_w > p_{w^c}$. Finally $\mathcal{H}_0$ is rejected for large values of $\Lambda(T)$.

This approach can be extended to the discrete bi-dimensional case. Let a rectangular region $[0, R_1] \times [0, R_2]$ and $\{X_{i,j}, 1 \leq i \leq R_1, 1 \leq j \leq R_2\}$ be a set of random independent variables with values in $\mathbb{N}$. For example $X_{i,j}$ corresponds to the number of events observed in the region
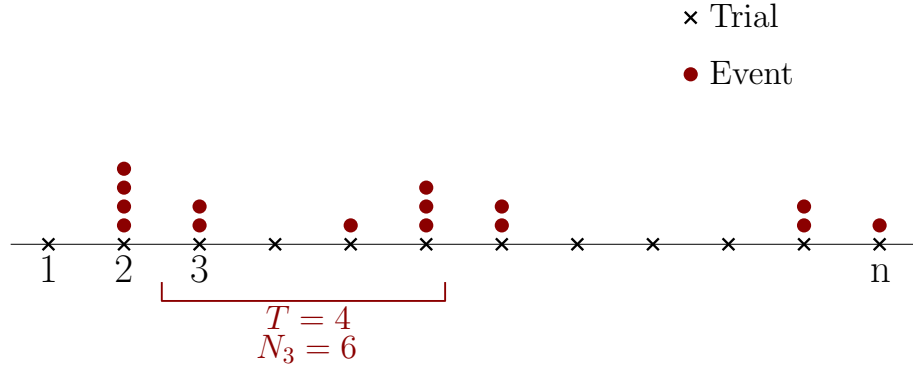
Figure 2.14: Example of the scanning window for the uni-dimensional scan statistic

$s_{i,j} = [i-1, i] \times [j-1, j]$. Then for a window of size $T_1 \times T_2 \in \mathbb{N}^* \times \mathbb{N}^*$, we note

$$N_{i,j} = \sum_{i'=i}^{i+T_1-1} \sum_{j'=j}^{j+T_2-1} X_{i',j'}$$

the number of events in the window $[i-1, i+T_1-1] \times [j-1, j+T_2-1]$ (see Figure 2.15) and the discrete bi-dimensional scan statistic is then defined by

$$\Lambda(T_1, T_2) = \max_{\substack{1 \le i \le R_1-T_1+1 \\ 1 \le j \le R_2-T_2+1}} N_{i,j}.$$

The example presented in Figure 2.15 leads to $\Lambda(2,3) = 7$. By considering a set of rectangular potential clusters

$$\mathcal{W}(T_1, T_2) = \{w_{i,j} = [i, i+T_1-1] \times [j, j+T_2-1], 1 \le i \le R_1-T_1+1, 1 \le j \le R_2-T_2+1\},$$

we have

$$\Lambda(T_1, T_2) = \max_{w_{i,j} \in \mathcal{W}(T_1, T_2)} N^{(w_{i,j})}$$

where $N^{(w_{i,j})} = \sum_{(i',j') \in w_{i,j}} X_{i',j'}$ is a concentration index associated with the potential cluster $w_{i,j}$. In this context, $\mathcal{H}_0$ and $\mathcal{H}_1$ can be rewritten respectively as $\mathcal{H}_0 : \forall(i,j) \in [1, R_1] \times [1, R_2], X_{i,j} \sim \mathcal{F}(p)$ and $\mathcal{H}_1 : \exists w \in \mathcal{W}(T_1, T_2)$ such that $\forall(i,j) \in w, X_{i,j} \sim \mathcal{F}(p_w), \forall(i,j) \in w^c, X_{i,j} \sim \mathcal{F}(p_{w^c}), p_w > p_{w^c}$. The null hypothesis is rejected for large values of $\Lambda(T_1, T_2)$.

Note that here we have considered only rectangular potential clusters and the discrete case but this work can also be extended to the case of a continuous location space (Naus, 1966) and different window shapes such as circles or ellipses (Alm, 1997; Anderson and Titterington, 1997).

Spatial scan statistics extend the bi-dimensional case of scan statistics. Kulldorff and Nagarwalla (1995) and Kulldorff (1997) were the first to propose a method to detect spatial clusters and to evaluate their statistical significance by a Monte-Carlo approach without any pre-selection bias by considering a circular scanning window of varying size. This approach can be decomposed into two steps. The first one corresponds to the detection of the most likely cluster (MLC) among the set of potential circular clusters of variable sizes $\mathcal{W}$ and is explained in the paragraphs "Scanning window" and "Concentration index". The second step consists in determining the statistical significance of the MLC and is detailed in the paragraph "Determining the statistical significance of the MLC".

It should be noted that the vast majority of spatial scan statistics models proposed in the literature (including Kulldorff and Nagarwalla's ones) are based on the assumption of
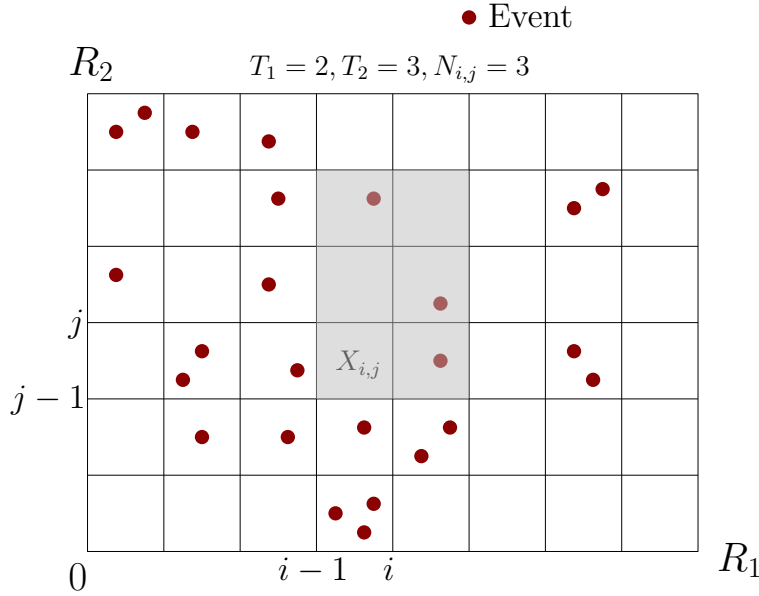
Figure 2.15: Example of the scanning window for the bi-dimensional scan statistic

independence of spatial observations. Although this assumption is too simplistic and does not respect Tobler's first law of geography, the authors argue that spatial dependence is taken into account in the scanning process. More recently, authors have studied the impact of spatial dependence on the performance of spatial scan statistics, showing that it leads to a significant inflation of the type I error. This topic is detailed in section "Taking into account the spatial autocorrelation".

Scanning window.   The scanning process uses a scanning window of variable size (and potentially shape) that moves throughout the study region. By varying its location, its size, and possibly its shape, we obtain a set $\mathcal{W}$ of windows $w \subset S$ called potential clusters. In particular, Kulldorff (1997) proposed to consider potential clusters containing at most 50% of the population at risk. In general, a common approach (Kulldorff et al., 2009) is to consider potential clusters containing at most 50% of the observations. Indeed, as stated by Kulldorff et al. (2009), when a cluster contains more than 50% of the observations, it is more appropriate to consider the remaining observations as a spatially disconnected cluster. It should be noted, however, that it is possible to set a limit lower than 50% of the observations, based on expert knowledge. For example, in environmental data such as pollution data, the clusters detected must be relatively small because of the local nature of the pollutants. Note however that this biases the statistical inference and that it is preferable to use an *a posteriori* filtering on the size of the detected clusters. More details on this subject are given in Chapter 6. Many authors consider the circular scanning window of variable size introduced by Kulldorff (1997). In this context, the set of potential clusters $\mathcal{W}$ associated with n spatial units $s_1, \ldots, s_n$ in which we observe a realization of a random variable X is defined as the set of discs centered on a location and passing through another location:

$$\mathcal{W} = \left\{ w_{i,j} \text{ such that } |w_{i,j}| \leq \frac{n}{2} \right\}$$

where $w_{i,j}$ is the disc centered on s_i that passes through s_j and $|w_{i,j}|$ is the number of observations in $w_{i,j}$ (Figure 2.16). However, others proposed different shapes of clusters such as elliptical clusters (Kulldorff et al., 2006), or clusters of arbitrary shape (Tango and Takahashi, 2005; Assunção et al., 2006; Dematteï et al., 2007).

Concentration index.   Kulldorff and Nagarwalla (1995) proposed to compute a concentration index $C^{(w)}$ (a likelihood ratio) to compare the distribution of the observations in each potential
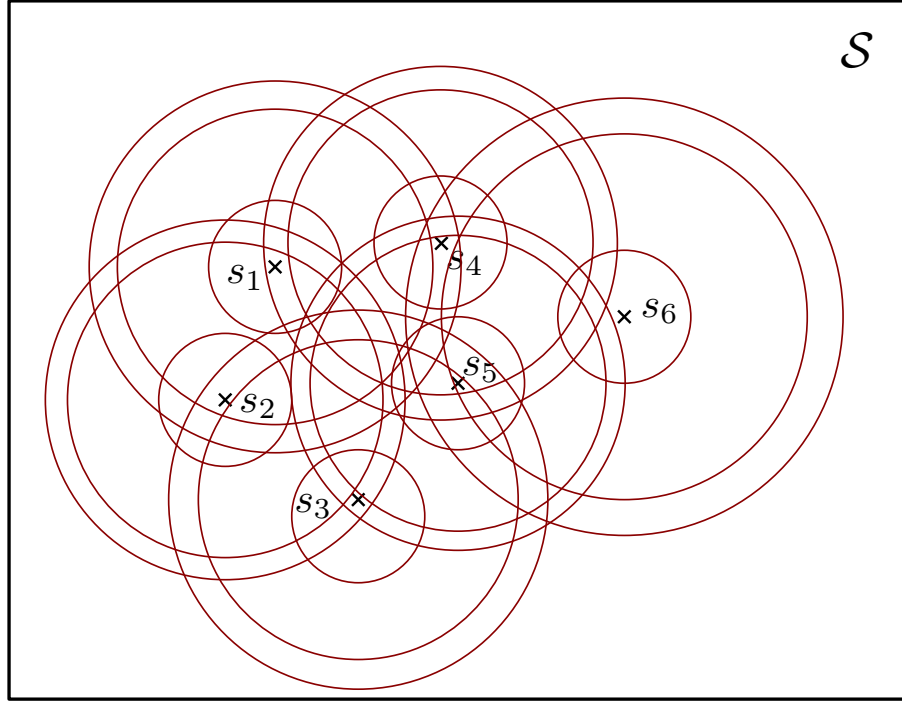
Figure 2.16: Example of circular potential clusters containing between 1 and 50% of the six spatial locations

cluster w with their distribution outside (in w^c). Later on, other concentration indices have been proposed: Some authors consider parametric concentration indexes while others use nonparametric concentration indexes instead. A review of the literature will be given in the following sections.

Then, the spatial scan statistic Λ is defined as the maximum of $C^{(w)}$ on the set of potential clusters $\mathcal{W}$:

$$\Lambda = \max_{w \in \mathcal{W}} C^{(w)}.$$

The most likely cluster (MLC) is defined as the cluster maximizing the concentration index among the potential clusters of $\mathcal{W}$:

$$\mathrm{MLC} = \arg\max_{w \in \mathcal{W}} C^{(w)}.$$

Determining the statistical significance of the MLC. Once the MLC is detected, its statistical significance still needs to be determined. A major issue in spatial scan statistics is that we can rarely obtain the distribution of the test statistic Λ under $\mathcal{H}_0$ because of the overlap between the different potential clusters. Thus, this inference step is often achieved (Kulldorff, 1997; Kulldorff et al., 2009; Cucala et al., 2017) through a Monte-Carlo approach (Dwass, 1957): the idea is either to permute M times the observations in the different spatial units (when the spatial distribution of the data is unknown), or to generate M data sets under the null hypothesis $\mathcal{H}_0$ otherwise. Then for each data set obtained m, the spatial scan statistic $\Lambda^{(m)}$ can be computed. Finally, the p-value associated with Λ can be estimated by

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

This approach can be very costly from a computational point of view, especially if the precision required for the estimation of the p-value associated with the cluster (and therefore the number of Monte-Carlo simulations) is high, since the computation time is multiplied by $M + 1$.

Alternatives to the Monte-Carlo method have been proposed. For example, Abrams et al. (2010) proposed a method for estimating the p-value from replications under $\mathcal{H}_0$, which is more accurate than the Monte-Carlo method for the same number of replications. Briefly, the approach consists in fitting a Gumbel distribution on the spatial scan statistics obtained on the replications under $\mathcal{H}_0$. This allows to approximate the distribution of the spatial scan statistic under $\mathcal{H}_0$ and to deduce an estimate of the p-value. The advantage of this method is its accuracy which allows to decrease the number of replications under $\mathcal{H}_0$ and thus to reduce the computation time. Finally, Bayesian spatial scan statistics exist and do not use a Monte-Carlo approach.

Secondary clusters. It is often interesting to detect not only the most likely cluster, but also secondary clusters, i.e., statistically significant clusters co-existing with the MLC. Kulldorff and Nagarwalla (1995) proposed to define a secondary cluster as a cluster that also presents a high concentration index and does not overlap with the MLC. The statistical significance of this cluster is calculated as if the cluster was itself the MLC, i.e., the associated p-value is estimated by comparing the value of its concentration index to the value of the spatial scan statistic obtained on each permuted data set or on each data set generated under $\mathcal{H}_0$. This approach is conservative since the concentration index of the secondary clusters is compared with the maximum of the concentration indexes of each data set under $\mathcal{H}_0$ which implies that the secondary clusters must reject $\mathcal{H}_0$ "by their own strength".

Zhang et al. (2010) proposed an alternative to the conservative approach of Kulldorff and Nagarwalla (1995) with a sequential approach. The latter consists, once the MLC is detected and declared statistically significant, in removing the data from the MLC (as if there were no spatial units there) and in determining the MLC on the new data thus obtained. If it is statistically significant it will be considered as a secondary cluster and the approach is repeated until no cluster is statistically significant. Other approaches were proposed (Li et al., 2011; Xu and Gangnon, 2016; Lin et al., 2016).

Spatial scan statistics for univariate data

A first spatial scan statistic was proposed by Kulldorff and Nagarwalla (1995) for count data. The approach, based on a likelihood ratio and a Bernoulli model was then extended to the Poisson model by Kulldorff (1997).

Bernoulli and Poisson models. In the Bernoulli model (Kulldorff and Nagarwalla, 1995) the event is modeled by a Bernoulli distribution with parameter $0 < p < 1$. For example, one can consider a case-control study in a sample of the population in which one seeks to detect clusters in which the frequency of cases is abnormally high.
Let X be the associated spatial counting process. Then for $B \subset S$, $X(B)$ counts the number of events in B: $X(B) \sim \mathcal{B}(\mu(B), p)$ where $\mu(B)$ corresponds to the at-risk population in B.
Then, in the context of spatial scan statistics the null hypothesis $\mathcal{H}_0$ (absence of cluster) and the alternative hypothesis associated with a potential cluster w ($\mathcal{H}_1^{(w)}$) can be defined as $\mathcal{H}_0 : \forall i, s_i \in S, X(s_i) \sim \mathcal{B}(\mu(s_i), p_S)$ and $\mathcal{H}_1^{(w)} : \forall i, s_i \in w, X(s_i) \sim \mathcal{B}(\mu(s_i), p_w)$ and $\forall i, s_i \in w^{\mathsf{c}}, X(s_i) \sim \mathcal{B}(\mu(s_i), p_{w^{\mathsf{c}}}), p_w \neq p_{w^{\mathsf{c}}}$.

Then the concentration index proposed by Kulldorff and Nagarwalla (1995) is $C_{\mathcal{B}}^{(w)} = \dfrac{\mathcal{L}_{\mathcal{H}_1^{(w)}}}{\mathcal{L}_{\mathcal{H}_0}}$ where $\mathcal{L}_{\mathcal{H}_1^{(w)}}$ and $\mathcal{L}_{\mathcal{H}_0}$ correspond respectively to the likelihood under $\mathcal{H}_1^{(w)}$ and under $\mathcal{H}_0$. Thus

we can show that the spatial scan statistic $\Lambda_{\mathcal{B}}$ is

$$\Lambda_{\mathcal{B}} = \max_{w \in \mathcal{W}} \frac{\left(\frac{X(w)}{\mu(w)}\right)^{X(w)} \left(1 - \frac{X(w)}{\mu(w)}\right)^{\mu(w)-X(w)} \left(\frac{X(w^c)}{\mu(w^c)}\right)^{X(w^c)} \left(1 - \frac{X(w^c)}{\mu(w^c)}\right)^{\mu(w^c)-X(w^c)}}{\left(\frac{X(S)}{\mu(S)}\right)^{X(S)} \left(1 - \frac{X(S)}{\mu(S)}\right)^{\mu(S)-X(S)}}.$$

The Poisson model (Kulldorff, 1997) considers a Poisson distribution on $X(B)$: $X(B) \sim \mathcal{P}(p\mu(B))$ where $\mu(B)$ corresponds to the at-risk population in B. This model is particularly adapted to studies on disease registers (exhaustive studies) in which one focuses on detecting clusters of over-incidence of a disease by adjusting on the underlying population.

In this context the null hypothesis $\mathcal{H}_0$ is $\mathcal{H}_0 : \forall i, s_i \in S, X(s_i) \sim \mathcal{P}(p_S \mu(s_i))$ and the alternative hypothesis $\mathcal{H}_1^{(w)}$ associated with the potential cluster w is $\mathcal{H}_1^{(w)} : \forall i, s_i \in w, X(s_i) \sim \mathcal{P}(\mu(s_i)p_w)$ and $\forall i, s_i \in w^c, X(s_i) \sim \mathcal{P}(\mu(s_i)p_{w^c}), p_w \neq p_{w^c}$. Then the concentration index proposed by Kulldorff (1997) is $C_{\mathcal{P}}^{(w)} = \frac{\mathcal{L}_{\mathcal{H}_1^{(w)}}}{\mathcal{L}_{\mathcal{H}_0}}$ and the spatial scan statistic $\Lambda_{\mathcal{P}}$ is

$$\Lambda_{\mathcal{P}} = \max_{w \in \mathcal{W}} \frac{\left(\frac{X(w)}{\mu(w)}\right)^{X(w)} \left(\frac{X(w^c)}{\mu(w^c)}\right)^{X(w^c)}}{\left(\frac{X(S)}{\mu(S)}\right)^{X(S)}}.$$

Since the work of Kulldorff and Nagarwalla (1995) and Kulldorff (1997), many authors have proposed to develop spatial scan statistics methods for other distribution models in frequentist or Bayesian frameworks, or even nonparametric spatial scan statistics. This is the subject of the following paragraphs.

Other model-based spatial scan statistics. In the case of continuous data, Kulldorff et al. (2009) proposed a spatial scan statistic based on a Gaussian model. Here $\mathcal{H}_0$ can be defined as $\mathcal{H}_0$: $\forall w \in \mathcal{W}, \mu_X(w) = \mu_X(w^c)$ and the alternative hypothesis associated with the potential cluster w is $\mathcal{H}_1^{(w)} : \mu_X(w) \neq \mu_X(w^c)$, where $\mu_X(w)$ and $\mu_X(w^c)$ correspond respectively to the mean of X in w and outside w, $|w|$ and $|w^c|$ are the number of spatial units in w and outside w.

We assume $\forall i \in [\![1, n]\!], X_i \sim \mathcal{N}(\alpha_i, \tau_i^2)$ where X_i is the realization of X in the spatial location s_i. Thus under $\mathcal{H}_0, \alpha_i = \mu$ and $\tau_i^2 = \sigma^2$, and under $\mathcal{H}_1^{(w)}, \alpha_i = \begin{cases} \mu_w & \text{if } s_i \in w \\ \mu_{w^c} & \text{otherwise} \end{cases} (\mu_w \neq \mu_{w^c})$ and $\tau_i^2 = \sigma_w^2$.

Then Kulldorff et al. (2009) defined the spatial scan statistic for Gaussian data as the maximum of the likelihood ratio on the set of potential clusters $\mathcal{W}$ and used it to detect clusters of abnormally low birth weight in New York City. Huang et al. (2009) developed a similar approach with heteroskedasticity: $X_i \sim \mathcal{N}(\alpha_i, \tau_i^2)$ where under $\mathcal{H}_0$, $\alpha_i = \mu$ and $\tau_i^2 = \dfrac{\sigma^2}{\delta_i}$ and under $\mathcal{H}_1^{(w)}$, $\alpha_i = \begin{cases} \mu_w & \text{if } s_i \in w \\ \mu_{w^c} & \text{otherwise} \end{cases} (\mu_w \neq \mu_{w^c})$ and $\tau_i^2 = \dfrac{\sigma_w^2}{\delta_i}$ where δ_i is a known weight. This approach presents the advantage of taking into account the heteroskedasticity of the data that is naturally present when the data are aggregated by the mean within each spatial unit s_i from n_i unobserved individual data: a suitable weighting is then $\delta_i = n_i$.

Note that both approaches are respectively equivalent to the homoskedastic and heteroskedastic approaches proposed by Cucala (2014).

Other distribution models were studied. For example, Jung et al. (2007) and Jung et al. (2010) respectively developed spatial scan statistics for ordinal and multinomial data. The first one

must be considered to detect clusters on ordered categorical data, e.g., to detect clusters in which the most severe stages of a cancer are statistically more numerous. On the contrary, the spatial scan statistic for multinomial data is devoted to unordered categorical data, e.g., to detect clusters in which the variants of a disease are distributed differently from elsewhere. Other researchers also considered a compound Poisson distribution (Rosychuk et al., 2006) to detect clusters when the individuals may experience the event of interest several times, or a Quasi-Poisson model (Zhang et al., 2012) allowing to take into account overdispersion.

In the case of rare diseases, models for data counting with excess zeros are particularly relevant. Thus Cançado et al. (2014) proposed a spatial scan statistic based on a zero-inflated Poisson model (ZIP). Then, de Lima et al. (2015) proposed an extension of this model by considering a ZIP model taking into account a possible overdispersion that would persist in strictly positive values.

The approaches presented above have all been developed in the frequentist framework. In the Bayesian framework, Gangnon and Clayton (2003) proposed a hierarchical scan statistic based on a Poisson distribution. This approach allows to get the posterior distribution of the number of clusters as well as an estimated probability that a spatial unit belongs to a cluster. Moreover, Neill et al. (2005) noticed that the approach of Kulldorff (1997) presents a problem in case of misspecification and is expensive in computation time which can make it inapplicable in practice for very large data sets. He proposed a Bayesian version *via* a Gamma-Poisson model. A major difference between the approach of Gangnon and Clayton (2003) and that of Neill et al. (2005) is that Gangnon and Clayton (2003) supposed a prior distribution of the clusters such that some clusters are more likely than others whereas Neill et al. (2005) assumed that all clusters have the same probability.

In the same way, Cançado et al. (2017) proposed a Bayesian version to the ZIP spatial scan statistic with a Bayesian zero-inflated binomial model. Nonparametric approaches for spatial scan statistics have also been proposed, this is the subject of the following paragraph.

Nonparametric spatial scan statistics. In addition to these parametric methods, nonparametric spatial scan statistics have also been proposed for univariate data. Here the null hypothesis can be written as $\mathcal{H}_0$: $\forall w \in \mathcal{W}, F_w = F_{w^c}$ (absence of cluster) and the alternative hypothesis is $\mathcal{H}_1$: exists w such that $F_w(x) = F_{w^c}(x - \Delta), \Delta \neq 0$ where F_w and F_{w^c} are the cumulative distribution functions of X in w and w^c respectively. When $\Delta > 0$, X takes higher values in w than in w^c whereas when $\Delta < 0$, X takes lower values in w than in w^c (Figure 2.17). Jung and Cho (2015) proposed to consider a Wilcoxon-Mann-Whitney test statistic. For this purpose, a value R_i corresponding to the rank of X_i in $(X_1, \ldots, X_n)$ is assigned to each spatial location s_i: $R_i = \sum_{j=1}^{n} \mathbb{1}_{X_j \leq X_i}$. By noting $W_w = \sum_{i, s_i \in w} R_i$ and $T_w = \dfrac{W_w - \mathbb{E}_{\mathcal{H}_0}(W_w)}{\sqrt{\mathbb{V}_{\mathcal{H}_0}(W_w)}}$, we have $T_w \underset{\mathcal{H}_0}{\approx} \mathcal{N}(0, 1)$ for $|w| \geq 10$ and $|w^c| \geq 10$.

Then the authors proposed to consider $1 - \Phi(T_w)$ as a concentration index if $\Delta > 0$, and $\Phi(T_w)$ if $\Delta < 0$, where Φ is the cumulative distribution function of a $\mathcal{N}(0, 1)$, that is to consider the p-value associated with T_w: $p^{(w)}$ as a concentration index. When $|w| < 10$ or $|w^c| < 10$, they proposed to use an exact computation of the p-value $p^{(w)}$ (Lehmann and D'Abrera, 2006).

Then the nonparametric spatial scan statistic for univariate data proposed by Jung and Cho (2015) is $\Lambda = \min_{w \in \mathcal{W}} p^{(w)}$.

It remains to compute $\mathbb{E}_{\mathcal{H}_0}(W_w)$ and $\mathbb{V}_{\mathcal{H}_0}(W_w)$. However it is easy to show that $\mathbb{E}_{\mathcal{H}_0}(W_w) = \dfrac{|w|(n+1)}{2}$ and $\mathbb{V}_{\mathcal{H}_0}(W_w) = |w||w^c|\dfrac{n+1}{12}$.

Note that Cucala (2016) proposed a similar approach by considering more simply the statistic
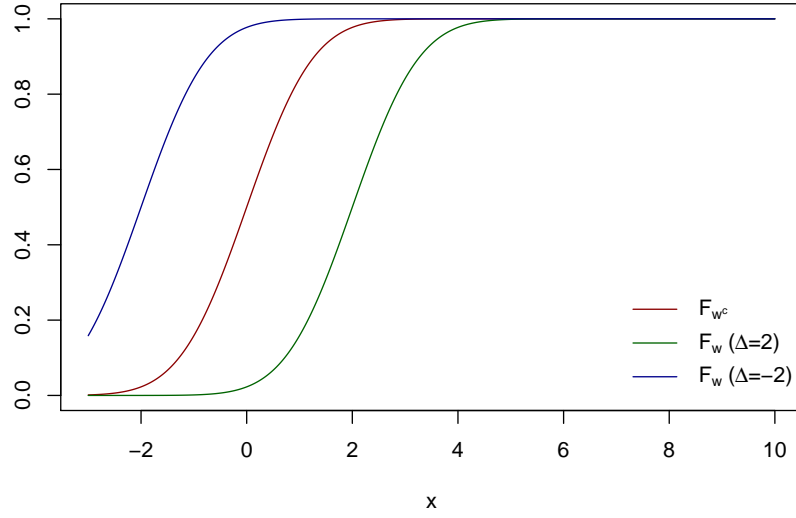
Figure 2.17: F_w and F_{w^c} for $\Delta = 2$ and $\Delta = -2$ where F_{w^c} is the cumulative distribution function of a $\mathcal{N}(0,1)$

$$\max_{w \in \mathcal{W}} |T_w|.$$

Spatial scan statistics for high-dimensional and survival data

In practice it happens that several variables are measured in the spatial units. Typically, in the context of environmental surveillance, researchers seek to determine if there exist abnormally polluted areas (with several pollutants). Applying the previous approaches on each variable separately allows the detection of clusters characterized by a single variable, but does not allow the detection of clusters by taking into account the information available in all the variables. In this context Kulldorff et al. (2007) developed a spatial scan statistic for multivariate data. However he supposed that the variables are independent which simplifies the log-likelihood in a sum of log-likelihoods. This assumption is rather simplistic and rarely verified in practice. For example, in environmental science researchers are interested in detecting spatial clusters of pollutants, and it is well known that these are often correlated (Kumar and Joseph, 2006; Wu et al., 2016). Thus Cucala et al. (2017) proposed a spatial scan statistic for multivariate data based on a multivariate Gaussian distribution allowing to take into account the potential correlations between the variables. Cucala et al. (2019) also proposed a nonparametric rank-based spatial scan statistic for multivariate data based on a Wilcoxon-Mann-Whitney test statistic for multivariate data (Oja and Randles, 2004).

With the advance of sensing and data storage capacity, data are increasingly being measured continuously over time. This led to the introduction of functional data and functional data analysis (see Section 1.1 for more details). In domains where data presents a spatial component, the emergence of functional data has naturally given rise to spatial functional data, characterized by the observation of one or more curves in each spatial location. In this context, Smida et al. (2022) proposed a first nonparametric spatial scan statistic for univariate functional data. By denoting P_w and P_{w^c} the probability measures of X respectively in a potential cluster w and outside w, the authors proposed to test $\mathcal{H}_0 : \forall w \in \mathcal{W}, P_w = P_{w^c}$ (absence of cluster) against the alternative hypothesis $\mathcal{H}_1$: exists w such that P_w and P_{w^c} differ by a shift $\Delta_w \neq 0$. For this purpose, they considered the Wilcoxon-Mann-Whitney test

statistic for high dimensional data proposed by Chakraborty and Chaudhuri (2014):

$$T_w = \frac{1}{\sqrt{|w||w^c|n}} \sum_{i,s_i \in w} \sum_{j,s_j \in w^c} \frac{X_j - X_i}{||X_j - X_i||}.$$

More precisely, they used $||T_w||$ as a concentration index to define the nonparametric spatial scan statistic for univariate functional data as $\Lambda = \max_{w \in \mathcal{W}} ||T_w||$.

However, to our knowledge no parametric spatial scan statistics method has been developed in this context. Thus, this thesis will address methodological developments in this framework.

In the context of spatial epidemiology of survival data, the detection of spatial clusters in which survival times are longer or shorter than elsewhere allows researchers to make hypotheses about risk factors or, on the contrary, protective factors associated with the survival. The identification of these clusters also allows targeted actions to be carried out in an attempt to improve survival. In this context, several methods of spatial scan statistics have been developed. For example, Huang et al. (2007) and Bhatt and Tiwari (2014) proposed parametric spatial scan statistics based on exponential and Weibull distributions respectively. Then Bhatt and Tiwari (2016) extended these models to any density function for the survival time of the form

$$f(t; \gamma, a, b, c) = \frac{c t^{ac-1} \exp\left[-\dfrac{t^c}{\gamma^b}\right]}{\gamma^{ab}\Gamma(a)}, t > 0$$

where $a, b, c > 0$ are known and $\gamma > 0$ is unknown. In particular this model allows to consider exponential, Weibull and Rayleigh distributions. Finally, Usman and Rosychuk (2018) proposed a log-Weibull model. A first method based on a semi-parametric Cox model has also been proposed (Cook et al., 2007).

Note that contrary to the other spatial scan statistics, in these methods the observations are individual and each individual is associated with a spatial location (for example, his place of residence or a less precise location for privacy reasons; see Section 2.2 for more details).

The previously mentioned approaches of spatial scan statistics for survival data all assume the independence of the observations. However, as explained in Section 2.2, on the one hand the survival times of the individuals in the same spatial unit can be correlated, and on the other hand spatial units can be spatially dependent. Recently, several authors have investigated the impact of not taking spatial correlation into account in spatial scan statistics. They have shown that this leads to an increase in the type I error (more details are given in the next section). This thesis will therefore develop methodologies to take into account both the intra-spatial unit correlations and the spatial dependence between the spatial units for spatial cluster detection of survival data.

Taking into account the spatial autocorrelation

Spatial scan statistics classically assume that the observations are independent. However, this assumption is somewhat simplistic in reality since the spatial nature of the data naturally leads to a potential spatial autocorrelation as specified by the first law of geography stated by Tobler (1969): "Everything is related to everything else. But near things are more related than distant things". Thus Loh and Zhu (2007) showed that the presence of spatial correlation leads to an increase of the type I error and proposed a spatial scan statistic based on a spatial generalized linear mixed model to take into account the possible spatial correlation. For his part, Lin (2014) proposed a method based on a quasi-likelihood, allowing the detection of clusters by simultaneously adjusting for covariates and a possible spatial correlation. Lee et al. (2019) also considered this question and considered that the geographical area could be divided

into K regions containing several spatial units. He then proposed to decompose the variable of interest into a mean term, an error term specific to the region and an error term specific to the spatial unit. Finally, Ahmed et al. (2021b) proposed to consider an autoregressive spatial model (SAR) on the variable of interest to perform a filtering in order to make the variable independent, and Rogerson (2021) studied the impact of spatial dependence on the power and type I error, and the effect of a "pre-whitening" step using the APLE estimator proposed by Li et al. (2007, 2012) and an autoregressive spatial model.

Note that the model of Lee et al. (2019) allows to consider a spatial dependence between the spatial units of a same region but not between the K regions, while the methods proposed by Loh and Zhu (2007); Lin (2014); Rogerson (2021); Ahmed et al. (2021b) allow to model the spatial dependence between all the spatial units.

The interested reader can find a more complete review of the literature on spatial scan statistics in Abolhassani and Prates (2021). We inform the reader that in the following chapters, for the purpose of clarity, some concepts may be repeated.

The following chapters develop new approaches to spatial scan statistics, both in the context of functional data, but also for survival data.

First, since no parametric spatial scan statistic was developed for univariate functional data, we propose two new spatial scan statistics for univariate functional data in Chapter 3. In the context of multivariate functional data, it seems that the approach proposed by Smida et al. (2022) can be easily adapted. However, this has never been investigated in this framework. Thus in Chapter 4, we study the performances of the adaptation of this approach to multivariate functional data and we propose three new spatial scan statistics in this context. Then, Chapter 6 presents the R package **HDSpatialScan** developed to allow the easy use of these spatial scan statistics.

In the context of spatial survival data, our contribution (Chapter 5) consists in developing a new spatial scan statistic based on a Cox model with a spatially structured shared frailty allowing to model the possible dependencies, both between individuals of the same spatial unit and between spatial units.

Chapter 3

Spatial scan statistics for univariate functional data

Contents

1   Introduction

Technological advances have resulted in data collection in large volumes over time. This led to the introduction of functional data analysis (FDA) by Ramsay and Silverman (2005b). In domains in which data naturally present a spatial component, the emergence of functional data thus led to the development of spatial functional data (Fernández de Castro and Manteiga, 2008; Delicado et al., 2010; Menafoglio et al., 2013). In this framework, data interpolation methods (such as kriging (Giraldo et al., 2010; Bohorquez et al., 2016) or cokriging (Monestiez et al., 2008)), regression (Ternynck, 2014; Zhang et al., 2016; Ahmed et al., 2021a), or clustering (Giraldo et al., 2012; Vandewalle et al., 2022) approaches have been proposed.

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, leading to a sharp loss of power and a significant increase of false-positive rate (Cucala et al., 2017, 2019). Recently, Smida et al. (2022) developed a nonparametric approach based on a functional Wilcoxon-Mann-Whitney test statistic. In a simulation study they showed the efficiency of their method in comparison with a univariate approach. However, to the best of our knowledge, no formal definition of a spatial cluster in the context of functional data has been proposed. Moreover, 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 propose definitions of spatial clusters in the context of spatial functional data. Then, 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 those of the method of Smida et al. (2022). Our methods' application to a real data set is presented in Section 4. Lastly, the results of this work are discussed in Section 5.

Note that the developments presented in this chapter have been published in *Spatial Statistics* (December 2021) in collaboration with Mohamed-Salem Ahmed (University of Lille), Matthieu Marbac (University of Rennes, ENSAI) and Michaël Genin (University of Lille). Moreover, they are implemented in the package **HDSpatialScan** available in the CRAN repository (Frévent et al., 2022).

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

$$\mathbb{E}[X_i \mid s_i \in w] = \mathbb{E}[X_i \mid s_i \notin w] + \Delta \text{ where } \Delta \neq 0.$$

In the same way, in the multivariate case ($X \in \mathbb{R}^p$, $p \geq 2$), 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 $\Delta = (\Delta_1, \ldots, \Delta_p)^\top \neq 0$, and for all $j \in [\![1, p]\!]$, $\Delta_j \geq 0$ or for all $j \in [\![1, p]\!]$, $\Delta_j \leq 0$.
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 and 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:

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

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:

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

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, we shall focus on variable-size circular clusters. Hence in line with Kulldorff's work (Kulldorff, 1997), the set of potential circular clusters $\mathcal{W}$ can be defined by:

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

where $w_{i,j}$ is the disc centered on s_i that passes through s_j: a cluster cannot cover more than 50% of the studied region. $|w_{i,j}|$ denotes the number of sites in $w_{i,j}$.
It should be noted that researchers have suggested many other shapes, such as elliptical clusters (Kulldorff et al., 2006), rectangular clusters (Chen and 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 $\mathcal{L}^2(\mathcal{T}, \mathbb{R})$ of real valued, square-integrable functions on $\mathcal{T}$.
Cuevas et al. (2004) and Górecki and Smaga (2015) adapted the classical ANOVA F-statistic for $\mathcal{L}^2$ processes. Without loss of generality, and considering two independent samples of trajectories drawn from two $\mathcal{L}^2$ 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:

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

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, $F_n^{(w)}$ 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):

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

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

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

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 $\mathcal{H}_0 : \forall w \in \mathcal{W}, \mu_w(t) = \mu_{w^c}(t) = \mu_S(t)$ (where $\mu_g(t) = \mathbb{E}[X_g(t)]$, $g \in \{w, w^c\}$) 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 $\bar{X}_g(t) = \frac{1}{|g|} \sum_{i,s_i \in g} X_i(t)$ are empirical estimators of μ_g ($g \in \{w, w^c\}$) and

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

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, 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):

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

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

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

2.4   Computing the statistical 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 data sets 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 data set and $\Lambda^{(1)}, \ldots, \Lambda^{(M)}$ be the scan statistics observed in the simulated data sets. According to Dwass (1957) the p-value for Λ with real data is estimated by

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

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. (2022).

As stated by Smida et al. (2022), 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. (2022) 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 Appendix A.

3.1   Design of the simulation study

Artificial data sets were generated by using the geographic locations of the 94 French *départements* (county-type administrative areas; Figure 3.1). The location of the i-th *département* was defined by its centroid s_i, $i \in [\![1, 94]\!]$. A cluster w (composed of eight *départements* in the Paris region; the red area, see Figure 3.1) was defined and simulated for each artificial data set.

3.1.1   Generation of the artificial data sets

At each location s_i ($i \in [\![1, 94]\!]$), the artificial data sets X_i were generated using the following model (see Qiu et al., 2021, for more details) and measured at 101 equally spaced times on
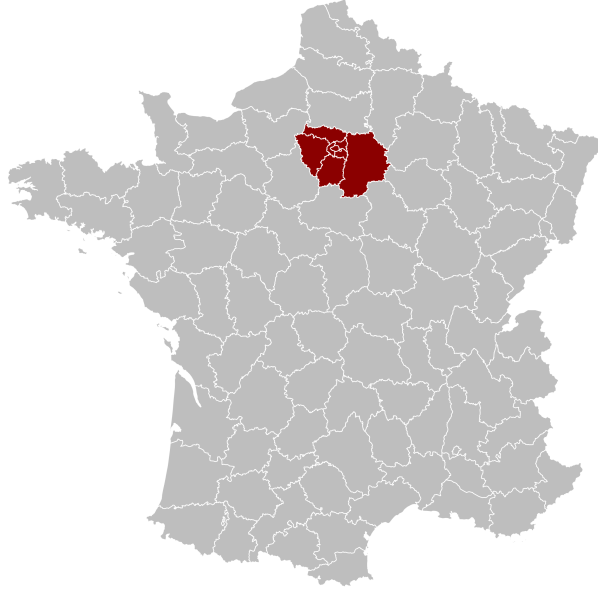
Figure 3.1: The 94 French *départements* and the true cluster (in red) simulated for each artificial data set.

$[0, 1]$:

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

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

$$\Psi_k(t) = \begin{cases} 1 & \text{if } k = 1 \\ \sqrt{2}\sin\left[k\pi t\right] & \text{if } k \text{ even} \\ \sqrt{2}\cos\left[(k-1)\pi t\right] & \text{if } k \text{ odd and } k > 1 \end{cases}, \text{ and } b_i(t) \sim \mathcal{N}(0, 0.2^2).$$

Four distributions for the $v_{i,j,k}$, $j = 1, 2$ were considered: (i) $v_{i,j,k} \sim \mathcal{N}(0, 1)$, (ii) $v_{i,j,k} \sim t(4)/\sqrt{2}$, (iii) $v_{i,j,k} \sim (\chi^2(4) - 4)/(2\sqrt{2})$ and (iv) $v_{i,j,k} \sim (\mathcal{E}(0.5) - 2)/2$.

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 (characterized by α) 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 presented in Figure 3.2. Since the data had been generated with noise $b_i(t)$, we applied a smoothing method with a cubic B-spline basis (Figure 3.3).

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 data sets. 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 respective performances of the methods 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 data sets yielding to the rejection of $\mathcal{H}_0$, the true positive
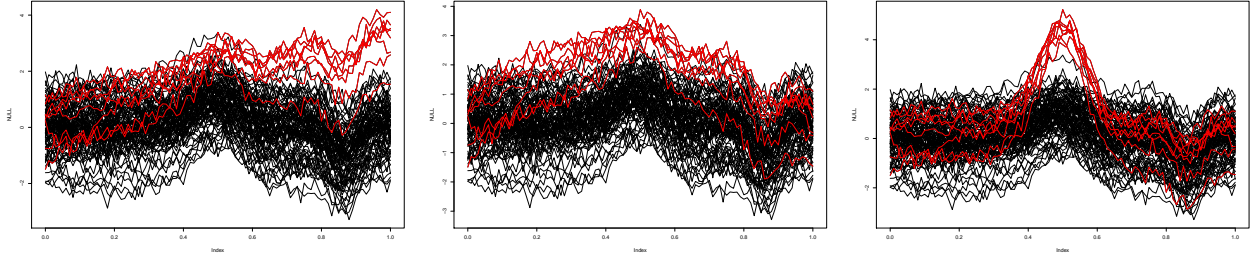
Figure 3.2: The simulation study: an example of the data generated for the Gaussian process before smoothing, with $\Delta(t) = \Delta_1(t) = 3t$ (left panel), $\Delta(t) = \Delta_2(t) = 8t(1-t)$ (middle panel) and $\Delta(t) = \Delta_3(t) = 10\exp\left[-100(t-0.5)^2\right]/3$ (right panel). The red curves correspond to the observations in the cluster.
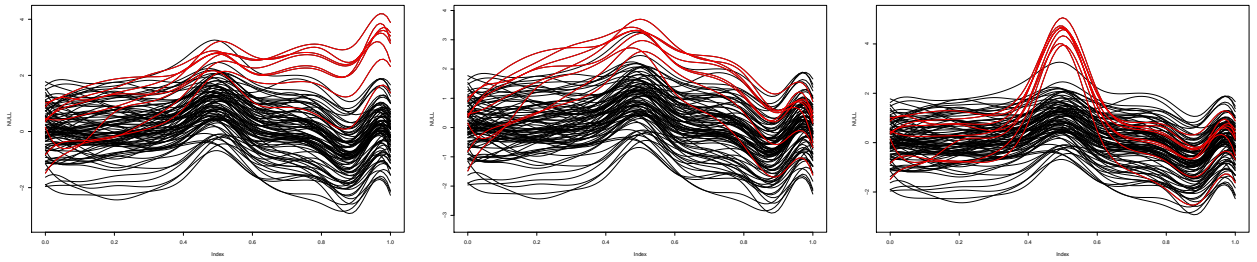
Figure 3.3: The simulation study: an example of the data generated for the Gaussian process after smoothing, with $\Delta(t) = \Delta_1(t) = 3t$ (left panel), $\Delta(t) = \Delta_2(t) = 8t(1-t)$ (middle panel) and $\Delta(t) = \Delta_3(t) = 10\exp\left[-100(t-0.5)^2\right]/3$ (right panel). The red curves correspond to the observations in the cluster.

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.

3.2   The results of the simulation study

The results of the simulation are shown in Figures 3.4, 3.5 and 3.6.

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 Figures 3.4, 3.5 and 3.6).

Smoothing the data slightly improves the performance of the pointwise approach (DFFSS).

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. 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, a Chi-square or an Exponential 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 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 3.7, 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 3.7, 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.

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. (2022) for details) and that did not cover the MLC (Kulldorff, 1997). For all methods, the set of potential clusters $\mathcal{W}$ is the one described in Section 2. After smoothing the data with a cubic B-spline basis, 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 PFSS, the DFFSS and the NPFSS methods each detected two statistically significant spatial clusters (Figure 3.8 and Table 3.1). The PFSS and the DFFSS identified the same two statistically significant spatial clusters of elevated unemployment rates: the MLC (7 *départements*, $\hat{p}_{\text{DFFSS}} = 0.005$, $\hat{p}_{\text{PFSS}} = 0.011$) was located in south-eastern France and the secondary cluster (9 *départements*, $\hat{p}_{\text{DFFSS}} = 0.027$, $\hat{p}_{\text{PFSS}} = 0.039$) 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 statistically significant large (40-*départements*) MLC ($\hat{p}_{\text{NPFSS}} = 0.013$) with low unemployment rates in the center of France. The unemployment rate curves were heterogeneous, and the MLC included some *départements* with an unemployment rate above the national average (Figure 3.8). The secondary cluster (9 *départements*, $\hat{p}_{\text{NPFSS}} = 0.039$) detected by the NPFSS was exactly the same as that detected by the DFFSS.

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
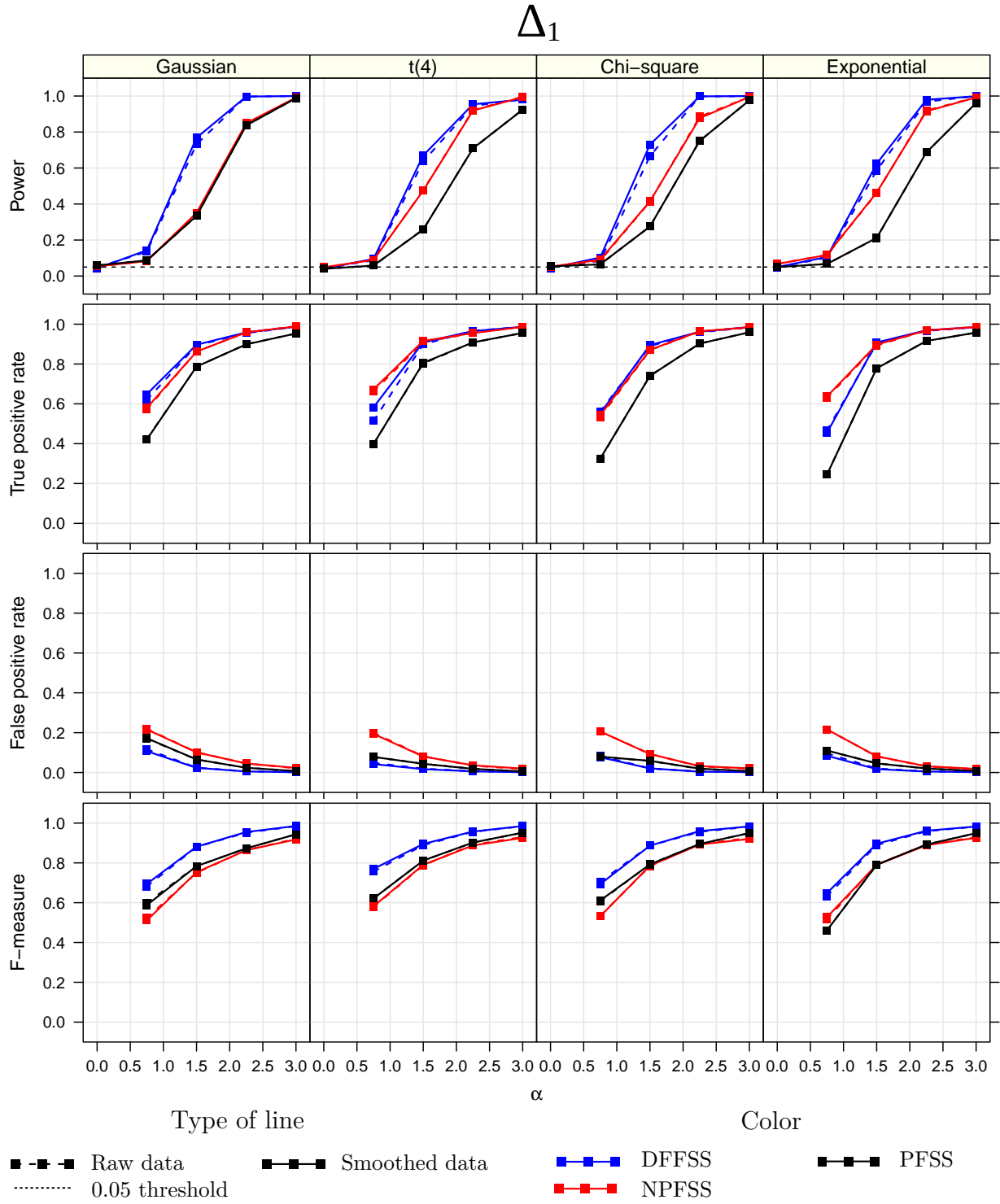
Figure 3.4: The simulation study: comparison of the NPFSS, PFSS and DFFSS methods for the shift $\Delta_1(t) = \alpha t$. For each method, the power curves, the true-positive and false-positive rates, and the F-measure values for detection of the true cluster as the MLC are shown. α is the parameter that controls the cluster intensity.
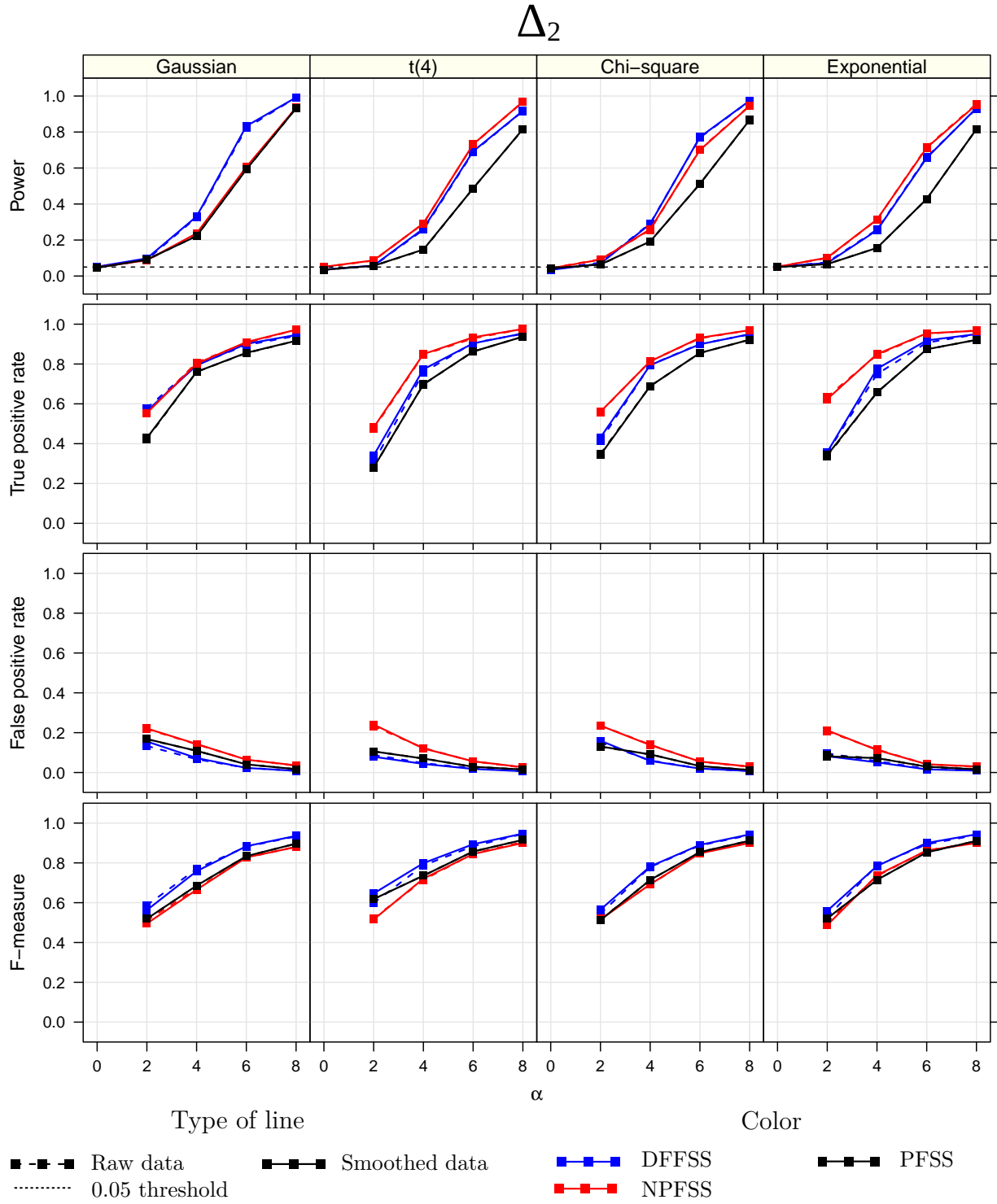
Figure 3.5: The simulation study: comparison of the NPFSS, PFSS and DFFSS methods for the shift $\Delta_2(t) = \alpha t(1-t)$. For each method, the power curves, the true-positive and false-positive rates, and the F-measure values for detection of the true cluster as the MLC are shown. α is the parameter that controls the cluster intensity.
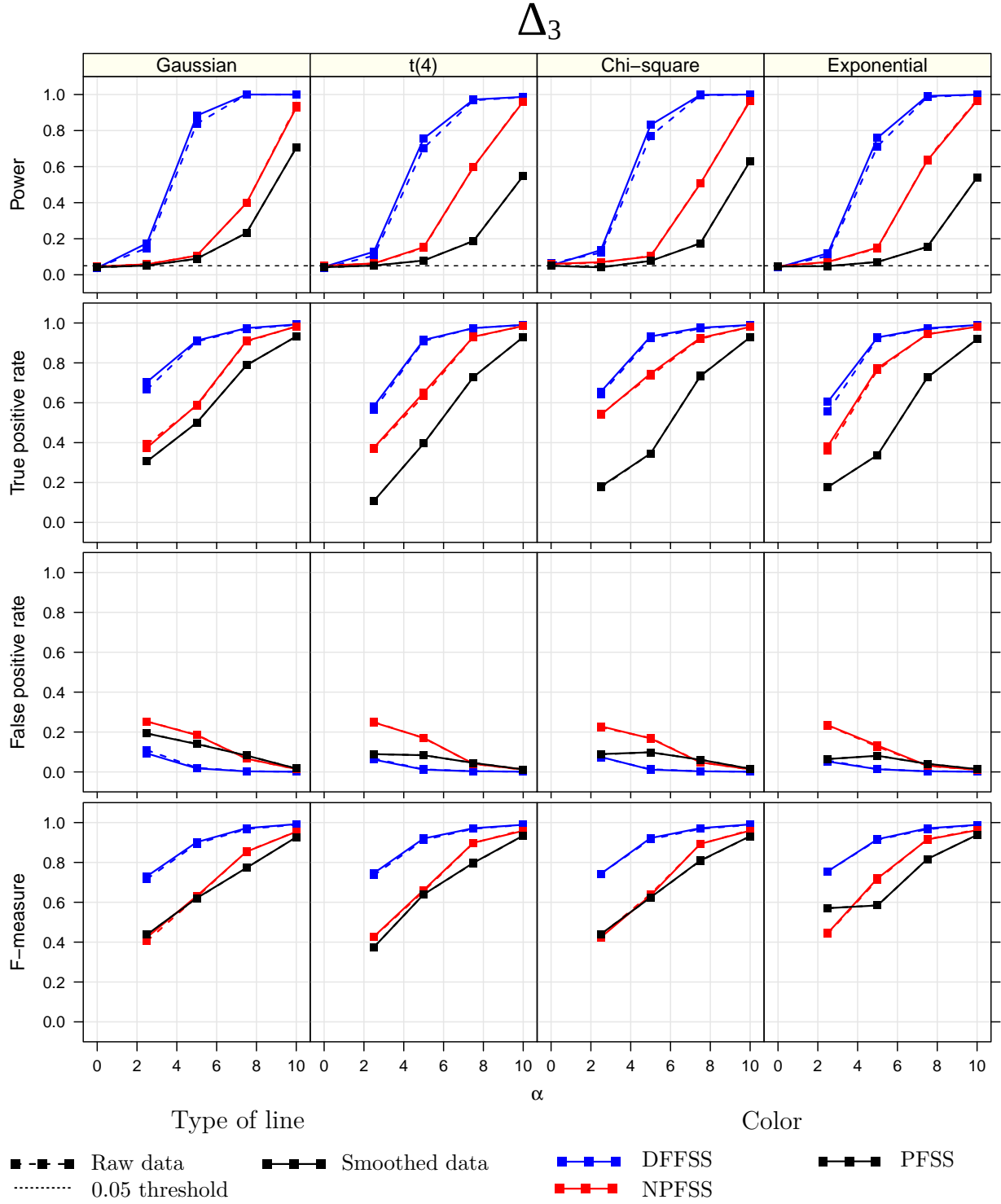
Figure 3.6: The simulation study: comparison of the NPFSS, PFSS and DFFSS methods for the shift $\Delta_3(t) = \alpha \exp\left[-100(t - 0.5)^2\right]/3$. For each method, the power curves, the true-positive and false-positive rates, and the F-measure values for detection of the true cluster as the MLC are shown. α is the parameter that controls the cluster intensity.
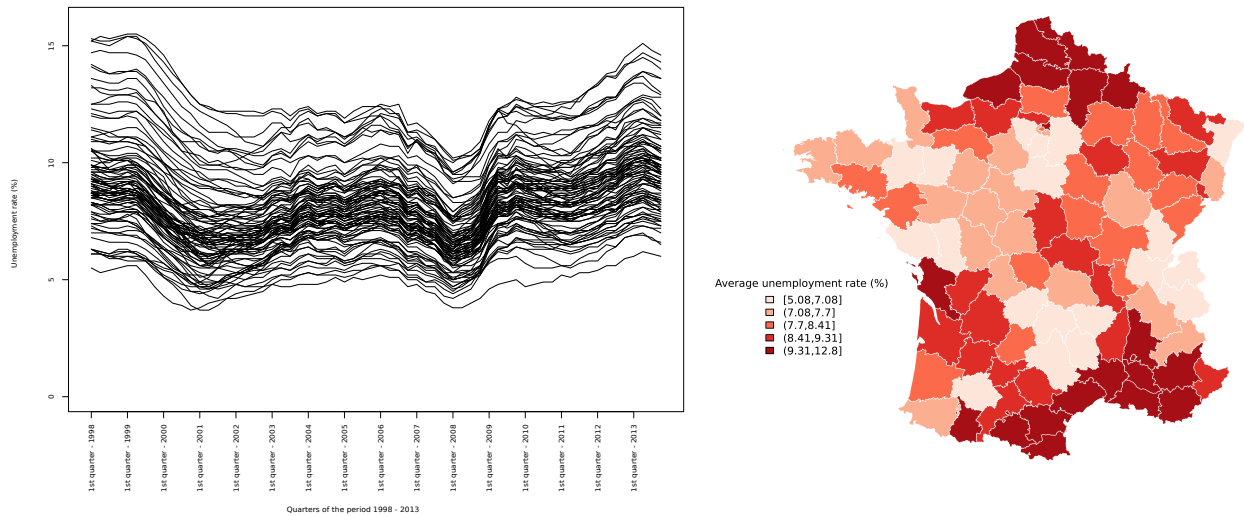
Figure 3.7: Unemployment rate curves from 1998 to 2013 in each of the 94 French *départements* (left panel), and the spatial distribution of the mean unemployment rates over the period (right panel).

Table 3.1: 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
DFFSS	1	7	0.005
	2	9	0.027
PFSS	1	7	0.011
	2	9	0.039

developed by Cuevas et al. (2004) and the "distribution-free" scan statistic of Cucala (2014) as a concentration index for spatial scan statistics.

In a simulation study, we compared our new methods with the NPFSS developed by Smida et al. (2022). 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).

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 center 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 PFSS and the DFFSS 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 data sets. 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 and Silverman (2005b)).

The spatial scan statistic methods we proposed are based on the assumption that the observations are considered independent, which is a classical assumption in the field of spatial scan statistics. However, in the presence of spatial autocorrelation in the data, this assumption of independence between the observations is no longer verified and can lead to an inflation of the type I error in the random permutation procedure, as shown by Lee et al. (2019), Lin (2014), Loh and Zhu (2007) and Ahmed et al. (2021b). Taking into account spatial autocorrelation in univariate functional scan statistics appears to be a complex subject that requires further development. Furthermore, it should be noted that the DFFSS method relies on the homoscedasticity hypothesis of the variance function of the observations. However, the DFFSS scan statistic can be adapted to the heteroscedasticity of the variance function by using a weighted version of the pointwise statistic $I^{(w)}(t)$ as proposed by Cucala (2014) in the univariate case.

Although the PFSS and the DFFSS 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 statistically 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 and Smaga (2017). It should be noted that the functional Wilcoxon-Mann-Whitney test developed by Chakraborty and Chaudhuri (2014) can be adapted for use with multivariate functional data by taking account of a suitable scalar product for calculation of the sign function.
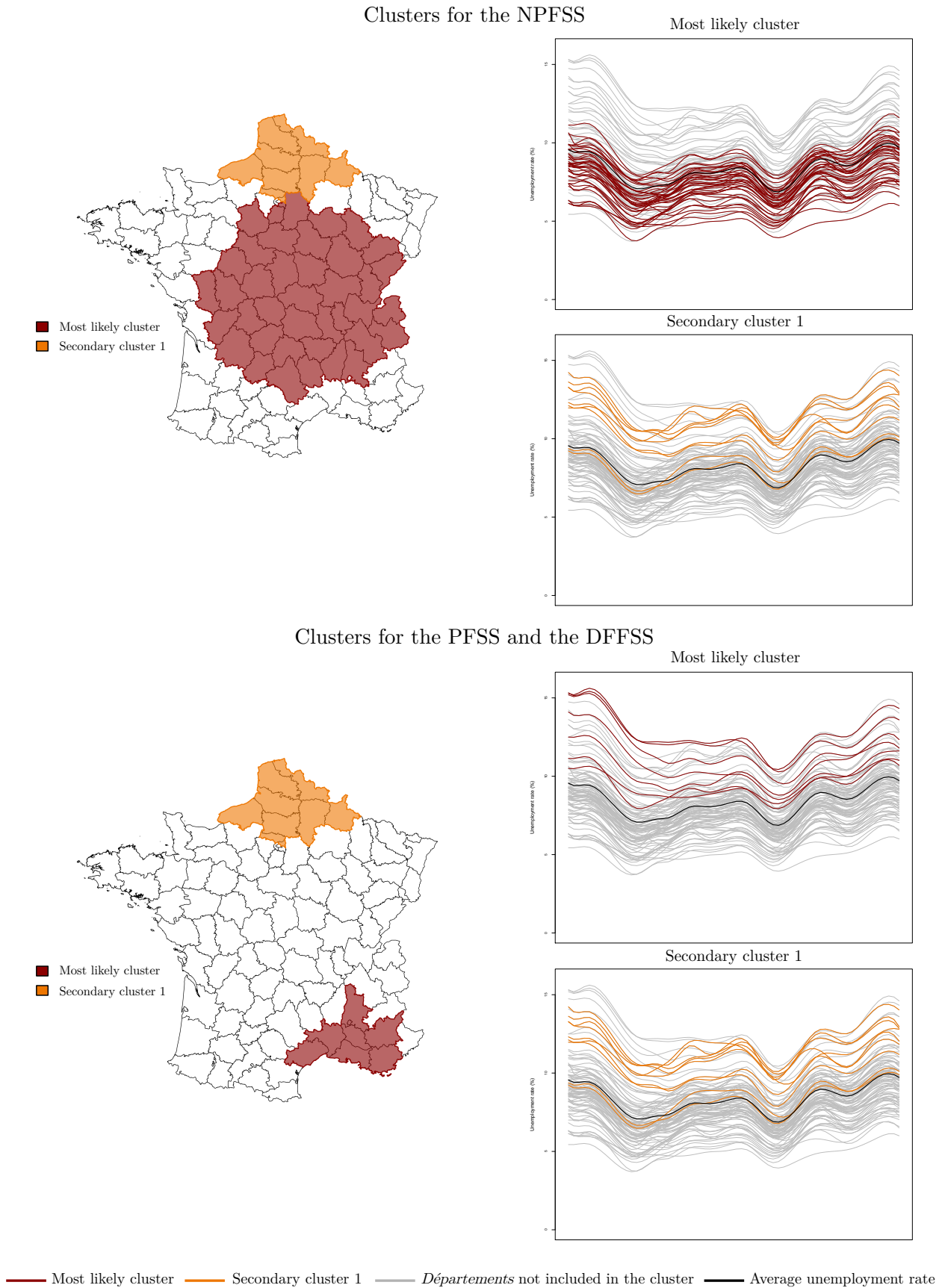
Figure 3.8: Statistically significant spatial clusters of high or low unemployment rates detected by the NPFSS (top panel) and the DFFSS and the PFSS (bottom panel). For each method, the MLC is shown in red and the secondary cluster is shown in orange. The unemployment rate curves (from 1998 to 2013) for the *département* in the MLC and in the secondary cluster are presented as correspondingly colored lines in the graphs. The black curve is the national average unemployment rate.

Chapter 4

Spatial scan statistics for multivariate functional data

Contents

1   Introduction

Environmental science has demonstrated clearly that exposure to pollutants is associated with adverse health effects. Many studies have shown that air pollution exposure is related to elevated mortality (Schwartz and Dockery, 1992; Anderson et al., 1996; Di et al., 2017; Huang et al., 2022) and a higher risk of respiratory tract disease (such as asthma and lung cancer) or cardiovascular disease, for example (Dockery et al., 1993; Brook et al., 2010; Sava and Carlsten, 2012; Hoek et al., 2013; Loomis et al., 2013; Newby et al., 2015; Liu et al., 2019). Since the 1960s, the authorities in many countries have been trying to reduce exposure to

pollutants (Zou et al., 2014). However, the literature data show that exposure remains high in some areas and often varies within the same locality: people living near busy roads or factories are much more exposed to certain pollutants than people in more affluent neighbourhoods. The World Health Organization (World Health Organization, 2016) reported that in 2012, air pollution caused the death of three million people worldwide. These findings emphasize the need to detect areas of high exposure to air pollutants so that the authorities can act to reduce pollutant concentrations, prevent certain pollution-related diseases, and improve public health more generally.

In many fields, researchers want to determine whether the aggregation of events in space results from chance or not. For example, in the field of epidemiology, the detection of statistically significant spatial clusters of high disease incidence allows: (i) researchers to hypothesize about the causes of disease, and (ii) to provide accurate information to public health actors to implement local policies. In environmental monitoring, scientists look for environmental black spots, defined as geographical areas with elevated pollutant concentrations, by analyzing differences in pollution curves for a given geographical area (i.e. spatial multivariate functional data). This allows local authorities to take localized actions to reduce pollution in these areas.

Thanks to progress in sensing and data storage, data are increasingly being measured continuously over time. This led to the emergence of univariate and multivariate functional data and then the popularization of functional data analysis (FDA) by Ramsay and Silverman (2005b). In addition to a recent trend towards multivariate functional analysis (Berrendero et al., 2011; Jacques and Preda, 2014b; Schmutz et al., 2020; Golovkine et al., 2022), in domains in which data naturally involve a spatial component (demography, environmental science, and agricultural science (Hung, 2016)), the emergence of functional data has led to the introduction of spatial functional data (Delicado et al., 2010); these are characterized by the observation of one or more curves in each spatial location. In particular, environmental scientists are increasingly studying this type of data in spatiotemporal (Carrera-Hernández and Gaskin, 2007; Brunet et al., 2007; Fan et al., 2020) or spatial functional frameworks (Ballari et al., 2018; Wang et al., 2019).

In the context of cluster detection, we previously developed in Chapter 3 two spatial scan statistics for univariate spatial functional data respectively based on an analysis of variance (ANOVA) for functional data and a pointwise approach based on a Student's t-test, whereas Smida et al. (2022) developed a nonparametric spatial scan statistic for univariate functional data.

As mentioned above, pollution cluster detection is of great importance in environmental monitoring. In the present study, we analyzed data on the concentrations (in $\mu g.m^{-3}$) of four pollutants: nitrogen dioxide (NO_2), ozone (O_3), and particulate matter with a diameter below $10\mu m$ or below $2.5\mu m$ (PM_{10} and $PM_{2.5}$, respectively). Using daily average pollutant data provided by the PREV'AIR French national air quality forecasting system (www.prevair.org) from October 1 to October 31, 2021 (i.e. 31 values for each variable) aggregated into grid cells of about 2 km by 2 km (located by their center of gravity) in the *Nord-Pas-de-Calais* region of northern France, Figure 4.1 shows the average concentrations in each geographical area over the study period.

Figure 4.1 emphasizes the spatial heterogeneity of these data. High concentrations of O_3 tend to aggregate in the areas of Montreuil and Calais, whereas high concentrations of the other pollutants tend to aggregate in the urban areas of Dunkerque and Lille. Moreover, the daily concentration curves vary markedly over time. We reasoned that functional spatial scan statistics might be of value for detecting spatial clusters of pollutant concentrations over the period from October 1 to October 31, 2021. We also observed that the spatial distributions for NO_2, PM_{10} and $PM_{2.5}$ tended to be similar (especially for the PM), whereas the spatial distributions for NO_2 and O_3 were diametrically opposed; this is in line with the literature data
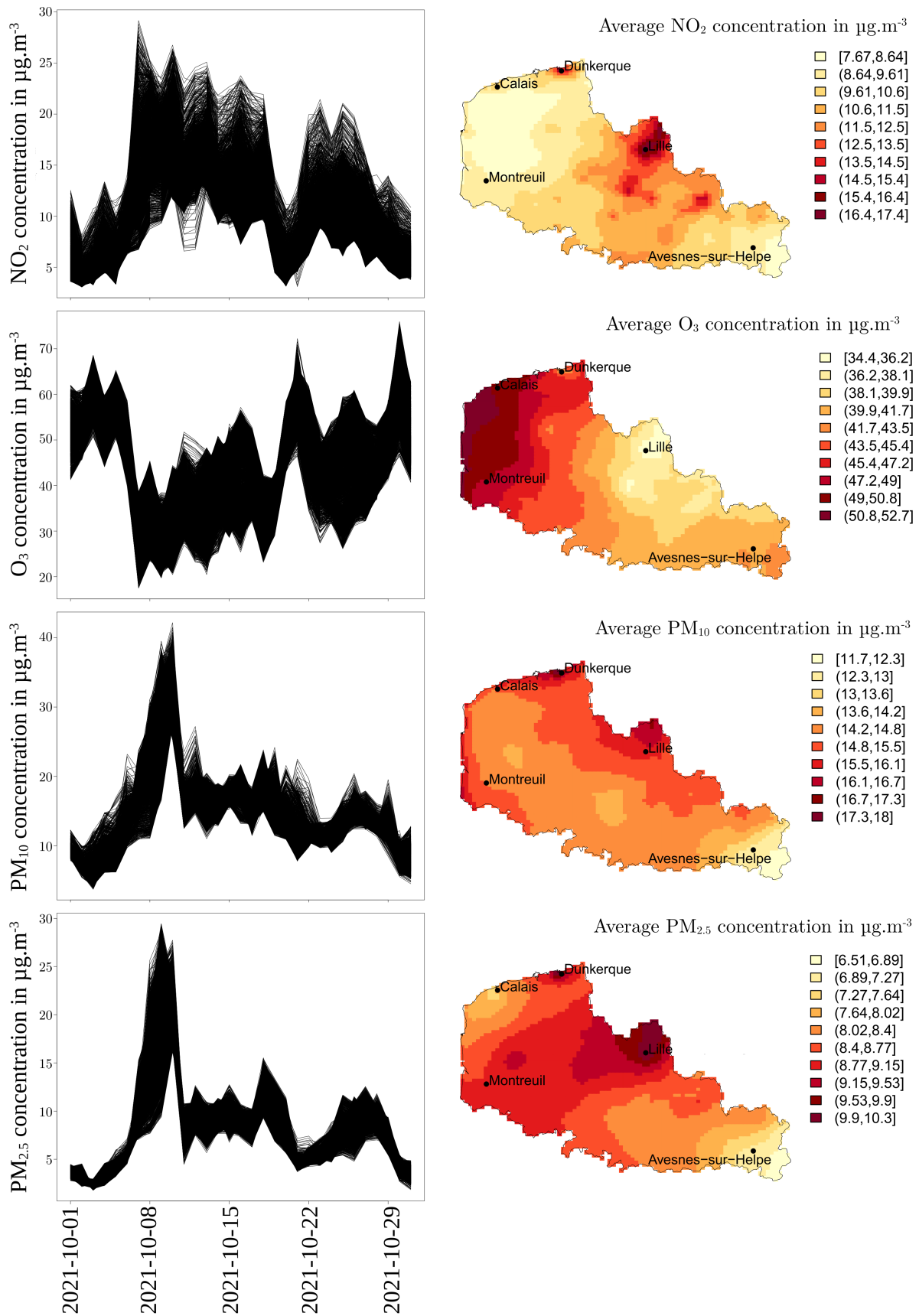
Figure 4.1: Daily NO_2, O_3, PM_{10} and $PM_{2.5}$ concentration curves (from October 1 to October 31, 2021) in the *Nord-Pas-de-Calais* region of northern France, at a spatial resolution of approximately 2 km by 2 km (left panels), together with the spatial distributions of the mean concentrations for each pollutant over the period (right panels)

on correlations between these pollutants (Kumar and Joseph, 2006; Wu et al., 2016). Spatial scan statistics for multivariate functional data take account of possible correlations between variables and so might be relevant in this situation. However, to the best of our knowledge and despite the large number of potential applications, spatial scan statistics for multivariate functional data have not previously been described.

There are several possible ways of detecting spatial clusters of atypical pollutant concentrations. Firstly, one could summarize the concentrations of each pollutant at each site (in terms of their respective averages over time) and then apply a parametric multivariate spatial scan statistic (Cucala et al., 2017) or the nonparametric spatial scan statistic developed by Cucala et al. (2019). However, the loss of information would be huge (Tarpey et al., 2021). Secondly, one could apply a spatial scan statistic for univariate functional data to each pollutant separately. However, this approach (i) would only detect clusters of individual pollutants and not clusters of multipollutant exposure, and (ii) would not take account of correlations between the four pollutants. It is therefore necessary to develop a spatial scan statistic for multivariate spatial functional data, while taking account of possible dependence between the variables.

We reasoned that by choosing a suitable scalar product, the nonparametric spatial scan statistic for functional data developed by Smida et al. (2022) could be extended to multivariate processes. To the best of our knowledge, Smida et al.'s statistic has not been considered in this context, and a parametric scan statistic for multivariate functional data has not previously been developed.

Hence, the objective of the present study was to develop three new spatial scan statistics for multivariate functional data, based on statistical tests for comparisons of multivariate functional samples. Recently Górecki and Smaga (2017) and Qiu et al. (2021) respectively developed a multivariate analysis of variance (MANOVA) test statistic and a functional Hotelling T^2-test statistic for multivariate functional data. We also suggest using the multivariate extension of the Wilcoxon rank-sum test (developed by Oja and Randles (2004)) as a pointwise test statistic. Using these statistics, we (i) adapted the parametric and the distribution-free spatial scan statistics for univariate functional data (developed in Chapter 3) for use in a multivariate functional framework, and (ii) investigated a new nonparametric multivariate functional method based on the observations' ranks at each time point.

Below, Section 2 describes the parametric multivariate functional scan statistic, the multivariate version of the distribution-free functional spatial scan statistic developed in Chapter 3, and a new rank-based spatial scan statistic for multivariate functional data. Section 3 describes a simulation study and a comparison with our adaptation of Smida et al.'s method to a multivariate functional context. These approaches are applied to an air pollution data set in Section 4. Lastly, we discuss our results in Section 5.

It should be noted that the developments presented in this chapter are currently under minor revision in the *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (August 2022) in collaboration with Mohamed-Salem Ahmed (University of Lille), Sophie Dabo-Niang (University of Lille) and Michaël Génin (University of Lille). Moreover, all the methods presented here are implemented in the package **HDSpatialScan** available in the CRAN repository (Frévent et al., 2022).

2 Methodology

2.1 General principle

Let $\{X(t), \ t \in \mathcal{T}\}$ be a p-dimensional vector-valued stochastic process where $\mathcal{T}$ is an interval of $\mathbb{R}$. 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 $s_1, \ldots, s_n$. In practice, the process is measured (possibly with an error) only at discrete times of $\mathcal{T}$. These observation times might vary in number and in value, both from one site to another but also from one variable to another. $\{X_{i,1}^{(j)}, \ldots, X_{i,m_{i,j}}^{(j)}\}$ denotes the effective observations of the jth variable of X in s_i at $m_{i,j}$ time points of $\mathcal{T}$. Next, $X_i^{(j)}$ can be approximated from the longitudinal observations $\{X_{i,1}^{(j)}, \ldots, X_{i,m_{i,j}}^{(j)}\}$. In brief, one can suppose that $X_i^{(j)}$ can be written as a linear combination of K_j basis functions $\varphi_1^{(j)}, \ldots, \varphi_{K_j}^{(j)}$ (where $K_j \leq \min\{m_{1,j}, \ldots, m_{n,j}\}$): for $i \in [\![1, n]\!]$, for $j \in [\![1, p]\!]$, $X_i^{(j)}(t) = \sum_{k=1}^{K_j} a_{i,k,j} \varphi_k^{(j)}(t)$. This is equivalent to writing X_i as

$$X_i(t) = \varphi(t) a_i$$

where $a_i = (a_{i,1,1}, \ldots, a_{i,K_1,1}, a_{i,1,2}, \ldots, a_{i,K_2,2} \ldots, a_{i,1,p}, \ldots, a_{i,K_p,p})^{\top}$ is computed using an interpolation method (if the observations are assumed to be errorless) or, with an ordinary or penalized least squares method (if some observations may be erroneous), which smooths the data,

and $\varphi(t) = \begin{pmatrix} \varphi_1^{(1)}(t) & \ldots & \varphi_{K_1}^{(1)}(t) & 0 & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & \varphi_1^{(2)}(t) & \ldots & \varphi_{K_2}^{(2)}(t) & 0 & \ldots & 0 \\ & & & & \ldots & & & & \\ 0 & \ldots & \ldots & \ldots & \ldots & 0 & \varphi_1^{(p)}(t) & \ldots & \varphi_{K_p}^{(p)}(t) \end{pmatrix}$.

The choice of the basis φ depends on the nature of the data considered: a Fourier basis is appropriate for periodic data, whereas it is possible to use a polynomial, spline or wavelet basis for non-periodic data (see Ramsay and Silverman, 2005b, for more details).

Hereafter, the observations X_i are considered to be independent, which is a classical assumption in scan statistics. Spatial scan statistics are designed to detect spatial clusters and 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). We previously defined the notion of a cluster in a univariate functional framework. These definitions can be easily extended to the multivariate functional context by defining a *multivariate magnitude cluster* w as follows:

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

where $\Delta(t) = (\Delta_1(t), \ldots, \Delta_p(t))^{\top}$, and exists $j \in [\![1, p]\!]$ such that Δ_j is of constant sign and non-zero over at least one sub-interval of $\mathcal{T}$. In the same way, a *multivariate shape cluster* can be defined as follows:

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

where $\Delta(t) = (\Delta_1(t), \ldots, \Delta_p(t))^{\top}$ and exists $j \in [\![1, p]\!]$ such that Δ_j is not constant almost everywhere.

Ever since Cressie's publication (Cressie, 1977), a scan statistic has been defined by the maximum value of a concentration index over a set of potential clusters $\mathcal{W}$. In the following and without loss of generality, we focus on variable-size circular clusters (as introduced by Kulldorff and Nagarwalla, 1995). Next, the set of potential clusters $\mathcal{W}$ is the set of discs centered on a location and passing through another location:

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

where $w_{i,j}$ is the disc centered on s_i that passes through s_j and $|w_{i,j}|$ is the number of sites in $w_{i,j}$. Thus, a cluster cannot cover more than 50% of the studied region, as recommended

by Kulldorff and Nagarwalla (1995). It should be noted that other possibilities have been described, including elliptical clusters (Kulldorff et al., 2006), rectangular clusters (Chen and Glaz, 2009), and graph-based clusters (Cucala et al., 2013).

Below, we present a parametric scan statistic in subsection 2.2, a distribution-free scan statistic in subsection 2.3, and a new rank-based scan statistic for multivariate functional data in subsection 2.4.

2.2 A parametric spatial scan statistic for multivariate functional data

In this subsection, the process X is supposed to take values in 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 product $\langle X, Y \rangle = \int_{\mathcal{T}} X(t)^\top Y(t) \, \mathrm{d}t$.

Starting from a functional ANOVA, we developed a parametric scan statistic for univariate functional data in Chapter 3. A classical MANOVA (the Lawley–Hotelling trace test (Oja and Randles, 2004)) was adapted by Górecki and Smaga (2017) for $\mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$ processes: by considering two groups g_1 and g_2 of independent random observations of two p-dimensional stochastic processes X_{g_1} and X_{g_2} taking values in $\mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$, the MANOVA tests the equality of the two mean vector-valued functions μ_{g_1} and μ_{g_2}, where $\mu_{g_i}(t) = \mathbb{E}[X_{g_i}(t)] \in \mathbb{R}^p$, $i = 1, 2$, $t \in \mathcal{T}$.

For the cluster detection problem, the null hypothesis $\mathcal{H}_0$ (the absence of a cluster) can be defined as $\mathcal{H}_0 : \forall w \in \mathcal{W}$, $\mu_w = \mu_{w^c} = \mu_S$, 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 $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$. Thus, we can use the functional MANOVA to compare the mean functions in w and w^c.

Górecki and Smaga (2017) adapted different MANOVAs for use with functional frameworks. However, the Wilks lambda test statistic, the Lawley-Hotelling trace test statistic, and the Pillai trace test statistic had similar performance levels. Furthermore, the three statistics often outperformed (in terms of power) the tests that use random projections. We therefore decided to assess use of the Lawley-Hotelling trace test for cluster detection by using the following statistic:

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

where

$$H_w = |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)][\bar{X}_w(t) - \bar{X}(t)]^\top \, \mathrm{d}t + |w^c| \int_{\mathcal{T}} [\bar{X}_{w^c}(t) - \bar{X}(t)][\bar{X}_{w^c}(t) - \bar{X}(t)]^\top \, \mathrm{d}t$$

and

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

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

Next, $\mathrm{LH}^{(w)}$ is considered to be a concentration index and can be maximized over the set of potential clusters $\mathcal{W}$, which results in the following definition of a multivariate parametric functional spatial scan statistic (MPFSS):

$$\Lambda_{\mathrm{MPFSS}} = \max_{w \in \mathcal{W}} \mathrm{LH}^{(w)}.$$

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

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

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

Chapter 3 developed a distribution-free spatial scan statistic for univariate functional data by combining the distribution-free scan statistic for non-functional data developed by Cucala (2014) (based on a Student's t-test), and the generalization of a pointwise test over the time (Lin et al., 2021).

Very recently, Qiu et al. (2021) developed a version of this pointwise test for p-dimensional functional data ($p \geq 2$) and used it to compare the mean functions of X in two groups.

We supposed that for each time t, $\mathbb{V}[X_i(t)] = \Sigma(t, t)$ for all $i \in [\![1, n]\!]$, where Σ is a $p \times p$ covariance matrix function.

Thus, as mentioned above, in the context of cluster detection, the null hypothesis $\mathcal{H}_0$ can be defined as: $\mathcal{H}_0 : \forall w \in \mathcal{W}$, $\mu_w = \mu_{w^c} = \mu_S$, 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: $\mathcal{H}_1^{(w)} : \mu_w \neq \mu_{w^c}$. Next, Qiu et al. (2021) developed the following statistic and used it to compare the mean function μ_w in w with the mean function μ_{w^c} in w^c:

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

where $T_n(t)^{(w)}$ 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 in subsection 2.2, 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.

Next, $T_{n,\max}^{(w)}$ is considered to be a concentration index and is maximized over the set of potential clusters $\mathcal{W}$, yielding the following multivariate distribution-free functional spatial scan statistic (MDFFSS):

$$\Lambda_{\text{MDFFSS}} = \max_{w \in \mathcal{W}} T_{n,\max}^{(w)}.$$

The MLC is therefore

$$\text{MLC} = \arg\max_{w \in \mathcal{W}} T_{n,\max}^{(w)}.$$

2.4   A new rank-based spatial scan statistic for multivariate functional data

Oja and Randles (2004) developed a p-dimensional ($p \geq 2$) extension of the classical Wilcoxon rank-sum test using multivariate ranks. Following on from Oja and Randles's definitions, we

define the notion of "pointwise multivariate ranks" as follows for each time $t \in \mathcal{T}$:

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

where $\text{sgn}(\cdot)$ is the spatial sign function

$$\text{sgn} \quad : \quad \mathbb{R}^p \quad \to \quad \mathbb{R}^p$$
$$x \quad \mapsto \quad \begin{cases} ||x||_2^{-1}x & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

and $A_X(t)$ is a pointwise data-based transformation matrix that makes the pointwise multivariate ranks behave as though they were spherically distributed in the unit p-sphere:

$$\frac{p}{n} \sum_{i=1}^{n} R_i(t)R_i(t)^\top = \frac{1}{n} \sum_{i=1}^{n} R_i(t)^\top R_i(t) \mathrm{I}_p.$$

Note that this matrix is similar to Tyler's transformation matrix and can be easily computed using an iterative procedure.

Oja and Randles (2004) compared the cumulative distribution functions of real multivariate observations in two groups. In the context of multivariate functional data, their statistic can be considered to be a pointwise test statistic for each time t; the pointwise multivariate extension of the Wilcoxon rank-sum test statistic is defined as

$$W(t)^{(w)} = \frac{pn}{\displaystyle\sum_{i=1}^{n} R_i(t)^\top R_i(t)} \left[|w| \, ||\bar{R}_w(t)||_2^2 + |w^c| \, ||\bar{R}_{w^c}(t)||_2^2 \right]$$

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

Next, as above, we suggest generalizing the information over the time with

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

Then, 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 $\mathcal{H}_1^{(w)}$: $\exists t \in \mathcal{T}$, $F_{w,t}(x) = F_{w^c,t}(x - \Delta_t)$, $\Delta_t \neq 0$.

Next, $W^{(w)}$ can be considered to be a concentration index and is maximized over the set of potential clusters $\mathcal{W}$, so that the multivariate rank-based functional spatial scan statistic (MRBFSS) is defined as follows:

$$\Lambda_{\text{MRBFSS}} = \max_{w \in \mathcal{W}} W^{(w)}.$$

Thus, the MLC is

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

2.5   Computing the statistical significance of the MLC

Once the MLC has been detected, its statistical significance must be evaluated. The distribution of the scan statistic Λ (Λ_{MPFSS}, Λ_{MDFFSS} or Λ_{MRBFSS}) is intractable under $\mathcal{H}_0$, due to the dependence between $\mathcal{S}^{(w)}$ and $\mathcal{S}^{(w')}$ if $w \cap w' \neq \emptyset$ ($\mathcal{S} = \text{LH}, T_{n,\max}$ or W). We therefore chose to obtain a large set of simulated data sets by randomly permuting the observations X_i in the spatial locations. This technique (called "random labelling") has already been used in spatial scan statistics (Kulldorff et al., 2009; Cucala et al., 2017).

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

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

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

2.6   Secondary clusters

It is also possible to detect so-called "secondary clusters", using several strategies. Here, we consider the approach of Kulldorff (1997), which defines secondary clusters as geographical areas corresponding to a 2$^{\text{nd}}$ maximum of the concentration index, a 3$^{\text{rd}}$ maximum, and so on. Their p-value is calculated by comparing the values of the concentration index with the value of the test statistic $\Lambda^{(m)}$ observed for each set of permutations m. Note that this approach is rather conservative because for a secondary cluster to be declared statistically significant, it must be able to reject $\mathcal{H}_0$ per se - as if it were the MLC itself. Lastly, only statistically significant secondary clusters that do not overlap with a more likely cluster were considered. Although there are other approaches for detecting secondary clusters (Zhang et al., 2010), we will not consider them here.

3   Computational analysis in a simulation study

In a simulation study, we compare the performance levels of the MPFSS, MDFFSS, and MRBFSS. Smida et al. (2022) developed a nonparametric scan statistic for univariate functional data (NPFSS); although it can be extended to the multivariate functional framework, it has not previously been studied in this context. Thus, we also decided to include Smida et al.'s approach in the simulation by adapting it for multivariate functional data; to this end, we considered the inner product $\langle X, Y \rangle = \int_{\mathcal{T}} X(t)^{\top} Y(t) \, \mathrm{d}t$ which allowed to redefine the NPFSS in the multivariate functional context as $\Lambda_{\text{NPFSS}} = \max_{w \in \mathcal{W}} U^{(w)}$ where

$$U^{(w)} = \sqrt{\frac{|w||w^{\mathsf{c}}|}{n}} \left\| \frac{1}{|w||w^{\mathsf{c}}|} \sum_{i, s_i \in w} \sum_{j, s_j \in w^{\mathsf{c}}} \frac{X_j - X_i}{||X_j - X_i||} \right\|$$

and $||X_j - X_i|| = \sqrt{\int_{\mathcal{T}} (X_j(t) - X_i(t))^{\top}(X_j(t) - X_i(t)) \, \mathrm{d}t}$.

3.1 Design of the simulation study

Artificial data sets were generated by using the geographical locations of the 94 French *départements* (county-type administrative areas), as shown in Figure 4.2. The location of each *département* was defined by its centroid. For each artificial data set, we simulated a spatial cluster w (composed of eight *départements* in the Paris region, i.e. the red area in Figure 4.2).
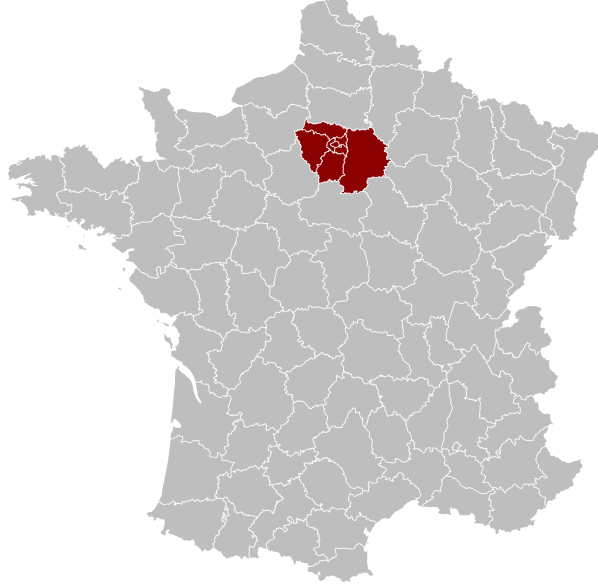
Figure 4.2: The 94 French *départements* and the spatial cluster (in red) simulated for each artificial data set.

3.1.1 Generation of the artificial data sets

The X_i were simulated according to the following model, with $p = 2$ (see Qiu et al. (2021); Martino et al. (2019) for more details):

for each $i \in [\![1, 94]\!]$, $X_i(t) = (\sin [2\pi t^2]^5, 1 + 2.3t + 3.4t^2 + 1.5t^3)^\top + \Delta(t)\mathbb{1}_{s_i \in w} + \varepsilon_i(t), \ t \in [0, 1],$

where $\varepsilon_i(t) = \sum_{k=1}^{100} Z_{i,k} \sqrt{1.5 \times 0.2^k}\theta_k(t) + b_i(t)$, with $\theta_k(t) = \begin{cases} 1 & \text{if } k = 1 \\ \sqrt{2}\sin[k\pi t] & \text{if } k \text{ even} \\ \sqrt{2}\cos[(k-1)\pi t] & \text{if } k \text{ odd and } k > 1 \end{cases}$

and $b_i(t) \sim \mathcal{N}(0, 0.2^2 \mathrm{I}_2)$.

The functions X_i were measured at 101 equally spaced times on $[0, 1]$.

The covariance matrix of the $Z_{i,k}$ is denoted as $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, and three distributions of the $Z_{i,k}$ were considered: (i) a normal distribution: $Z_{i,k} \sim \mathcal{N}(0, \Sigma)$, (ii) a standardized Student distribution: $Z_{i,k} = U_{i,k} \left(\frac{V_{i,4}}{4}\right)^{-0.5}$ where the $U_{i,k}$ are independent $\mathcal{N}(0, \Sigma/2)$ variables and the $V_{i,4}$ are independent $\chi_2(4)$ variables, and (iii) a standardized exponential distribution: $Z_{i,k} = \left[U_{i,k} - \binom{2}{2}\right]/2$ where the $U_{i,k}$ are independent and $U_{i,k} \sim \mathcal{E}(\frac{1}{2}, \Sigma)$.

Note that ρ corresponds to the correlation between the two components of $X(t)$ at each time.

Three values of ρ (0.2, 0.5 and 0.8) were tested, and three types of clusters whose intensity was controlled by a parameter $\alpha > 0$ were studied: $\Delta_1(t) = \alpha(t, t)^\top$, $\Delta_2(t) = \alpha(t(1-t), t(1-t))^\top$

and $\Delta_3(t) = \alpha(\exp\left[-100(t-0.5)^2\right]/3, \exp\left[-100(t-0.5)^2\right]/3)^\top$. Since all the clusters vary over time and are positive and non-zero over $\mathcal{T} = [0,1]$ (except possibly for $t = 0$ or $t = 1$), they correspond to multivariate magnitude clusters and multivariate shape clusters.

Different values of the parameter α were considered for each Δ: $\alpha \in \{0,\ 0.375,\ 0.75,\ 1.125,\ 1.5\}$ for Δ_1, $\alpha \in \{0,\ 1,\ 2,\ 3,\ 4\}$ for Δ_2 and $\alpha \in \{0,\ 1.25,\ 2.5,\ 3.75,\ 5\}$ for Δ_3. Note that $\alpha = 0$ was also tested, in order to evaluate the conservation of the nominal type I error. An example of the data for $\rho = 0.2$ and for the Gaussian distribution for the $Z_{i,k}$ is given in Figure 4.3. Since the data had been generated with noise $b_i(t)$, we applied a smoothing method with a cubic B-spline basis using an ordinary least squares method (Figure 4.4).
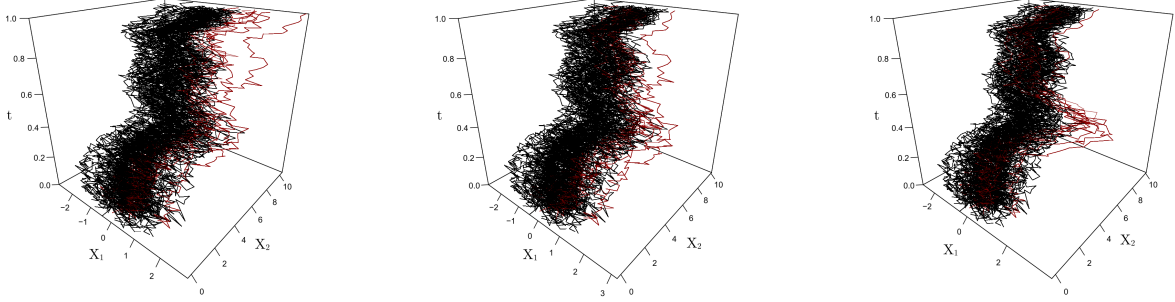
Figure 4.3: The simulation study: the two components of the data generated for the Gaussian process and $\rho = 0.2$, with $\Delta(t) = \Delta_1(t) = 1.5(t,t)^\top$ (left panel), $\Delta(t) = \Delta_2(t) = 4(t(1-t), t(1-t))^\top$ (middle panel) and $\Delta(t) = \Delta_3(t) = 5(\exp\left[-100(t-0.5)^2\right]/3, \exp\left[-100(t-0.5)^2\right]/3)^\top$ (right panel) without smoothing. The red curves correspond to the observations in the cluster.
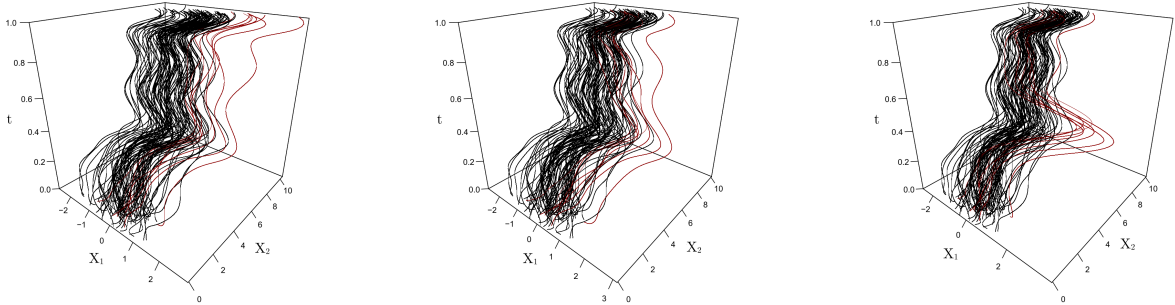
Figure 4.4: The simulation study: the two components of the data generated for the Gaussian process and $\rho = 0.2$, with $\Delta(t) = \Delta_1(t) = 1.5(t,t)^\top$ (left panel), $\Delta(t) = \Delta_2(t) = 4(t(1-t), t(1-t))^\top$ (middle panel) and $\Delta(t) = \Delta_3(t) = 5(\exp\left[-100(t-0.5)^2\right]/3, \exp\left[-100(t-0.5)^2\right]/3)^\top$ (right panel) (Figure 4.3) after smoothing. The red curves correspond to the observations in the cluster.

3.1.2   Comparison of the methods

For each $Z_{i,k}$ distribution, each Δ, each ρ, and each value of α, 1000 artificial data sets were simulated. A total of 999 samples were generated by random permutations of the data, and the type I error was set to 5%. The methods' respective performance levels were compared

with regard to four criteria: the power, the true positive rate, the false positive rate, and the F-measure. We also compared the performance levels before and after smoothing.

The power was estimated as the proportion of simulations leading to the rejection of $\mathcal{H}_0$, depending on the type I error. Among the simulated data sets leading to the rejection of $\mathcal{H}_0$, the true positive rate is the average proportion of sites correctly detected among the sites in w, the false positive rate is the average 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 (the positive predictive value) and the true positive rate.

3.2 Results of the simulation study

The results of the simulation are presented in Figures 4.5, 4.6 and 4.7.
For $\alpha = 0$, all methods appeared to yield a type I error of 0.05, regardless of the type of process, the type of Δ and the level of correlation ρ (see the power curves in Figures 4.5, 4.6 and 4.7).

For all methods, the performance level fell slightly when the correlation ρ increased. The performances of both pointwise methods (MDFFSS and MRBFSS) improved when the data were smoothed. Given that smoothing has little or no effect on the MPFSS and the NPFSS, we mainly analyzed the performance levels after smoothing.
The NPFSS and the MPFSS showed similar powers for the Gaussian distribution and the shifts Δ_1 and Δ_2. For non-Gaussian distributions of the $Z_{i,k}$ or the shift Δ_3, however, the NPFSS yielded higher powers than the MPFSS. The MDFFSS presented the highest powers in the Gaussian case. However, its level of performance also fell when the data were not distributed normally. In the latter case, the MRBFSS performed best in terms of power (except for Δ_2, although the MRBFSS still performed well in this instance). For a Gaussian distribution, the MRBFSS performed better (in terms of power) than the NPFSS (except with regard to Δ_2) and the MPFSS.
The MRBFSS almost always gave the highest true positive rate (except for Δ_2). The true positive rates for the MDFFSS were also very high for normal data but were lower for non-normal data. The MPFSS gave the lowest true positive rates but also very low false positive rates. The MDFFSS gave the lowest false positive rates. The MRBFSS often showed higher false positive rates, although these were very close to the NPFSS's false positive rates for the shifts Δ_1 and Δ_2 and were much lower than the latter for Δ_3. As a result, the MDFFSS showed the highest F-measures, followed by the MRBFSS (except for Δ_2). For Δ_1 and Δ_2, the MPFSS and the NPFSS yielded similar F-measures for the Gaussian distribution, whereas the F-measures were lower for the MPFSS for non-normal distributions. Lastly, for the local shift Δ_3, the F-measures for the NPFSS and the MPFSS were much lower than those of the MDFFSS and the MRBFSS.
Through this simulation study, the MDFFSS and the MRBFSS appear to be the preferred methods for cluster detection in Gaussian and non-Gaussian data respectively, especially for the detection of localized clusters in time such as those characterized by the shift Δ_3.

4 Application to real data: air pollution in the *Nord-Pas-de-Calais*

We applied the spatial scan statistics for multivariate functional data to the air pollution data set provided by PREV'AIR (as described in Section 1), in order to detect statistically significant spatial clusters of atypical pollutant concentrations. The data set consisted of the concentrations
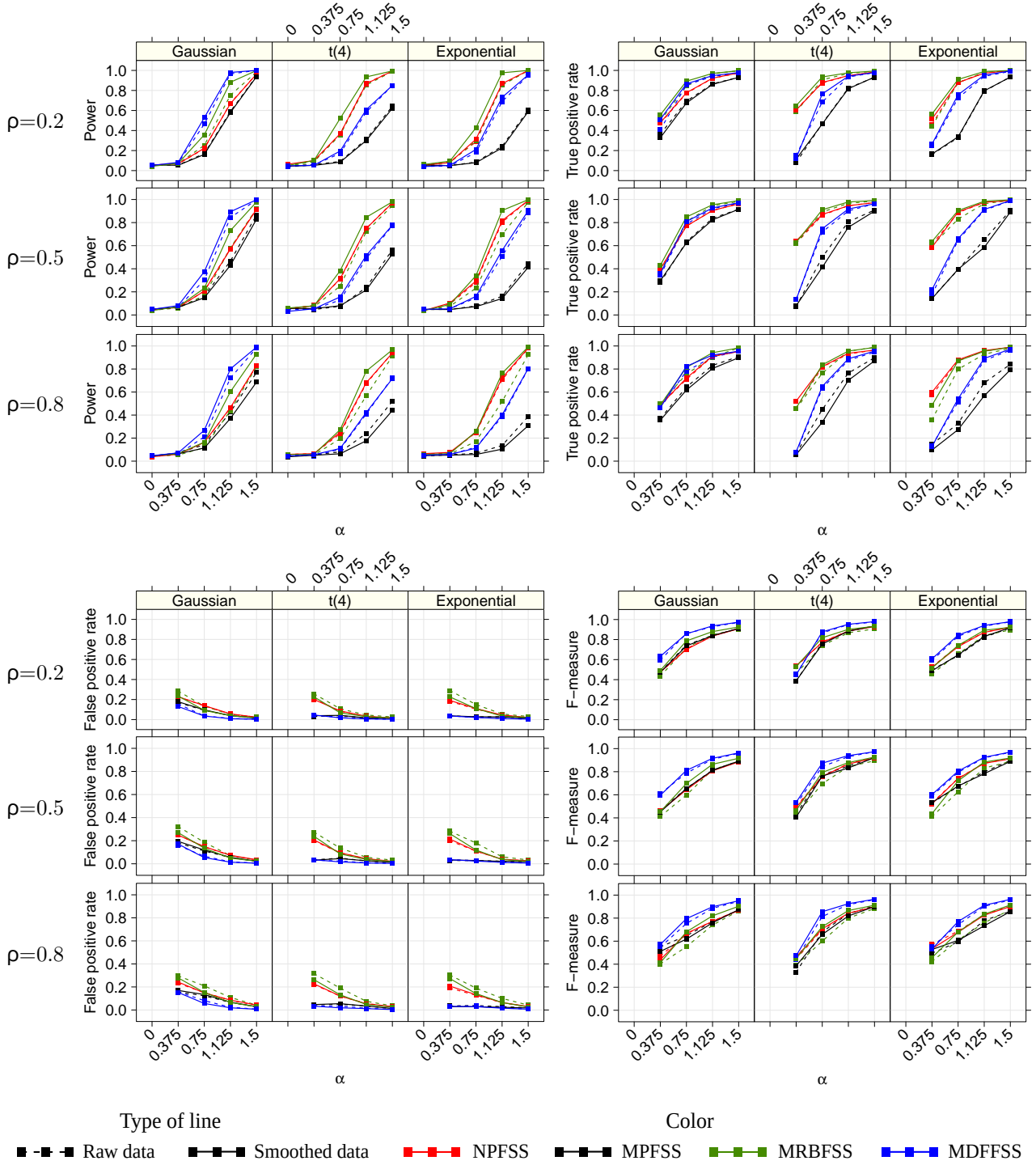
Figure 4.5: The simulation study: a comparison of the NPFSS, MDFFSS, MRBFSS and MPFSS methods for the shift $\Delta_1(t) = (\alpha t, \alpha t)^\top$. For each method and each level of correlation ρ, the Figure shows the power curves, the true positive and false positive rates, and the F-measure values for detection of the spatial cluster as the MLC. α is the parameter that controls the cluster intensity.

(in $\mu g.m^{-3}$) of NO_2, O_3, PM_{10} and $PM_{2.5}$ from October 1 to 31, 2021, in the *Nord-Pas-de-Calais* region. The data provided by PREV'AIR were already aggregated to 3231 2 km $\times$ 2 km grid cells located by their center of gravity.
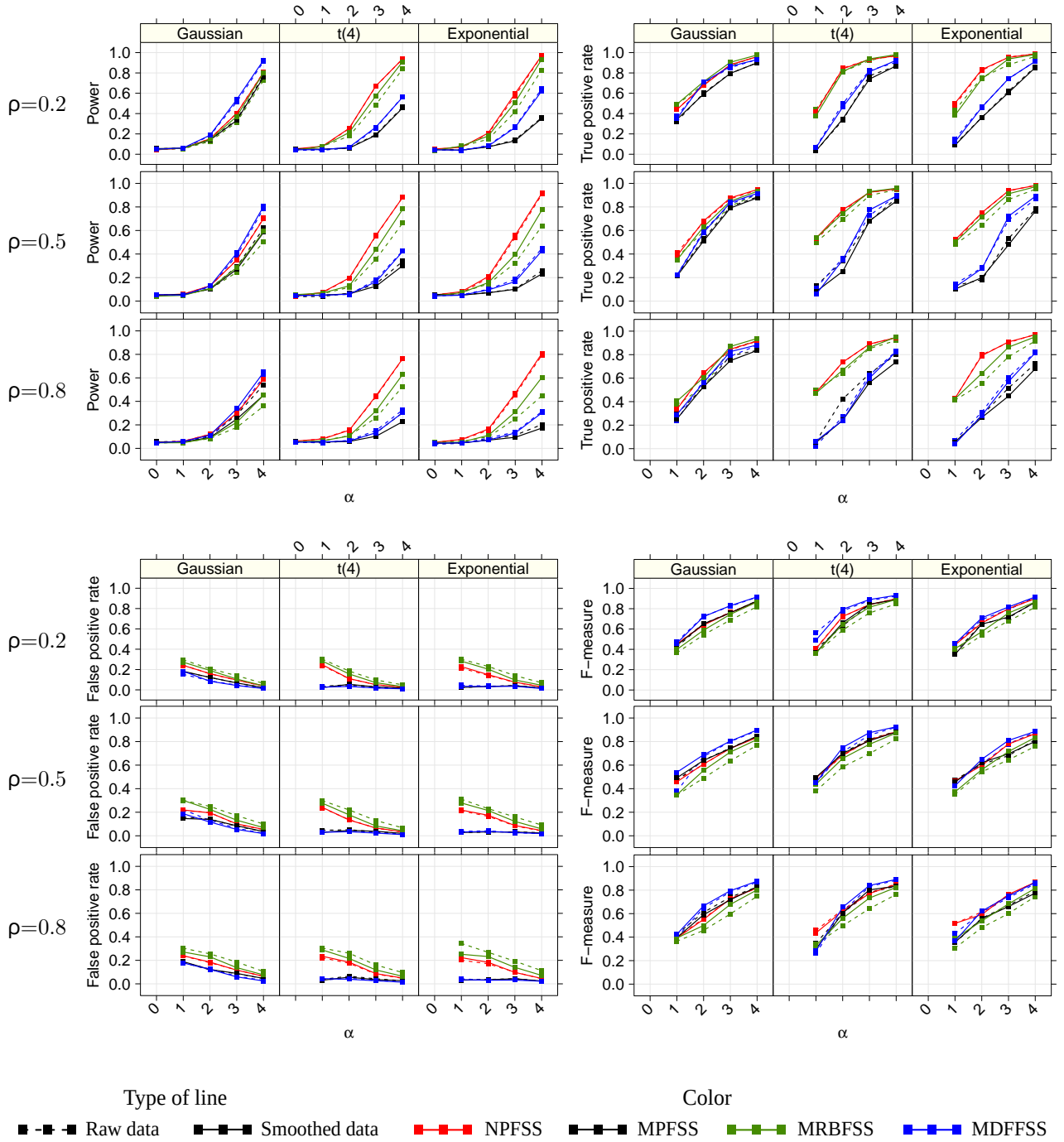
Figure 4.6: The simulation study: a comparison of the NPFSS, MDFFSS, MRBFSS and MPFSS methods for the shift $\Delta_2(t) = (\alpha t(1 - t), \alpha t(1 - t))^\top$. For each method and each level of correlation ρ, the Figure shows the power curves, the true positive and false positive rates, and the F-measure values for detection of the spatial cluster as the MLC. α is the parameter that controls the cluster intensity.

4.1 Data processing

Pollutant data are potentially noisy; in this case, the simulation study showed that smoothing the data improved the spatial scan statistics' level of performance. We therefore applied a cubic B-spline smoothing step using an ordinary least squares method. This smoothing did not yield the pollution peaks that correspond (for example) to particular phenomena on certain days of the week. However, we also used nonparametric locally estimated scatterplot
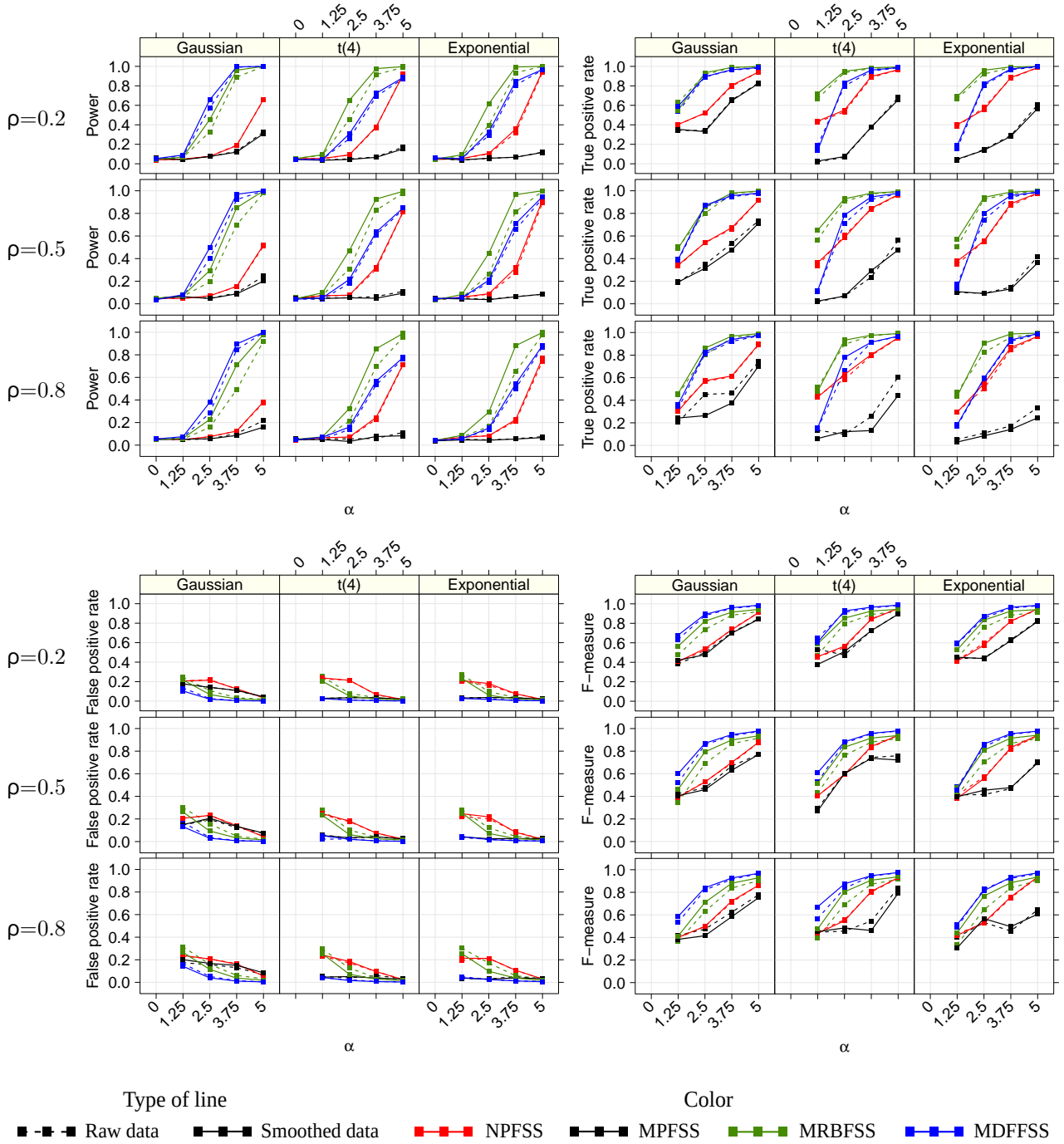
Figure 4.7: The simulation study: a comparison of the NPFSS, MDFFSS, MRBFSS and MPFSS methods for the shift $\Delta_3(t) = (\alpha \exp\left[-100(t-0.5)^2\right]/3, \alpha \exp\left[-100(t-0.5)^2\right]/3)^\top$. For each method and each level of correlation ρ, the Figure shows the power curves, the true positive and false positive rates, and the F-measure values for detection of the spatial cluster as the MLC. α is the parameter that controls the cluster intensity.

smoothing, which yielded the peaks. The curves obtained with each type of smoothing and the average concentration maps are shown in Figures 4.8 and 4.9. The maps highlighted spatial heterogeneity, with aggregates of high concentrations of NO_2, PM_{10} and $PM_{2.5}$ in Dunkerque and Lille and high concentrations of O_3 in the area between Montreuil and Calais. In general, NO_2 and O_3 had opposite spatial distributions, PM_{10} and $PM_{2.5}$ had very similar spatial distributions, and the latter two were quite similar to the spatial distribution for NO_2. In environmental science, it is well known that these pollutants are correlated (Kumar and

Joseph, 2006; Wu et al., 2016). Here, we computed the correlation surfaces for the pollution data (Figure 4.10). We observed a strong negative correlation between NO_2 and O_3, a very strong positive correlation (over the same given observation time) between PM_{10} and $PM_{2.5}$, and high positive correlations (again over the same observation time) between NO_2 and PM_{10} and between NO_2 and $PM_{2.5}$.

Lastly, the pollutant concentration curves (Figures 4.8 and 4.9) revealed marked variability over time. The presence of spatial heterogeneity (with aggregates of high pollutant concentrations), marked variability over time, and correlations between pollutants prompted us to apply the spatial scan statistics developed for multivariate functional data. This allows to detect clusters of simultaneous exposure to pollutants while taking account of correlations between pollutants.

4.2   Spatial cluster detection

We sought to determine whether there were statistically significant pollution clusters in the data from October 1 to October 31, 2021, after two types of smoothing.
We considered a round scanning window with a maximum radius of 10 km; small clusters of pollution are more relevant because the sources of pollutants are very localized (Hagler et al., 2021; Müller et al., 2022). The main source of NO_2 is road traffic, and $PM_{2.5}$ is mainly emitted in urban areas (by heating systems and road traffic) or industrial areas. More precisely, we considered the set of potential circular clusters with a maximum radius such that a potential cluster would never contain more than 50% of the total number of spatial locations (the approach recommended by Kulldorff and Nagarwalla (1995)). Next, we examined the MLC and the statistically significant secondary clusters detected with each method (see Section 2.6 for more details) and selected the cluster with the highest concentration index and a radius of less than 10 km. This type of filtering (rather than a set of potential clusters with a maximum radius of 10 km) ensures unbiased statistical inferences (Kulldorff, 2006). The statistical significance of the MLC was evaluated *via* 999 Monte-Carlo permutations, and the threshold for statistical significance was set to 0.05.

4.3   Results for the data from October 1 to October 31, 2021, after nonparametric smoothing

The detected clusters are shown in Figures 4.11 and 4.12.
The MDFFSS detected a statistically significant cluster ($\hat{p} = 0.001$) around the area of Lille; it was characterized by high concentrations of NO_2, PM_{10} and $PM_{2.5}$, and low concentrations of O_3. Lille is an urban area situated at crossroads between several major European road transportation routes; traffic (notably heavy good vehicles) is the main source of air emissions in general and NO_2 in particular (Citepa, a). Particle emissions (especially $PM_{2.5}$) are also influenced by urban heating in this densely populated area (Citepa, b,c).

Both the MRBFSS and the MPFSS detected a statistically significant cluster ($\hat{p} = 0.001$) in the very rural area of Avesnes-sur-Helpe. The cluster was characterized by low NO_2, $PM_{2.5}$ and PM_{10} concentrations. In rural areas, PM_{10} might be present at higher levels during harvest periods (due to harvesting, fertilizer spreading and ploughing), and so October is probably not the best month for judging "normal" peaks of PM_{10}.

Lastly, the NPFSS detected a statistically significant cluster ($\hat{p} = 0.001$) to the southwest of Calais. The fact that this coastal area has large areas of grasslands and crops and almost no urban activity might explain the lower observed NO_2 concentrations. Sea breezes and solar radiation might result in more O_3 formation in this area than in inland areas. O_3 is a secondary
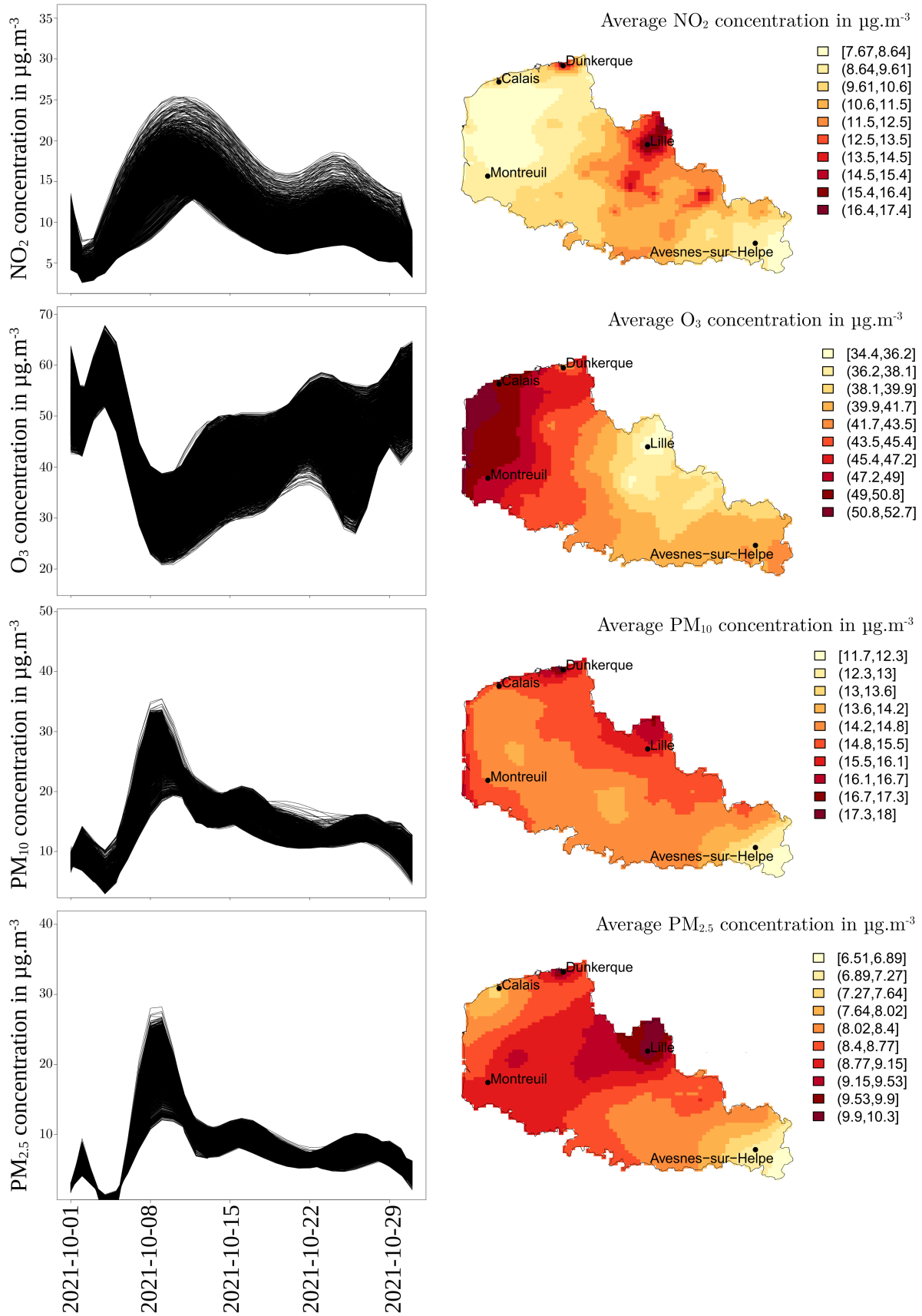
Figure 4.8: Daily smoothed concentration curves of NO_2, O_3, PM_{10} and $PM_{2.5}$ (from October 1 to October 31, 2021) in the *Nord-Pas-de-Calais* region of northern France at a spatial resolution of approximately 2 km by 2 km (left panels), together with the spatial distributions of the average smoothed concentrations for each pollutant over the period (right panels) using a B-spline smoothing.
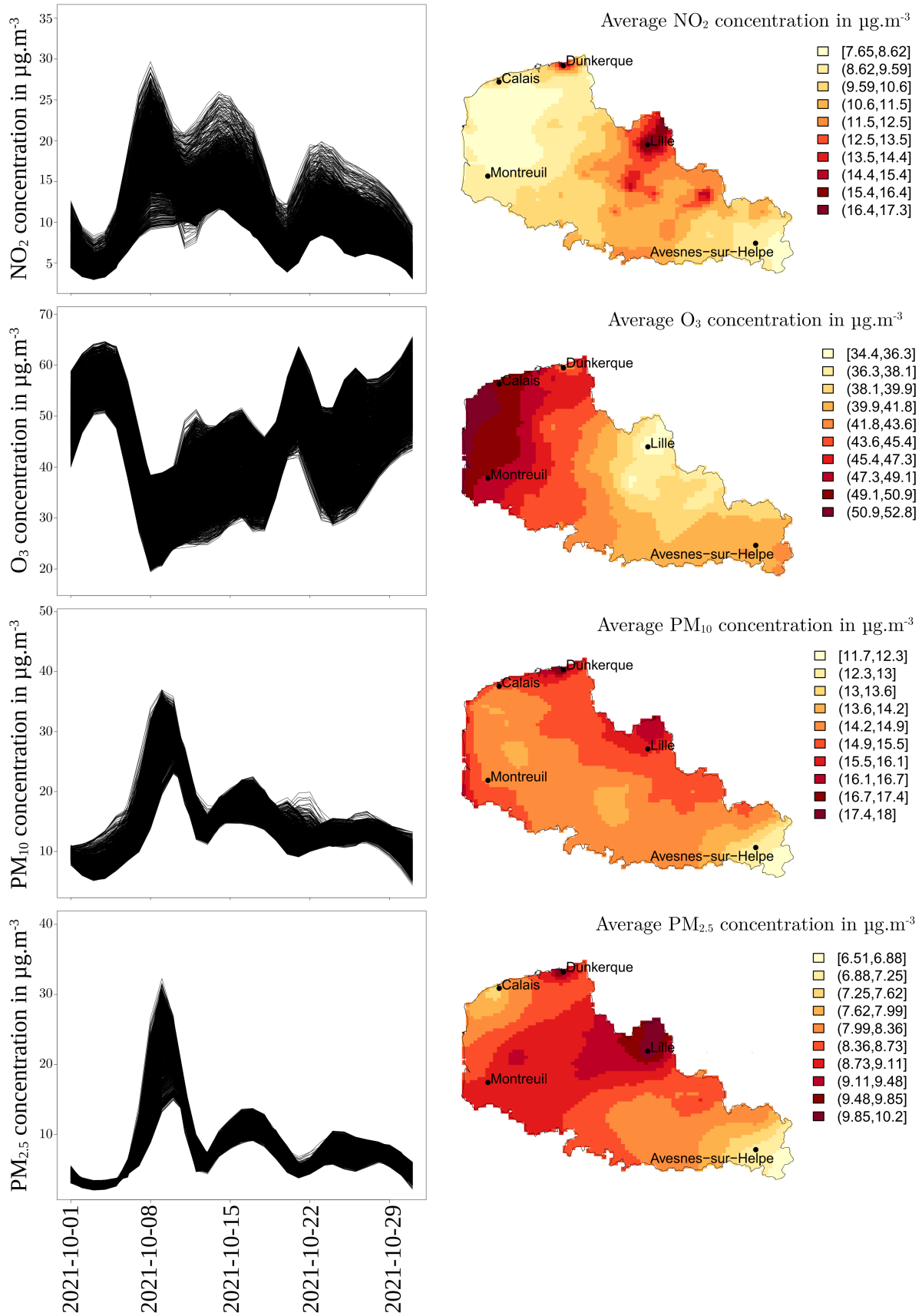
Figure 4.9: Daily smoothed concentration curves of NO_2, O_3, PM_{10} and $PM_{2.5}$ (from October 1 to October 31, 2021) in the *Nord-Pas-de-Calais* region of northern France at a spatial resolution of approximately 2 km by 2 km (left panels), together with the spatial distributions of the average smoothed concentrations for each pollutant over the period (right panels) using a nonparametric smoothing.
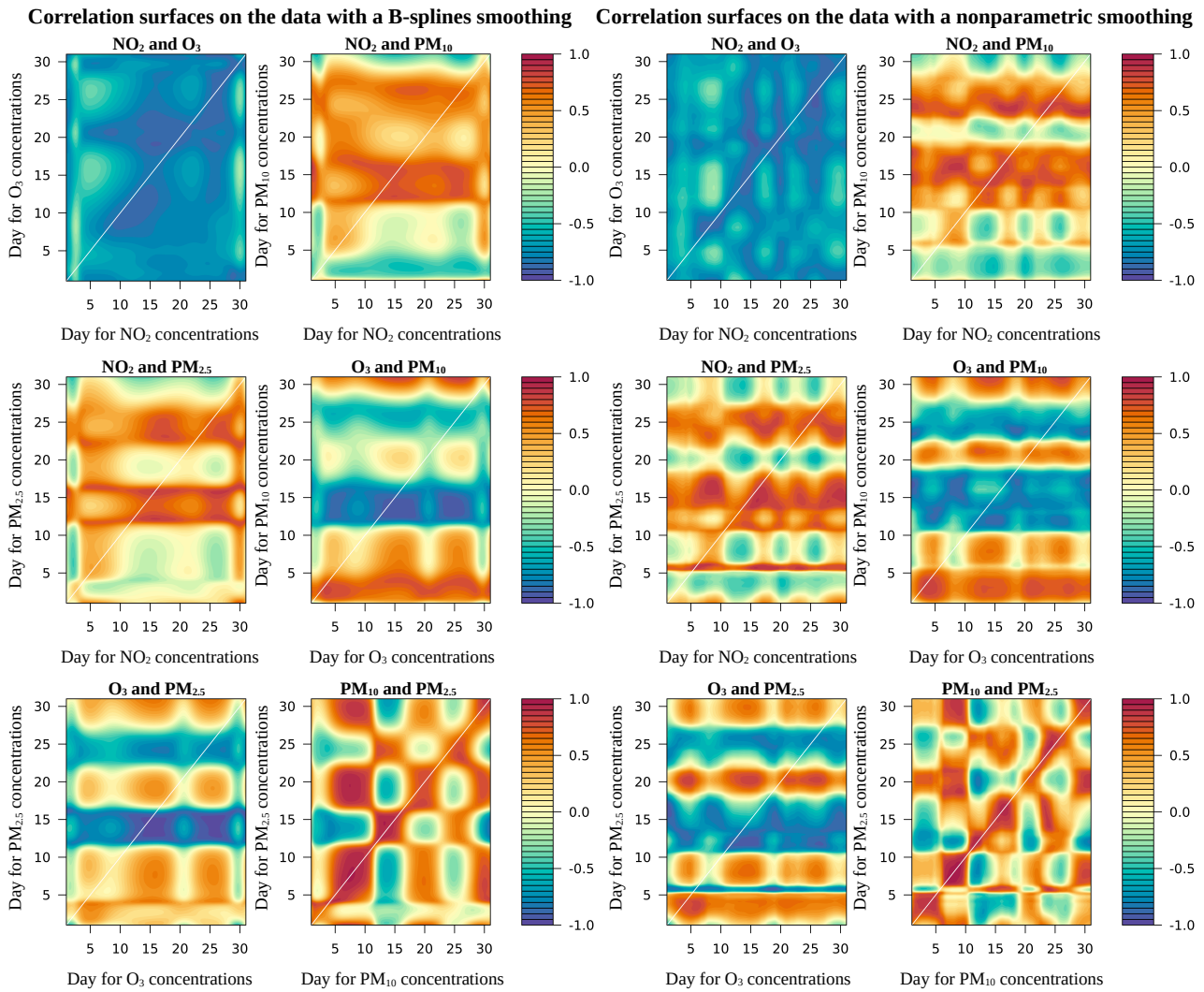
Figure 4.10: Correlation surfaces for the pollutant data from October 1 to October 31, 2021, after B-spline smoothing (left panels) and nonparametric smoothing (right panels). A white diagonal line indicates the same observation time for the two pollutants.
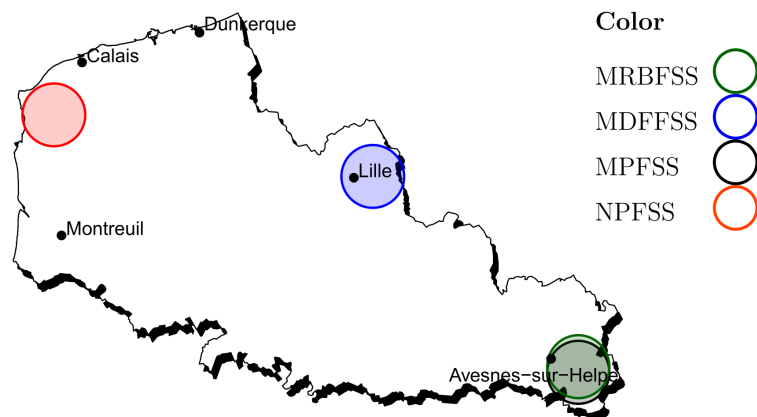
Figure 4.11: Pollutant MLCs (NO_2, O_3, PM_{10} and $PM_{2.5}$) detected by the MRBFSS, the NPFSS, the MDFFSS and the MPFSS, after nonparametric smoothing of the pollutant concentration data.
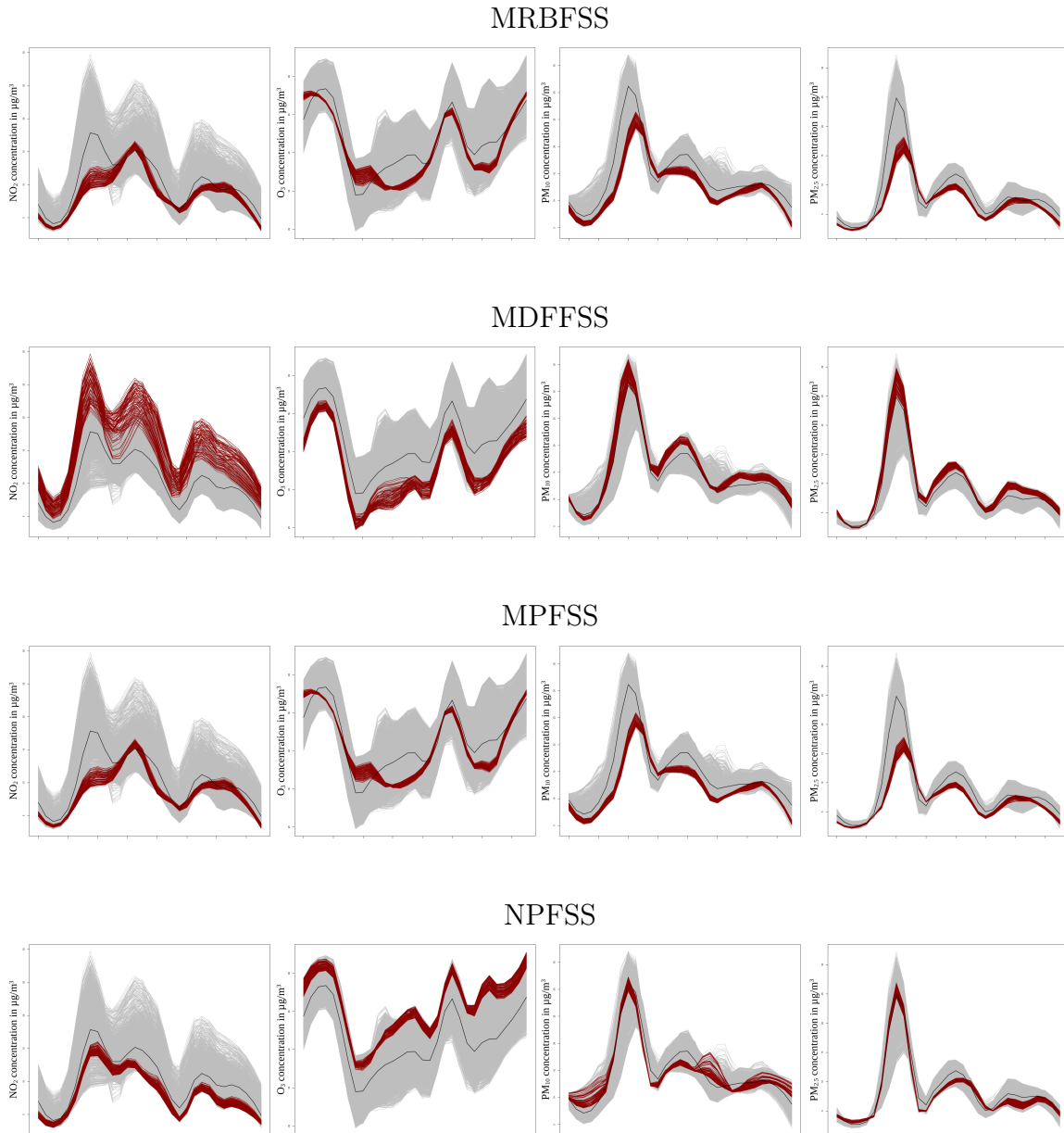
Figure 4.12: The daily concentration curves (after nonparametric smoothing) for the pollutants (NO_2, O_3, PM_{10} and $PM_{2.5}$) from October 1 to October 31, 2021, in each 2 km by 2 km grid cell within the clusters detected by the MRBFSS, the NPFSS, the MDFFSS and the MPFSS (colored lines). The black curves correspond to the daily average concentration curves in the *Nord-Pas-de-Calais* region.

pollutant formed from primary pollutants (particularly nitrogen oxides) in photochemical reactions (Richards, 1956). Levels of O_3 are higher in peri-urban and rural areas than in urban areas, which agrees with the high concentrations of O_3 observed in this rural cluster and the low concentrations observed in the urban cluster of Lille (Figure 4.12). However, it should be noted that the O_3 concentrations in the rural cluster are much lower than the daily threshold of $100 \mu g.m^{-3}$ given by the World Health Organization (World Health Organization (WHO)); hence, one cannot consider that this cluster is affected by O_3 pollution. The cluster also contains France's largest quarry, which light account for the moderate PM_{10} concentrations.

The difference between the MLCs detected by the various statistics might appear surprising. However, the difference can be explained by the secondary clusters, as set out in Section 2.6.

Table 4.1 shows that the MLCs detected by a given method are secondary clusters ($\hat{p} = 0.001$) detected by the other methods.

Method	Avesnes-sur-Helpe	Lille	Southwest of Calais
MRBFSS	MLC	Secondary cluster 1	Secondary cluster 3
MPFSS	MLC	Secondary cluster 1	Secondary cluster 3
MDFFSS	Secondary cluster 2	MLC	Secondary cluster 3
NPFSS	Secondary cluster 5	Secondary cluster 1	MLC

Table 4.1: MLCs and secondary clusters ($\hat{p} = 0.001$) for the four methods (MRBFSS, MPFSS, MDFFSS and NPFSS) in the Avesnes-sur-Helpe area, the Lille area, and the southwest of Calais.

4.4   Results for the data from October 1 to October 31, 2021, after cubic B-spline smoothing

The detected clusters are shown in Figures 4.13 and 4.14.
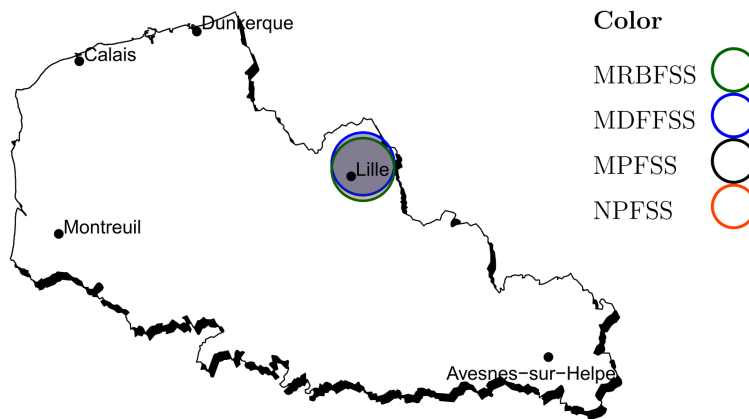
Figure 4.13: Pollutant MLCs (NO_2, O_3, PM_{10} and $PM_{2.5}$) detected by the MRBFSS, the NPFSS, the MDFFSS and the MPFSS after cubic B-spline smoothing.

After cubic B-spline smoothing (which is stronger than nonparametric smoothing), all methods detect the same MLC: the cluster in the Lille area ($\hat{p} = 0.001$) was characterized by high values of NO_2, PM_{10} and $PM_{2.5}$, and low values of O_3. As explained above, this is consistent with the literature data; Lille is a highly urban area, and levels of NO_2, PM_{10} and $PM_{2.5}$ are usually very high in such areas.

5   Discussion

By developing an MPFSS, an MDFFSS, and an MRBFSS, we were able to detect spatial clusters of abnormal values in multivariate functional data indexed in space. Our analysis took account of all the available time-domain information and possible correlations between the variables. In the context of environmental monitoring, our objective was to enable the detection of abnormal pollutant values so that alerts can be issued in areas with exposure to several pollutants. The detection of areas with excessive exposure to pollutants is of major importance in public health; the adverse health impact and synergistic effects of air pollutants
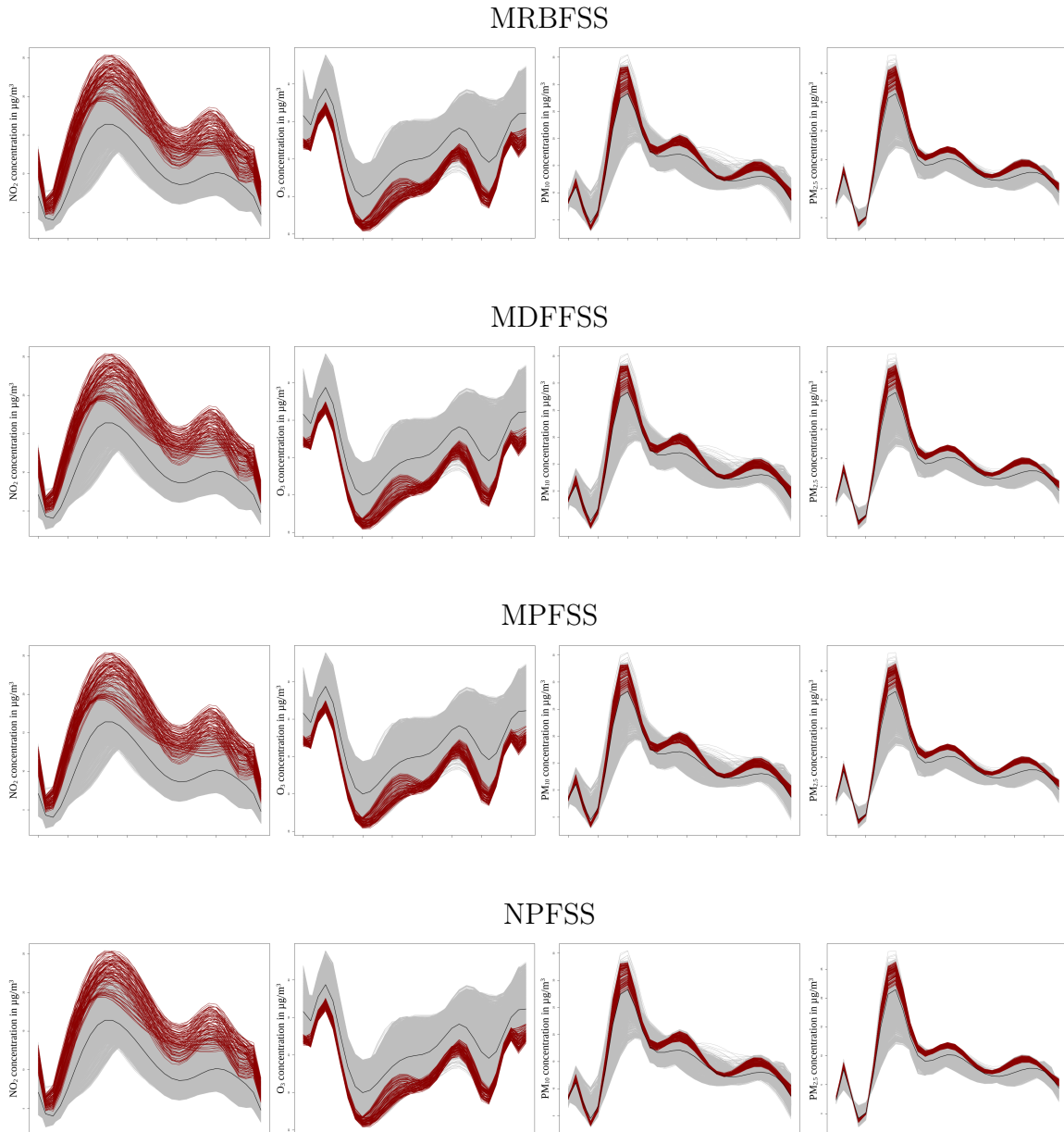
Figure 4.14: Smoothed daily concentration curves (using cubic B-splines) for NO_2, O_3, PM_{10} and $PM_{2.5}$ from October 1 to October 31, 2021, in each 2 km by 2 km grid cell within the clusters detected by the MRBFSS, NPFSS, MDFFSS and MPFSS (colored lines). The black curves correspond to the daily average concentration curves for the *Nord-Pas-de-Calais* region as a whole.

(Huang et al., 2022) are well known and have prompted a large number of field studies and modelling studies (Shaddick and Wakefield, 2002; Finazzi et al., 2013).

Although Smida et al. (2022) developed the NPFSS approach in a univariate functional framework, this statistic can be extended to the multivariate functional case. We compared the MDFFSS, the MPFSS, the MRBFSS, and the NPFSS in a simulation study. The results showed that (i) the performance levels of all methods decreased as the correlation between the variables increased, and (ii) smoothing the data improved the performance level for the pointwise approaches (the MDFFSS and MRBFSS). The MRBFSS was more powerful than the MPFSS and the NPFSS, regardless of the data distribution and the correlation between the variables. The same was true for the MDFFSS with a Gaussian distribution. The MPFSS

and the MDFFSS gave the lowest false positive rates. However, the MRBFSS often gave the highest true positive rate and the MPFSS often gave the lowest true positive rate; this resulted in very high F-measures for the MRBFSS and the MDFFSS. When the data distribution was far from normal, the MPFSS and the MDFFSS performed less well but still gave very low false positive rates. For a local shift, the pointwise approaches (the MDFFSS and MRBFSS) performed better than the NPFSS and the MPFSS in terms of the power, the true positive rate, the false positive rate, and the F-measures.

Next, we used the four approaches to detect clusters of abnormal pollutant concentration values in the *Nord-Pas-de-Calais* at a spatial resolution of 2 km by 2 km. First, we have smoothed the data slightly in order to reach the pollution peaks. Both the MRBFSS and the MPFSS detected a statistically significant MLC with low pollutant levels for NO_2, PM_{10} and $PM_{2.5}$ in the rural area of Avesnes-sur-Helpe. The MDFFSS detected a statistically significant MLC with high pollutant levels for NO_2, PM_{10} and $PM_{2.5}$ in the Lille urban area. This is consistent with the literature data because levels of PM_{10} and especially NO_2 and $PM_{2.5}$ tend to be higher in urban areas. The NPFSS detected a statistically significant MLC (characterized by low NO_2 and high O_3 concentrations) in a rural area to the southwest of Calais. Since solar radiation is generally more intense in coastal areas, this increases the formation of O_3 and leads to higher concentrations of this pollutant.
Although the MLCs differed from one method to another, it should be noted that each cluster was statistically significant in all four approaches. Greater smoothing "erased" the pollution peaks and led all four methods to detect a statistically significant MLC in the Lille area.

It should be noted that we considered circular clusters only. On pollutant maps, the high mean concentration areas have an elongated shape (especially for O_3 on the coastline), which suggests that other shapes of clusters might be relevant. By way of an example, Tango and Takahashi (2005) assessed irregularly shaped clusters with a predetermined maximum size by considering all sets of sites connected to each other. However it should be noted that this approach generated many more potential clusters (and therefore required much more computing time) than the approach developed by Kulldorff (1997). The same disadvantage applies to the elliptical cluster approach developed by Kulldorff et al. (2006). However, these problems can be overcome with the graph-based clusters developed by Cucala et al. (2013). Another possible approach was developed by Lin et al. (2016); the researchers suggested group the circular clusters together to form clusters with arbitrary shapes.

It should also be noted that in this type of approach, the variables must be measured on the same spatial scale. However, this is not always possible in practice. If some variables have a lattice profile and others have a geostatistical profile, the change of support problem (Gelfand et al., 2001) arises. In some cases, the spatial units correspond to the regions of a country; the latter sometimes change for administrative or political reasons. In such a case, the modified areal unit problem (Larsen, 2000) arises. Furthermore, some variables may be observed on a finer scale than others. In such a case, there are two possible solutions. The first consists in aggregating some variables until an identical spatial scale is obtained for all the variables. However this upscaling technique (Bierkens et al., 2000) causes information to be lost. Secondly, the spatial scale can be reduced by downscaling (Bierkens et al., 2000; Shaddick et al., 2018) with (for example) kriging methods. The latter can be used to predict geostatistical data at sites at which they are lacking (Giraldo et al., 2011).

It should be borne in mind that our functional approaches do not have the same objective as Kulldorff's spatiotemporal approach (Kulldorff et al., 1998). In fact, our goal here was to detect spatial clusters by taking into account the information available over the entire

observation period. In contrast, Kulldorff et al. (1998) sought to identify spatiotemporal clusters, i.e. geographical areas in which something atypical happens during a time subinterval.

The spatial scan statistics presented here took account of possible correlations between variables. However, we assumed that the observations at the various spatial locations were independent; this is a very strong assumption, albeit a classical one in spatial scan statistics. For pollution data, it is not surprising that sites located close to each other tend to have similar pollutant concentration values. Several research studies have tried to take account of possible spatial dependence in scan statistics (Loh and Zhu, 2007; Lin, 2014; Lee et al., 2019; Ahmed et al., 2021b). However, these studies are quite recent and were applied to rather simple models. It would be very useful to be able to take account of possible spatial dependence in the spatial scan statistics developed for multivariate functional data. This would be quite challenging, however, given the complexity of the data and the models and possible differences in spatial autocorrelation from one variable to another.

Chapter 5

Spatial scan statistics for survival data

Contents

1 Introduction

In the field of spatial epidemiology, the study of the spatial distribution of time-to-event data can identify areas in which the survival time of patients is different from the rest of the geographical area (i.e. the survival time is longer or shorter). From an epidemiological point of view, the identification of these areas of unusual survival time is particularly useful for identifying local risk factors that condition survival. Moreover, this information can help public health decision-makers to develop and implement targeted, specific local policies. In the context of spatial cluster detection in time-to-event data, Huang et al. (2007) and Bhatt and Tiwari (2014) developed spatial scan statistics based on an exponential model and a Weibull model, respectively. More recently, Usman and Rosychuk (2018) developed a parametric model considering a log-Weibull distribution. Although these methods are widely used in practice to detect spatial clusters of time-to-event data (Gregorio et al., 2007; Henry et al., 2009; Wan et al., 2012), they are totally parametric. A first semi-parametric method (using a Cox model) was developed by Cook et al. (2007).

Unlike other spatial scan statistics models, the above-mentioned exponential, Weibull, log-Weibull and Cox models consider data measured at the individual level. However,

in health data studies, the patient's exact geographic location is rarely known (e.g. for reasons of anonymity), and patients are located through an administrative spatial unit (e.g. municipalities). In this context, the above-mentioned methods are based on the strong assumption of independence between observations - a classical assumption in the field of spatial scan statistics. This assumption is associated with two major drawbacks. Firstly, the methods do not take account of the potential correlation between the survival times of the individuals within the same spatial unit, namely the intra-spatial unit correlation. The latter can be induced by characteristics of the spatial units that have not been measured in the study (e.g. healthcare supply) but that affect the patients' survival (Austin, 2017). Secondly, the methods do not take account of potential spatial dependence between spatial units. However, one can logically expect geographically close units to be more strongly dependent than distant ones (Li, 2009). Furthermore, it has been shown that ignoring spatial correlation when using spatial scan statistics leads to a significant increase in the type I error (Loh and Zhu, 2007). Since then, several researchers have developed methods that take spatial correlations into account (Loh and Zhu, 2007; Lin, 2014; Lee et al., 2019; Ahmed et al., 2021b). However, none of these methods was designed for time-to-event data and the adjustment for intra-spatial unit correlations.

In the analysis of time-to-event data, various models have been developed to take account of unobserved factors common to groups of individuals; for example, members of the same family share genetic factors, and patients in the same hospital often receive much the same care. One way of taking this intra-group homogeneity into account involves introducing a random effect common to all individuals in a group, namely shared frailty (Clayton, 1978; Liang et al., 1995; Hougaard, 2000). The shared frailties are assumed to be independent between groups (Liang et al., 1995). However, when the groups correspond to spatial units, this assumption is unrealistic because close spatial units tend to be dependent (Arlinghaus, 1995). To this end, Li and Ryan (2002) extended shared frailty models to the case of spatially correlated frailty, which take account of not only intra-spatial unit correlation but also possible spatial dependence between spatial units. This approach has since been widely applied (Banerjee et al., 2003; Ojiambo and Kang, 2013; Aswi et al., 2020).

Here, we present a new spatial scan statistic for time-to-event data based on a Bayesian semi-parametric Cox model with spatially correlated shared frailties. Section 2 describes the methodological aspects of the scan statistic model. Section 3 presents both the design and the results of simulation studies evaluating (i) the performance of conventional methods on data sets with intra-spatial unit correlation and (ii) the performance of our approach on data sets with both intra-spatial unit correlation and spatial dependence between spatial units. Section 4 describes the application of our method to epidemiological data and the detection of spatial clusters of mortality in patients with end-stage renal disease in northern France. Lastly, we discuss of results in Section 5.

2 Methodology

2.1 General principle

Let us consider K non-overlapping spatial locations $s_1, \ldots, s_k, \ldots, s_K$ of an observation domain $S \subset \mathbb{R}^2$ and let $i_1^{(k)}, \ldots, i_n^{(k)}, \ldots, i_{N_k}^{(k)}$ be N_k individuals at spatial location s_k. The total number of individuals in S is defined as $N = \sum_{k=1}^{K} N_k$. Here, we are interested in the time-to-event data measured on individuals: $T_{i_n^{(k)}}$ and $\delta_{i_n^{(k)}}$ are respectively the observation time of the i_nth individual in spatial location s_k and the censoring indicator, which is equal to 0 if the individual $i_n^{(k)}$ is censored and 1 otherwise. In the following, we only considered the cases of right censoring (i.e., the event of interest could not have occurred before the beginning of the study). Censoring

was assumed to be uninformative, and the event times were assumed to be independent of the censoring times.

We sought to test for the presence of spatial clusters in which individuals have shorter (or longer) survival times than other individuals in the rest of S. In this context, spatial scan statistics are designed to detect spatial clusters and to test their statistical significance by testing 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 time-to-event values). According to Cressie (1977), a spatial scan statistic is the maximum of a concentration index over a set of potential clusters $\mathcal{W}$. In the following and without loss of generality, we focused on variable-size circular clusters. Hence, in line with Kulldorff (1997), the set of potential circular clusters $\mathcal{W}$ can be defined as: $\mathcal{W} = \{w_{k,l}/1 \leq |w_{k,l}| \leq \frac{N}{2}, 1 \leq k, l \leq K\}$, where $w_{k,l}$ is the disc centered on s_k that passes through s_l and $|w_{k,l}|$ is the number of individuals in $w_{k,l}$: a cluster comprises at most 50% of the study population (i.e., $N/2$) (Kulldorff and Nagarwalla, 1995). It should be noted that other cluster shapes have been described in the literature, such as elliptical clusters (Kulldorff et al., 2006), rectangular clusters (Chen and Glaz, 2009) or arbitrarily shaped clusters (Tango and Takahashi, 2005; Zhou et al., 2015; Yin and Mu, 2018).

2.2   The model

We assume that the instantaneous hazard rate at time t for the individual $i_n^{(k)}$ is

$$\lambda_{i_n^{(k)}}(t|\mathbf{Z}_{i_n^{(k)}}, \varphi_k) = \lambda_0(t) \exp\left[\boldsymbol{\beta}^\top \mathbf{Z}_{i_n^{(k)}} + \varphi_k\right],$$

where $\mathbf{Z}_{i_n^{(k)}} = (Z_{i_n^{(k)},1}, \ldots, Z_{i_n^{(k)},p})^\top$ is a vector of p covariates associated with the individual $i_n^{(k)}$, and φ_k is the shared frailty associated with the spatial location s_k. The presence of a spatial cluster in the data results in an effect on the survival times in the spatial units involved. Hence, the effect of this cluster has been incorporated within the shared frailty: for each potential cluster w, φ_k can be decomposed into α_w (a cluster effect) and X_k (effect specific to the spatial location s_k). Thus, the shared frailties φ_k associated with the potential cluster w can be rewritten as $\varphi_k^{(w)} = \alpha_w \mathbb{1}_{s_k \in w} + X_k$ where $\mathbb{E}[X_k] = 0$. In this context, the test hypotheses can be rewritten as $\mathcal{H}_0 : \forall w \in \mathcal{W}, \alpha_w = 0$ (the absence of a cluster), and the alternative hypothesis associated with the potential cluster w is $\mathcal{H}_1^{(w)} : \alpha_w \neq 0$ (the presence of a cluster w, in which the individuals present atypical survival times).

Moreover, the spatial nature of the data requires one to take account of a possible spatial dependence between the spatial locations s_k, and thus a possible correlation between the X_k. This makes it possible to distinguish the effect of the cluster from the spatial correlation of unobserved factors on the scale of the spatial unit. Thus, we considered the conditional autoregressive (CAR) model developed by Leroux et al. (2000) for the distribution of the X_k:

$$X_k | X_{-k} \sim \mathcal{N}\left(\frac{\rho \sum_{l=1}^{K} v_{k,l} X_l}{\rho \sum_{l=1}^{K} v_{k,l} + 1 - \rho}, \frac{\sigma_X^2}{\rho \sum_{l=1}^{K} v_{k,l} + 1 - \rho}\right)$$

where $X_{-k} = \{X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K\}$, $v_{k,l} = 1$ if s_k and s_l are adjacent (i.e. they share a common boundary) and 0 if not, and $\rho \in [0, 1]$ is the spatial correlation parameter. It should be noted that in the absence of a spatial correlation, the X_k are independent and identically distributed (i.i.d.) according to a normal distribution $\mathcal{N}(0, \sigma_X^2)$. Conversely, if $\rho = 1$ (i.e., complete spatial correlation between the spatial units), the X_k are distributed according to an intrinsic CAR (ICAR) model (Besag et al., 1991).

The method comprises two steps. The first step (Section 2.2.1) consists in estimating the shared frailties φ_k and their spatial correlation parameter ρ. In a second step (Section 2.2.2), a scan procedure is developed and applied to the estimates of the shared frailties, in order to identify clusters of spatial units in which the φ_k are significantly higher (corresponding to a higher risk) or significantly lower (corresponding to a lower risk) than elsewhere. Lastly, the procedure for determining the statistical significance of the identified spatial clusters is described in Section 2.2.3.

2.2.1   Estimation of the φ_k and ρ

This first step consists in estimating the φ_k and ρ in a Bayesian framework by using the integrated nested Laplace approximation (INLA) (see Rue et al. (2009) for details).

The φ_k are considered under both the $\mathcal{H}_0$ and $\mathcal{H}_1$ hypotheses. However, it should be noted that (i) neither X_k nor ρ depend on the clustering assumptions, since they depend only on the spatial structure of the data, and (ii) only a single vector of φ_k needs to be estimated in order to best fit the observed data. Therefore, the φ_k must be estimated under the true hypothesis among $\mathcal{H}_0$ and the set of alternative hypotheses $\mathcal{H}_1^{(w)}$, i.e., the hypothesis under which the observations have been generated. In line with the approach developed by Ahmed et al. (2021b), we need to determine the "best model" among the candidate hypotheses ($\mathcal{H}_0$ and $\mathcal{H}_1^{(w)}$). To this end, for each potential cluster $w \in \mathcal{W}$, we considered the Bayes factor $\mathrm{BF}^{(w)}$, defined as the marginal likelihood ratio between the model under $\mathcal{H}_1^{(w)}$ ($\mathcal{M}_1^{(w)}$) and the model under $\mathcal{H}_0$ ($\mathcal{M}_0$):

$$\mathrm{BF}^{(w)} = \frac{\mathbb{P}\left[\left\{T_{i_n^{(k)}}, \delta_{i_n^{(k)}}, \boldsymbol{Z}_{i_n^{(k)}}, \mathbb{1}_{i_n^{(k)} \in w}\right\} \middle| \mathcal{M}_1^{(w)}\right]}{\mathbb{P}\left[\left\{T_{i_n^{(k)}}, \delta_{i_n^{(k)}}, \boldsymbol{Z}_{i_n^{(k)}}\right\} \middle| \mathcal{M}_0\right]}.$$

Next, considering all the models under $\mathcal{H}_1^{(w)}$ we used the above criterion to select the "best model" $\mathcal{M}_1^{(w^*)}$, i.e., the one associated with the potential cluster w maximizing $\mathrm{BF}^{(w)}$. Lastly, to decide whether the estimates should be kept under $\mathcal{H}_0$ or under $\mathcal{H}_1^{(w^*)}$, we followed the rule of thumb developed by Jeffreys (1961): if $\mathrm{BF}^{(w^*)} \geq 30$, then we keep the estimates (using the posterior mean) under $\mathcal{H}_1^{(w^*)}$; otherwise, we keep the estimates (using the posterior mean) under $\mathcal{H}_0$. This threshold of 30 corresponds to "very strong" evidence for $\mathcal{H}_1^{(w^*)}$. Note that if the selected model is $\mathcal{M}_1^{(w^*)}$, the chosen estimate of φ_k is $\varphi_k^* = \hat{\alpha}_{w^*}\mathbb{1}_{s_k \in w^*} + \hat{X}_k$ and if the selected model is $\mathcal{M}_0$, $\varphi_k^* = \hat{X}_k$.

2.2.2   Scan procedure

Here, we present a scan procedure on the φ_k^* that identifies spatial clusters of spatial units in which the φ_k^* are significantly higher (corresponding to a higher risk) or significantly lower (corresponding to a lower risk) than elsewhere. Thus, the hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1^{(w)}$ are redefined in terms of the distribution of the φ_k^*, as follows:

$$\mathcal{H}_0 : \boldsymbol{\varphi}^* \sim \mathcal{N}(\alpha\mathbb{1}, \sigma^{2(0)}A^{-1}) \text{ and}$$

$$\mathcal{H}_1^{(w)} : \boldsymbol{\varphi}^* \sim \mathcal{N}(\alpha_w\mathbb{1}_w + \alpha_{w^c}\mathbb{1}_{w^c}, \sigma^{2(w)}A^{-1}), \ \alpha_w \neq \alpha_{w^c}$$

where $\boldsymbol{\varphi}^* = (\varphi_1^*, \ldots, \varphi_K^*)^\top$, $\mathbb{1}$ is the column vector composed only of 1, $\mathbb{1}_w$ and $\mathbb{1}_{w^c}$ are the column indicator vectors of w and w^c respectively, and $A = \rho^*R + (1 - \rho^*)\mathrm{I}_K$ in which R is the square matrix composed of the elements

$$R_{k,l} = \begin{cases} \sum_{j=1}^{K} v_{k,j} & \text{if } k = l \\ -v_{k,l} & \text{otherwise} \end{cases}.$$

Note that these assumptions are equivalent to considering the same variance-covariance structure under $\mathcal{H}_0$ and $\mathcal{H}_1^{(w)}$, as with the CAR model considered above (see the proof in Section 1 of Appendix C). Since $w \cap w^{\mathsf{c}} = \emptyset$, one assumes under $\mathcal{H}_1^{(w)}$ that the frailty means in w and w^{c} are different (α_w and $\alpha_{w^{\mathsf{c}}}$, respectively).

The unknown parameters α, $\sigma^{2(0)}$, α_w, $\alpha_{w^{\mathsf{c}}}$ and $\sigma^{2(w)}$ are estimated by their maximum likelihood estimators (for proofs, see Appendix C):

$$\hat{\alpha} = \frac{\mathbb{1}^{\top} A \boldsymbol{\varphi}^*}{\mathbb{1}^{\top} A \mathbb{1}},$$

$$\widehat{\sigma^{2(0)}} = \frac{1}{K}[\boldsymbol{\varphi}^{*\top} A \boldsymbol{\varphi}^* - 2\hat{\alpha} \mathbb{1}^{\top} A \boldsymbol{\varphi}^* + \hat{\alpha}^2 \mathbb{1}^{\top} A \mathbb{1}],$$

$$\hat{\alpha}_{w^{\mathsf{c}}} = \left[\mathbb{1}_{w^{\mathsf{c}}}^{\top} A \mathbb{1}_{w^{\mathsf{c}}} - \frac{\mathbb{1}_w^{\top} A \mathbb{1}_{w^{\mathsf{c}}} \mathbb{1}_w^{\top} A \mathbb{1}_{w^{\mathsf{c}}}}{\mathbb{1}_w^{\top} A \mathbb{1}_w}\right]^{-1} \left[\mathbb{1}_{w^{\mathsf{c}}}^{\top} A \boldsymbol{\varphi}^* - \frac{\mathbb{1}_w^{\top} A \boldsymbol{\varphi}^* \mathbb{1}_w^{\top} A \mathbb{1}_{w^{\mathsf{c}}}}{\mathbb{1}_w^{\top} A \mathbb{1}_w}\right],$$

$$\hat{\alpha}_w = \frac{\mathbb{1}_w^{\top} A \boldsymbol{\varphi}^* - \hat{\alpha}_{w^{\mathsf{c}}} \mathbb{1}_w^{\top} A \mathbb{1}_{w^{\mathsf{c}}}}{\mathbb{1}_w^{\top} A \mathbb{1}_w} \quad \text{and}$$

$$\widehat{\sigma^{2(w)}} = \frac{1}{K}[\boldsymbol{\varphi}^* - \hat{\alpha}_w \mathbb{1}_w - \hat{\alpha}_{w^{\mathsf{c}}} \mathbb{1}_{w^{\mathsf{c}}}]^{\top} A [\boldsymbol{\varphi}^* - \hat{\alpha}_w \mathbb{1}_w - \hat{\alpha}_{w^{\mathsf{c}}} \mathbb{1}_{w^{\mathsf{c}}}].$$

The log-likelihood function under $\mathcal{H}_0$ is then expressed as follows:

$$\ell_{\mathcal{H}_0}(\hat{\alpha}, \widehat{\sigma^{2(0)}}) = -\frac{K}{2} \ln[2\pi] - \frac{1}{2} \ln|A^{-1}| - \frac{K}{2} \ln[\widehat{\sigma^{2(0)}}] - \frac{K}{2},$$

while the log-likelihood function associated with $\mathcal{H}_1^{(w)}$ can be expressed as:

$$\ell_{\mathcal{H}_1}(\hat{\alpha}_w, \hat{\alpha}_{w^{\mathsf{c}}}, \widehat{\sigma^{2(w)}}) = -\frac{K}{2} \ln[2\pi] - \frac{1}{2} \ln|A^{-1}| - \frac{K}{2} \ln[\widehat{\sigma^{2(w)}}] - \frac{K}{2}.$$

Thus, the log-likelihood ratio associated with the potential cluster w is

$$\begin{aligned} LLR^{(w)} &= \ell_{\mathcal{H}_1}(\hat{\alpha}_w, \hat{\alpha}_{w^{\mathsf{c}}}, \widehat{\sigma^{2(w)}}) - \ell_{\mathcal{H}_0}(\hat{\alpha}, \widehat{\sigma^{2(0)}}) \\ &= \frac{K}{2}\left[\ln \frac{\widehat{\sigma^{2(0)}}}{\widehat{\sigma^{2(w)}}}\right]. \end{aligned}$$

Lastly, the spatial scan statistic can be defined as

$$\Lambda = \max_{w \in \mathcal{W}} LLR^{(w)}.$$

The most likely cluster (MLC) is then defined as

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

2.2.3 Statistical significance

Once the MLC has been detected, its statistical significance must be evaluated. However, the distribution of Λ does not have a closed form under $\mathcal{H}_0$. In the literature, this distribution is usually approximated with a Monte-Carlo procedure (Dwass, 1957). Two main methods can be distinguished, depending on the presence (or not) of a distributional hypothesis for the data. The first method consists in generating data sets under $\mathcal{H}_0$, which thus requires a distributional hypothesis (Kulldorff, 1997). The second method (random labelling) consists in randomly permuting the observations among the spatial locations (Kulldorff et al., 2009). In the present case, random labelling is not applicable because the permutation of the observations would change the spatial correlations. Therefore, we used the first method to approximate the

distribution of Λ under $\mathcal{H}_0$: since we had assumed a distribution for φ_k, we can generate M data sets under $\mathcal{H}_0$ *via* $\hat{\alpha}$ and $\widehat{\sigma^{2(0)}}$, which correspond respectively to the estimators of the mean and variance of the φ_k under $\mathcal{H}_0$. For each generated data set m ($1 \leq m \leq M$), one computes the associated spatial scan statistic $\Lambda^{(m)}$, giving an approximation of the distribution of Λ under $\mathcal{H}_0$. Lastly, the p-value associated with the MLC is estimated as

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

3 Simulation studies

Huang et al. (2007) and Cook et al. (2007) developed spatial scan statistics for time-to-event data indexed in space. However, they supposed that the individuals are independent, which is a strong and not very realistic hypothesis because the observations of individuals located in the same spatial unit can be correlated. Thus, in a first simulation study (Section 3.1), we investigated the impact of the presence of intra-spatial unit correlation on the type I error of the methods developed by Huang et al. (2007) and Cook et al. (2007).

In Section 3.2, we conducted two simulation studies. The first one (Section 3.2.2) evaluated our method's ability to correctly estimate both the spatial correlation parameter and the cluster effect. The second (Section 3.2.3) evaluated the performance of our approach in the context of cluster detection, with spatial correlation and also compared it with the particular i.i.d. ($\rho = 0$) and ICAR ($\rho = 1$) versions.

Lastly, Section 3.3 investigates the performance of our approach in the presence of different levels of censoring of time-to-event data.

3.1 The impact of intra-spatial unit correlation on the type I error in standard methods

In this simulation study, we evaluated the type I errors of conventional spatial scan statistics for cluster detection in survival data (namely the exponential model developed by Huang et al. (2007) and the method based on a log-rank test developed by Cook et al. (2007)), in the presence of intra-spatial unit correlation.

3.1.1 Design of the simulation study

We considered 1690 individuals distributed in 169 spatial units; the latter corresponded to administrative sub-divisions in northern France and were located by their centroid. We defined a spatial cluster w (characterized by α) composed of 135 individuals located in 14 contiguous spatial units (the green area in Figure 5.1).

We considered the following simulation model for the individual $i_n^{(k)}$ in the spatial unit s_k:

$$\lambda_{i_n^{(k)}}(t|\varphi_k) = \lambda_0(t) \exp\left[\varphi_k\right],$$

with $\lambda_0(t) = \frac{1}{2}$ which results in an exponential model. The event times were simulated by inverse transform sampling: for each individual $i_n^{(k)}$, we generated a uniformly distributed random number $u_{i_n^{(k)}}$ on $[0, 1]$, which then allowed us to generate a survival time $T_{i_n^{(k)}}$ by

$$T_{i_n^{(k)}} = \inf_{t > 0} 1 - S_{i_n^{(k)}}(t) > u_{i_n^{(k)}}.$$
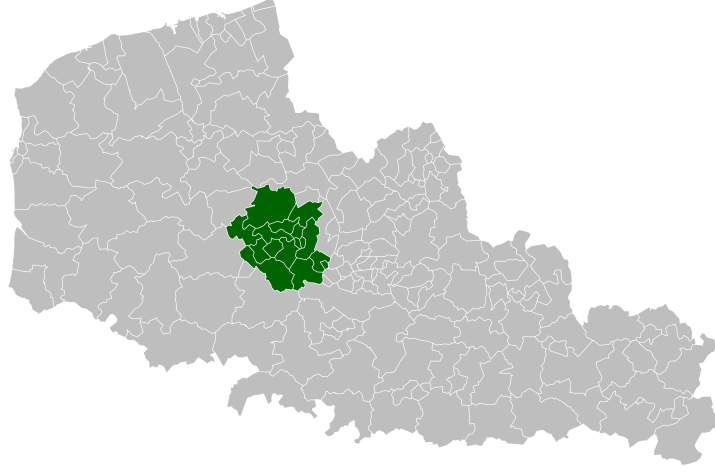
Figure 5.1: Simulated cluster (in green) in 169 administrative subdivisions of northern France.

Note that this results in $T_{i_n^{(k)}} = -2\ln\left[1 - u_{i_n^{(k)}}\right]\exp\left[-\varphi_k\right]$.

The φ_k were defined as the vector $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_K)^\top$, such that

$$\boldsymbol{\varphi} \sim \mathcal{N}(\alpha\mathbb{1}_w, \sigma^2[\rho R + (1-\rho)I_K]^{-1}),$$

where $\mathbb{1}_w$ is the column indicator vector of w.

Here, we focused our analysis on the type I errors ($\alpha = 0$) in the exponential model (Huang et al., 2007) and the log-rank test method (Cook et al., 2007) in the presence of a non-spatially correlated ($\rho = 0$) shared frailty for the frailty variance σ^2, which ranged from 0.001 to 0.101 in increments of 0.010.

For each value of σ^2, 100 data sets were simulated. The statistical significance of the MLC was evaluated in the same way as in the original publications, i.e., by using 999 permutations of the data. The type I error was set to 0.05.

3.1.2   Results

Figure 5.2 shows the type I error as a function of σ^2. One can note that the type I error increases with σ^2, showing that the nominal level is not maintained. This can be explained by the fact that under hypothesis $\mathcal{H}_0$ (the absence of a cluster), the increase in σ^2 leads directly to an increase in the variance of X_k. Since the two standard models do not incorporate a shared frailty, the identification of false-positive spatial clusters is essentially due to the intra-spatial unit correlation (i.e., the variance of X_k).

3.2   Evaluation of the method's performance

Here, two simulation studies were conducted. The first (Section 3.2.2) assessed the ability of our method to accurately estimate both the spatial correlation parameter and the cluster effect. The second (Section 3.2.3) evaluated the performance of our method in the context of cluster detection and compared it with two particular versions of the model in the presence of spatial correlation: one assuming no spatial correlation (the i.i.d. frailty model) and one assuming complete spatial correlation (the ICAR frailty model).

3.2.1   Design of the simulation studies

The designs of these simulation studies are very similar to that presented in Section 3.1. However, given that we wanted to investigate the impact of spatial correlation on cluster
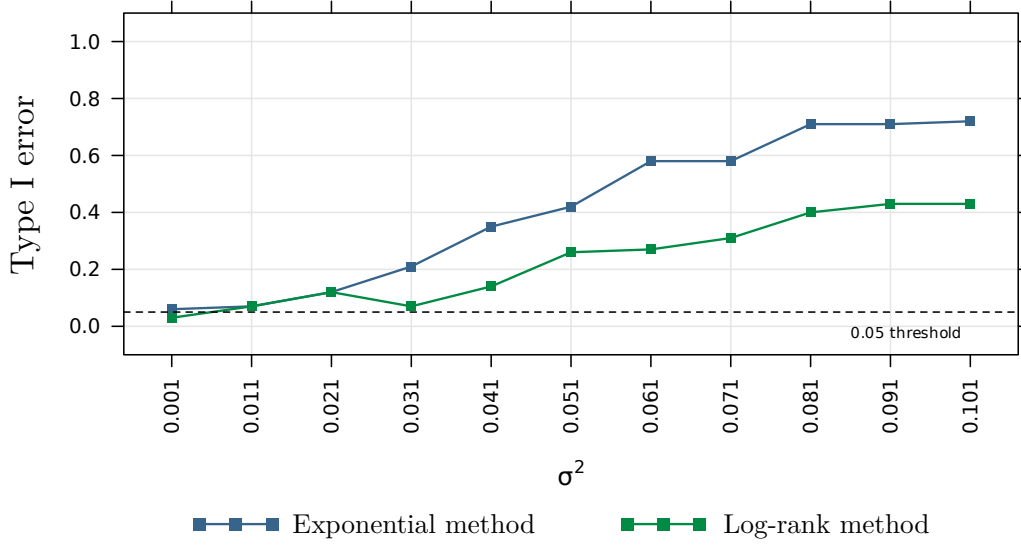
Figure 5.2: Type I error in the exponential method (Huang et al., 2007) and the log-rank test method (Cook et al., 2007), as a function of the degree of intra-spatial unit correlation (characterized by the simulated values of the shared frailty variance σ^2).

detection, we set σ^2 to 1 and considered several values for the parameters controlling the spatial correlation $\rho \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and the cluster effect $\alpha \in \{0, 0.5, 1, 1.5, 2\}$. Note that $\alpha = 0$ was considered in order to evaluate the maintenance of the type I error.

For each value of the spatial correlation parameter ρ and each value of α, 100 data sets were simulated. The statistical significance of the MLC was evaluated through 999 generations of the data under $\mathcal{H}_0$ (see Section 2.2.3 for more details), and the type I error was set to 0.05.

The performances were measured through four criteria: the power, the true positive rate, the false positive rate, and the positive predictive value. The power was estimated as the proportion of simulations leading to the rejection of $\mathcal{H}_0$, depending on the type I error. Using the simulated data sets leading to the rejection of $\mathcal{H}_0$, the true positive rate was defined as the mean proportion of individuals correctly detected among the individuals in w, the false positive rate was defined as the mean proportion of individuals in w^{c} that were included in the detected cluster, and the positive predictive value corresponded to the mean proportion of individuals in w within the detected cluster.

Since the estimations of the φ_k and ρ were performed in a Bayesian framework, we considered the following Leroux CAR prior for X_k: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2[\rho R + (1-\rho)I_K]^{-1})$, with a $\boldsymbol{\beta}(1,1)$ prior for the spatial correlation parameter ρ and a $\Gamma(10^{-3}, 10^{-3})$ prior for the precision $1/\sigma^2$. For α_w, we chose a non-informative prior $\mathcal{N}(0, 10^3)$.

Lastly, for the baseline hazard λ_0, the observation times were divided into n_T time intervals. Here, n_T was set to the number of unique times divided by 20. Next, λ_0 was assumed to be constant in each time interval, and for each interval I we assumed that $\lambda_0 = \exp(c_I)$. We chose a Gaussian prior on the c_I increments with a precision τ such that $\tau \sim \Gamma(10^{-3}, 10^{-3})$: $\Delta c_I = c_I - c_{I-1} \sim \mathcal{N}(0, \tau^{-1})$.

3.2.2   Evaluation of the estimates of ρ and α_w

Section 2.2.1 presented the estimation of the φ_k. Briefly, it consisted in choosing either the estimates under the best hypothesis $\mathcal{H}_1^{(w)}$: $\mathcal{H}_1^{(w^*)}$ (in this case, the estimates are $\varphi_k^* = \hat{\alpha}_{w^*}\mathbb{1}_{s_k \in w^*} + \hat{X}_k$) or the estimates under $\mathcal{H}_0$ ($\varphi_k^* = \hat{X}_k$). The present section focuses on the bias of the estimates obtained for the spatial correlation parameter (ρ^*) and for the cluster effect ($\hat{\alpha}_{w^*}$). Note that for the cluster effect, we only considered simulations that did not retain $\mathcal{H}_0$

(otherwise, an estimate $\hat{\alpha}_{w^*}$ was unavailable). Thus, the estimates obtained were compared with the true values of the spatial correlation parameter and the cluster effect.

Figure 5.3 shows the selected ρ^* as a function of the parameters ρ and α, and the estimations $\hat{\alpha}_{w^*}$ with the INLA method when one selects $\mathcal{H}_1$ according to the Bayes factor criterion.

(a)
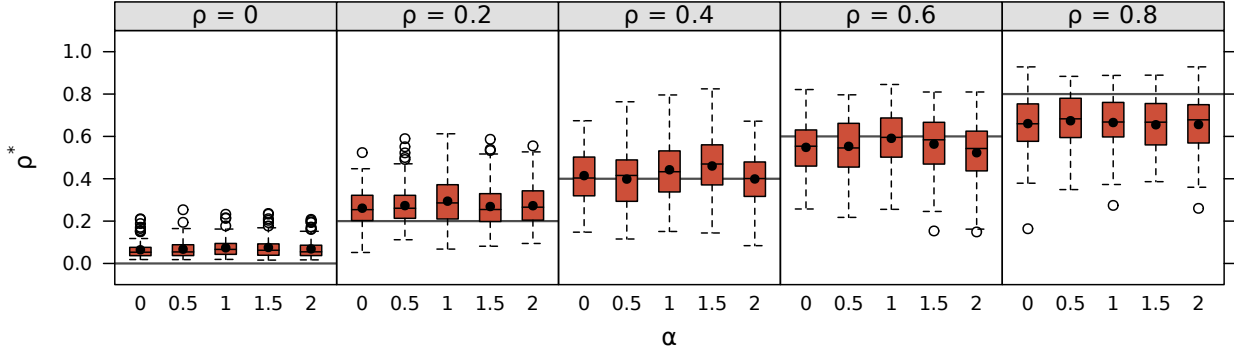
(b)
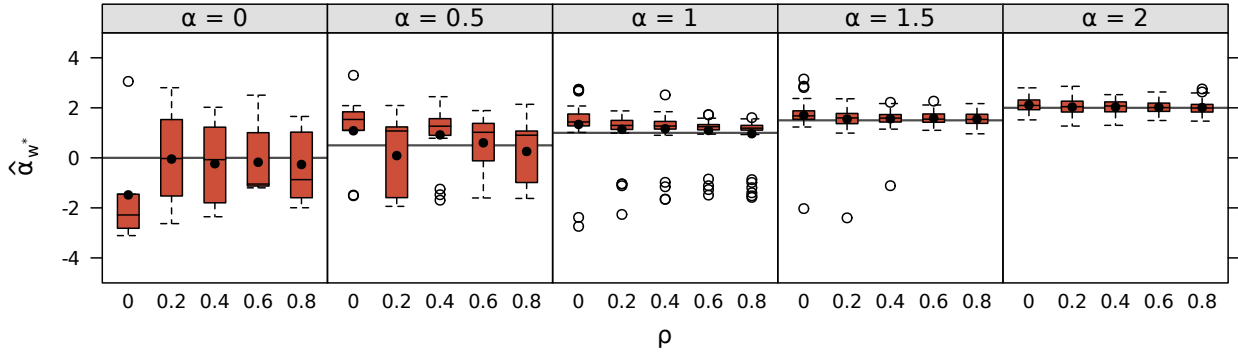
Figure 5.3: Simulation study: the selected values of ρ^* according to the parameters ρ and α (panel (a)) and $\hat{\alpha}_{w^*}$ obtained with the INLA method when we selected $\mathcal{H}_1$ according to the Bayes factor criterion (panel (b)). The horizontal lines in panels (a) and (b) correspond respectively to the true values of the parameters ρ and α, and the black points represent the mean estimates obtained.

Our approach estimates the cluster effect well when the simulated values of α are 1, 1.5 and 2. Although the cluster effect might appear to be poorly estimated for α values of 0 and 0.5, this was because our approach rarely selected $\mathcal{H}_1$ for these values and so few estimates were made. The parameter ρ was well estimated generally but was slightly overestimated for $\rho = 0$ and slightly underestimated for $\rho = 0.8$.

3.2.3   The impact of spatial correlation on cluster detection

We next evaluated the performance of our new method in the context of cluster detection. Two particular versions of the method were also considered, in order to investigate the impact of taking account of potential intra-spatial unit correlation but not spatial dependence between spatial units (the i.i.d. model with $\rho = 0$) or taking account of spatial dependence between spatial units without adjusting its intensity (i.e. by considering it to be complete: the ICAR model, $\rho = 1$).

Note that the ICAR model with $\rho = 1$ leads to a non-invertible matrix A, and so it was not possible to generate data under $\mathcal{H}_0$ to estimate the p-value associated with the most likely

cluster (see Section 2.2.3 for more details). To overcome this problem, the value of the spatial correlation was set to 0.999 (instead of 1) in the scan procedure (Section 2.2.2) for the ICAR model.

Figure 5.4 shows the type I error, the power curves, true positive rates, false positive rates, and positive predictive values obtained with our method and with its two special cases (i.i.d. and ICAR).

For the Leroux CAR model, the performances were relatively stable as a function of ρ, although the type I error was slightly above the 5% threshold. This was not the case for $\rho = 0$ because then ρ^* slightly overestimates ρ (Figure 5.3), which makes the method quite conservative). It should be noted that the true positive rates, the false positive rates and the positive predictive values appear to be less stable when $\alpha = 0.5$. This was because these indicators are only computed for simulations that lead to the rejection of $\mathcal{H}_0$, which were not numerous when $\alpha = 0.5$.

The i.i.d. model failed to maintain a reasonable type I error as ρ increased. Moreover, the power as a function of ρ was less stable than that of the CAR model.

The ICAR model tended to absorb the cluster effect into the spatial correlation parameter ρ. This was particularly the case when the true value of ρ was low. Thus, the type I errors remain reasonable but the power tended to decrease as ρ decreased.

The false positive rates were very low for the three approaches. However, the true positive rates and the positive predictive values were lower for the i.i.d. and the ICAR models than for the CAR model.

We also investigated the performance of our method with other thresholds for the Bayes factor (i.e., 3, 10 and 100, which correspond respectively to "substantial", "strong" and "decisive" levels of evidence for $\mathcal{H}_1^{(w^*)}$ (Jeffreys, 1961)). The results are presented in Figure C.1 in Appendix C.

3.3   The influence of censoring

Here, we describe the simulation study that was designed to evaluate the performance of our approach in the presence of different levels of data censoring.

3.3.1   Design of the simulation study

Due to computational time constraints, the simulation's design differed slightly from those of the previous studies: we considered 940 individuals distributed in the 94 French *départements* (counties) located by their centroid. The simulated cluster contains 73 individuals in the 8 *départements* of the Île-de-France region (the green area in Figure 5.5).

The data were generated in the same way as in Section 3.2, except that different proportions of the observations were censored (10%, 20%, 30% and 40%). Administrative censoring was considered according to Montez-Rath et al.'s (2017) method. Briefly, the end of the study was determined so that the intended proportion of censoring was achieved.

For each value of the spatial correlation parameter ρ, each value of α, and each censoring percentage, 100 data sets were simulated. The statistical significance of the MLC was evaluated through 999 generations of the data under $\mathcal{H}_0$ (see Section 2.2.3 for more details), and the type I error was set to 0.05.

The performances were measured through the same four criteria as in Section 3.2: the power, the true positive rate, the false positive rate, and the positive predictive value.

Note that in this simulation study, the *a priori* distributions were the same as in Section 3.2.
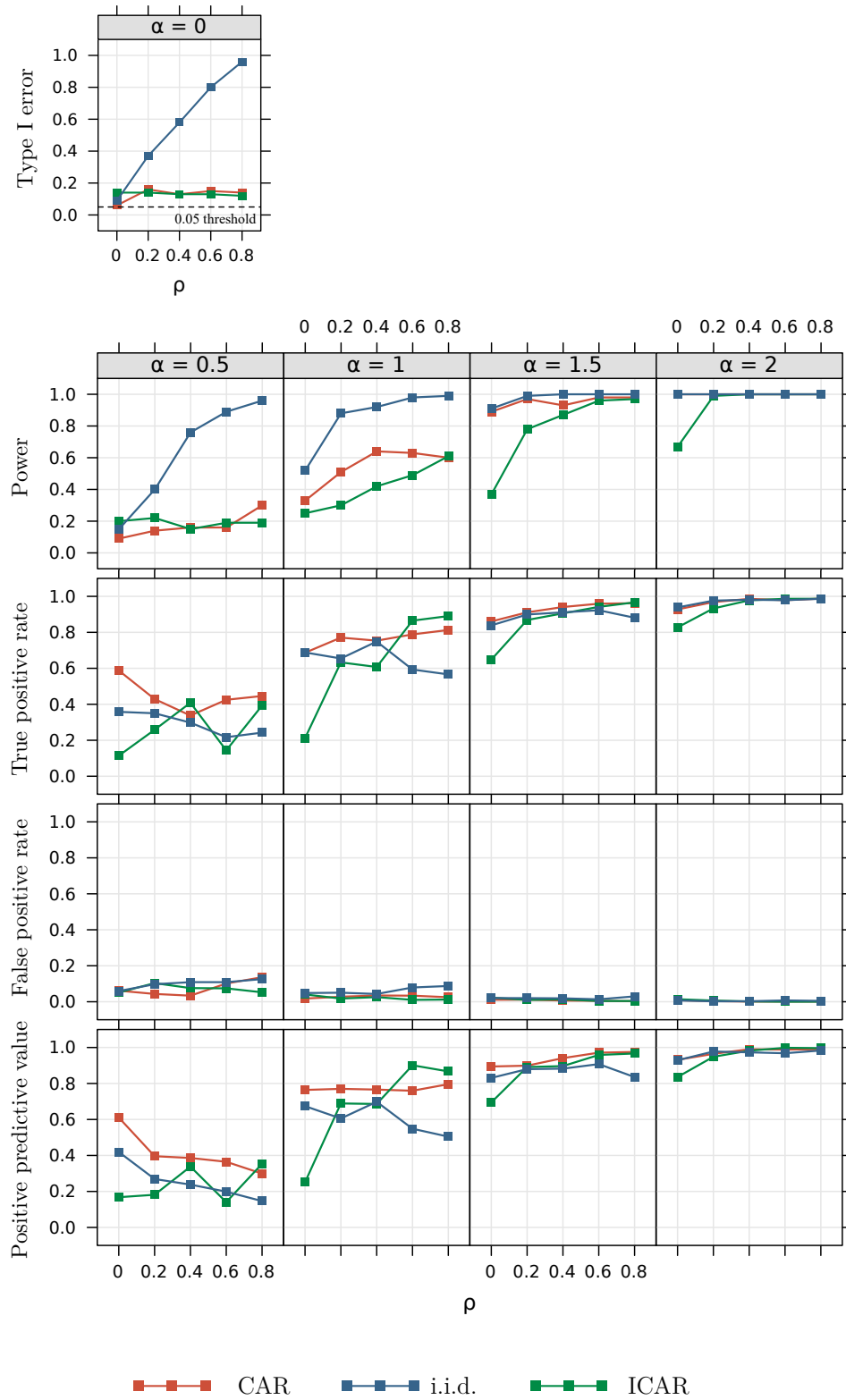
Figure 5.4: Simulation study: Comparison of the type I error, power curves, true positive rates, false positive rates, and positive predictive values for the CAR, ICAR and i.i.d. models. α is the parameter that controls the cluster intensity and ρ controls the spatial correlation.
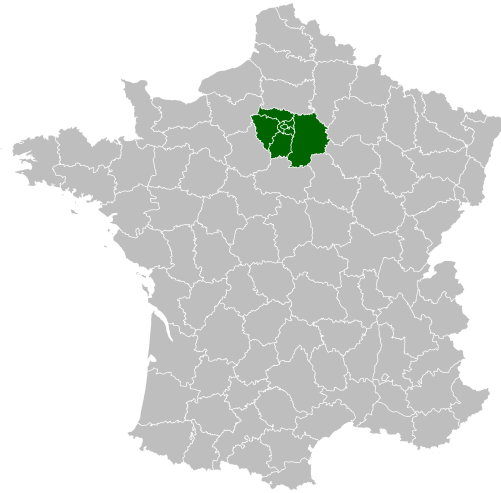
Figure 5.5: Simulated cluster (in green) in the 94 *départements* (counties) of France.

3.3.2 Results

The results of the simulation study are shown in Figure 5.6. Firstly, we found that the power of our method increases as the proportion of censoring increases. This was also the case for the type I error. Although the type I error remained stable and close to the nominal value (whatever the values of ρ) when 10% of the observations were censored, it tended to move away from the nominal value as the censoring percentage increased.

The false positive rates remained very low, regardless of the censoring rate. However, the true positive rates and positive predictive values decreased as the censoring rate increased. Lastly, the impact of censoring on the performance indicators decreased when the cluster's intensity α increased.

4 Application to epidemiological data

4.1 ESRD mortality and related confounding factors

We considered data provided by the French renal epidemiology and information network (REIN) registry on end-stage renal disease (ESRD) in northern France between 2004 and 2020. The methodology of the REIN registry has been described elsewhere (Couchoud et al., 2005). Here, we focused on the analysis of mortality, measured by the survival time after the initiation of dialysis in ESRD patients aged 70 and over. This patient population is characterized by (i) a high mortality rate, thus leading to a high number of observed deaths and, (ii) a low frequency of kidney transplantation, thus minimizing the effect of this known competing risk of death among ESRD patients (Hallan et al., 2012; Ayav et al., 2016). The data covered 6071 individuals but the exact time to survival after the initiation of dialysis was not known in 17% of cases. These censored observations are either patients still alive at the end of the study (15.7%), patients lost to follow-up (0.7%), or patients having received a kidney transplant (in which case, the censoring time corresponds to the date of transplantation; 0.6%). The geographical region studied (the *Nord-Pas-de-Calais* region of northern France) is divided into 80 *cantons* (a French administrative subdivision), and each individual's stated place of residence was linked to the corresponding *canton*.

We also considered 18 variables measured at the individual level and that are known to be confounders of survival in patients with ESRD (Couchoud et al., 2015; Fu et al., 2021). Thus, spatial cluster detection was adjusted by introducing the following confounders into each model as covariates: age (in years), sex, body mass index (in kg/m^2), the type of nephropathy
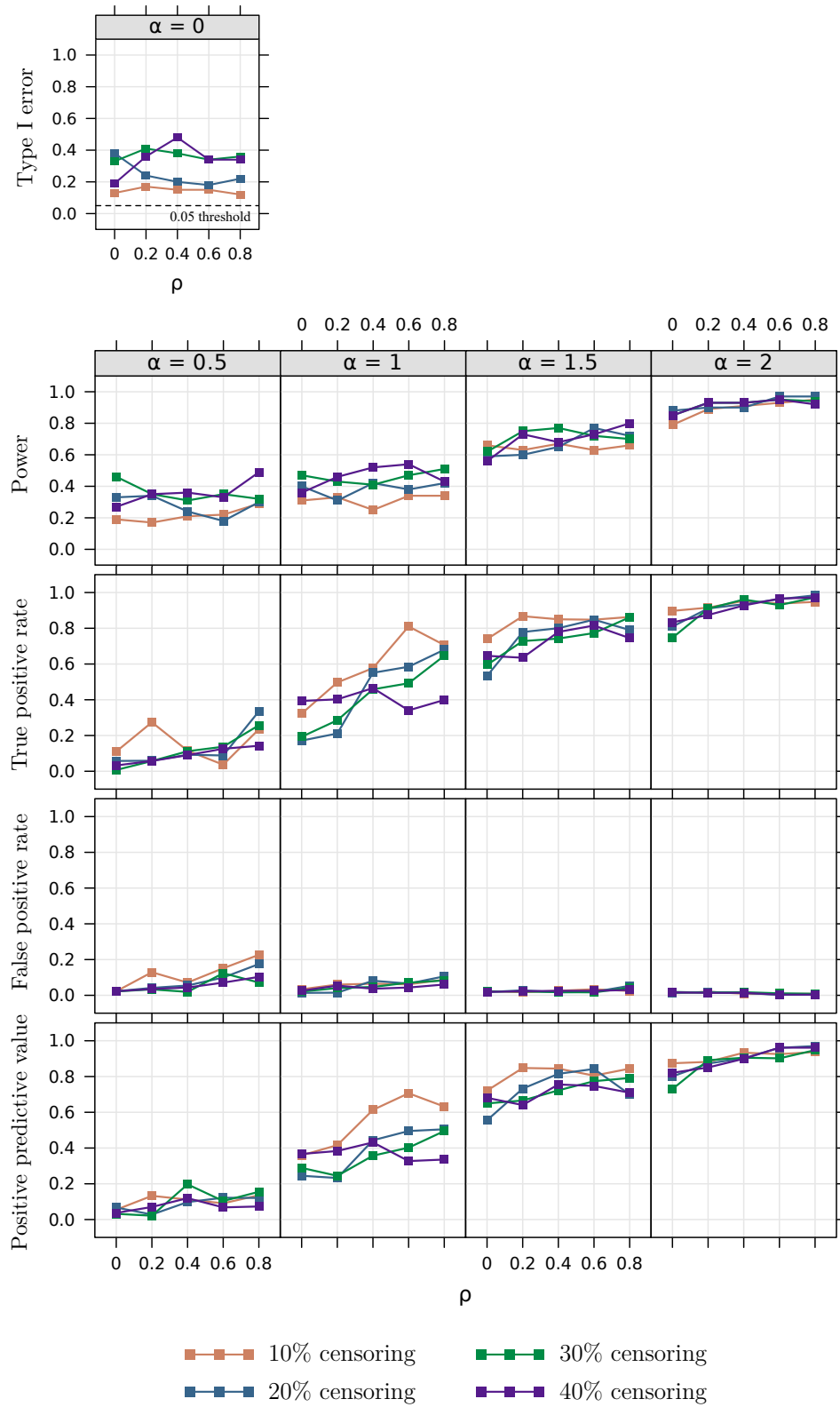
Figure 5.6: Simulation study: Comparison of the power curves, true positive rates, false positive rates, and positive predictive values according to the percentage of censored observations. α is the parameter that controls the cluster intensity, and ρ controls the spatial correlation.

(polycystic, primitive glomerulonephritis, hypertension or vascular, diabetic, pyelonephritis, other), the number of cardiovascular comorbidities (none, one, two or more), mobility (independent walking, need for help from a third party, total disability), the blood hemoglobin level (in g/dL), the serum albumin level (in g/dL), the dialysis method (hemodialysis or peritoneal dialysis), the glomerular filtration rate (below 7, between 7 and 10 or over 10 mL/min/1.73m^2), the period of treatment initiation (2004-2009, 2010-2015 or 2016-2020), whether or not the treatment was initiated urgently, and the presence or absence of diabetes, chronic respiratory disease, respiratory assistance, cirrhosis, severe behavioral disorder, or active malignant cancer. Details of these confounding factors are available in Appendix C.

4.2 Spatial cluster detection

In order to detect spatial clusters of atypical (shorter or longer) survival times among patients with ESRD, five models were considered: the exponential model (Model 1) developed by Huang et al. (2007), the log-rank method developed by Cook et al. (2007) (Model 2) and versions of the Cox-model-based method presented here for considering three types of shared frailty: i.i.d. ($\rho = 0$) (Model 3), CAR ($\rho \in]0, 1[$) (Model 4), and ICAR ($\rho = 1$) (Model 5).

Each model was used to detect spatial clusters of atypical survival times among the patients with ESRD when compared with patients in the rest of the region studied. To adjust survival times for the confounders in Model 1, we used an exponential regression method as proposed by Huang et al. (2007). Regarding Model 2, we adopted the approach developed by Jung (2009) and Ahmed and Genin (2020), which consists in estimating the coefficients associated with the confounders in the model under $\mathcal{H}_0$ and then setting their value in the scan statistic developed by Cook et al. (2007). Regarding Models 3 to 5 (the shared frailty models), we also adopted this approach by setting (under each alternative hypothesis $\mathcal{H}_1^{(w)}$) the coefficients associated with the confounding factors to the values estimated in the model under $\mathcal{H}_0$ in the φ_k estimation step (Section 2.2.1).

In order to provide an indicator of the cluster-associated risk that is independent of the model considered, we estimated the hazard ratio (HR) associated with each cluster in a conventional Cox model adjusted for the confounding factors.

The MLC was considered, as were secondary clusters that had a high Λ value and did not cover the MLC (Kulldorff, 1997). The statistical significance of the detected spatial clusters was evaluated by performing 999 Monte-Carlo simulations, with a type I error of 0.05.

4.3 Results

The spatial clusters detected by each of the five models (exponential, log-rank, i.i.d., CAR, and ICAR frailty) are presented in Figure 5.7. Detailed information on the spatial clusters is presented in Table 5.1.

Both the exponential model (Model 1, panel (a) in Figure 5.7) and the method based on the log-rank test (Model 2, panel (b) in Figure 5.7) identified the same two statistically significant spatial clusters. The MLC (located in the northeast of the region, shown in green) had similar levels of statistical significance in the two models ($\hat{p} = 0.004$ and $\hat{p} = 0.005$, respectively) and had longer survival times than in the rest of the geographical area studied (HR = 0.84 for both models). The first secondary cluster (located in the western part of the region, shown in red) also had similar levels of statistical significance in the two models ($\hat{p} = 0.025$ and $\hat{p} = 0.043$, respectively) and was characterized by shorter survival times (HR=1.13 for both methods).

The i.i.d. frailty model (Model 3, panel (c)) identified the same statistically significant MLC as the exponential model and the method based on the log-rank test ($\hat{p} = 0.006$). The CAR model (Model 4, panel (d)) and ICAR model (Model 5, panel (e)) both detected the same MLC, which contained three more spatial units than the MLC detected by the other models. This MLC was characterized by longer survival times (HR=0.86). The MLC was statistically significant for the CAR model but not for the ICAR model ($\hat{p} = 0.011$ and $\hat{p} = 0.178$ respectively). The first secondary cluster detected by the three frailty models is the same as that detected by the exponential model and the method based on the log-rank test. However, it was not statistically significant for any of the shared frailty models ($\hat{p} = 0.138$ for the i.i.d. frailty model, $\hat{p} = 0.083$ for the CAR frailty model, and $\hat{p} = 0.949$ for the ICAR frailty model).

The small differences between the conventional spatial scan statistics methods (Huang et al., 2007; Cook et al., 2007) and the three shared frailty models developed here can be explained by the low variance of the shared frailties (see Figure 5.8 for the posterior distribution of σ^2 with each model).

Table 5.1: Description of the statistically significant spatial clusters detected by the method developed by Huang et al. (2007) (Model 1 (exponential)), the method of Cook et al. (2007) (Model 2 (log-rank)) and those detected by the shared frailty models (Model 3 (i.i.d.), Model 4 (CAR) and Model 5 (ICAR)), after adjustment for confounding factors.

Model	Cluster	p-value	Number of spatial units	Number of individuals	Number of events	Hazard ratio[1]
Model 1 (Exponential)	MLC	0.004	10	1091	890	0.84
	Secondary cluster 1	0.025	43	2632	2163	1.13
Model 2 (Log-rank)	MLC	0.005	10	1091	890	0.84
	Secondary cluster 1	0.043	43	2632	2163	1.13
Model 3 (i.i.d. frailty)	MLC	0.006	10	1091	890	0.84
	Secondary cluster 1	0.138	43	2632	2163	1.13
Model 4 (CAR frailty)	MLC	0.011	13	1346	1094	0.86
	Secondary cluster 1	0.083	43	2632	2163	1.13
Model 5 ICAR frailty	MLC	0.178	13	1346	1094	0.86
	Secondary cluster 1	0.949	43	2632	2163	1.13

[1] The hazard ratio was computed using a Cox model with adjustment for confounders.

5 Discussion

Here, we developed a new spatial scan statistic for survival data indexed in space. It allows one to (i) take account of both potential intra-spatial unit correlation and spatial dependence between spatial units, and (ii) adjust the cluster detection for confounding factors. This method is based on a Cox model that includes spatially structured shared frailty distributed according to a Leroux CAR model.

In a simulation study, we showed that in the presence of intra-spatial unit correlation, the existing methods (Cook et al., 2007; Huang et al., 2007) are confronted by a huge increase in the type I error. Thereafter, the performance of the CAR model was evaluated in the context of cluster detection and compared with two particular versions of it: the i.i.d. frailty model and the ICAR frailty model. The CAR model presented the best performances in the presence of spatial correlation, which thus demonstrated good-quality adjustment. In the last simulation
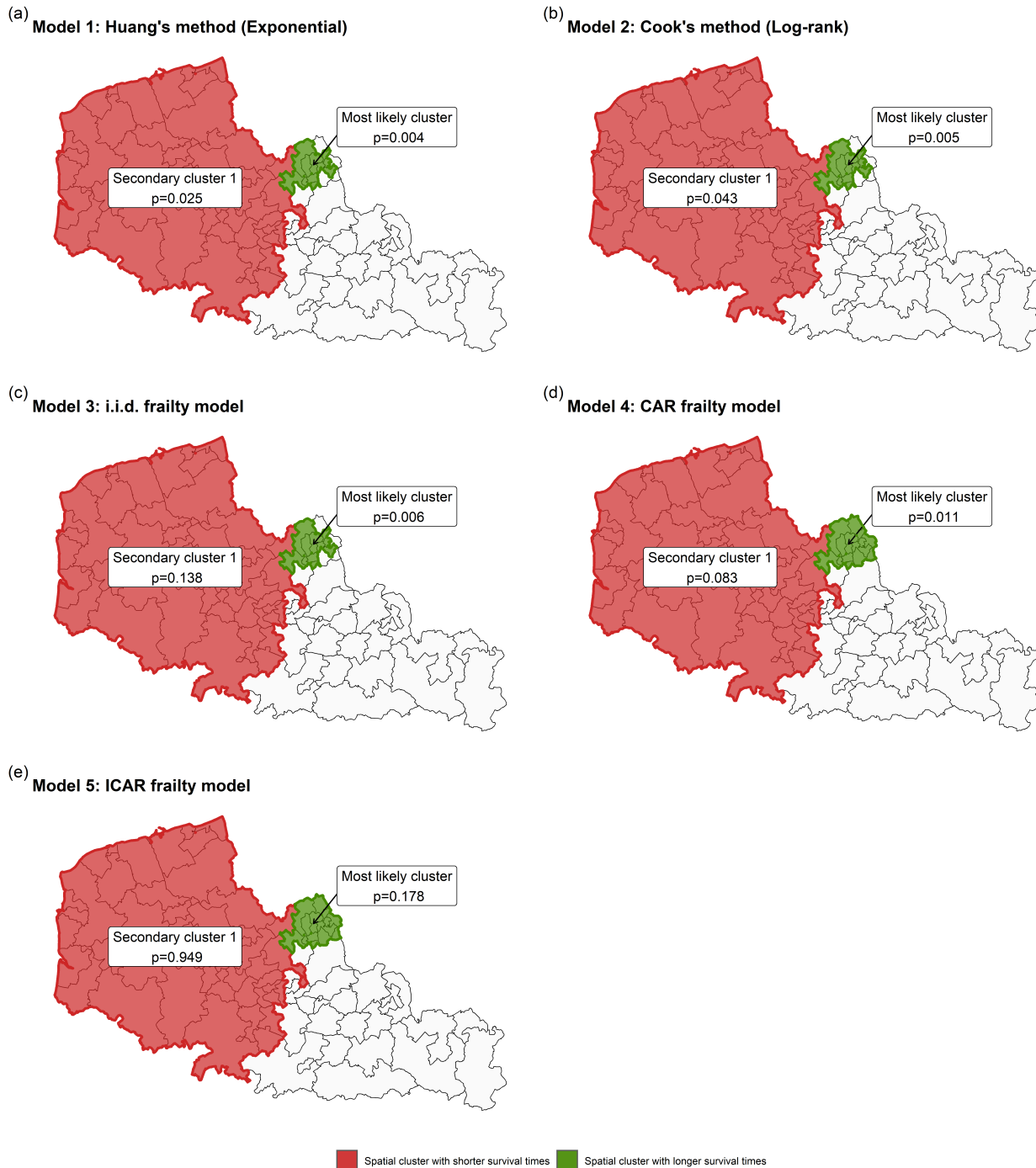
Figure 5.7: Spatial clusters detected by the method developed by Huang et al. (2007) (Model 1 (exponential), panel(a)), the method developed by Cook et al. (2007) (Model 2 (Log-rank), panel (b)) and those highlighted by the shared frailty models (Model 3 (i.i.d.), panel (c); Model 4 (CAR), panel (d); Model 5 (ICAR), panel (e)) after adjustment for confounding factors. Spatial clusters in green indicate longer survival times for patients with ESRD, compared with the rest of the region studied. Conversely, spatial clusters in red indicate shorter survival times for patients with ESRD.

study, we showed that the performance of the CAR model is adequate as long as the percentage of censored observations does not exceed 20%.

These approaches were then applied to epidemiological data, i.e. the detection of clusters of

(a)   (b)

(c)   (d)
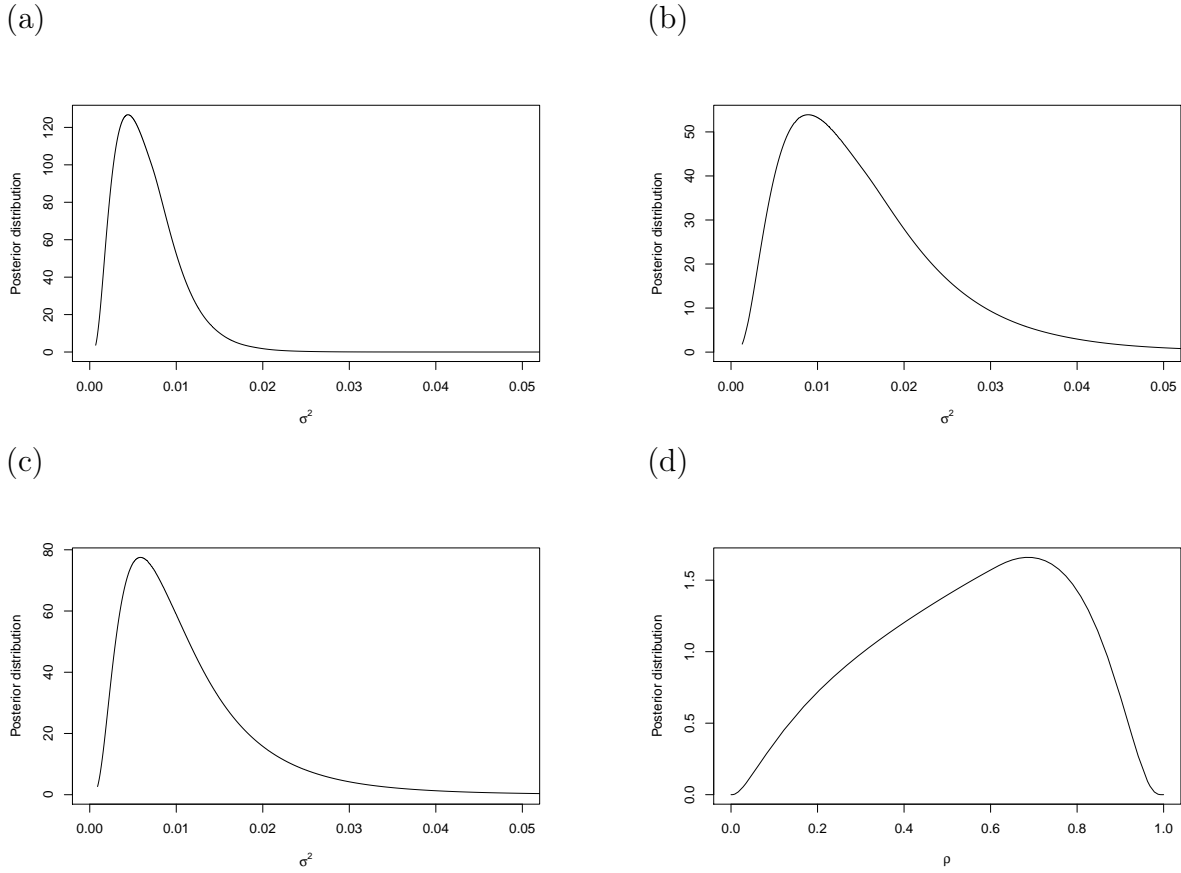
Figure 5.8: Posterior distribution of the frailty variance with the i.i.d. (panel (a)), CAR (panel (b)) and ICAR (panel (c)) models, and the posterior distribution of the spatial correlation parameter ρ obtained with the CAR model (panel (d)).

abnormally low or high survival times in elderly patients with ESRD in northern France during the period 2004-2020. The conventional approaches (Cook et al., 2007; Huang et al., 2007) detected two statistically significant clusters: one in the northeast of the region (corresponding to longer survival times, i.e. a lower risk than elsewhere) and the other containing the whole western part of the region (corresponding to lower survival times, i.e. a higher risk). The i.i.d. shared frailty model only detected the cluster in the northeast of the region as being statistically significant. Assuming a complete spatial correlation, the ICAR model also identified a MLC in the northeast of the region but this was not statistically significant. When we considered the CAR frailty model that allowed flexibility of the spatial correlation, a statistically significant cluster was detected in the northeast of the region. The cluster's p-value was slightly higher than that provided by the i.i.d. shared frailty model; this can be explained by the fact that the CAR model takes account of spatial correlation. These results are consistent with those of the simulation study.

In both the simulation study and the application to epidemiological data, circular potential clusters were considered. However, other cluster shapes (e.g. elliptical clusters (Kulldorff et al., 2006)) could be considered, since the shape of the scanning window has an impact on the power of cluster detection. It should be noted that other scanning window shapes can be easily implemented in our method because this only changes the set of potential clusters $\mathcal{W}$.

In the population of elderly patients with ESRD, only a low percentage had received a kidney transplant. However, this percentage is higher in the general population (Couchoud et al., 2015). It is well known that kidney transplantation is a competing risk for death in patients with ESRD, and failure to take account of this risk in the analysis might bias the estimate of survival (Hallan et al., 2012). In this context, the method developed here should be modified

to account for competing risks by considering (for example) the model developed by Fine and Gray (1999).

Here, the spatial correlation parameter ρ was assumed to be constant over the whole study area. This assumption may be too simplistic because this coefficient can vary spatially (Crawford, 2009). However, the integration of a varying spatial correlation coefficient would be challenging because it is necessary to clearly distinguish between the effect of spatial correlation and the effects of spatial clusters in the data. Adapting the method developed here to this context could be the subject of future research.

In our model, we included covariates as fixed effects. However, it is possible to consider them as random effects that may have a spatial correlation. One way of taking these spatially structured effects into account involves the use of conditional autoregressive models for the prior distributions of the coefficients associated with the covariates.

Lastly, our spatial scan statistic can be extended to deal with recurrent events. For example, one might be interested in the time until an asthma attack in patients treated for asthma, and a patient might experience several asthma attacks during the study period. One possible approach is to consider shared frailty at the individual level, making it possible to take account of unobserved, subject-specific factors (Kleinbaum and Klein, 2012). However, the time to an asthma attack might also exhibit an intra-spatial unit correlation, due (for example) to environmental factors. In this context, one approach would be to consider a nested frailty model (Rondeau, 2010), i.e., a model with both shared frailties at the level of spatial units with a potential spatial correlation, and shared frailties at the level of individuals, in order to take account of unobserved factors that are specific to spatial units (e.g. air quality) and those that are specific to individuals (e.g. tobacco consumption).

Chapter 6

Software: the R package HDSpatialScan

1 Introduction

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., 2003a). 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, 2017a) and the R package **SpatialEpiApp** (Moraga, 2017b) 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 scanning 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** (Dematteï et al., 2006) or **scanstatistics** (Allévius, 2018a,b) for spatial or space-time data. 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 chapter is organized as follows: Section 2 presents the different models implemented in the R package **HDSpatialScan**. Section 3 describes the implementation of the methods, and, in Section 4, examples of use of the package are given. Finally the chapter is concluded in Section 5.

It should be noted that the content of this chapter has been published in *The R Journal*, in collaboration with Mohamed-Salem Ahmed (University of Lille), Julien Soula (University of Lille), Zaineb Smida (University of Montpellier), Lionel Cucala (University of Montpellier),

Sophie Dabo-Niang (University of Lille) and Michaël Genin (University of Lille).

2   Models

The R package **HDSpatialScan** implements spatial scan statistics for multivariate and functional data. First, it provides the spatial scan statistics for multivariate data taking account of the potential correlations between the variables and developed by Cucala et al. (2017) and Cucala et al. (2019). Then, it implements the two spatial scan statistics for univariate functional data presented in Chapter 3, as well as the nonparametric spatial scan statistics for univariate functional data developed by Smida et al. (2022). In the context of multivariate functional data, it provides the three spatial scan statistics developed in Chapter 4 as well as our adaptation of the nonparametric spatial scan statistic for univariate functional data (Smida et al., 2022) to the multivariate functional context.

2.1   Spatial scan statistics for multivariate data

In this subsection we consider the case where several continuous variables are simultaneously observed in each spatial location: $X = (X^{(1)}, \ldots, X^{(p)})^\top$ 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 6.1 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 circumference 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).

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

Then we can compute the maximum likelihood estimates (MLE) of $\mu, \mu_w, \mu_{w^c}, \Sigma$ and Σ_{w,w^c}: $\hat{\mu}, \widehat{\mu}_w, \widehat{\mu}_{w^c}, \widehat{\Sigma}$ and $\widehat{\Sigma}_{w,w^c}$, and we can show that the log-likelihood ratio between these two
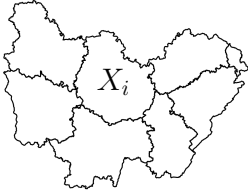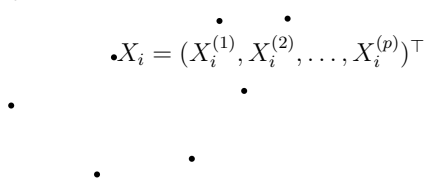
Data	$X_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)})^\top$		$X_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)})^\top$	
Application example	Lattice data: • Unemployment rate and fraction of the population that has not graduated from high school	Geostatistical data: • Temperature and air pressure	Point pattern: • 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:
	  • Multivariate Gaussian spatial scan statistic
	  • "MG" argument in the scan function of the package
	Non-Gaussian data:
	  • 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 6.1: In this figure the data is composed of multivariate vectors X_i. Several applications are proposed for lattice, geostatistical, and point data. For each of these 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.

hypotheses is

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

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

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

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 $\mathcal{H}_0 : X_1, \ldots, X_n$ are identically distributed, whatever the associated location.

Let

$$\begin{aligned} \text{sgn} \quad : \quad & \mathbb{R}^p \quad \rightarrow \quad \mathbb{R}^p \\ & x \quad \mapsto \quad \begin{cases} ||x||_2^{-1} x & \text{if } x \neq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad ,$$

then the multivariate ranks R_i are defined by $R_i = \dfrac{1}{n} \sum\limits_{j=1}^{n} \text{sgn}(A_X(X_i - X_j))$ where the matrix A_X makes the ranks such that $\dfrac{p}{n} \sum\limits_{i=1}^{n} R_i R_i^\top = \dfrac{1}{n} \sum\limits_{i=1}^{n} R_i^\top R_i I_p$. 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

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

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 $\lambda_{\text{MNP}} = \max\limits_{w \in \mathcal{W}} U^2(w)$.

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 Cucala (2016) (UNP).

2.2 Spatial scan statistics for univariate functional data

This subsection considers 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. Here we briefly recall the spatial scan statistics for univariate functional data developed in Chapter 3 (PFSS and DFFSS), and the approach proposed by Smida et al. (2022) (NPFSS). We also propose another spatial scan statistic based on pointwise ranks (namely the "univariate rank-based functional spatial scan statistic" (URBFSS)), which corresponds to the adaptation of the MRBFSS presented in Chapter 4 to the univariate functional context.

Figure 6.2 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, 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).
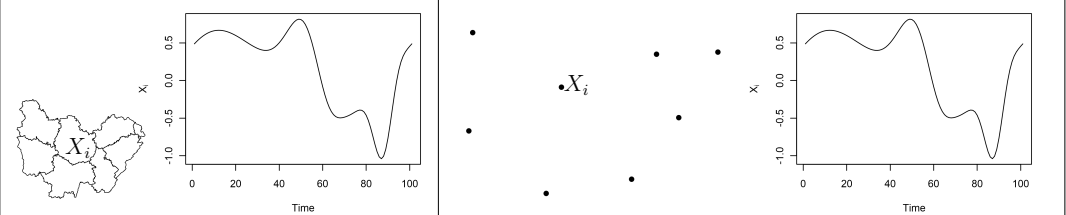
Data			
Application example	Lattice data: • Unemployment rate over time	Geostatistical data: • Temperature over time	Point pattern: • Circumference of trees over time
Question	Is there a statistically significant cluster of high or low unemployment rate curves?	Is there a statistically significant cluster of high or low temperature curves?	Is there a statistically significant cluster of trees with high or low circumference curves?
Methods	Gaussian data: • Distribution-free functional spatial scan statistic • "DFFSS" argument Non-Gaussian data: • Univariate rank-based functional spatial scan statistic • "URBFSS" argument		
Interpretation	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.		

Figure 6.2: In this figure the data is a set of curves X_i. Several applications are proposed for lattice, geostatistical, and point data. For each of these data, the table indicates the question that can be asked for the detection of clusters on the 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.

2.2.1 A parametric and a distribution-free spatial scan statistics for univariate functional data

In Chapter 3 we developed a parametric functional spatial scan statistic (PFSS)

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

and a distribution free functional spatial scan statistic (DFFSS)

$$\Lambda_{\text{DFFSS}} = \max_{w \in \mathcal{W}} \sup_{t \in \mathcal{T}} I^{(w)}(t),$$

where

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

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

and $||.||_2$ is defined as $||f||_2^2 = \int_{\mathcal{T}} f^2(t)\,\mathrm{d}t$.

The reader is invited to refer to Chapter 3 for more details about these methods.

2.2.2 A nonparametric spatial scan statistic for functional data

The R package **HDSpatialScan** also implements the nonparametric spatial scan statistic (NPFSS) proposed by Smida et al. (2022) in the context of univariate functional data and presented in Chapter 2: $\Lambda_{\text{NPFSS}} = \max_{w \in \mathcal{W}} U^{(w)}$, where

$$U^{(w)} = \left\|\frac{1}{\sqrt{|w||w^c|}n}\sum_{i,s_i \in w}\sum_{j,s_j \in w^c}\frac{X_j - X_i}{||X_j - X_i||_2}\right\|_2.$$

2.2.3 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 MRBFSS developed in Chapter 4 to this context.

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 $\{X_1(t), \ldots, X_n(t)\}$, 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 respectively 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 similarly to Cucala (2016) 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 $\mathcal{H}_1^{(w)}$: $\exists t \in \mathcal{T}$, $F_{w,t}(x) = F_{w^c,t}(x - \Delta_t)$, $\Delta_t \neq 0$.

Then, 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): $\Lambda_{\text{URBFSS}} = \max_{w \in \mathcal{W}} T^{(w)}$.

2.3   Spatial scan statistics for multivariate functional data

This subsection considers 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. Here, we recall the spatial scan statistics for multivariate functional data developed in Chapter 4.

Figure 6.3 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. To detect spatial clusters, in the case of Gaussian data we will prefer the multivariate distribution-free functional spatial scan statistic (MDFFSS) and for non-Gaussian data we will use the pointwise rank-based approach (MRBFSS).

Chapter 4 developed three new spatial scan statistics for multivariate functional data. The first one is a parametric spatial scan statistic for multivariate functional data (MPFSS) based on a functional MANOVA Lawley–Hotelling trace test (Górecki and Smaga, 2017):

$$\Lambda_{\mathrm{MPFSS}} = \max_{w \in \mathcal{W}} \mathrm{Trace}(H_w E_w^{-1}),$$

where

$$H_w = |w| \int_{\mathcal{T}} \left[\bar{X}_w(t) - \bar{X}(t) \right] \left[\bar{X}_w(t) - \bar{X}(t) \right]^{\top} \, \mathrm{d}t + |w^{\mathsf{c}}| \int_{\mathcal{T}} \left[\bar{X}_{w^{\mathsf{c}}}(t) - \bar{X}(t) \right] \left[\bar{X}_{w^{\mathsf{c}}}(t) - \bar{X}(t) \right]^{\top} \, \mathrm{d}t \text{ and}$$

$$E_w = \sum_{j, s_j \in w} \int_{\mathcal{T}} \left[X_j(t) - \bar{X}_w(t) \right] \left[X_j(t) - \bar{X}_w(t) \right]^{\top} \mathrm{d}t + \sum_{j, s_j \in w^{\mathsf{c}}} \int_{\mathcal{T}} \left[X_j(t) - \bar{X}_{w^{\mathsf{c}}}(t) \right] \left[X_j(t) - \bar{X}_{w^{\mathsf{c}}}(t) \right]^{\top} \mathrm{d}t.$$

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 $\mathrm{LH}^{(w)} = \mathrm{Trace}(H_w E_w^{-1})$, (2) the Pillai trace test statistic $\mathrm{P}^{(w)} = \mathrm{Trace}(H_w (H_w + E_w)^{-1})$, (3) the Roy's largest root test statistic $\mathrm{R}^{(w)} = \lambda_{\max}(H_w E_w^{-1})$ where $\lambda_{\max}(H_w E_w^{-1})$ is the maximum eigenvalue of $H_w E_w^{-1}$ and (4) the Wilks lambda test statistic $\mathrm{W}^{(w)} = \dfrac{\det(E_w)}{\det(H_w + E_w)}$.

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:

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

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

Chapter 4 also proposed a distribution-free spatial scan statistic for multivariate functional data (MDFFSS): $\Lambda_{\mathrm{MDFFSS}} = \max_{w \in \mathcal{W}} \sup_{t \in \mathcal{T}} T_n(t)^{(w)}$ where

$$T_n(t)^{(w)} = \frac{|w||w^{\mathsf{c}}|}{n} \left(\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t) \right)^{\top} \hat{\Sigma}(t, t)^{-1} \left(\bar{X}_w(t) - \bar{X}_{w^{\mathsf{c}}}(t) \right) \text{ and}$$
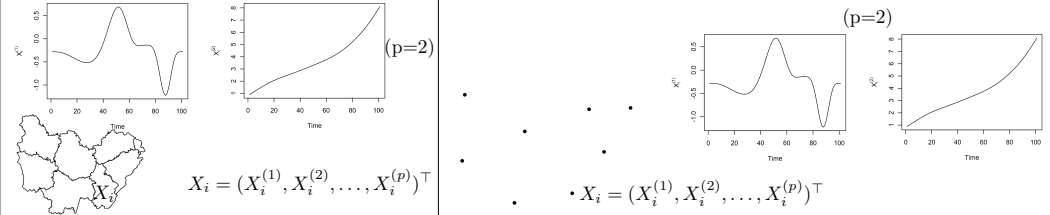
Data	$X_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)})^\top$	(p=2) $\cdot\ X_i = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)})^\top$	
Application example	Lattice data: • Unemployment rate and fraction of the population that has not graduated from high school over time	Geostatistical data: • Temperature and air pressure over time	Point pattern: • Circumference and height of trees over time
Question	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?	Is there a statistically significant cluster of high temperature and low air pressure curves?	Is there a statistically significant cluster of trees with high circumference and height curves?
Methods	Gaussian data: • Multivariate distribution-free functional spatial scan statistic • "MDFFSS" argument in the scan function of the package Non-Gaussian data: • Multivariate rank-based functional spatial scan statistic • "MRBFSS" argument in the scan function of the package		
Interpretation	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.		

Figure 6.3: In this figure the data is a set of multivariate curves X_i (here $X_i = (X_i^{(1)}, X_i^{(2)})^\top$ is composed of two curves). Examples of applications are proposed for lattice, geostatistical, and point data. For each of these data, the table indicates the question that can be asked for the detection of clusters of the 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.

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

Finally it also developed a rank-based spatial scan statistic for multivariate functional data (MRBFSS):

$$\Lambda_{\text{MRBFSS}} = \max_{w \in \mathcal{W}} \sup_{t \in \mathcal{T}} W(t)^{(w)}$$

where

$$W(t)^{(w)} = \frac{pn}{\sum_{i=1}^n R_i(t)^\top R_i(t)} \left[|w|\ ||\bar{R}_w(t)||_2^2 + |w^c|\ ||\bar{R}_{w^c}(t)||_2^2 \right]$$

and $R_i(t)$ are the pointwise multivariate ranks.

It should be noted 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 in Chapter 4. In this context, $\Lambda_{\text{NPFSS}} = \max_{w \in \mathcal{W}} U^{(w)}$, where

$$U^{(w)} = \left|\left| \frac{1}{\sqrt{|w||w^c|n}} \sum_{i,s_i \in w} \sum_{j,s_j \in w^c} \frac{X_j - X_i}{||X_j - X_i||_2} \right|\right|_2$$

and $||.||_2$ is defined as $||f||_2^2 = \int_{\mathcal{T}} f(t)^\top f(t)\ \mathrm{d}t$.

The reader is invited to refer to Chapter 4 for more details.

2.4   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 (Cucala, 2016) 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 MDFFSS and the MRBFSS in the multivariate one. 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 MDFFSS respectively. Table 6.1 summarizes the methods and their performances. The symbols ✓ and ✗ 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 6.2 gives an idea of the computation time of the different scanning methods proposed by the package (see Appendix B for more details). It should be noted that the computation time of the different spatial scan statistics methods is dependent on the size of the data sets, and in particular a function of the number of sites, the number of observation times and the number of variables considered.

3   Software

3.1   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"`, `"DFFSS"` and `"URBFSS"` apply the parametric, the distribution-free and the new rank-based functional approaches on univariate functional data, and `"MPFSS"`, `"MDFFSS"`, 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.

3.1.1   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

Table 6.1: 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 URBFSS) and multivariate functional data (MPFSS, MDFFSS, 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			*Univariate data*		
UG[a]	✓	✓	✓	X	X	✓
UNP[a]	✓	✓		✓	✓	
	Multivariate data			*Multivariate data*		
MG[b]	✓	✓	✓	X	X	✓
MNP[b]	✓	✓		✓	✓	
	Functional univariate data			*Functional univariate data*		
	Non localized clusters in time			Non localized clusters in time		
PFSS[c]			✓	X		✓
DFFSS[d]	✓	✓	✓	✓	✓	✓
NPFSS[e]		✓			✓	
URBFSS[f]	✓	✓		✓	✓	
	Localized clusters in time			Localized clusters in time		
PFSS[c]	X	X		X	X	
DFFSS[d]	✓	✓	✓	✓	✓	✓
NPFSS[e]	X			X		
URBFSS[f]	✓	✓	✓	✓	✓	✓
	Functional multivariate data			*Functional multivariate data*		
	Non localized clusters in time			Non localized clusters in time		
MPFSS[c]			✓	X	X	✓
MDFFSS[d]	✓	✓	✓			✓
NPFSS[e]		✓		✓	✓	
MRBFSS[f]	✓	✓		✓	✓	
	Localized clusters in time			Localized clusters in time		
MPFSS[c]	X	X		X	X	✓
MDFFSS[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 (MDFFSS) 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

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, 2005b), by using for example the R package **fda** (Ramsay et al., 2020).

3.1.2 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`,

Table 6.2: 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
DFFSS[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
MDFFSS[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 (DFFSS) and the Multivariate Distribution-free Functional (MDFFSS) 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

`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 statistically 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.

3.1.2.1   *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. 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 `nb_CPU`). 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).

3.1.2.2   *A posteriori* filtering

Sometimes after that the p-value of each potential cluster has been computed, the user may want to retrieve only the statistically 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.

3.1.3   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 statistically 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 statistically 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"`, `"ResScanOutputUniFunct"` or `"ResScanOutputMultiFunct"`. The objects of class `"ResScanOutputMulti"` and `"ResScanOutputMultiFunct"` also include the element `variable_names`, and the objects of class `"ResScanOutputUniFunct"` and
`"ResScanOutputMultiFunct"` include the element `time`.

3.2   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 `spobject` 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.

4.1   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 $\mu g.m^{-3}$ 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 $s_1, \ldots, s_{169}$: nitrogen dioxide (NO_2), ozone (O_3) and fine particles PM_{10}

and PM$_{2.5}$ corresponding respectively to particles whose diameter is less than $10\mu m$ and $2.5\mu m$. 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 6.4 and 6.5). 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.
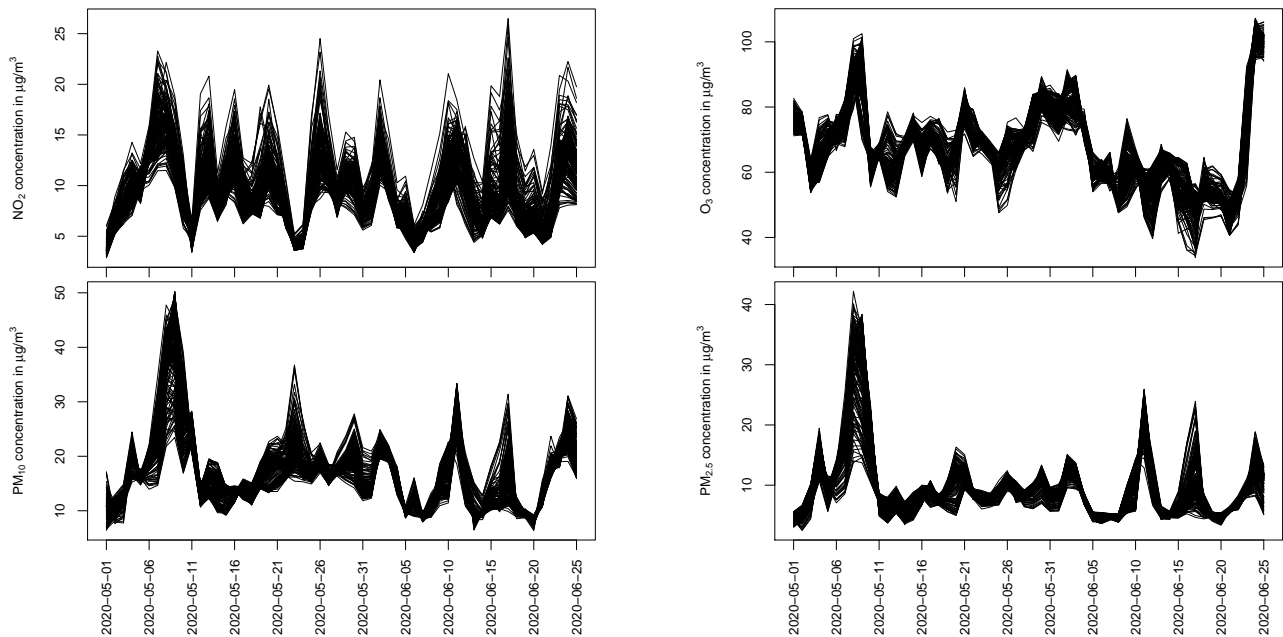
Figure 6.4: Daily concentration curves of NO$_2$, O$_3$, PM$_{10}$ and PM$_{2.5}$ (from May 1, 2020 to June 25, 2020) in each of the 169 *cantons* of *Nord-Pas-de-Calais* (a region in northern France).

The maps in Figure 6.5 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 6.4 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.
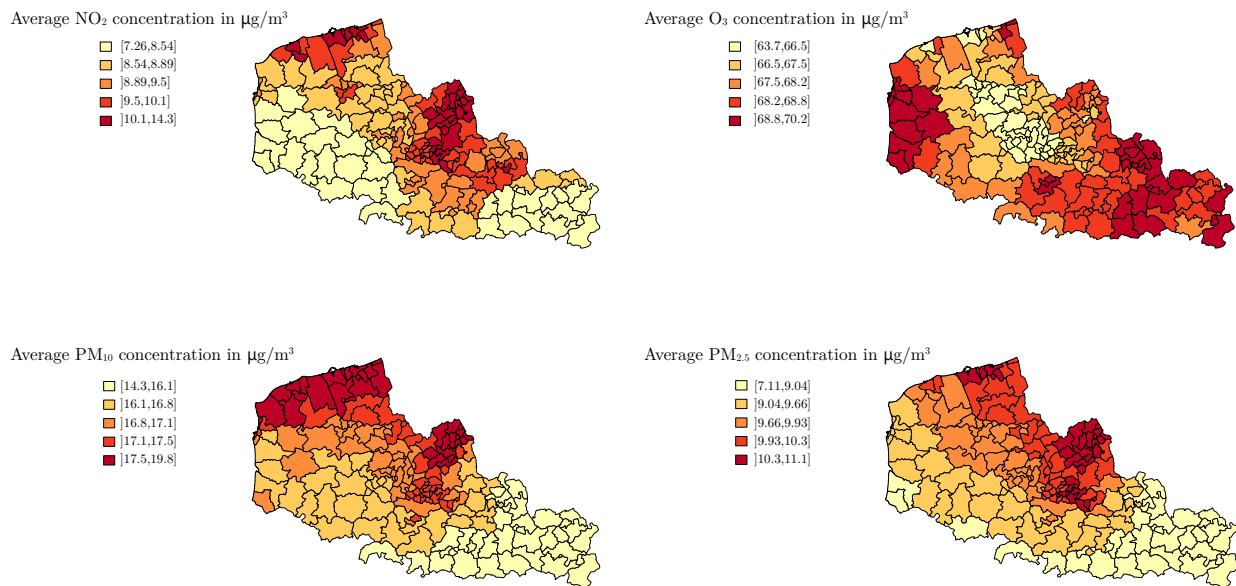
Figure 6.5: Spatial distributions of the average concentrations of NO_2, O_3, PM_{10} and $PM_{2.5}$ over the period from May 1, 2020 to June 25, 2020.

4.2   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 of multiple exposure to pollutants or, on the contrary, areas with little pollution. We first check the normality of each variable, 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 subsection 3.1, 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 subsection 4.1, 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 statistically 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 the sake of brevity we choose here and in the following to 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 6.6).

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)
```
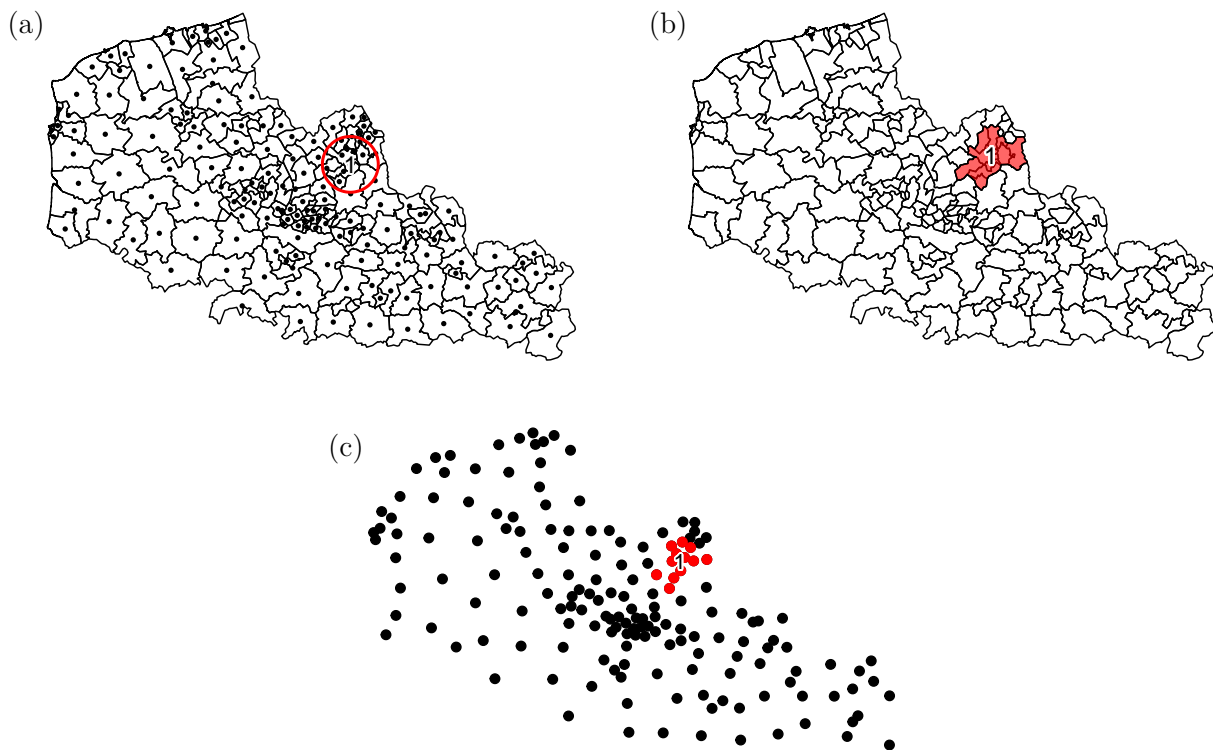

Figure 6.6: Visualization of the most likely cluster with the function `plot` with the types `"map"` (panel a), `"map2"` (panel b) and `"schema"` (panel c) for the MNP scan procedure computed with the function `SpatialScan`

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
## 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 6.7).

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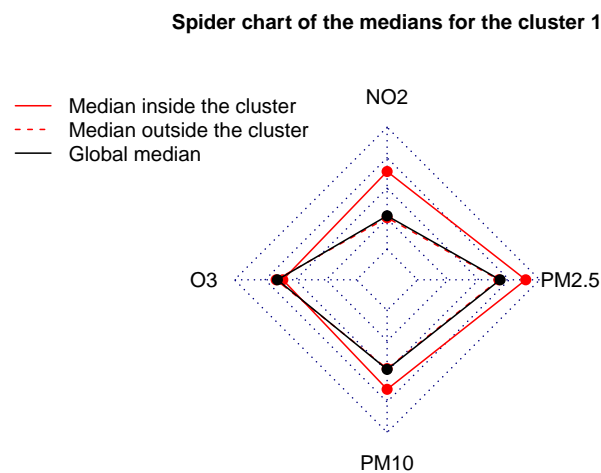

Figure 6.7: Spider chart for the most likely cluster detected by the MNP scan procedure,
obtained with the function `plotSummary`

The MLC is located in the area of Lille. The summary and Figure 6.7 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 $PM_{2.5}$ 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 6.4 present a marked
temporal variability during the study period. Thus it could be interesting to apply functional
spatial scan statistics.

4.3   A univariate functional spatial scan statistic

In this subsection we only consider the pollutant NO_2. Applying a spatial scan statistic for
univariate functional data will thus allow to highlight areas where the NO_2 concentration curves
are abnormally high or, on the contrary abnormally low. We choose to use the URBFSS scan
procedure since it often presents higher powers and true positive rates than the other univariate
functional methods as its multivariate counterpart MRBFSS. As mentioned in subsection 4.2

we decide to use the set of potential clusters recommended by Kulldorff and Nagarwalla (1995), which corresponds 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_urbfss <- 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_urbfss, type = "map2", spobject = map_sites, only.MLC = TRUE)
```
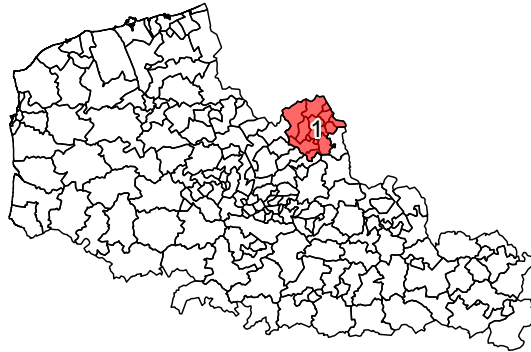

Figure 6.8: Visualization of the most likely cluster for the URBFSS scan procedure with the function `plot` with `type = "map2"`

Again the MLC is located in the area of Lille (Figure 6.8). 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 6.9): this is a cluster of overexposure to NO_2, 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)
```

4.4   A multivariate functional 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 URBFSS scan procedure, we use the MRBFSS in this context, with the same restrictions *a priori* and *a posteriori* as for the MNP and the URBFSS 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)
```
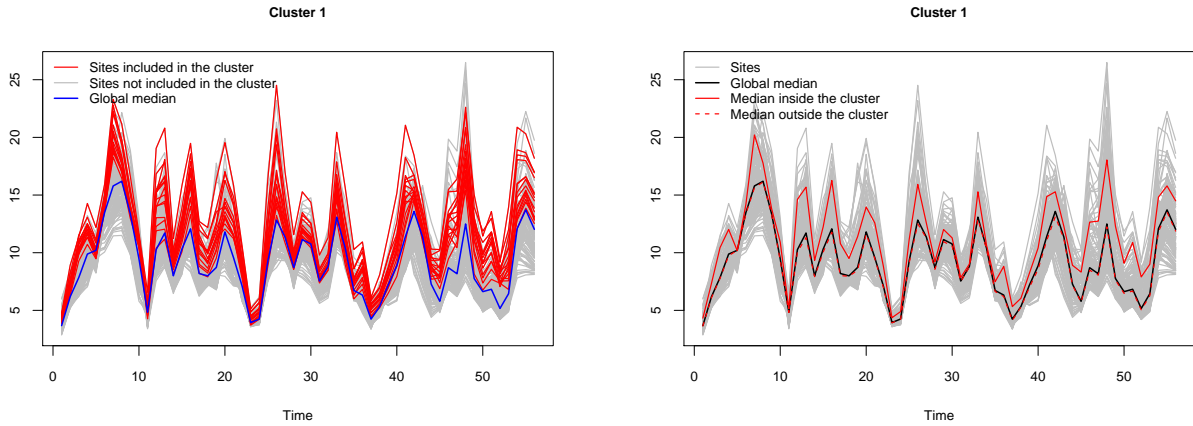
Figure 6.9: Characterization of the most likely cluster for the URBFSS scan approach in the context of univariate functional data with the functions `plotCurves` (left panel) and `plotSummary` (right panel)
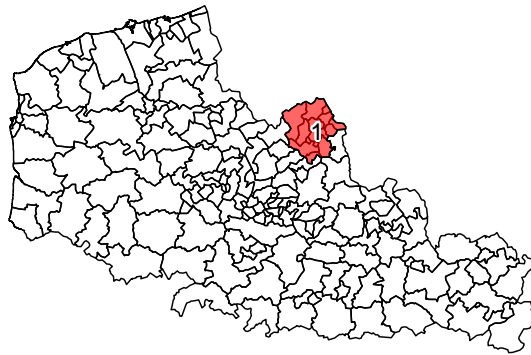
Figure 6.10: Visualization of the most likely cluster for the MRBFSS scan procedure with the function `plot` with `type = "map2"`

The detected cluster is exactly the same as before and is therefore located in the urban area of Lille (Figure 6.10).

Again we will display the curves in the cluster by adding the curve of the global median (Figure 6.11), 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 $PM_{2.5}$ (Figure 6.12). As mentioned in subsection 4.2, in environmental science it is well-known that NO_2 and $PM_{2.5}$ 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)
```

5   Conclusion

Here 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),
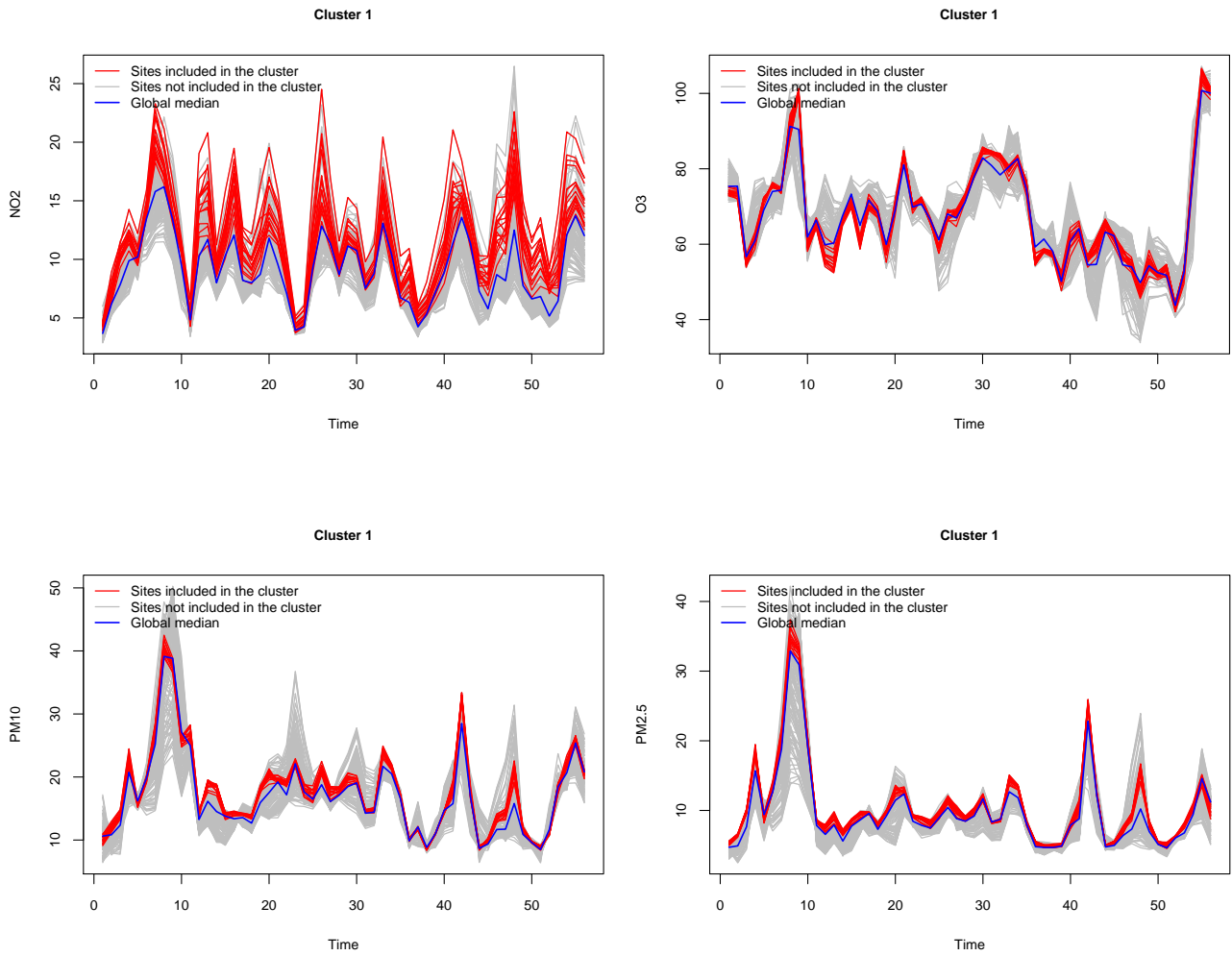
Figure 6.11: Characterization of the most likely cluster for the MRBFSS scan approach in the context of multivariate functional data with the function `plotCurves`

and the new rank-based scan statistic for univariate functional data presented in Section 2. 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.
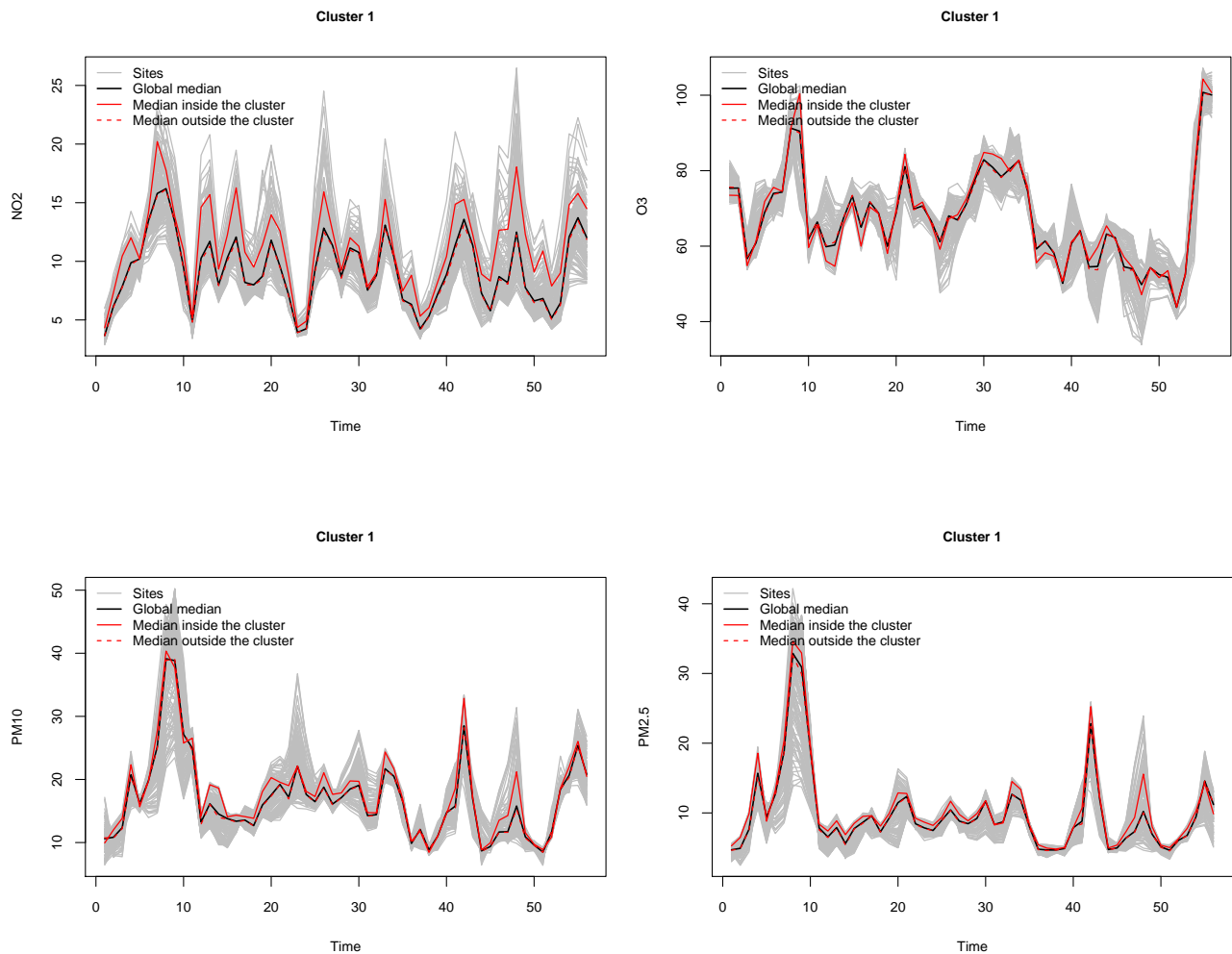
Figure 6.12: Characterization of the most likely cluster for the MRBFSS scan approach in the context of multivariate functional data with the function `plotSummary`

7

Chapter

General conclusion and perspectives

Contents

1 Conclusion

In this thesis, we considered the spatial analysis of high-dimensional and survival data. More precisely, we were interested in spatial scan statistics in order to detect statistically significant clusters on functional and survival data. We first proposed new spatial scan statistics for univariate functional data (Chapter 3). We then extended this approach to multivariate functional data (Chapter 4) and developed an R package to make them easily usable by practitioners (Chapter 6). Lastly, we proposed a new spatial scan statistic for survival data based on a Bayesian approach (Chapter 5).

In Chapter 3, we proposed two new spatial scan statistics for univariate functional data. These are based on an ANOVA test statistic for functional data and a Student's pointwise test statistic. After having evaluated the performance of the approaches on simulated data and compared them with the nonparametric method proposed by Smida et al. (2022), we applied them on economic data corresponding to the unemployment rate curves in France. Our methods detected two statistically significant clusters of higher unemployment rates than elsewhere. The work presented in this chapter was published in *Spatial Statistics*.

Chapter 4 developed three spatial scan statistics for cluster detection on multivariate functional data. These methods are based on a MANOVA test statistic for multivariate functional data, a pointwise Hotelling's T^2-test statistic and a pointwise Wilcoxon-Mann-Whitney test statistic for multivariate data (Oja and Randles, 2004). We also proposed an adaptation of the method of Smida et al. (2022) to the multivariate functional framework in order to compare the performance of our approaches with the latter. Finally, we sought to determine the presence of statistically significant (low or high) pollution clusters in the north of France by applying the methods to the daily concentration of the pollutants NO_2, O_3, PM_{10} and $PM_{2.5}$. This work is currently under minor revision in the *Journal of the Royal Statistical Society: Series C*

(Applied Statistics).

Chapter 6 presented the R package **HDSpatialScan**, available on the CRAN, which implements the spatial scan statistics for univariate and multivariate functional data presented in Chapters 3 and 4, and the approach proposed by Smida et al. (2022). Note that this package also implements other spatial scan statistics proposed in the literature, e.g., the approaches for multivariate data proposed by Cucala et al. (2017) and Cucala et al. (2019). This development was published in *The R Journal*.

Chapter 5 developed a new spatial scan statistic for survival data. The proposed method presents the particularity of taking into account the possible correlations, both between individuals of the same spatial unit, but also between spatial units, by considering a Cox model with spatially structured frailties. After having evaluated the impact of the presence of a correlation between the survival times of the individuals of the same spatial unit on the approaches of the literature, we evaluated the performance of our approach in the presence of a dependence between the individuals of the same spatial unit and between spatial units. In this simulation study we also evaluated the impact of censored data on cluster detection. Finally we applied the method for detecting spatial clusters of abnormal survival times in elderly people with end-stage kidney failure in northern France.

2 Perspectives

In our contributions, some questions and remarks appear, leading to perspectives that will contribute to future research works. In particular we will distinguish mathematical developments both in the direct continuation of this thesis, and broader developments, in the context of spatial analysis, as well as applications in the context of various scientific projects. These perspectives are the subject of the following sections.

2.1 Mathematical developments

2.1.1 Direct perspectives in spatial scan statistics

In this thesis, the proposed spatial scan statistics methods for functional data do not allow to take into account the underlying at-risk population. An unequally distributed population across the territory leads to heteroskedasticity in the aggregated data at the scale of spatial units, which is not taken into account by our models. An idea to account for it could be to consider an approach similar to Cucala (2014) that proposed to use a weighting in the construction of the spatial scan statistic. This weighting allows in particular to adjust the cluster detection on the at-risk population.

Moreover, our proposed methods, as well as that of Smida et al. (2022), do not allow to adjust the cluster detection on confounding factors. In some applications, it would be interesting to adjust the cluster detection on such factors, as the population average age, or the gender (proportion of men for example). These adjustments avoid the detection of clusters that are related to the underlying structure of the population (gender or socio-economic level for example). An idea that we intend to investigate is to consider the following functional linear model (Antoch et al., 2008; Benatia et al., 2017):

$$Y_i(t) = \beta_0(t) + \int_{\mathcal{T}} \beta(s,t)^\top X_i(s) \, \mathrm{d}s + \alpha_w(t) \mathbb{1}_{s_i \in w} + \varepsilon_i(t), \tag{7.1}$$

where $X_i \in \mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$ is a multivariate functional covariate associated with $Y \in \mathcal{L}^2(\mathcal{T}', \mathbb{R})$. $\alpha_w \in \mathcal{L}^2(\mathcal{T}', \mathbb{R})$ is the effect associated with a potential cluster w, $\beta_0 \in \mathcal{L}^2(\mathcal{T}', \mathbb{R})$, $\beta \in \mathcal{L}^2(\mathcal{T} \times \mathcal{T}', \mathbb{R}^p)$ and the ε_i are independent and identically distributed (i.i.d.) random variables in $\mathcal{L}^2(\mathcal{T}', \mathbb{R})$ such that $\forall t \in \mathcal{T}', \mathbb{E}[\varepsilon_i(t)] = 0$.

Thus, in the context of cluster detection, the null hypothesis (the absence of a cluster) is $\mathcal{H}_0 : \forall w \in \mathcal{W}, \alpha_w = \tilde{0}$ and the alternative hypothesis associated with a potential cluster w can be defined as $\mathcal{H}_1^{(w)} : \alpha_w \neq \tilde{0}$, where $\tilde{0}$ denotes the zero function.

In the context of spatial survival data, we proposed a first method of spatial scan statistics to take into account a possible correlation between the survival times of the individuals in the same spatial unit, as well as a possible spatial dependence between spatial units. However, to the best of our knowledge, there is no spatial scan statistic available in the literature dealing with recurrent events, despite a large number of potential applications.

For example, one might be interested in the time until asthma attack in patients treated for asthma, during a period of study where a patient may experience several attacks. This is typically a recurrent event case.

One possible approach is to consider a model with individual frailty that allows to take into account the dependence between successive events of an individual that may be due to unobserved subject-specific factors (Kleinbaum and Klein, 2012). Moreover, the time to asthma attack may also exhibit an intra-spatial unit correlation (due to environmental factors for example), and the spatial units may also be spatially dependent (close spatial units may tend to present similar environmental conditions). Thus, we propose to consider a nested frailty model (Rondeau, 2010) allowing to consider both frailties at the level of individuals and frailties at the level of spatial units. The latter can also be spatially correlated and in this case we can consider a conditional autoregressive Leroux CAR model. Then, the instantaneous hazard rate at time t for the jth event of individual $i_n^{(k)}$ in spatial unit s_k is:

$$\lambda_{i_n^{(k)}, j}(t) = \lambda_0(t) \exp\left[\beta^\top Z_{i_n^{(k)}, j} + \nu_{i_n^{(k)}} + \varphi_k\right]$$

where $\nu_{i_n^{(k)}}$ is the frailty associated with the individual $i_n^{(k)}$ and φ_k is the frailty associated with the spatial unit s_k, $\nu_{i_n^{(k)}}$ and φ_k are independent and such that $\mathbb{E}[\nu_{i_n^{(k)}}] = 0$ and $\mathbb{E}[\varphi_k] = 0$. $Z_{i_n^{(k)}, j}$ is a vector of p covariates associated with the survival. Such a model will allow to take into account both unobserved covariates at the level of individuals (e.g. their tobacco consumption), and unobserved covariates at the level of spatial units (e.g. air pollution or presence of pollen). We intend to investigate this model in our future work.

Another possible direction is to develop spatial scan statistic models for survival data accounting for competing risks.

Indeed, in the application of Chapter 5, we considered epidemiological data of elderly patients with end-stage renal disease (ESRD). On these data we observed a small percentage (0.6%) of patients who had received a renal transplant. However, this percentage is higher in the general population (Couchoud et al., 2015) and it is known that renal transplantation is a competing event for survival of patients with ESRD. Not taking this into account leads to a bias in the estimation of the survival function (Hallan et al., 2012). One idea could be to modify the method proposed in Chapter 5 to take competing risks into account, for example by considering the model proposed by Fine and Gray (1999). We could also go further by considering multi-state models. Typically, when one is interested in modeling the time until cancer in patients, these patients will move through different states (Akwiwu et al., 2022). This type of modeling may allow the detection of spatial clusters in which patients move more quickly from one state to another, in other words, geographical areas in which the health status of patients worsens more quickly. Note that this example consists of a sequence of progressive states but multi-state models also allow the modeling of competing risks for example (Figure 7.1).
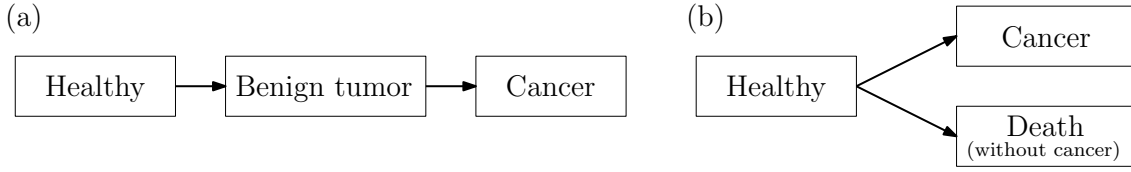
Figure 7.1: Examples of multi-states models. Model (a) refers to a progressive model whereas model (b) is a competing risk model

These developments will be the subject of future works in collaboration with Mohamed-Salem Ahmed, Sophie Dabo-Niang and Michaël Génin.

2.1.2 Broader perspectives in spatial data analysis

In this thesis we were interested in spatial data analysis in the context of cluster detection and more particularly in spatial scan statistics for high dimensional data and survival data. To go beyond the contributions in this area, we aim in the framework of cluster analysis, to develop a spatial clustering method for high dimensional data. Furthermore, in the context of spatial analysis, we would also like to investigate supervised learning methods. Thus, in order of priority, we will conduct the following work: (i) the development of a spatial regression model with a functional covariate and a real response variable, (ii) the extension of this approach for a response variable valued in $\mathbb{R}^p$ and (iii) the development of a spatial clustering method for high dimensional data. In a longer term perspective we would also like to adapt the approaches (i) and (ii) to non-linear models using for example kernel methods.

These broader perspectives require the definition of new concepts. Indeed, we consider in approaches (i) and (ii) the notion of signatures. These allow to solve the problem of large dimension of the functional random variables without supposing that these are square integrable. They are therefore applicable to a wider category of processes than the classical approaches consisting in projecting the functional data in a basis of functions in the space of square integrable functions. Moreover, the signatures allow to better capture the differences between some kind of curves (O'Rourke and Washington, 1985; Fermanian, 2022). Moreover, in these approaches we consider a spatial regression model, namely a spatial autoregressive model, allowing to model endogenous interaction effects on the response variable. Finally in the three approaches and more particularly in the proposed approach (iii), we use the notion of tensors. Briefly, these generalize matrices to higher dimensional data. Typically the observation of an image over time can be modeled by a three dimensional tensor (two dimensions for the image and one for the time). These concepts will be briefly recalled in these perspectives, although the reader can find more details in the Appendix D.

2.1.2.1 A supervised approach for the analysis of spatial functional data: A functional spatial autoregressive model

In the usual FSAR model, the relationship between a real response variable Y and a functional explanatory variable X ($X \in \mathcal{L}^2(\mathcal{T})$) is modelled by the following functional spatial autoregressive model (FSAR) with endogenous interactions:

$$Y_i = \rho_0 \sum_{j=1}^{n} v_{ij} Y_j + \int_{\mathcal{T}} X_i(t)\theta^*(t)dt + U_i, \qquad i = 1, \ldots, n, \ n = 1, 2, \ldots \qquad (7.2)$$

where the autoregressive parameter ρ_0 is in a compact space $\mathcal{R}$, $\theta^*(\cdot)$ is a parameter function assumed to belong to the space of functions $\mathcal{L}^2(\mathcal{T})$, and $(v_{ij})_{j=1,\ldots,n}$ is the ith row of a weight

matrix V_n. Here X_i, Y_i, $i = 1, \ldots, n$ are observations of X and Y at given spatial location s_i.

Here we assume that X is a continuous path of bounded variation taking values in $\mathbb{R}^p, p \geq 2$:

$$
\begin{aligned}
X : [0, 1] &\rightarrow \mathbb{R}^p \\
t &\mapsto (X_t^{(1)}, \ldots, X_t^{(p)})^\top
\end{aligned}
$$

with

$$
\sup_{(t_0, \ldots, t_k) \in I} \sum_{i=1}^{k} ||X_{t_i} - X_{t_{i-1}}|| < \infty
$$

where $I = \{(t_0, \ldots, t_k) | k > 0, 0 = t_0 < \cdots < t_k = 1\}$. Note that this assumption is less restrictive than $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R}^p)$.

Then the signature $Sig(X)$ of X over the interval $\mathcal{T}$ is

$$
Sig(X) = (1, \mathbf{X}^1, \ldots, \mathbf{X}^d, \ldots)
$$

where

$$
\mathbf{X}^d = \int \cdots \int_{0 \leq t_1 < \cdots < t_d \leq 1} dX_{t_1} \otimes \ldots \otimes dX_{t_d} \in (\mathbb{R}^p)^{\otimes d}.
$$

The truncated signature of order D is $Sig^D(X) = (1, \mathbf{X}^1, \ldots, \mathbf{X}^D)$, for every integer $D \geq 1$. Let the tensor algebra space

$$
T((\mathbb{R}^p)) := \{(a_0, a_1, \ldots, a_d, \ldots) | \forall d \geq 0, a_d \in (\mathbb{R}^p)^{\otimes d}\}
$$

and the dth truncated tensor algebra space

$$
T^d((\mathbb{R}^p)) := \bigoplus_{i=0}^{d} (\mathbb{R}^p)^{\otimes i},
$$

by convention $(\mathbb{R}^p)^{\otimes 0} = \mathbb{R}$.
Then the signature $Sig(X)$ is valued in $T((\mathbb{R}^p))$ while the truncated version of order D is an element of $T^D((\mathbb{R}^p))$.

Let $(e_i)_{i=1}^{p}$ be the canonical basis of $\mathbb{R}^p$, then $(e_{i_1} \otimes \cdots \otimes e_{i_d})_{(i_1, \ldots, i_d) \in [\![1, p]\!]^d}$ form a basis of $(\mathbb{R}^p)^{\otimes d}$ (see Remark 2.5 from Levin et al. (2016)) and the signature of X can be rewritten as

$$
Sig(X) = 1 + \sum_{d=1}^{\infty} \sum_{(i_1, \ldots, i_d)} \mathcal{S}_{(i_1, \ldots, i_d)}(X) e_{i_1} \otimes \cdots \otimes e_{i_d} \in T((\mathbb{R}^p)), \tag{7.3}
$$

where $\mathcal{S}_{(i_1, \ldots, i_d)}(X)$ is the signature coefficient of X indexed by $(i_1, \ldots, i_d)$:

$$
\mathcal{S}_{(i_1, \ldots, i_d)}(X) = \int \cdots \int_{0 \leq t_1 < \cdots < t_d \leq 1} dX_{t_1}^{(i_1)} \ldots dX_{t_d}^{(i_d)} \in \mathbb{R}.
$$

In the following we consider the signature functional linear regression model:

$$
Y_i = \rho_0 \sum_{j=1}^{n} v_{ij} Y_j + \langle Sig(\theta^*), Sig(X_i) \rangle + U_i, \tag{7.4}
$$

where the autoregressive parameter ρ_0 is in a compact space $\mathcal{R}$, $\theta^*(\cdot)$ is a parameter function assumed to belong to $\mathcal{C}_1(\mathcal{T}, \mathbb{R}^p)$, the set of continuous path of bounded variation taking values

in $\mathbb{R}^p$, so is X_i. We assume that $v_{ij} = O(h_n^{-1})$ uniformly in all i, j, where the rate sequence h_n can be bounded or divergent, such that $h_n = o(n)$.

The disturbances $\{U_i, \ i = 1, \ldots, n, \ n = 1, 2, \ldots\}$ are assumed to be independent random variables such that $\mathbb{E}[U_i] = 0$ and $\mathbb{E}[U_i^2] = \sigma_0^2$. They are supposed to be independent of $\{X_i(t), t \in \mathcal{T}, \ i = 1, \ldots, n, \ n = 1, 2 \ldots\}$.

Let $\mathbf{X}_n(\theta^*)$ be the $n \times 1$ vector of ith element $\langle Sig(\theta^*), Sig(X_i)\rangle$; then, one can rewrite (7.4) as

$$S_n \mathbf{Y}_n = \mathbf{X}_n(\theta^*) + \mathbf{U}_n, \qquad n = 1, 2, \ldots$$

where $S_n = (I_n - \rho_0 V_n)$, $\mathbf{Y}_n$ and $\mathbf{U}_n$ are two $n \times 1$ vectors of elements Y_i and U_i, $i = 1, \ldots, n$ respectively, and I_n denotes the $n \times n$ identity matrix.

Let $S_n(\rho) = I_n - \rho V_n$, so the conditional quasi log-likelihood function of the vector $\mathbf{Y}_n$ given $\{Sig(X_i), \ i = 1, \ldots, n, \ n = 1, 2 \ldots\}$ is given by:

$$\ell_n(\rho, \theta(.), \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) + \log|S_n(\rho)|$$
$$- \frac{1}{2\sigma^2}[S_n(\rho)\mathbf{Y}_n - \mathbf{X}_n(\theta)]^\top [S_n(\rho)\mathbf{Y}_n - \mathbf{X}_n(\theta)]. \qquad (7.5)$$

The quasi-maximum likelihood estimates of ρ_0, $Sig(\theta^*)$ and σ_0^2 are the parameters ρ, $Sig(\theta)$, and σ^2 that maximize (7.5). However this likelihood cannot be maximized without addressing the difficulty produced by the infinite dimension of the signatures $Sig(\theta^*)$ and $Sig(X)$. To solve this problem, we use the truncated signature.

Let $\{(\varphi_{(i_1,\ldots,i_d)})_{(i_1,\ldots,i_d) \in [\![1,p]\!]^d}, \varphi_{(i_1,\ldots,i_d)} = e_{i_1} \otimes \cdots \otimes e_{i_d}\}$. Using an expansion on this orthonormal basis of $(\mathbb{R}^p)^{\otimes d}$, we can write $Sig(X)$ and $Sig(\theta^*)$ as follows:

$$Sig(X) = 1 + \sum_{d=1}^{\infty}\sum_{(i_1,\ldots,i_d)} \mathcal{S}_{(i_1,\ldots,i_d)}(X)\varphi_{(i_1,\ldots,i_d)} \text{ and } Sig(\theta^*) = 1 + \sum_{d=1}^{\infty}\sum_{(i_1,\ldots,i_d)} \mathcal{S}_{(i_1,\ldots,i_d)}(\theta^*)\varphi_{(i_1,\ldots,i_d)}.$$

Let D_n be a positive sequence of integers that increase asymptotically as $n \to \infty$, then

$$\langle Sig(X), Sig(\theta^*)\rangle = 1 + \sum_{d=1}^{D_n}\sum_{(i_1,\ldots,i_d)} \mathcal{S}_{(i_1,\ldots,i_d)}(X)\mathcal{S}_{(i_1,\ldots,i_d)}(\theta^*) + \sum_{d=D_n+1}^{\infty}\sum_{(i_1,\ldots,i_d)} \mathcal{S}_{(i_1,\ldots,i_d)}(X)\mathcal{S}_{(i_1,\ldots,i_d)}(\theta^*).$$
$$(7.6)$$

Let the coefficient-signature of X be the sequence of all signature coefficients, denoted by $\mathcal{S}(X)$,

$$\mathcal{S}(X) = (1, \mathcal{S}_{(1)}(X), \ldots, \mathcal{S}_{(p)}(X), \mathcal{S}_{(1,1)}(X), \mathcal{S}_{(1,2)}(X), \ldots, \mathcal{S}_{(i_1,\ldots,i_d)}(X), \ldots)$$

and the truncated coefficient-signature at order D be

$$\mathcal{S}^D(X) = (1, \mathcal{S}_{(1)}(X), \mathcal{S}_{(2)}(X), \ldots, \underbrace{\mathcal{S}_{(p,\ldots,p)}(X)}_{D \text{ terms}}).$$

Then, the strategy consists of approximating $\langle Sig(X), Sig(\theta^*)\rangle$ in (7.6) using only the first term of the right-hand side and the corresponding truncated coefficient-signature:

$$\langle Sig(X), Sig(\theta_n^*)\rangle \approx \mathcal{S}^{D_n}(X)^\top \mathcal{S}^{D_n}(\theta^*).$$

This is possible when the second term of the right-hand side vanishes asymptotically. In the following we assume that this assumption is satisfied and we consider the smallest $D^* = D_n^*$

verifying this.

Under this truncation strategy, and by noting $s_p(D) = \sum_{d=0}^{D} p^d = \frac{p^{D+1}-1}{p-1}$ the dimension of the truncated coefficient-signature $\mathcal{S}^D(X)$, $\mathbf{X}_n(\theta^*)$ may be approximated by $\xi_{D_n}\theta^*$, where $\theta^* = (\theta_1^*, \ldots, \theta_{s_p(D_n)}^*)^\top = \mathcal{S}^{D_n}(\theta^*)^\top$ $(\theta_1^* = 1)$ and ξ_{D_n} is an $n \times s_p(D_n)$ matrix whose i^{th} line is given by

$$\varepsilon^{(i)} = \mathcal{S}^{D_n}(X_i), \qquad i = 1, \ldots, n.$$

We suppose that the $\mathcal{S}(X_i)$ and so the $\varepsilon^{(i)}$ are centered (which can be easily satisfied by subtracting their average).

Therefore the truncated conditional quasi log-likelihood function is

$$
\begin{aligned}
\tilde{\ell}_n(\rho, \theta, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) + \log|S_n(\rho)| \\
&\quad - \frac{1}{2\sigma^2}\left[S_n(\rho)\mathbf{Y}_n - \xi_{D_n}\theta\right]^\top\left[S_n(\rho)\mathbf{Y}_n - \xi_{D_n}\theta\right].
\end{aligned}
\tag{7.7}
$$

For a fixed ρ, (7.7) is maximized at

$$\hat{\theta}_{n,\rho} = (\xi_{D_n}^\top\xi_{D_n})^{-1}\xi_{D_n}^\top S_n(\rho)\mathbf{Y}_n = (\hat{\theta}_{j,n,\rho})_{j=1,\ldots,s_p(D_n)}$$

and

$$
\begin{aligned}
\hat{\sigma}_{n,\rho}^2 &= \frac{1}{n}\left(S_n(\rho)\mathbf{Y}_n - \xi_{D_n}\hat{\theta}_{n,\rho}\right)^\top\left(S_n(\rho)\mathbf{Y}_n - \xi_{D_n}\hat{\theta}_{n,\rho}\right) \\
&= \frac{1}{n}\mathbf{Y}_n^\top S_n^\top(\rho)M_n S_n(\rho)\mathbf{Y}_n,
\end{aligned}
$$

where $M_n = I_n - \xi_{D_n}(\xi_{D_n}^\top\xi_{D_n})^{-1}\xi_{D_n}^\top$.

The concentrated truncated conditional quasi log-likelihood function of ρ is:

$$\tilde{\ell}_n(\rho) = -\frac{n}{2}[\log(2\pi) + 1] - \frac{n}{2}\log(\hat{\sigma}_{n,\rho}^2) + \log|S_n(\rho)|.$$

Then the estimator of ρ_0 is $\hat{\rho}_n$, which maximizes $\tilde{\ell}_n(\rho)$, and those of the vector θ^* and variance σ_0^2 are, respectively, $\hat{\theta}_{n,\hat{\rho}_n}$ and $\hat{\sigma}_{n,\hat{\rho}_n}^2$. The corresponding estimator of the signature parameter $Sig(\theta^*)$ is:

$$Sig(\hat{\theta}_n) = 1 + \sum_{d=1}^{D_n}\sum_{(i_1,\ldots,i_d)}\hat{\theta}_{(i_1,\ldots,i_d),n,\hat{\rho}_n}\varphi_{(i_1,\ldots,i_d)},$$

where $\hat{\theta}_{(i_1,\ldots,i_d),n,\hat{\rho}_n}$ is the element of the estimated truncated signature $\hat{\theta}_{n,\hat{\rho}_n}$ related to $(i_1,\ldots,i_d)$.

In the continuation of this work we intend to show the following asymptotic convergence theorems, under mild assumptions:

Theorem 1. $\hat{\rho}_n$ is consistent and satisfies

$$\sqrt{\frac{n}{h_n}}(\hat{\rho}_n - \rho_0) \underset{n\to\infty}{\to} \mathcal{N}(0, s_\rho^2),$$

with $s_\rho^2 = \lim\limits_{n\to\infty}\frac{s_n^2 h_n}{n}\left\{\frac{h_n}{n}\left[\Delta_n + \sigma_0^2\mathrm{tr}((G_n^\top + G_n)G_n)\right]\right\}^{-2}$, where

$$s_n^2 = (\mu_4 - 2\sigma_0^4)\sum_{i=1}^{n}G_{ii}^2 + \frac{1}{n}\mathrm{tr}^2(G_n)(\sigma_0^4 - \mu_4 - \sigma_0^2\theta^{*\top}\Gamma_{D_n}\theta^*) + \mathrm{tr}(G_n G_n^\top)(\sigma_0^4 + \sigma_0^2\theta^{*\top}\Gamma_{D_n}\theta^*),$$

$G_n = V_n S_n^{-1}$, $\Delta_n = n \left(\text{tr} \left(\frac{G_n^\top G_n}{n} \right) - \text{tr}^2 \left(\frac{G_n}{n} \right) \right) \theta^{*\top} \Gamma_{D_n} \theta^*$, and $\Gamma_{D_n} = \mathbb{E} \left(\frac{1}{n} \xi_{D_n}^\top \xi_{D_n} \right)$.

Theorem 2. $\hat{\sigma}_n^2$ is a consistent estimator of σ_0^2 and satisfies

$$\sqrt{n}(\hat{\sigma}_{n,\hat{\rho}_n}^2 - \sigma_0^2) \underset{n\to\infty}{\to} \mathcal{N}(0, s_\sigma^2),$$

with

$$s_\sigma^2 = \mu_4 - \sigma_0^4 + 4s_\rho^2 \lim_{n\to\infty} h_n \left[\frac{\text{tr}(G_n)}{n} \right]^2.$$

Theorem 3.

$$\frac{n \left(\hat{\theta}_{n,\hat{\rho}_n} - \theta^* \right)^\top \Gamma_{D_n} \left(\hat{\theta}_{n,\hat{\rho}_n} - \theta^* \right) - \sigma_0^2 s_p(D_n)}{\sqrt{2 s_p(D_n)}} \underset{n\to\infty}{\to} \mathcal{N}(0, \sigma_0^4).$$

2.1.2.2 A supervised approach for the analysis of spatial functional data: A multivariate functional spatial autoregressive model

We previously proposed a new functional spatial autoregressive model (FSAR) with a scalar response $Y \in \mathbb{R}$ and a functional covariate X. Its specificity is to consider the signature of X instead of using a projection of X in a basis of functions.

We are also currently working on a multivariate spatial autoregressive model (MSAR) (Yang and Lee, 2017; Zhu et al., 2020) with functional covariates, modeling functional covariates by their signatures with Issa-Mbenard Dabo (University of Bordeaux). Briefly, the MSAR model proposed by Yang and Lee (2017); Zhu et al. (2020) considers a response variable $Y \in \mathbb{R}^Q$ and an explanatory variable $X \in \mathbb{R}^P$. $(Y, X) \in \mathbb{R}^Q \times \mathbb{R}^P$ is observed in n spatial units $s_1, \ldots, s_n$. The authors considered the following model:

$$\boldsymbol{Y} = V\boldsymbol{Y}R + \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y} = (Y_1 | \ldots | Y_n)^\top \in \mathbb{R}^{n\times Q}$, $\boldsymbol{X} = (X_1 | \ldots | X_n)^\top \in \mathbb{R}^{n\times P}$, and $\boldsymbol{\varepsilon} = (\varepsilon_1 | \ldots | \varepsilon_n)^\top \in \mathbb{R}^{n\times Q}$ and $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}^Q$ are i.i.d. random variables such that $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{V}[\varepsilon_i] = \Sigma$, $\forall i \in [\![1, n]\!]$. V is a spatial weights $n \times n$ matrix and R is a $Q \times Q$ matrix such that the diagonal elements represent the own-variable spatial effects $(\rho_{q,q})$ and the off-diagonal elements represent the cross-variable spatial effects $(\rho_{q,q'}, q \neq q')$ (Yang and Lee, 2017).

Here we propose a multivariate spatial autoregressive model with functional covariates: we suppose $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R}^P)$ and we note $\boldsymbol{X}(t) = (X_1(t) | \ldots | X_n(t))^\top \in \mathbb{R}^{n\times P}, \forall t \in \mathcal{T}$. Then we propose the following multivariate spatial autoregressive model with functional covariates:

$$\boldsymbol{Y} = V\boldsymbol{Y}R + \int_{\mathcal{T}} \boldsymbol{X}(t)\beta(t) \, \mathrm{d}t + \boldsymbol{\varepsilon}.$$

Then we propose not to consider a projection of X in a basis of functions but rather to model its effect through its signature. Let $S^m(X_i)$ be the truncated signature of $X_i(.)$ at order m: $S^m(X_i) \in \mathbb{R}^{s_P(m)}$.

The model then becomes

$$\boldsymbol{Y} = V\boldsymbol{Y}R + \boldsymbol{S^m}(\boldsymbol{X})\beta_m + \boldsymbol{\varepsilon}$$

where $\boldsymbol{S^m}(\boldsymbol{X}) = (S^m(X_1) | \ldots | S^m(X_n))^\top \in \mathbb{R}^{n\times s_P(m)}$, $\beta_m \in \mathbb{R}^{s_P(m)}$.
And it can be rewritten as

$$\boldsymbol{Y} - V\boldsymbol{Y}R = \boldsymbol{S^m}(\boldsymbol{X})\beta_m + \boldsymbol{\varepsilon}.$$

Note that the use of signatures allows in fact to assume only that X is a continuous path of bounded variation taking values in $\mathbb{R}^P$. The continuation of this work is in progress.

In future work we would like to extend these spatial regression models for functional data to non linear models using kernel approaches.

2.1.2.3   An unsupervised approach to cluster spatial high-dimensional data

Nowadays, with the increase of storage and computational capacities, more and more high-dimensional data are available. These can be modeled by tensors, which are a generalization of matrices for high-dimensional space. This led to the development of new analysis methods for tensors such as regression (Li and Zhang, 2017; Zhou et al., 2021) or clustering (Sun and Li, 2019; Cai et al., 2021). In the context of spatial data analysis, spatial clustering methods allow to partition data into groups by taking into account both the proximity of observations and the proximity of their spatial locations. In the context of prevention campaigns, this allows for example to adapt the campaigns locally according to the behavior or socio-economic characteristics of individuals. However, to our knowledge, there exists no spatial clustering method for high-dimensional data modeled by tensors. In this context, our objective is to develop a spatial clustering method for tensors. This is the purpose of this section.

Method proposed by Cai et al. (2021). It concerns a non spatial clustering approach of tensors that we briefly describe here. Let $\chi_1, \ldots, \chi_n \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_M}$ be n M-dimensional tensors. It is assumed that these are coming from the following mixture model:

$$Z \sim \text{Multinomial}(\pi_1^*, \ldots, \pi_K^*)$$
$$\chi | Z = k \sim \mathcal{N}_T(U_k^*, \underline{\Sigma}_k^*)$$

where $\mathcal{N}_T$ is the tensor normal distribution. Then

$$\chi_1, \ldots, \chi_n \overset{i.i.d.}{\sim} \pi_1^* \mathcal{N}_T(U_1^*, \underline{\Sigma}_1^*) + \cdots + \pi_K^* \mathcal{N}_T(U_K^*, \underline{\Sigma}_K^*).$$

The aim of this clustering approach is to find the groups of the χ_i and to estimate the parameters U_k^* and $\underline{\Sigma}_k^*$.

However in general, since the tensors are high dimensional, there are more parameters to estimate than observations. The idea is therefore to make assumptions to reduce the dimension:

Assumption 1: the U_k^* are low-rank: $U_k^* = \displaystyle\sum_{r=1}^{R} w_{k,r}^* \beta_{k,r,1}^* \circ \beta_{k,r,2}^* \circ \cdots \circ \beta_{k,r,M}^*.$

Assumption 2: the U_k^* are sparse: the sparsity is supposed on the $\beta_{k,r,m}^*$.

Assumption 3: the $\underline{\Sigma}_k^*$ are separable: the precision $\underline{\Omega}_k^*$ is such that $\underline{\Omega}_k^* = \Omega_{k,M}^* \otimes \cdots \otimes \Omega_{k,2}^* \otimes \Omega_{k,1}^*$. Then $\underline{\Sigma}_k^* = \Sigma_{k,M}^* \otimes \cdots \otimes \Sigma_{k,2}^* \otimes \Sigma_{k,1}^*$ where $\Omega_{k,m}^* = \Sigma_{k,m}^{*-1}$.

Assumption 4: the $\underline{\Omega}_k^*$ are sparse: the $\Omega_{k,m}^*$ are sparse.

Let $\Theta = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)^\top$ where
$\theta_k = (\beta_{k,1,1}^\top, \ldots, \beta_{k,1,M}^\top, w_{k,1}, \ldots, \beta_{k,R,1}^\top, \ldots, \beta_{k,R,M}^\top, w_{k,R}, vec(\Omega_{k,1})^\top, \ldots, vec(\Omega_{k,M})^\top).$

Then the log-likelihood for (χ, Z) is

$$\ell(\Theta | \chi, Z) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{Z_i = k} [\log (f_k(\chi_i | \theta_k)) + \log (\pi_k)].$$

The parameters are then estimated by a conditional expectation-maximization method for high dimensional data (*via* a penalization of the different parameters to ensure sparsity). The idea is to compute at step (t+1)

$$Q_n(\Theta|\Theta^{(t)}) = \mathbb{E}_{Z|\chi,\Theta^{(t)}}[\ell(\Theta|\chi, Z)]$$

where

$$\mathbb{E}_{Z|\chi,\Theta^{(t)}}[\ell(\Theta|\chi, Z)] = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{E}_{Z|\chi,\Theta^{(t)}}[\mathbb{1}_{Z_i=k}(\log(f_k(\chi_i|\theta_k)) + \log(\pi_k))]$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{P}(Z_i = k|\chi_i, \Theta^{(t)})[\log(f_k(\chi_i|\theta_k)) + \log(\pi_k)].$$

Let

$$\tau_{i,k}(\Theta^{(t)}) = \mathbb{P}(Z_i = k|\chi_i, \Theta^{(t)})$$

$$= \frac{f_k(\chi_i|\Theta^{(t)})\mathbb{P}(Z_i = k|\Theta^{(t)})}{\sum_{k'=1}^{K} f_{k'}(\chi_i|\Theta^{(t)})\mathbb{P}(Z_i = k'|\Theta^{(t)})}$$

$$= \frac{f_k(\chi_i|\theta_k^{(t)})\pi_k^{(t)}}{\sum_{k'=1}^{K} f_{k'}(\chi_i|\theta_{k'}^{(t)})\pi_{k'}^{(t)}}.$$

Then $Q_n(\Theta|\Theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{i,k}(\Theta^{(t)})[\log(f_k(\chi_i|\theta_k)) + \log(\pi_k)].$

The parameters are then estimated iteratively by maximizing

$$Q_n(\Theta|\Theta^{(t)}) - \sum_{k,r,m} \lambda_0^{(t+1)}||\beta_{k,r,m}||_1 - \sum_{k,m} \lambda_m^{(t+1)}||\Omega_{k,m}||_{1,\text{off}}$$

where $||A||_{1,\text{off}} = \sum_{i\neq j}|A_{i,j}|$ and $||x||_1 = \sum_i |x_i|$.

Algorithm 2 briefly presents the clustering algorithm steps.

Our proposition. We propose to adapt the method of Cai et al. (2021) for spatial clustering.

By noting Z_i the group of χ_i, we suppose $\chi_i|Z_i = k \sim \mathcal{N}_T(U_k^*, \Sigma_k^*)$.
We also assume a joint probability function of the Potts model for $Z_1, \ldots, Z_n$: $\pi(Z_1, \ldots, Z_n) \propto \exp\left[\sum_{j>i} w_{i,j}\mathbb{1}_{Z_i=Z_j}\right]$ (Sugasawa and Murakami, 2021), where $w_{i,j}$ allows to model the spatial dependence. $w_{i,j}$ can be a function of the distance between s_i and s_j, or a function of their adjacency for example.

Then the likelihood can be written as

$$\ell(\Theta|\chi, Z) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}_{Z_i=k}\left[\log(f_k(\chi_i|\theta_k)) + \sum_{j>i} w_{i,j}\mathbb{1}_{Z_j=k}\right].$$

Algorithm 2: Algorithm proposed by Cai et al. (2021) for the update of the parameters in the clustering approach of tensors

---

Input: $\chi_1, \ldots, \chi_n$, K, R, T, $\lambda_0^{(t)}, \lambda_1^{(t)}, \ldots, \lambda_M^{(t)}$ for $t \in [\![1, T]\!]$
Initialization: $\pi_k^{(0)}, \beta_{k,r,m}^{(0)}, w_{k,r}^{(0)}, \Omega_{k,m}^{(0)}$
for $t \in [\![1, T]\!]$ **do**

     1. Compute $\tau_{i,k}(\Theta^{(t)})$ $\forall i, k$

     2. Compute $\pi_k^{(t+1)} = \dfrac{1}{n} \sum_{i=1}^{n} \tau_{i,k}(\Theta^{(t)})$, $\forall k$

     3. Compute $\beta_{k,r,m}^{(t+1)}$ $\forall k, r, m$ with the updated values along k, r, m

     4. Compute $w_{k,r}^{(t+1)}$ $\forall k, r$ using the values $\beta_{k,r,m}^{(t+1)}$ for the $\beta_{k,r,m}$

     5. Compute $\Omega_{k,m}^{(t+1)}$ using the values $\beta_{k,r,m}^{(t+1)}$ and $w_{k,r}^{(t+1)}$ and the updated values of $\Omega_{k',m'}$ as they are obtained

     if convergence **then**
        $\lfloor$ stop

Here

$$
\mathbb{E}_{Z|\underline{\chi},\Theta^{(t)},Z_1^{(t)},\ldots,Z_n^{(t)}}[\ell(\Theta|\chi, Z)] = \left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{Z|\underline{\chi},\Theta^{(t)},Z_1^{(t)},\ldots,Z_n^{(t)}} \left[\mathbb{1}_{Z_i=k} \left(\log\left(f_k(\chi_i|\theta_k)\right) + \sum_{j \neq i} w_{i,j} \mathbb{1}_{Z_j=k} \right) \right] \right\}
$$

$$
= \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ \mathbb{E}_{Z|\underline{\chi},\Theta^{(t)},Z_1^{(t)},\ldots,Z_n^{(t)}}[\mathbb{1}_{Z_i=k}] \log\left(f_k(\chi_i|\theta_k)\right) \right.
$$

$$
\left. + \sum_{j \neq i} w_{i,j} \mathbb{E}_{Z|\underline{\chi},\Theta^{(t)},Z_1^{(t)},\ldots,Z_n^{(t)}}[\mathbb{1}_{Z_i=k,Z_j=k}] \right\}
$$

$$
= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}[Z_i = k|\underline{\chi},\Theta^{(t)}, Z_1^{(t)}, \ldots, Z_n^{(t)}] \log\left(f_k(\chi_i|\theta_k)\right)
$$

$$
+ \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j \neq i} w_{i,j} \mathbb{P}[Z_i = k, Z_j = k|\underline{\chi},\Theta^{(t)}, Z_1^{(t)}, \ldots, Z_n^{(t)}].
$$

Since $\mathbb{P}[Z_i = k, Z_j = k|\underline{\chi},\Theta^{(t)}, Z_1^{(t)}, \ldots, Z_n^{(t)}]$ does not depend on θ_k, the idea is to maximize (with a penalization of the parameters)

$$
\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}[Z_i = k|\underline{\chi},\Theta^{(t)}, Z_1^{(t)}, \ldots, Z_n^{(t)}] \log\left(f_k(\chi_i|\theta_k)\right),
$$

that is

$$
\sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{i,k}(\Theta^{(t)}) \log\left(f_k(\chi_i|\theta_k)\right),
$$

with

$$\tau_{i,k}(\Theta^{(t)}) = \mathbb{P}[Z_i = k | \chi_i, \Theta^{(t)}, Z_1^{(t)}, \ldots, Z_n^{(t)}] = \frac{f_k(\chi_i | \theta_k^{(t)}) \exp\left[\sum_{j \neq i} w_{i,j} \mathbb{1}_{Z_j^{(t)} = k}\right]}{\sum_{k''=1}^{K} f_{k''}(\chi_i | \theta_{k''}^{(t)}) \exp\left[\sum_{j \neq i} w_{i,j} \mathbb{1}_{Z_j^{(t)} = k''}\right]}.$$

Our future goal is first to adapt the conditional expectation-maximization algorithm proposed by Cai et al. (2021) to this new model. Then, we intend to show that the proposed methodology succeeds to retrieve the correct number of groups K^* and that for $K = K^*$, the parameters estimators are consistent.

2.2 Applications in health

Having carried out this thesis within a public health research team, I was led to collaborate with several clinical researchers, environmentalists and epidemiologists through several multidisciplinary research projects in the field of health. These different projects and my contributions are briefly presented in the following.

IPACOS project (2021-2024). The project IPACOS (Impact du contexte Pandémique de COVID-19 sur les COnduites Suicidaires et leur prise en charge (Impact of the COVID-19 Pandemic on Suicidal Behavior and its Treatment), financed with 90,000€ from the DREES (Direction de la Recherche, des études, de l'évaluation et des statistiques)) is headed by Professor Guillaume Vaiva (Univ. Lille, Inserm, CHU Lille) and aims to study the impact of the pandemic context on the incidence of suicide attempts and deaths. It consists of three main objectives: (i) to analyze the impact of the pandemic and health measures on suicide attempts and deaths by suicide in the general population and among people with a previous suicide attempt, (ii) to study the impact of the pandemic and health measures on suicidal ideation, and (iii) to analyze the difficulties in managing suicide attempts specific to the pandemic context.

Within the framework of the first objective, we first conducted a study with Maëlle Baillet (data science intern in ILIS), Michaël Génin (assistant professor and hospital practitioner), Antoine Lamer (assistant professor and data scientist) and Marielle Wathelet (hospital practitioner in public health) on the spatio-temporal modeling of the association between the incidence of suicide attempts and the incidence of COVID-19 in France. For this purpose, we considered spatio-temporal hierarchical Bayesian ecological regressions and data from VigilanS (a system for recontacting people who have attempted suicide), as well as data on hospitalizations for suicide attempts (*Programme de Médicalisation des Système d'Information (PMSI)* national database) over a 3-year period starting 1 year before the 1st confinement. This first work is well advanced and will be the subject of an article to be submitted soon.

Then, as an extension of this work, we would like to apply the spatial scan statistics for multivariate functional data developed in Chapter 4 in order to determine if there exist statistically significant clusters of co-occurrence of suicide attempts and COVID-19 in France. This study will be conducted in collaboration with Michaël Génin, Antoine Lamer and Marielle Wathelet.

French-Irish INHALE project (2023). In parallel, in the framework of the INHALE project (Ireland and France Need Healthier Air for healthier Lungs - the Evidence, 5000€ from the Hubert Curien Partnership "Ulysses" and the Irish-French cooperation), I will collaborate with other researchers of the University of Lille (Sophie Dabo-Niang (university professor), Michaël Génin (assistant professor and hospital practitioner) and Mohamed-Salem Ahmed (research scientist)) and the University College Dublin (Ireland) (Michelle Carey (assistant

professor), Uche Mbaka (PhD student) and Thiago Americo Da Silva Cardoso (PhD student)). This project, which is scheduled to begin in January 2023, aims at analyzing the effects of air pollution (NO_2, O_3, PM_{10} and $PM_{2.5}$) on respiratory health (lung cancer, asthma, respiratory failure and acute bronchiolitis) in Ireland and France and to produce highly accurate estimates of air pollution in Ireland and France by accounting for the anisotropy and non-stationarity of the spatial field.

Renal disease and air pollution (2022-2023). In the context of survival data we have applied in Chapter 5 our methods on the French renal epidemiology and information network (REIN) registry in the north of France. This first exploratory study laid the groundwork for a national project to (i) analyze the spatial distribution of mortality in patients with end-stage kidney disease and, (ii) assess the relationship between mortality and outdoor air pollution. This study is based both on data from the national REIN registry, from 2009 to 2020, including 116,724 patients, and on pollution data produced by the European Environment Agency. Moreover, this project has been validated by the scientific council of the Agence de Biomédecine and has been financed with 10,000€.
The study on the national base will be realized in collaboration with Michaël Génin (assistant professor and hospital practitioner), Aghiles Hamroun (university hospital assistant in public health) and Sébastien Gomis (clinical research associate in charge of the REIN registry).

Spatial epidemiology of hip and proximal humerus fractures (2022-2023). The objective of this study is to analyze the spatial distribution of hip and proximal humerus fracture co-occurrence on the French territory based on data from the French hospital discharge database (PMSI national database) in the population of people aged 50 years and older, between 2016 and 2020. Several studies considering these pathologies separately have shown a spatial heterogeneity in France (Maravic et al., 2005; Barbier et al., 2009; Héquette-Ruz et al., 2020). Our objective is to determine if there exist spatial clusters of co-occurrence of these two pathologies by using the spatial scan statistics for multivariate functional data.
This study will be realized in collaboration with Michaël Génin (assistant professor and hospital practitioner), Jean-Baptiste Beuscart (university professor and hospital practitioner in geriatrics), and Antoine Lamer (assistant professor and data scientist).

2.3   Softwares

In order to make our methods accessible to a large number of researchers and practitioners, and easily usable in many domains such as health, environment or industry, we developed the R package "HDSpatialScan" (Frévent et al., 2022) available on the CRAN repository and implementing all the spatial scan statistics for functional data presented in this thesis. In the near future we would like to improve this package by slightly modifying the implementation of the non pointwise methods (NPFSS, PFSS and MPFSS) in order to allow observations not equally spaced in time. We would also like to integrate the possibility for the user to smooth the functional data before applying the spatial scan statistics (currently if the user wishes to perform this step he needs to do it outside the package). We also plan to add the spatial scan statistic for functional data based on the functional linear model presented in Equation 7.1.

In the context of the spatial analysis of survival data, we proposed a spatial scan statistic allowing to take into account both the possible correlation between the survival times of the individuals of the same spatial unit and the possible spatial dependence between spatial units. Thus, in order to make this approach easily accessible to practitioners, we would also like to propose it in an R package. This implementation is made possible by the recent publication

of Van Niekerk et al. (2022) introducing a new formulation of the INLA methodology which shows a considerably reduced computation time. The package may later implement the other spatial scan statistics approaches for survival data presented in these perspectives.

Finally, we also plan to develop an R package implementing the functional spatial autoregressive models with functional covariates presented in these perspectives as well as the nonparametric kernel spatial regression for functional data when the work will be completed.

Bibliography

Abolhassani, A. and Prates, M. O. (2021) An up-to-date review of scan statistics. *Statistics Surveys*, **15**, 111–153.

Abrams, A. M., Kleinman, K. and Kulldorff, M. (2010) Gumbel based p-value approximations for spatial scan statistics. *International Journal of Health Geographics*, **9**, 1–12.

Ahmed, M.-S., Broze, L., Dabo-Niang, S. and Gharbi, Z. (2021a) Quasi-maximum likelihood estimators for functional linear spatial autoregressive models. *Geostatistical Functional Data Analysis*, 286–328.

Ahmed, M.-S., Cucala, L. and Genin, M. (2021b) Spatial autoregressive models for scan statistic. *Journal of Spatial Econometrics*, **2**, 1–20.

Ahmed, M.-S. and Genin, M. (2020) A functional-model-adjusted spatial scan statistic. *Statistics in medicine*, **39**, 1025–1040.

Akwiwu, E. U., Klausch, T., Jodal, H. C., Carvalho, B., Løberg, M., Kalager, M., Berkhof, J. and H Coupé, V. M. (2022) A progressive three-state model to estimate time to cancer: a likelihood-based approach. *BMC medical research methodology*, **22**, 1–16.

Allard, D. and Guillot, G. (2000) Clustering geostatistical data. In *Proceedings of the sixth geostatistical conference*.

Allévius, B. (2018a) *scanstatistics: Space-Time Anomaly Detection using Scan Statistics*. URL: https://CRAN.R-project.org/package=scanstatistics.

— (2018b) scanstatistics: Space-time anomaly detection using scan statistics. *Journal of Open Source Software*, **3**.

Alm, S. E. (1997) On the distributions of scan statistics of a two-dimensional poisson process. *Advances in Applied Probability*, **29**, 1–18.

Anderson and Henderson (1986) Cancer incidence in census tracts with broadcasting towers in Honolulu, Hawaii. *Honolulu City Council. Honolulu, HI: Environmental and Epidemiology Program, Hawaii, cité dans Goldsmith JR, "Epidemiological studies of radio-frequency radiation: current status and areas of concern" : The Science of the Total Environment, 1996.*

Anderson, H. R., de Leon, A. P., Bland, J. M., Bower, J. S. and Strachan, D. P. (1996) Air pollution and daily mortality in London: 1987-92. *Bmj*, **312**, 665–669.

Anderson, N. and Titterington, D. (1997) Some methods for investigating spatial clustering, with epidemiological applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **160**, 87–105.

Anselin, L. (1988) *Spatial econometrics: methods and models*, vol. 4. Springer Science & Business Media.

— (1995) Local indicators of spatial association—lisa. *Geographical analysis*, **27**, 93–115.

— (2009) Spatial regression. *The SAGE handbook of spatial analysis*, **1**, 255–276.

Antoch, J., Prchal, L., Rosa, M. R. D. and Sarda, P. (2008) Functional linear regression with functional response: application to prediction of electricity consumption. In *Functional and Operatorial Statistics*, 23–29. Springer.

Arlinghaus, S. (1995) *Practical handbook of spatial statistics*. CRC press.

Arribas, I. P. (2018) Derivatives pricing using signature payoffs. *arXiv preprint arXiv:1809.09466*.

Assunção, R., Costa, M., Tavares, A. and Ferreira, S. (2006) Fast detection of arbitrarily shaped disease clusters. *Statistics in medicine*, **25**, 723–742.

Aswi, A., Cramb, S., Duncan, E., Hu, W., White, G. and Mengersen, K. (2020) Bayesian spatial survival models for hospitalisation of Dengue: A case study of Wahidin hospital in Makassar, Indonesia. *International Journal of Environmental Research and Public Health*, **17**, 878.

Austin, P. C. (2017) A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review*, **85**, 185–203.

Ayav, C., Beuscart, J.-B., Briançon, S., Duhamel, A., Frimat, L. and Kessler, M. (2016) Competing risk of death and end-stage renal disease in incident chronic kidney disease (stages 3 to 5): the EPIRAN community-based study. *BMC nephrology*, **17**, 1–13.

Bailly, A., Guesnier, B., Paelinck, J. and Sallez, A. (1995) *Stratégies spatiales. Comprendre et maîtriser l'espace*. RECLUS.

Balan, T. A. and Putter, H. (2020) A tutorial on frailty models. *Statistical methods in medical research*, **29**, 3424–3454.

Bali, J. L. and Boente, G. (2014) Robust functional principal component analysis. In *New Advances in Statistical Modeling and Applications*, 41–54. Springer.

Ballari, D., Giraldo, R., Campozano, L. and Samaniego, E. (2018) Spatial functional data analysis for regionalizing precipitation seasonality and intensity in a sparsely monitored region: Unveiling the spatio-temporal dependencies of precipitation in Ecuador. *International Journal of Climatology*, **38**, 3337–3354.

Banerjee, S., Wall, M. M. and Carlin, B. P. (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, **4**, 123–142.

Barbier, S., Ecochard, R., Schott, A.-M., Colin, C., Delmas, P., Jaglal, S. and Couris, C. (2009) Geographical variations in hip fracture risk for women: strong effects hidden in standardised ratios. *Osteoporosis international*, **20**, 371–377.

Belkhiri, L., Tiri, A. and Mouni, L. (2020) Spatial distribution of the groundwater quality using kriging and co-kriging interpolations. *Groundwater for Sustainable Development*, **11**, 100473.

Benatia, D., Carrasco, M. and Florens, J.-P. (2017) Functional linear regression with functional response. *Journal of econometrics*, **201**, 269–291.

Berrendero, J. R., Justel, A. and Svarc, M. (2011) Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, **55**, 2619–2634.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 192–236.

Besag, J. and Newell, J. (1991) The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **154**, 143–155.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, **43**, 1–20.

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

— (2016) A spatial scan statistic for survival data based on generalized life distribution. *Communications in Statistics-Theory and Methods*, **45**, 5730–5744.

Bidin, C. M., de La Fuente Marcos, R., de La Fuente Marcos, C. and Carraro, G. (2010) Not an open cluster after all: the NGC 6863 asterism in Aquila. *Astronomy & Astrophysics*, **510**, A44.

Bierkens, M., Finke, P. and De Willigen, P. (2000) *Upscaling and downscaling methods for environmental research*. Kluwer Academic.

Bithell, J. (1995) The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, **14**, 2309–2322.

Bivand, R. S. and Gómez-Rubio, V. (2021) Spatial survival modelling of business re-opening after Katrina: Survival modelling compared to spatial probit modelling of re-opening within 3, 6 or 12 months. *Statistical Modelling*, **21**, 137–160.

Bohorquez, M., Giraldo, R. and Mateu, J. (2016) Optimal sampling for spatial prediction of functional data. *Statistical Methods & Applications*, **25**, 39–54.

Bouayad Agha, S. and De Bellefon, M.-P. (2018) Spatial autocorrelation indices. *Handbook of Spatial Analysis: Theory Aplication with R*, 51–68.

Bourgault, G., Marcotte, D. and Legendre, P. (1992) The multivariate (co) variogram as a spatial weighting function in classification methods. *Mathematical Geology*, **24**, 463–478.

Bouveyron, C. and Brunet-Saumard, C. (2014) Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, **71**, 52–78.

Bouveyron, C., Jacques, J., Schmutz, A., Simoes, F. and Bottini, S. (2022) Co-clustering of multivariate functional data for the analysis of air pollution in the south of France. *The Annals of Applied Statistics*, **16**, 1400–1422.

Brook, D. (1964) On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, **51**, 481–483.

Brook, R. D., Rajagopalan, S., Pope III, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A. et al. (2010) Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. *Circulation*, **121**, 2331–2378.

Brunet, M., Jones, P. D., Sigró, J., Saladié, Ó., Aguilar, E., Moberg, A., Della-Marta, P. M., Lister, D., Walther, A. and López, D. (2007) Temporal and spatial temperature variability and change over Spain during 1850-2005. *Journal of Geophysical Research: Atmospheres*, **112**.

Cai, B., Zhang, J. and Sun, W. W. (2021) Jointly modeling and clustering tensors in high dimensions. *arXiv preprint arXiv:2104.07773*.

Cançado, A. L., Fernandes, L. B. and da Silva, C. Q. (2017) A Bayesian spatial scan statistic for zero-inflated count data. *Spatial Statistics*, **20**, 57–75.

Cançado, A. L., da Silva, C. Q. and da Silva, M. F. (2014) A spatial scan statistic for zero-inflated Poisson process. *Environmental and ecological statistics*, **21**, 627–650.

Cao, X., Wei, X., Han, Y. and Lin, D. (2014) Robust face clustering via tensor decomposition. *IEEE transactions on cybernetics*, **45**, 2546–2557.

Cardot, H., Crambes, C. and Sarda, P. (2007) Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. In *Statistical methods for biostatistics and related fields*, 221–243. Springer.

Carrera-Hernández, J. and Gaskin, S. (2007) Spatio temporal analysis of daily precipitation and temperature in the basin of Mexico. *Journal of Hydrology*, **336**, 231–249.

Carvalho, A. X. Y., Albuquerque, P. H. M., de Almeida Junior, G. R. and Guimaraes, R. D. (2009) Spatial hierarchical clustering. *Revista Brasileira de Biometria*, **27**, 411–442.

Fernández de Castro, B. and Manteiga, W. G. (2008) Boosting for real and functional samples: an application to an environmental problem. *Stochastic Environmental Research and Risk Assessment*, **22**, 27–37.

Chakraborty, A. and Chaudhuri, P. (2014) A Wilcoxon-Mann-Whitney type test for infinite dimensional data. *Biometrika*, **102**, 239–246.

Chen, C., Kim, A. Y., Ross, M. and Wakefield, J. (2018) *SpatialEpi: Methods and Data for Spatial Epidemiology*. URL: https://CRAN.R-project.org/package=SpatialEpi.

Chen, J. and Glaz, J. (2009) Approximations for two-dimensional variable window scan statistics. In *Scan Statistics* (eds. J. Glaz, V. Pozdnyakov and S. Wallenstein), 109 – 128. Birkhäuser Boston.

Chen, K.-T. (1957) Integration of paths, geometric invariants and a generalized Baker-Hausdorff formula. *Annals of Mathematics*, 163–178.

— (1977) Iterated path integrals. *Bulletin of the American Mathematical Society*, **83**, 831–879.

Chi, G. and Zhu, J. (2008) Spatial regression models for demographic analysis. *Population Research and Policy Review*, **27**, 17–42.

Chiles, J.-P. and Delfiner, P. (1999) *Geostatistics: modeling spatial uncertainty*. John Wiley & Sons.

Chiou, J.-M., Müller, H.-G. and Wang, J.-L. (2003) Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 405–423.

Chiou, J.-M., Yang, Y.-F. and Chen, Y.-T. (2016) Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis*, **146**, 301–312.

Citepa (a) Dioxyde d'azote (format secten). https://www.citepa.org/fr/2021-nox/ ; Accessed: 2022-01-21.

— (b) Particules inférieures à 10 μm (format secten). https://www.citepa.org/fr/2021-pm10/ ; Accessed: 2022-01-21.

— (c) Particules inférieures à 2.5 μm (format secten). https://www.citepa.org/fr/2021-pm2-5/ ; Accessed: 2022-01-21.

Clayton, D. G. (1978) A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65**, 141–151.

Cliff, A. and Ord, J. (1973) *Spatial Autocorrelation*. Monographs in spatial and environmental systems analysis. Pion.

Cook, A. J., Gold, D. R. and Li, Y. (2007) Spatial cluster detection for censored outcome data. *Biometrics*, **63**, 540–549.

Couchoud, C., Stengel, B., Landais, P., Aldigier, J.-C., de Cornelissen, F., Dabot, C., Maheut, H., Joyeux, V., Kessler, M., Labeeuw, M., Isnard, H. and Jacquelinet, C. (2005) The renal epidemiology and information network (REIN): a new registry for end-stage renal disease in France. *Nephrology Dialysis Transplantation*, **21**, 411–418.

Couchoud, C. G., Beuscart, J.-B. R., Aldigier, J.-C., Brunet, P. J. and Moranne, O. P. (2015) Development of a risk stratification algorithm to improve patient-centered care and decision making for incident elderly patients with end-stage renal disease. *Kidney international*, **88**, 1178–1186.

Cox, D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202.

Crawford, T. (2009) Scale analytical. In *International Encyclopedia of Human Geography* (eds. R. Kitchin and N. Thrift), 29–36. Oxford: Elsevier.

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

— (1988) Spatial prediction and ordinary kriging. *Mathematical geology*, **20**, 405–421.

— (1993) *Statistics for spatial data*. John Wiley & Sons.

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

— (2016) A Mann-Whitney scan statistic for continuous data. *Communications in Statistics-Theory and Methods*, **45**, 321–329.

Cucala, L., Demattei, C., Lopes, P. and Ribeiro, A. (2013) A spatial scan statistic for case event data based on connected components. *Computational Statistics*, **28**, 357–369.

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

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

Cuevas, A., Febrero, M. and Fraiman, R. (2002) Linear functional regression: the case of fixed design and functional response. *Canadian Journal of Statistics*, **30**, 285–300.

Cuevas, A., Febrero-Bande, M. and Fraiman, R. (2004) An ANOVA test for functional data. *Computational Statistics & Data Analysis*, **47**, 111–122.

Dabo-Niang, S., Ternynck, C. and Yao, A.-F. (2016) Nonparametric prediction of spatial multivariate data. *Journal of Nonparametric Statistics*, **28**, 428–458.

Dai, W. and Genton, M. G. (2019) Directional outlyingness for multivariate functional data. *Computational Statistics & Data Analysis*, **131**, 50–65.

Daniel, K., Onyango, N. O. and Sarguta, R. J. (2021) A spatial survival model for risk factors of under-five child mortality in Kenya. *International Journal of Environmental Research and Public Health*, **19**, 399.

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000) On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, **21**, 1324–1342.

Delaigle, A., Hall, P. and Pham, T. (2019) Clustering functional data into groups by using projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **81**, 271–304.

Delicado, P., Giraldo, R., Comas, C. and Mateu, J. (2010) Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, **21**, 224–239.

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

Dematteï, C., Molinari, N. and Daurès, J.-P. (2006) SPATCLUS: an R package for arbitrarily shaped multiple spatial cluster detection for case event data. *Computer Methods and Programs in Biomedicine*, **84**, 42–49.

Di, Q., Dai, L., Wang, Y., Zanobetti, A., Choirat, C., Schwartz, J. D. and Dominici, F. (2017) Association of short-term exposure to air pollution with mortality in older adults. *Jama*, **318**, 2446–2456.

Diday, E. (1971) Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. *Revue de statistique appliquée*, **19**, 19–33.

Diggle, P. J. and Chetwynd, A. G. (1991) Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 1155–1163.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., Ferris Jr, B. G. and Speizer, F. E. (1993) An association between air pollution and mortality in six US cities. *New England journal of medicine*, **329**, 1753–1759.

Durbeck, H., Greiling, D., Estberg, L., Long, A., Jacquez, G., Pallicaris, Y. and Hinton, S. (2012) *ClusterSeer: Software for the Detection and Analysis of Event Clusters, User Manual Book 2, Version 2.5.*

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

D'Urso, P. and Vitale, V. (2020) A robust hierarchical clustering for georeferenced data. *Spatial Statistics*, **35**, 100407.

Everitt, B. S., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis, Fifth Edition*. Wiley Series in Probability and Statistics. Wiley.

Fan, H., Zhao, C. and Yang, Y. (2020) A comprehensive analysis of the spatio-temporal variation of urban air pollution in China during 2014-2018. *Atmospheric Environment*, **220**, 117066.

Fermanian, A. (2021) *Learning time-dependent data with the signature transform*. Ph.D. thesis, Sorbonne Université.

— (2022) Functional linear regression with truncated signatures. *Journal of Multivariate Analysis*, 105031.

Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2011) Kernel regression with functional response. *Electronic Journal of Statistics*, **5**, 159–171.

Ferraty, F. and Vieu, P. (2006) *Nonparametric functional data analysis: theory and practice*, vol. 76. Springer.

Filippone, M., Camastra, F., Masulli, F. and Rovetta, S. (2008) A survey of kernel and spectral methods for clustering. *Pattern recognition*, **41**, 176–190.

Finazzi, F., Scott, E. M. and Fassò, A. (2013) A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **62**, 287.

Fine, J. P. and Gray, R. J. (1999) A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, **94**, 496–509.

Fouedjio, F. (2016a) A clustering approach for discovering intrinsic clusters in multivariate geostatistical data. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, 491–500. Springer.

— (2016b) Discovering spatially contiguous clusters in multivariate geostatistical data through spectral clustering. In *International Conference on Advanced Data Mining and Applications*, 547–557. Springer.

— (2016c) A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, **18**, 333–351.

— (2017) A spectral clustering approach for multivariate geostatistical data. *International Journal of Data Science and Analytics*, **4**, 301–312.

Friz, P. K. and Victoir, N. B. (2010) *Multidimensional stochastic processes as rough paths: theory and applications*, vol. 120. Cambridge University Press.

Frévent, C., Ahmed, M.-S., Soula, J., Smida, Z., Cucala, L., Dabo-Niang, S. and Genin, M. (2022) *HDSpatialScan: Multivariate and Functional Spatial Scan Statistics*. URL: https://CRAN.R-project.org/package=HDSpatialScan.

Fu, E. L., Evans, M., Carrero, J.-J., Putter, H., Clase, C. M., Caskey, F. J., Szymczak, M., Torino, C., Chesnaye, N. C., Jager, K. J. et al. (2021) Timing of dialysis initiation to reduce mortality and cardiovascular events in advanced chronic kidney disease: nationwide cohort study. *bmj*, **375**.

Gaetan, C. and Guyon, X. (2010) *Spatial statistics and modeling*, vol. 90. Springer.

Gangnon, R. E. and Clayton, M. K. (2003) A hierarchical model for spatially clustered disease rates. *Statistics in Medicine*, **22**, 3213–3228.

Gao, J., Zhang, Z., Hu, Y., Bian, J., Jiang, W., Wang, X., Sun, L. and 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**, 5431–44.

Gelfand, A. E., Zhu, L. and Carlin, B. P. (2001) On the change of support problem for spatio-temporal data. *Biostatistics*, **2**, 31–45.

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.

Getis, A. (2010) Spatial autocorrelation. In *Handbook of applied spatial analysis*, 255–278. Springer.

Giraldo, R., Delicado, P. and Mateu, J. (2010) Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of agricultural, biological, and environmental statistics*, **15**, 66–82.

— (2011) Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, **18**, 411–426.

— (2012) Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, **66**, 403–421.

Golovkine, S., Klutchnikoff, N. and Patilea, V. (2022) Clustering multivariate functional data using unsupervised binary trees. *Computational Statistics & Data Analysis*, **168**, 107376.

Gomez-Rubio, V., Serrano, P. E. M. and Rowlingson, B. (2020) *DClusterm: Model-Based Detection of Disease Clusters*. URL: https://CRAN.R-project.org/package=DClusterm.

Górecki, T. and Smaga, Ł. (2015) A comparison of tests for the one-way ANOVA problem for functional data. *Computational Statistics*, **30**, 987–1010.

— (2017) Multivariate analysis of variance for functional data. *Journal of Applied Statistics*, **44**, 2172–2189.

Goujon-Bellec, S., Demoury, C., Guyot-Goubin, A., Hémon, D. and Clavel, J. (2011) Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. *International journal of health geographics*, **10**, 1–12.

Gregorio, D. I., Huang, L., DeChello, L. M., Samociuk, H. and Kulldorff, M. (2007) Place of residence effect on likelihood of surviving prostate cancer. *Annals of epidemiology*, **17**, 520–524.

Greiling, D., Estberg, L., Long, A. and Jacquez, G. (2012) *ClusterSeer: Software for the Detection and Analysis of Event Clusters, User Manual Book 1, Version 2.5.*

Guhaniyogi, R., Qamar, S. and Dunson, D. B. (2017) Bayesian tensor regression. *The Journal of Machine Learning Research*, **18**, 2733–2763.

Gyurkó, L. G., Lyons, T., Kontkowski, M. and Field, J. (2013) Extracting information from the signature of a financial data stream. *arXiv preprint arXiv:1307.7244.*

Gómez-Rubio, V., Ferrándiz-Ferragud, J. and López-Quílez, A. (2015) *DCluster: Functions for the Detection of Spatial Clusters of Diseases.* URL: https://CRAN.R-project.org/package=DCluster.

Gómez-Rubio, V., Moraga, P., Molitor, J. and Rowlingson, B. (2019) DClusterm: Model-based detection of disease clusters. *Journal of Statistical Software*, **90**, 1–26.

Hagler, G., Birkett, D., Henry, R. C. and Peltier, R. E. (2021) Three years of high time-resolution air pollution monitoring in the complex multi-source harbor of New York and New Jersey. *Aerosol and Air Quality Research*, **21**.

Hallan, S. I., Matsushita, K., Sang, Y., Mahmoodi, B. K., Black, C., Ishani, A., Kleefstra, N., Naimark, D., Roderick, P., Tonelli, M. et al. (2012) Age and association of kidney measures with mortality and end-stage renal disease. *Jama*, **308**, 2349–2360.

Hanson, T. and Zhou, H. (2014) Spatial survival model. *Wiley StatsRef: Statistics Reference Online*, 1–8.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2003) Geoadditive survival models. *Corrado Lagazio, Marco Marchi (Eds)*, 116.

Henry, K. A., Niu, X. and Boscoe, F. P. (2009) Geographic disparities in colorectal cancer survival. *International journal of health geographics*, **8**, 1–13.

Héquette-Ruz, R., Beuscart, J.-B., Ficheur, G., Chazard, E., Guillaume, E., Paccou, J., Puisieux, F. and Genin, M. (2020) Hip fractures and characteristics of living area: a fine-scale spatial analysis in France. *Osteoporosis International*, **31**, 1353–1360.

Hitchcock, F. L. (1927) The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, **6**, 164–189.

Hoek, G., Krishnan, R. M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B. and Kaufman, J. D. (2013) Long-term air pollution exposure and cardio-respiratory mortality: a review. *Environmental health*, **12**, 1–16.

Holdaway, M. R. (1996) Spatial modeling and interpolation of monthly temperature using kriging. *Climate research*, **6**, 215–225.

Hougaard, P. (1986) Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

— (2000) *Shared frailty models*, 215–262. New York, NY: Springer New York.

Hsing, T. and Eubank, R. (2015) *Theoretical foundations of functional data analysis, with an introduction to linear operators*, vol. 997. John Wiley & Sons.

Huang, G., Brown, P. E., Fu, S. H. and Shin, H. H. (2022) Daily mortality/morbidity and air quality: Using multivariate time series with seasonally varying covariances. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

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

Huang, L., Tiwari, R. C., Zou, Z., Kulldorff, M. and Feuer, E. J. (2009) Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association*, **104**, 886–898.

Hung, M. (2016) *Applications of Spatial Statistics*. IntechOpen.

Jacques, J. and Preda, C. (2012) Model-based clustering of functional data. In *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.* Bruges, 459–464.

— (2014a) Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**, 231–255.

— (2014b) Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, **71**, 92–106.

Jeffreys, H. (1961) *Theory of probability (3rd Ed.)*. Oxford, UK: Oxford University Press.

Jiang, H. and Serban, N. (2012) Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, **54**, 108–119.

Jung, I. (2009) A generalized linear models approach to spatial scan statistics for covariate adjustment. *Statistics in medicine*, **28**, 1131–1143.

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

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

Jung, I., Kulldorff, M. and Richard, O. J. (2010) A spatial scan statistic for multinomial data. *Statistics in medicine*, **29**, 1910–1918.

Karandikar, R. L. (2006) On the Markov Chain Monte Carlo (MCMC) method. *Sadhana*, **31**, 81–104.

Khan, M. M., Roberson, S., Reid, K., Jordan, M. and Odoi, A. (2021) Geographic disparities and temporal changes of diabetes prevalence and diabetes self-management education program participation in Florida. *Plos one*, **16**.

Khanal, B., Picado, A., Bhattarai, N. R., Van Der Auwera, G., Das, M. L., Ostyn, B., Davies, C. R., Boelaert, M., Dujardin, J.-C. and Rijal, S. (2010) Spatial analysis of Leishmania Donovani exposure in humans and domestic animals in a recent Kala Azar focus in Nepal. *Parasitology*, **137**, 1597–1603.

Klein, J. P. (1992) Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 795–806.

Kleinbaum, D. G. and Klein, M. (2012) Recurrent event survival analysis. In *Survival Analysis*, 363–423. Springer.

Kleinbaum, D. G., Klein, M. et al. (2012) *Survival analysis: a self-learning text*, vol. 3. Springer.

Kleinman, K. (2015) *rsatscan: Tools, Classes, and Methods for Interfacing with SaTScan Stand-Alone Software.* URL: https://CRAN.R-project.org/package=rsatscan.

Knox, E. G. (1989) Detection of clusters. *In Elliott P (ed) Methodology of enquiries into disease clustering. Small Area Health Statistics Unit, London*, 17–20.

Kolda, T. G. and Bader, B. W. (2009) Tensor decompositions and applications. *SIAM review*, **51**, 455–500.

Kormilitzin, A., Saunders, K., Harrison, P., Geddes, J. and Lyons, T. (2016) Application of the signature method to pattern recognition in the CEQUEL clinical trial. *arXiv preprint arXiv:1606.02074*.

Kruskal, J. B. (1977) Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, **18**, 95–138.

— (1989) Rank, decomposition, and uniqueness for 3-way and n-way arrays. In *Multiway data analysis*, 7–18.

Kuechly, H. U., Kyba, C. C., Ruhtz, T., Lindemann, C., Wolter, C., Fischer, J. and Hölker, F. (2012) Aerial survey and spatial analysis of sources of light pollution in Berlin, Germany. *Remote Sensing of Environment*, **126**, 39–50.

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

— (1999) *Spatial Scan Statistics: Models, Calculations, and Applications*, 303–322. Birkhäuser, Boston, MA.

— (2006) *SaTScan^{TM} user guide*.

— (2018) *TreeScan User Guide*. URL: https://www.treescan.org/.

— (2021) *SaTScan User Guide for Version 9.7*. URL: https://www.satscan.org/.

Kulldorff, M., Athas, W. F., Feurer, E. J., Miller, B. A. and Key, C. R. (1998) Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. *American journal of public health*, **88**, 1377–1380.

Kulldorff, M., Fang, Z. and Walsh, S. J. (2003a) A tree-based scan statistic for database disease surveillance. *Biometrics*, **59**, 323–331.

Kulldorff, M., Huang, L. and Konty, K. (2009) A scan statistic for continuous data based on the normal probability model. *International journal of health geographics*, **8**, 1–9.

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

Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. and Platt, R. (2007) Multivariate scan statistics for disease surveillance. *Statistics in medicine*, **26**, 1824–1833.

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

Kulldorff, M., Tango, T. and Park, P. J. (2003b) Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, **42**, 665–684.

Kumar, R. and Joseph, A. E. (2006) Air pollution concentrations of $PM_{2.5}$, PM_{10} and NO_2 at ambient and kerbsite and their correlation in Metro City-Mumbai. *Environmental Monitoring and Assessment*, **119**, 191–199.

Lai, S., Jin, L. and Yang, W. (2017) Online signature verification using recurrent neural network and length-normalized path signature descriptor. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, vol. 1, 400–405. IEEE.

Larsen, J. L. (2000) *The modifiable areal unit problem: a problem or a source of spatial information?* The Ohio State University.

Lee, J., Sun, Y. and Chang, H. H. (2019) Spatial cluster detection of regression coefficients in a mixed-effects model. *Environmetrics*, **31**.

Lee, L.-F. (2004) Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, **72**, 1899–1925.

— (2007) GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, **137**, 489–514.

Legleiter, C. J. and Kyriakidis, P. C. (2008) Spatial prediction of river channel topography by kriging. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, **33**, 841–867.

Lehmann, E. and D'Abrera, H. (2006) *Nonparametrics: Statistical Methods Based on Ranks*. Springer. URL: <https://books.google.fr/books?id=Wy8nAQAAIAAJ>.

Leroux, B. G., Lei, X. and Breslow, N. (2000) Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, 179–191. Springer.

Levin, D., Lyons, T. and Ni, H. (2016) Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*.

Li, H., Calder, C. A. and Cressie, N. (2007) Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geographical analysis*, **39**, 357–375.

— (2012) One-step estimation of spatial dependence parameters: Properties and extensions of the APLE statistic. *Journal of Multivariate Analysis*, **105**, 68–84.

Li, J. and Tran, L. T. (2009) Nonparametric estimation of conditional expectation. *Journal of Statistical Planning and Inference*, **139**, 164–175.

Li, L. and Zhang, X. (2017) Parsimonious tensor response regression. *Journal of the American Statistical Association*, **112**, 1131–1146.

Li, X.-Z., Wang, J.-F., Yang, W.-Z., Li, Z.-J. and Lai, S.-J. (2011) A spatial scan statistic for multiple clusters. *Mathematical biosciences*, **233**, 135–142.

Li, Y. (2009) Modeling and analysis of spatially correlated data. *New Developments In Biostatistics And Bioinformatics*, 72–98.

Li, Y. and Lin, X. (2006) Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association*, **101**, 591–603.

Li, Y. and Ryan, L. (2002) Modeling spatial survival data using semiparametric frailty models. *Biometrics*, **58**, 287–297.

Liang, K.-Y., Self, S. G., Bandeen-Roche, K. J. and Zeger, S. L. (1995) Some recent developments for regression analysis of multivariate failure time data. *Lifetime data analysis*, **1**, 403–415.

de Lima, M. S., Duczmal, L. H., Neto, J. C. and Pinto, L. P. (2015) Spatial scan statistics for models with overdispersion and inflated zeros. *Statistica Sinica*, 225–241.

Lin, P.-S. (2014) Generalized scan statistics for disease surveillance. *Scandinavian Journal of Statistics*, **41**, 791–808.

Lin, P.-S., Kung, Y.-H. and Clayton, M. (2016) Spatial scan statistics for detection of multiple clusters with arbitrary shapes. *Biometrics*, **72**, 1226–1234.

Lin, S., Liu, F., Liu, Y. and Shen, L. (2019) Local feature tensor based deep learning for 3d face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–5. IEEE.

Lin, Z., Lopes, M. E. and Müller, H.-G. (2021) High-dimensional MANOVA via bootstrapping and its application to functional and sparse count data. *Journal of the American Statistical Association*.

Lind, J. (1768) *An Essay on Diseases Incidental to Europeans in Hot Climates*.

Liu, A. and Moitra, A. (2020) Tensor completion made practical. *Advances in Neural Information Processing Systems*, **33**, 18905–18916.

Liu, M., Jin, L. and Xie, Z. (2017) PS-LSTM: Capturing essential sequential online information with path signature and LSTM for writer identification. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1, 664–669. IEEE.

Liu, Y., Pan, J., Zhang, H., Shi, C., Li, G., Peng, Z., Ma, J., Zhou, Y. and Zhang, L. (2019) Short-term exposure to ambient air pollution and asthma mortality. *American journal of respiratory and critical care medicine*, **200**, 24–32.

Loche, R., Giron, B., Abrial, D., Cucala, L., CharrasGarrido, M. and De-Goer, J. (2016) *graphscan: Cluster Detection with Hypothesis Free Scan Statistic*. URL: https://CRAN.R-project.org/package=graphscan.

Loh, J. M. and Zhu, Z. (2007) Accounting for spatial correlation in the scan statistic. *The Annals of Applied Statistics*, **1**, 560–584.

Long, Z., Liu, Y., Chen, L. and Zhu, C. (2019) Low rank tensor completion for multiway visual data. *Signal processing*, **155**, 301–316.

Loomis, D., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H. and Straif, K. (2013) The carcinogenicity of outdoor air pollution. *Lancet Oncology*, **14**, 1262.

Lu, Z. and Chen, X. (2002) Spatial nonparametric regression estimation: Non-isotropic case. *Acta Mathematicae Applicatae Sinica*, **18**, 641–656.

— (2004) Spatial kernel regression estimation: weak consistency. *Statistics & probability letters*, **68**, 125–136.

Luquero, F., Banga, C., Remartínez, D., Urrutia, P. P. P., Baron, E. and Grais, R. (2011) Cholera epidemic in Guinea-Bissau (2008) : The importance of "place". *PloS one*, **6**.

Lyons, T. (2014) Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*.

Lyons, T. J. (1998) Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, **14**, 215–310.

Ma, G., He, L., Lu, C.-T., Yu, P. S., Shen, L. and Ragin, A. B. (2016) Spatio-temporal tensor analysis for whole-brain fMRI classification. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 819–827. SIAM.

Madhulatha, T. S. (2012) An overview on clustering methods. *IOSR Journal of Engineering*.

Mahanta, K. K., Hazarika, J., Barman, M. P. and Rahman, T. (2021) An application of spatial frailty models to recovery times of COVID-19 patients in India under Bayesian approach. *Journal of Scientific Research*, **65**, 150–155.

Mai, Q., Zhang, X., Pan, Y. and Deng, K. (2021) A doubly enhanced EM algorithm for model-based tensor clustering. *Journal of the American Statistical Association*, 1–15.

Maravic, M., Le Bihan, C., Landais, P. and Fardellone, P. (2005) Incidence and cost of osteoporotic fractures in France during 2001. A methodological approach by the national hospital database. *Osteoporosis international*, **16**, 1475–1480.

Marciano, L. H. S. C., de Faria Fernandes Belone, A., Rosa, P. S., Coelho, N. M. B., Ghidella, C. C., Nardi, S. M. T., Miranda, W. C., Barrozo, L. V. and Lastória, J. C. (2018) Epidemiological and geographical characterization of leprosy in a Brazilian hyperendemic municipality. *Cadernos de saude publica*, **34**.

Martino, A., Ghiglietti, A., Ieva, F. and Paganoni, A. M. (2019) A k-means procedure based on a Mahalanobis type distance for clustering multivariate functional data. *Statistical Methods & Applications*, **28**, 301–322.

Maskarinec, G., Cooper, J. and Swygert, L. (1994) Investigation of increased incidence in childhood leukemia near radio towers in Hawaii: preliminary observations. *J Environ Pathol Toxicol Oncol*.

Matheron, G. (1962) *Traité de géostatistique appliquée*. No. 14. Editions Technip.

— (1963) *Traité de géostatistique appliquée. 2. Le krigeage*. Editions Technip.

McGilchrist, C. and Aisbett, C. (1991) Regression with frailty in survival analysis. *Biometrics*, 461–466.

Melnykov, V., Michael, S. and Melnykov, I. (2015) *Recent Developments in Model-Based Clustering with Applications*, 1–39. Springer International Publishing.

Menafoglio, A., Secchi, P. and Dalla Rosa, M. (2013) A universal kriging predictor for spatially dependent functional data of a Hilbert space. *Electronic Journal of Statistics*, **7**, 2209–2240.

Michelozzi, P., Capon, A., Kirchmayer, U., Forastiere, F., Biggeri, A., Barca, A. and Perucci, C. A. (2002) Adult and childhood leukemia near a high-power radio station in Rome, Italy. *American Journal of Epidemiology*.

Minamisava, R., Nouer, S. S., de Morais Neto, O. L., Melo, L. K. and Andrade, A. L. S. (2009) Spatial clusters of violent deaths in a newly urbanized region of Brazil: highlighting the social disparities. *International journal of health geographics*, **8**, 1–10.

Minster, R., Viviano, I., Liu, X. and Ballard, G. (2021) CP decomposition for tensors via alternating least squares with QR decomposition. *arXiv preprint arXiv:2112.10855*.

Monestiez, P., Nerini, D., Dabo-Niang, S. and Ferraty, F. (2008) *A Cokriging Method for Spatial Functional Data with Applications in Oceanology*, 237–242.

Montez-Rath, M. E., Kapphahn, K., Mathur, M. B., Mitani, A. A., Hendry, D. J. and Desai, M. (2017) Guidelines for generating right-censored outcomes from a Cox model extended to accommodate time-varying covariates. *Journal of Modern Applied Statistical Methods*, **16**, 6.

Moraga, P. (2017a) SpatialEpiApp: A Shiny Web Application for the Analysis of Spatial and Spatio-Temporal Disease Data. *Spatial and Spatio-temporal Epidemiology*, **23**.

— (2017b) *SpatialEpiApp: A Shiny Web Application for the Analysis of Spatial and Spatio-Temporal Disease Data*. URL: https://cran.r-project.org/package=SpatialEpiApp.

Moran, P. A. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Morrill, J., Kormilitzin, A., Nevado-Holgado, A., Swaminathan, S., Howison, S. and Lyons, T. (2019) The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. In *2019 Computing in Cardiology (CinC)*, Page–1. IEEE.

Motarjem, K., Mohammadzadeh, M. and Abyar, A. (2019) Bayesian analysis of spatial survival model with non-gaussian random effect. *Journal of Mathematical Sciences*, **237**, 692–701.

Müller, I., Erbertseder, T. and Taubenböck, H. (2022) Tropospheric NO2: Explorative analyses of spatial variability and impact factors. *Remote Sensing of Environment*, **270**, 112839.

Nadaraya, E. A. (1964) On estimating regression. *Theory of Probability & Its Applications*, **9**, 141–142.

Naus, J. I. (1965) the distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, **60**, 532–538.

— (1966) Power comparison of two tests of non-random clustering. *Technometrics*, **8**, 493–517.

Neill, D., Moore, A. and Cooper, G. (2005) A Bayesian spatial scan statistic. *Advances in neural information processing systems*, **18**, 1003–1010.

Newby, D. E., Mannucci, P. M., Tell, G. S., Baccarelli, A. A., Brook, R. D., Donaldson, K., Forastiere, F., Franchini, M., Franco, O. H., Graham, I. et al. (2015) Expert position paper on air pollution and cardiovascular disease. *European heart journal*, **36**, 83–93.

Oja, H. and Randles, R. H. (2004) Multivariate nonparametric tests. *Statistical Science*, **18**, 598–605.

Ojiambo, P. and Kang, E. (2013) Modeling spatial frailties in survival analysis of cucurbit downy mildew epidemics. *Phytopathology*, **103**, 216–227.

Oliver, M. and Webster, R. (1989) A geostatistical basis for spatial weighting in multivariate classification. *Mathematical geology*, **21**, 15–35.

Openshaw, S., Charlton, M., Colin, W. and Craft, A. (1987) A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, **1**, 335–358.

O'Rourke, J. and Washington, R. (1985) Curve similarity via signatures. In *Machine Intelligence and Pattern Recognition*, vol. 2, 295–317. Elsevier.

Otani, T. and Takahashi, K. (2021) *rflexscan: The Flexible Spatial Scan Statistic*. URL: `https://CRAN.R-project.org/package=rflexscan`.

Pawitan, Y. and Huang, J. (2003) Constrained clustering of irregularly sampled spatial data. *Journal of Statistical Computation and Simulation*, **73**, 853–865.

PROJ contributors (2021) *PROJ coordinate transformation software library*. Open Source Geospatial Foundation. URL: `https://proj.org/`.

Qiu, Z., Chen, J. and Zhang, J.-T. (2021) Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis*, **157**.

Rabanser, S., Shchur, O. and Günnemann, S. (2017) Introduction to tensor decompositions and their applications in machine learning. *arXiv preprint arXiv:1711.10781*.

Ramsay, J. and Silverman, B. (2005a) From functional data to smooth functions. *Functional data analysis*, 37–58.

— (2005b) *Functional Data Analysis*. Springer Series in Statistics. Springer.

Ramsay, J. O., Graves, S. and Hooker, G. (2020) *fda: Functional Data Analysis*. URL: `https://CRAN.R-project.org/package=fda`.

Ramsay, J. O. and Ramsey, J. B. (2002) Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of econometrics*, **107**, 327–344.

Richards, L. M. (1956) The formation of ozone in polluted atmospheres. *Journal of the Air Pollution Control Association*, **5**, 216–246.

Ripley, B. D. (1981) *Spatial statistics*. John Wiley & Sons.

Robinson, P. M. (2011) Asymptotic theory for nonparametric regression with spatial data. *Journal of Econometrics*, **165**, 5–19.

Rogerson, P. A. (2021) Scan statistics adjusted for global spatial autocorrelation. *Geographical Analysis*.

Romano, E., Balzanella, A. and Verde, R. (2010) Clustering spatio-functional data: a model based approach. In *Classification as a Tool for Research*, 167–175. Springer.

— (2017) Spatial variability clustering for spatially dependent functional data. *Statistics and Computing*, **27**, 645–658.

Romano, E., Verde, R. and Cozza, V. (2011) Clustering spatial functional data: A method based on a nonparametric variogram estimation. In *New perspectives in statistical modeling and data analysis*, 339–346. Springer.

Romary, T., Ors, F., Rivoirard, J. and Deraisme, J. (2015) Unsupervised classification of multivariate geostatistical data: two algorithms. *Computers & geosciences*, **85**, 96–103.

Rondeau, V. (2010) Statistical models for recurrent events and death: Application to cancer events. *Mathematical and Computer modelling*, **52**, 949–955.

Rosychuk, R. J., Huston, C. and Prasad, N. G. (2006) Spatial event cluster detection using a compound Poisson distribution. *Biometrics*, **62**, 465–470.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series B (statistical methodology)*, **71**, 319–392.

Sava, F. and Carlsten, C. (2012) Respiratory health effects of ambient air pollution: an update. *Clinics in chest medicine*, **33**, 759–769.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W. and Lin, C.-T. (2017) A review of clustering techniques and developments. *Neurocomputing*, **267**, 664–681.

Schmutz, A., Jacques, J., Bouveyron, C., Cheze, L. and Martin, P. (2020) Clustering multivariate functional data in group-specific functional subspaces. *Computational Statistics*, **35**, 1101–1131.

Schwartz, J. and Dockery, D. W. (1992) Increased mortality in Philadelphia associated with daily air pollution concentrations. *Am Rev Respir Dis*, **145**, 600–604.

Scoggins, A., Kjellstrom, T., Fisher, G., Connor, J. and Gimson, N. (2004) Spatial analysis of annual air pollution exposure and mortality. *Science of the Total Environment*, **321**, 71–85.

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Dingenen, R. V., Dora, C. et al. (2018) Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**, 231–253.

Shaddick, G. and Wakefield, J. (2002) Modelling daily multivariate pollutant data at multiple sites. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 351–372.

Shangguan, P., Qiu, T., Liu, T., Zou, S., Liu, Z. and Zhang, S. (2020) Feature extraction of EEG signals based on functional data analysis and its application to recognition of driver fatigue state. *Physiological Measurement*, **41**, 125004.

Shi, G., Liu, J. and Zhong, X. (2021) Spatial and temporal variations of $PM_{2.5}$ concentrations in Chinese cities during 2015-2019. *International Journal of Environmental Health Research*, 1–13.

Smida, Z., Cucala, L., Gannoun, A. and Durif, G. (2022) A Wilcoxon-Mann-Whitney spatial scan statistic for functional data. *Computational Statistics & Data Analysis*, **167**, 107378.

Smirnov, O. and Anselin, L. (2001) Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics & Data Analysis*, **35**, 301–319.

Sørensen, H., Goldsmith, J. and Sangalli, L. M. (2013) An introduction with medical applications to functional data analysis. *Statistics in medicine*, **32**, 5222–5240.

Stone, R. A. (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in medicine*, **7**, 649–660.

Su, P.-F., Sie, F.-C., Yang, C.-T., Mau, Y.-L., Kuo, S. and Ou, H.-T. (2020) Association of ambient air pollution with cardiovascular disease risks in people with type 2 diabetes: a bayesian spatial survival analysis. *Environmental Health*, **19**, 1–12.

Sudakin, D. L., Horowitz, Z. and Giffin, S. (2002) Regional variation in the incidence of symptomatic pesticide exposures: applications of geographic information systems. *Journal of Toxicology: Clinical Toxicology*, **40**, 767–773.

Sugasawa, S. and Murakami, D. (2021) Spatially clustered regression. *Spatial Statistics*, **44**, 100525.

Sun, W. W. and Li, L. (2019) Dynamic tensor clustering. *Journal of the American Statistical Association*, **114**, 1894–1907.

Tait, P. A., McNicholas, P. D. and Obeid, J. (2020) Clustering higher order data: An application to pediatric multi-variable longitudinal data. *arXiv preprint arXiv:1907.08566*.

Takahashi, K. and Tango, T. (2005) A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, **4**.

Takahashi, K., Yokoyama, T. and Tango, T. (2010) *FleXScan v 3.1: Software for the Flexible Scan Statistic.*

Tan, X., Zhang, Y., Tang, S., Shao, J., Wu, F. and Zhuang, Y. (2012) Logistic tensor regression for classification. In *International Conference on Intelligent Science and Intelligent Data Engineering*, 573–581. Springer.

Tango, T. and Takahashi, K. (2005) A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, **4**, 1–15.

Tarpey, T., Petkova, E., Ciarleglio, A. and Ogden, R. T. (2021) Extracting scalar measures from functional data with applications to placebo response. *Statistics and its interface*, **14**, 255.

Taylor, B. M. and Rowlingson, B. S. (2017) Spatsurv: an r package for bayesian inference with spatial survival models. *Journal of Statistical Software*, **77**, 1–32.

Ternynck, C. (2014) Spatial regression estimation for functional data with spatial dependency. *Journal de la Société Française de Statistique*, **155**, 138–160.

Thamrin, S. A., Jaya, A. K., Mengersen, K. et al. (2021) Bayesian spatial survival modelling for dengue fever in makassar, indonesia. *Gaceta sanitaria*, **35**, S59–S63.

Therneau, T. M. and Grambsch, P. M. (2000) *Modeling Survival Data: Extending the Cox Model.* Springer New York, NY.

Tobler, W. R. (1969) Geographical filters and their inverses. *Geographical Analysis*, **1**, 234–253.

Tucker, L. R. (1966) Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311.

Usman, I. and Rosychuk, R. J. (2018) A log-Weibull spatial scan statistic for time to event data. *International Journal of Health Geographics*, **17**, 1–12.

Van Niekerk, J., Krainski, E., Rustand, D. and Rue, H. (2022) A new avenue for bayesian inference with inla. *arXiv preprint arXiv:2204.06797*.

Vandewalle, V., Preda, C. and Dabo-Niang, S. (2022) *Clustering Spatial Functional Data*, chap. 7, 155–174. John Wiley & Sons, Ltd.

Viviani, R., Grön, G. and Spitzer, M. (2005) Functional principal component analysis of fMRI data. *Human brain mapping*, **24**, 109–129.

Von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and computing*, **17**, 395–416.

Wackernagel, H. (2003) *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.

Waller, L. and Gotway, C. (2004) Applied spatial statistics for public health data. *Applied spatial statistics for public health data.*

Wan, L., Sun, Y., Lee, I., Zhao, W. and Xia, F. (2020) Industrial pollution areas detection and location via satellite-based IIoT. *IEEE Transactions on Industrial Informatics*, **17**, 1785–1794.

Wan, N., Zhan, F. B., Lu, Y. and Tiefenbacher, J. P. (2012) Access to healthcare and disparities in colorectal cancer survival in Texas. *Health & place*, **18**, 321–329.

Wang, D., Zhong, Z., Bai, K. and He, L. (2019) Spatial and temporal variabilities of $PM_{2.5}$ concentrations in China using functional data analysis. *Sustainability*, **11**, 1620.

Ward, M. D. and Gleditsch, K. S. (2018) *Spatial regression models*, vol. 155. Sage Publications.

Watson, G. S. (1964) Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.

World Health Organization (2016) Ambient air pollution: A global assessment of exposure and burden of disease.

World Health Organization (WHO) () WHO Air Quality Guidelines. https://www.c40knowledgehub.org/s/article/WHO-Air-Quality-Guidelines?language=en_US ; Accessed: 2022-03-15.

Wu, J., Xu, C., Wang, Q. and Cheng, W. (2016) Potential sources and formations of the $PM_{2.5}$ pollution in urban Hangzhou. *Atmosphere*, **7**, 100.

Xia, D., Yuan, M. and Zhang, C.-H. (2021) Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *The Annals of Statistics*, **49**.

Xu, J. and Gangnon, R. E. (2016) Stepwise and stagewise approaches for spatial cluster detection. *Spatial and spatio-temporal epidemiology*, **17**, 59–74.

Yang, K. and Lee, L.-f. (2017) Identification and QML estimation of multivariate and simultaneous equations spatial autoregressive models. *Journal of Econometrics*, **196**, 196–214.

Yang, W., Jin, L. and Liu, M. (2015) Chinese character-level writer identification using path signature feature, dropstroke and deep CNN. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 546–550. IEEE.

Yin, P. and Mu, L. (2018) A hybrid method for fast detection of spatial disease clusters in irregular shapes. *GeoJournal*, **83**, 693–705.

Zhang, L., Baladandayuthapani, V., Zhu, H., Baggerly, K. A., Majewski, T., Czerniak, B. A. and Morris, J. S. (2016) Functional CAR models for large spatially correlated functional datasets. *Journal of the American Statistical Association*, **111**, 772–786.

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

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

Zhou, H., Hanson, T. and Zhang, J. (2017) spbayessurv: Fitting bayesian spatial survival models using r. *arXiv preprint arXiv:1705.04584*.

Zhou, H., Li, L. and Zhu, H. (2013) Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, **108**, 540–552.

Zhou, J., Sun, W. W., Zhang, J. and Li, L. (2021) Partially observed dynamic tensor response regression. *Journal of the American Statistical Association*, 1–16.

Zhou, R., Shu, L. and Su, Y. (2015) An adaptive minimum spanning tree test for detecting irregularly-shaped spatial clusters. *Computational Statistics & Data Analysis*, **89**, 134–146.

Zhu, X., Huang, D., Pan, R. and Wang, H. (2020) Multivariate spatial autoregressive model for large scale social networks. *Journal of Econometrics*, **215**, 591–606.

Zou, B., Peng, F., Wan, N., Mamady, K. and Wilson, G. J. (2014) Spatial cluster detection of air pollution exposure inequities across the United States. *PLoS One*, **9**, e91917.

Appendix A

Reduction of the computation time for the NPFSS

Actually, Smida et al. (2022) defined a nonparametric scan statistic Λ for independent observations of a process X in a separable Banach space χ by

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

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 $X \in \mathcal{L}^2(\mathcal{T}, \mathbb{R}) \backslash \{0\}$. 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 $h \in \mathcal{L}^2(\mathcal{T}, \mathbb{R})$,

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

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

Reduction of the computation time

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

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

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 simulation study presented in Chapter 3 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 A.1. In this example, our method divided the computing time by a factor of approximately 100.

Table A.1: 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 $[0, 1]$ (see the simulation study presented in Chapter 3 for more details).

Method	Range	Mean (sd)
Basic method	[78, 120]	101 (8)
Improved method	[0.25, 0.87]	0.34 (0.06)

<div style="border:1px solid black">

Appendix **B**

</div>

Optimizing the computation time in the package HDSpatialScan

In addition to using vector and matrix operations, several methods are implemented in C++ in the package **HDSpatialScan** in order to reduce their computation time. This is the case for the PFSS, the MPFSS, the MRBFSS and the MDFFSS scan statistics.

The MPFSS also uses the following trick. The computation of the MPFSS requires the computation of two matrices H_w and E_w for each potential cluster w:

$$H_w = |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)][\bar{X}_w(t) - \bar{X}(t)]^\top \, \mathrm{d}t + |w^{\mathrm{c}}| \int_{\mathcal{T}} [\bar{X}_{w^{\mathrm{c}}}(t) - \bar{X}(t)][\bar{X}_{w^{\mathrm{c}}}(t) - \bar{X}(t)]^\top \, \mathrm{d}t$$

and

$$E_w = \sum_{j,s_j \in w} \int_{\mathcal{T}} [X_j(t) - \bar{X}_w(t)][X_j(t) - \bar{X}_w(t)]^\top \mathrm{d}t + \sum_{j,s_j \in w^{\mathrm{c}}} \int_{\mathcal{T}} [X_j(t) - \bar{X}_{w^{\mathrm{c}}}(t)][X_j(t) - \bar{X}_{w^{\mathrm{c}}}(t)]^\top \mathrm{d}t.$$

By highlighting constants in the calculation of H_w and E_w, the computation time of the MPFSS can be easily reduced.

We note $C_1 = \sum_{i=1}^{n} \int_{\mathcal{T}} X_i(t) X_i(t)^\top \, \mathrm{d}t$, $C_2 = n \int_{\mathcal{T}} \bar{X}(t) \bar{X}(t)^\top \, \mathrm{d}t$, $A_w = \int_{\mathcal{T}} \bar{X}_w(t) \bar{X}_w(t)^\top \, \mathrm{d}t$, $B_w = \int_{\mathcal{T}} \bar{X}_{w^{\mathrm{c}}}(t) \bar{X}_{w^{\mathrm{c}}}(t)^\top \, \mathrm{d}t$ and $C_w = \int_{\mathcal{T}} \bar{X}_w(t) \bar{X}_{w^{\mathrm{c}}}(t)^\top + \int_{\mathcal{T}} \bar{X}_{w^{\mathrm{c}}}(t) \bar{X}_w(t)^\top \, \mathrm{d}t$.

Then

$$H_w = |w| \int_{\mathcal{T}} [\bar{X}_w(t) - \bar{X}(t)][\bar{X}_w(t) - \bar{X}(t)]^\top \, \mathrm{d}t + |w^{\mathrm{c}}| \int_{\mathcal{T}} [\bar{X}_{w^{\mathrm{c}}}(t) - \bar{X}(t)][\bar{X}_{w^{\mathrm{c}}}(t) - \bar{X}(t)]^\top \, \mathrm{d}t$$

$$= |w| \int_{\mathcal{T}} \bar{X}_w(t) \bar{X}_w(t)^\top \, \mathrm{d}t - |w| \int_{\mathcal{T}} \bar{X}_w(t) \bar{X}(t)^\top \, \mathrm{d}t - |w| \int_{\mathcal{T}} \bar{X}(t) \bar{X}_w(t)^\top \, \mathrm{d}t + |w| \int_{\mathcal{T}} \bar{X}(t) \bar{X}(t)^\top \, \mathrm{d}t$$

$$+ |w^{\mathrm{c}}| \int_{\mathcal{T}} \bar{X}_{w^{\mathrm{c}}}(t) \bar{X}_{w^{\mathrm{c}}}(t)^\top \, \mathrm{d}t$$

$$- |w^{\mathrm{c}}| \int_{\mathcal{T}} \bar{X}_{w^{\mathrm{c}}}(t) \bar{X}(t)^\top \, \mathrm{d}t - |w^{\mathrm{c}}| \int_{\mathcal{T}} \bar{X}(t) \bar{X}_{w^{\mathrm{c}}}(t)^\top \, \mathrm{d}t + |w^{\mathrm{c}}| \int_{\mathcal{T}} \bar{X}(t) \bar{X}(t)^\top \, \mathrm{d}t.$$

Since $\bar{X}(t) = \frac{1}{n}(|w|\bar{X}_w(t) + |w^{\mathrm{c}}|\bar{X}_{w^{\mathrm{c}}}(t))$, we get

(1) $\int_{\mathcal{T}} \bar{X}_w(t) \bar{X}(t)^\top \, \mathrm{d}t = \frac{|w|}{n} A_w + \frac{|w^{\mathrm{c}}|}{n} \int_{\mathcal{T}} \bar{X}_w(t) \bar{X}_{w^{\mathrm{c}}}(t)^\top$

(2) $\int_{\mathcal{T}} \bar{X}(t) \bar{X}_w(t)^\top \, \mathrm{d}t = \frac{|w|}{n} A_w + \frac{|w^{\mathrm{c}}|}{n} \int_{\mathcal{T}} \bar{X}_{w^{\mathrm{c}}}(t) \bar{X}_w(t)^\top \, \mathrm{d}t$

$$(3) \quad \int_{\mathcal{T}} \bar{X}_{w^c}(t)\bar{X}(t)^\top \, \mathrm{d}t = \frac{|w|}{n} \int_{\mathcal{T}} \bar{X}_{w^c}(t)\bar{X}_w(t)^\top \, \mathrm{d}t + \frac{|w^c|}{n} B_w$$

$$(4) \quad \int_{\mathcal{T}} \bar{X}(t)\bar{X}_{w^c}(t)^\top \, \mathrm{d}t = \frac{|w|}{n} \int_{\mathcal{T}} \bar{X}_w(t)\bar{X}_{w^c}(t)^\top \, \mathrm{d}t + \frac{|w^c|}{n} B_w.$$

Thus $H_w = \left(|w| - 2\dfrac{|w|^2}{n}\right) A_w - \dfrac{2|w||w^c|}{n} C_w + \left(|w^c| - \dfrac{2|w^c|^2}{n}\right) B_w + C_2$.

And

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

$$= \sum_{j, s_j \in w} \left[\int_{\mathcal{T}} X_j(t)X_j(t)^\top \, \mathrm{d}t - \int_{\mathcal{T}} X_j(t)\bar{X}_w(t)^\top \, \mathrm{d}t - \int_{\mathcal{T}} \bar{X}_w(t)X_j(t)^\top \, \mathrm{d}t + \int_{\mathcal{T}} \bar{X}_w(t)\bar{X}_w(t)^\top \, \mathrm{d}t \right]$$

$$+ \sum_{j, s_j \in w^c} \left[\int_{\mathcal{T}} X_j(t)X_j(t)^\top \, \mathrm{d}t - \int_{\mathcal{T}} X_j(t)\bar{X}_{w^c}(t)^\top \, \mathrm{d}t - \int_{\mathcal{T}} \bar{X}_{w^c}(t)X_j(t)^\top \, \mathrm{d}t + \int_{\mathcal{T}} \bar{X}_{w^c}(t)\bar{X}_{w^c}(t)^\top \, \mathrm{d}t \right]$$

$$= C_1 - |w|A_w - |w^c|B_w.$$

And so the MPFSS only requires the computation of the quantities C_1, C_2, A_w, B_w and C_w where C_1 and C_2 do not depend on the potential cluster w.

Finally, the NPFSS uses in its implementation the calculation trick developed in Appendix A.

Appendix C

Supplementary materials of Chapter 5

1   The Leroux CAR prior

The Leroux CAR prior is defined as

$$
X_k|X_{-k} \sim \mathcal{N}\left(\frac{\rho \sum\limits_{l=1}^{K} v_{k,l}X_l}{\rho \sum\limits_{l=1}^{K} v_{k,l} + 1 - \rho}; \frac{\sigma_X^2}{\rho \sum\limits_{l=1}^{K} v_{k,l} + 1 - \rho}\right) \text{ where } v_{k,l}=1 \text{ if } s_k \text{ and } s_l \text{ are adjacent and}
$$

0 otherwise.

Let us show (Besag (1974)) that this is equivalent to $\boldsymbol{X} \sim \mathcal{N}(0, \sigma_X^2[\rho R + (1-\rho)I_K]^{-1})$, where

$$
R_{k,l} = \begin{cases} \sum\limits_{j=1}^{K} v_{k,j} & \text{if } k = l \\ -v_{k,l} & \text{otherwise} \end{cases},
$$

by applying Brook's lemma (Brook (1964)) on $\boldsymbol{X}$:

$$
\frac{\mathbb{P}(\boldsymbol{X})}{\mathbb{P}(\boldsymbol{0})} = \prod_{k=1}^{K} \frac{\mathbb{P}(X_k|X_1, \dots X_{k-1}, 0_{k+1}, \dots, 0_K)}{\mathbb{P}(0_k|X_1, \dots X_{k-1}, 0_{k+1}, \dots, 0_K)}.
$$

Let a_k and $b_{k,l}$ be defined as $a_k = \dfrac{\sigma_X^2}{\rho \sum\limits_{j=1}^{K} v_{k,j} + 1 - \rho}$ and $b_{k,l} = \dfrac{v_{k,l}}{\rho \sum\limits_{j=1}^{K} v_{k,j} + 1 - \rho}$, then

$$
\begin{aligned}
\frac{\mathbb{P}(\boldsymbol{X})}{\mathbb{P}(\boldsymbol{0})} &= \prod_{k=1}^{K} \frac{\exp\left[-\dfrac{1}{2a_k}\left(X_k - \rho \sum\limits_{l=1}^{k-1} b_{k,l}X_l\right)^2\right]}{\exp\left[-\dfrac{1}{2a_k}\left(0_k - \rho \sum\limits_{l=1}^{k-1} b_{k,l}X_l\right)^2\right]} \\
&= \prod_{k=1}^{K} \exp\left[-\frac{1}{2a_k}\left(X_k^2 - 2\rho \sum\limits_{l=1}^{k-1} b_{k,l}X_kX_l\right)\right] \\
&= \prod_{k=1}^{K} \exp\left[-\frac{1}{2\sigma_X^2}(\rho R_{k,k} + 1 - \rho)\left(X_k^2 + 2\rho \sum\limits_{l=1}^{k-1} \frac{R_{k,l}}{\rho R_{k,k} + 1 - \rho}X_kX_l\right)\right]
\end{aligned}
$$

$$= \prod_{k=1}^{K} \exp\left[-\frac{1}{2\sigma_X^2} \left((\rho R_{k,k} + 1 - \rho)X_k^2 + 2\rho \sum_{l=1}^{k-1} R_{k,l}X_k X_l \right) \right].$$

Let $A = \rho R + (1-\rho)I_K$, then

$$\frac{\mathbb{P}(\boldsymbol{X})}{\mathbb{P}(\boldsymbol{0})} = \prod_{k=1}^{K} \exp\left[-\frac{1}{2\sigma_X^2} \left(A_{k,k}X_k^2 + 2 \sum_{l=1}^{k-1} A_{k,l}X_k X_l \right) \right]$$

$$= \exp\left[-\frac{1}{2\sigma_X^2} \left(\sum_{k=1}^{K} A_{k,k}X_k^2 + 2 \sum_{k=1}^{K} \sum_{l=1}^{k-1} A_{k,l}X_k X_l \right) \right].$$

Moreover,

$$2\sum_{k=1}^{K}\sum_{l=1}^{k-1} A_{k,l}X_k X_l = \sum_{k=1}^{K}\sum_{l=1}^{k-1} A_{k,l}X_k X_l + \sum_{k=1}^{K}\sum_{l=1}^{k-1} A_{l,k}X_k X_l \text{ because } A \text{ is symmetric}$$

$$= \sum_{k=1}^{K}\sum_{l=1}^{k-1} A_{k,l}X_k X_l + \sum_{l=1}^{K}\sum_{k=l+1}^{K} A_{l,k}X_k X_l$$

$$= \sum_{k=1}^{K}\sum_{l=1}^{k-1} A_{k,l}X_k X_l + \sum_{k=1}^{K}\sum_{l=k+1}^{K} A_{k,l}X_l X_k$$

$$= \sum_{k=1}^{K}\sum_{\substack{l=1 \\ l\neq k}}^{K} A_{k,l}X_k X_l.$$

So

$$\frac{\mathbb{P}(\boldsymbol{X})}{\mathbb{P}(\boldsymbol{0})} = \exp\left[-\frac{1}{2\sigma_X^2} \left(\sum_{k=1}^{K} A_{k,k}X_k^2 + \sum_{k=1}^{K}\sum_{\substack{l=1 \\ l\neq k}}^{K} A_{k,l}X_k X_l \right) \right]$$

$$= \exp\left[-\frac{1}{2\sigma_X^2} \sum_{k=1}^{K}\sum_{l=1}^{K} A_{k,l}X_k X_l \right]$$

$$= \exp\left[-\frac{1}{2\sigma_X^2} \boldsymbol{X}^\top A \boldsymbol{X} \right].$$

Then $\mathbb{P}(\boldsymbol{X}) \propto \exp\left[-\frac{1}{2\sigma_X^2} \boldsymbol{X}^\top A \boldsymbol{X} \right]$ and we recognize the normal distribution $\mathcal{N}(0, \sigma_X^2 A^{-1})$.

2 Maximum likelihood estimators of α, $\sigma^{2(0)}$, α_w, α_{w^c} and $\sigma^{2(w)}$

2.1 Estimation under $\mathcal{H}_0$

Under $\mathcal{H}_0$, the likelihood is

$$\mathcal{L}_{\mathcal{H}_0} = \frac{1}{(2\pi)^{K/2}|A^{-1}|^{1/2}\sqrt{\sigma^{2K(0)}}} \exp\left[-\frac{1}{2\sigma^{2(0)}} [\boldsymbol{\varphi}^* - \alpha\mathbb{1}]^\top A [\boldsymbol{\varphi}^* - \alpha\mathbb{1}] \right].$$

Next, the log-likelihood is defined as

$$\ell_{\mathcal{H}_0} = -\frac{K}{2}\ln[2\pi] - \frac{1}{2}\ln|A^{-1}| - \frac{K}{2}\ln[\sigma^{2(0)}] - \frac{1}{2\sigma^{2(0)}}[\varphi^* - \alpha\mathbb{1}]^\top A[\varphi^* - \alpha\mathbb{1}]$$

$$= -\frac{K}{2}\ln[2\pi] - \frac{1}{2}\ln|A^{-1}| - \frac{K}{2}\ln[\sigma^{2(0)}] - \frac{1}{2\sigma^{2(0)}}[\varphi^{*\top}A\varphi^* - 2\alpha\mathbb{1}^\top A\varphi^* + \alpha^2\mathbb{1}^\top A\mathbb{1}].$$

$$\frac{\partial\ell_{\mathcal{H}_0}}{\partial\alpha} = -\frac{1}{2\sigma^{2(0)}}[-2\mathbb{1}^\top A\varphi^* + 2\alpha\mathbb{1}^\top A\mathbb{1}].$$

Thus $\dfrac{\partial\ell_{\mathcal{H}_0}}{\partial\alpha} = 0 \iff \alpha\mathbb{1}^\top A\mathbb{1} = \mathbb{1}^\top A\varphi^* \iff \alpha = \dfrac{\mathbb{1}^\top A\varphi^*}{\mathbb{1}^\top A\mathbb{1}}.$

$$\frac{\partial\ell_{\mathcal{H}_0}}{\partial\sigma^{2(0)}} = -\frac{K}{2\sigma^{2(0)}} + \frac{1}{2\sigma^{4(0)}}[\varphi^{*\top}A\varphi^* - 2\alpha\mathbb{1}^\top A\varphi^* + \alpha^2\mathbb{1}^\top A\mathbb{1}].$$

Then $\dfrac{\partial\ell_{\mathcal{H}_0}}{\partial\sigma^{2(0)}} = 0 \iff \sigma^{2(0)} = \dfrac{1}{K}[\varphi^{*\top}A\varphi^* - 2\alpha\mathbb{1}^\top A\varphi^* + \alpha^2\mathbb{1}^\top A\mathbb{1}].$

Ultimately, $\hat\alpha = \dfrac{\mathbb{1}^\top A\varphi^*}{\mathbb{1}^\top A\mathbb{1}}$ and $\widehat{\sigma^{2(0)}} = \dfrac{1}{K}[\varphi^{*\top}A\varphi^* - 2\hat\alpha\mathbb{1}^\top A\varphi^* + \hat\alpha^2\mathbb{1}^\top A\mathbb{1}].$

2.2   Estimation under $\mathcal{H}_1^{(w)}$

Under $\mathcal{H}_1^{(w)}$, the likelihood is

$$\mathcal{L}_{\mathcal{H}_1^{(w)}} = \frac{1}{(2\pi)^{K/2}|A^{-1}|^{1/2}\sqrt{\sigma^{2K(w)}}}\exp\left[-\frac{1}{2\sigma^{2(w)}}[\varphi^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}]^\top A[\varphi^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}]\right].$$

Thus, the log-likelihood can be computed:

$$\ell_{\mathcal{H}_1^{(w)}} = -\frac{K}{2}\ln[2\pi] - \frac{1}{2}\ln|A^{-1}| - \frac{K}{2}\ln[\sigma^{2(w)}] - \frac{1}{2\sigma^{2(w)}}[\varphi^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}]^\top A[\varphi^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}]$$

$$= -\frac{K}{2}\ln[2\pi] - \frac{1}{2}\ln|A^{-1}| - \frac{K}{2}\ln[\sigma^{2(w)}]$$

$$- \frac{1}{2\sigma^{2(w)}}[\varphi^{*\top}A\varphi^* - 2\alpha_w\mathbb{1}_w^\top A\varphi^* + \alpha_w^2\mathbb{1}_w^\top A\mathbb{1}_w + 2\alpha_w\alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c} - 2\alpha_{w^c}\mathbb{1}_{w^c}^\top A\varphi^* + \alpha_{w^c}^2\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c}].$$

$$\frac{\partial\ell_{\mathcal{H}_1^{(w)}}}{\partial\alpha_w} = -\frac{1}{2\sigma^{2(w)}}[-2\mathbb{1}_w^\top A\varphi^* + 2\alpha_w\mathbb{1}_w^\top A\mathbb{1}_w + 2\alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}].$$

Then $\dfrac{\partial\ell_{\mathcal{H}_1^{(w)}}}{\partial\alpha_w} = 0 \iff \alpha_w\mathbb{1}_w^\top A\mathbb{1}_w = \mathbb{1}_w^\top A\varphi^* - \alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}.$

$$\frac{\partial\ell_{\mathcal{H}_1^{(w)}}}{\partial\alpha_{w^c}} = -\frac{1}{2\sigma^{2(w)}}[-2\mathbb{1}_{w^c}^\top A\varphi^* + 2\alpha_{w^c}\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c} + 2\alpha_w\mathbb{1}_w^\top A\mathbb{1}_{w^c}].$$

Then $\dfrac{\partial\ell_{\mathcal{H}_1^{(w)}}}{\partial\alpha_{w^c}} = 0 \iff \alpha_{w^c}\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c} = \mathbb{1}_{w^c}^\top A\varphi^* - \alpha_w\mathbb{1}_w^\top A\mathbb{1}_{w^c}.$

We deduce:

$$\begin{cases}\dfrac{\partial\ell_{\mathcal{H}_1^{(w)}}}{\partial\alpha_w} = 0 \\[2mm] \dfrac{\partial\ell_{\mathcal{H}_1^{(w)}}}{\partial\alpha_{w^c}} = 0\end{cases} \iff \begin{cases}\alpha_w\mathbb{1}_w^\top A\mathbb{1}_w = \mathbb{1}_w^\top A\varphi^* - \alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c} \\ \alpha_{w^c}\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c} = \mathbb{1}_{w^c}^\top A\varphi^* - \alpha_w\mathbb{1}_w^\top A\mathbb{1}_{w^c}\end{cases}$$

$$\iff \begin{cases} \alpha_w = \dfrac{\mathbb{1}_w^\top A\boldsymbol{\varphi}^* - \alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w} \\[3mm] \alpha_{w^c}\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c} = \mathbb{1}_{w^c}^\top A\boldsymbol{\varphi}^* - \dfrac{\mathbb{1}_w^\top A\boldsymbol{\varphi}^* - \alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w}\mathbb{1}_w^\top A\mathbb{1}_{w^c} \end{cases}$$

$$\iff \begin{cases} \alpha_w = \dfrac{\mathbb{1}_w^\top A\boldsymbol{\varphi}^* - \alpha_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w} \\[3mm] \alpha_{w^c} = \left[\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c} - \dfrac{\mathbb{1}_w^\top A\mathbb{1}_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w}\right]^{-1}\left[\mathbb{1}_{w^c}^\top A\boldsymbol{\varphi}^* - \dfrac{\mathbb{1}_w^\top A\boldsymbol{\varphi}^*\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w}\right] \end{cases}.$$

$$\frac{\partial \ell_{\mathcal{H}_1^{(w)}}}{\partial \sigma^{2(w)}} = -\frac{K}{2\sigma^{2(w)}} + \frac{1}{2\sigma^{4(w)}}[\boldsymbol{\varphi}^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}]^\top A[\boldsymbol{\varphi}^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}].$$

Then $\dfrac{\partial \ell_{\mathcal{H}_1^{(w)}}}{\partial \sigma^{2(w)}} = 0 \iff \sigma^{2(w)} = \dfrac{1}{K}[\boldsymbol{\varphi}^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}]^\top A[\boldsymbol{\varphi}^* - \alpha_w\mathbb{1}_w - \alpha_{w^c}\mathbb{1}_{w^c}].$

Ultimately, $\hat{\alpha}_{w^c} = \left[\mathbb{1}_{w^c}^\top A\mathbb{1}_{w^c} - \dfrac{\mathbb{1}_w^\top A\mathbb{1}_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w}\right]^{-1}\left[\mathbb{1}_{w^c}^\top A\boldsymbol{\varphi}^* - \dfrac{\mathbb{1}_w^\top A\boldsymbol{\varphi}^*\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w}\right],$

$\hat{\alpha}_w = \dfrac{\mathbb{1}_w^\top A\boldsymbol{\varphi}^* - \hat{\alpha}_{w^c}\mathbb{1}_w^\top A\mathbb{1}_{w^c}}{\mathbb{1}_w^\top A\mathbb{1}_w}$ and $\widehat{\sigma^{2(w)}} = \dfrac{1}{K}[\boldsymbol{\varphi}^* - \hat{\alpha}_w\mathbb{1}_w - \hat{\alpha}_{w^c}\mathbb{1}_{w^c}]^\top A[\boldsymbol{\varphi}^* - \hat{\alpha}_w\mathbb{1}_w - \hat{\alpha}_{w^c}\mathbb{1}_{w^c}].$

3 Supplementary materials of the simulation study: influence of the threshold chosen for the Bayes factor
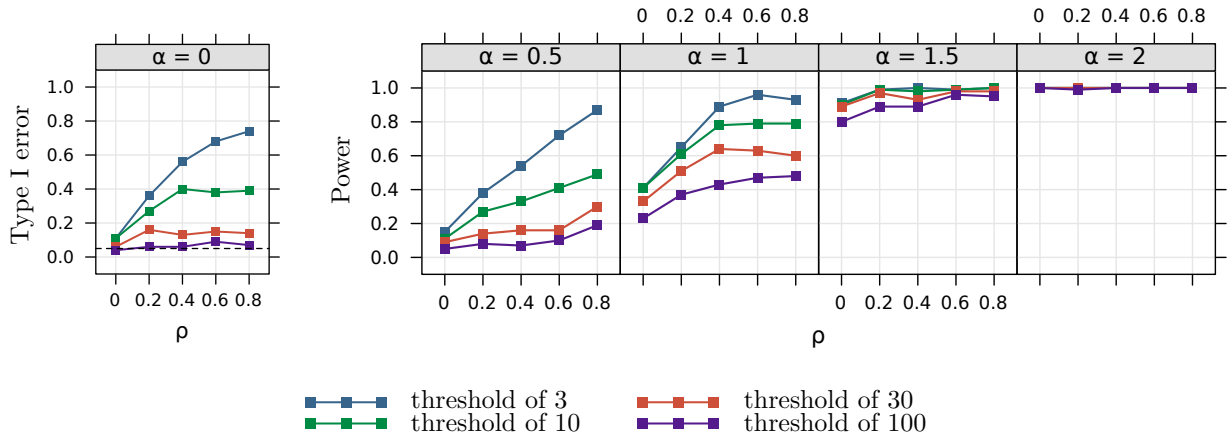

Figure C.1: Simulation study: type I error and power curves as a function of the chosen threshold for the Bayes factor used to select the frailty values under $\mathcal{H}_0$ or under $\mathcal{H}_1^{(w^*)}$. α is the parameter that controls the cluster intensity, and ρ controls the spatial correlation.

Figure C.1 shows the type I error and the power curves of the method developed here, with several thresholds for the Bayes factor (3, 10, 30 and 100). The thresholds of 3 and 10 do not maintain a stable type I error as a function of ρ. The threshold of 100 maintains the type I error very well but is very conservative. The threshold of 30 seems to be a good compromise.

However, we decided to look at whether the estimates of the parameters were better (i.e. less biased) with a threshold of 100 than with a threshold of 30. Figure C.2 shows that this was not the case.

(a)
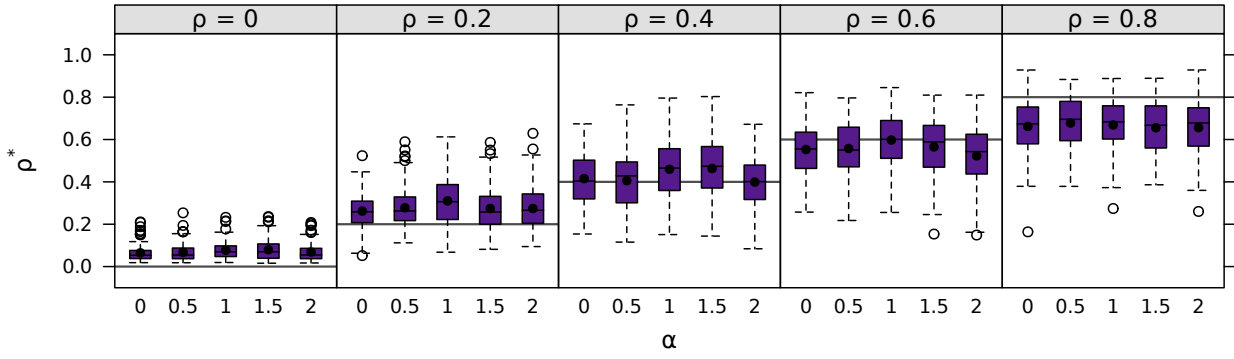
(b)
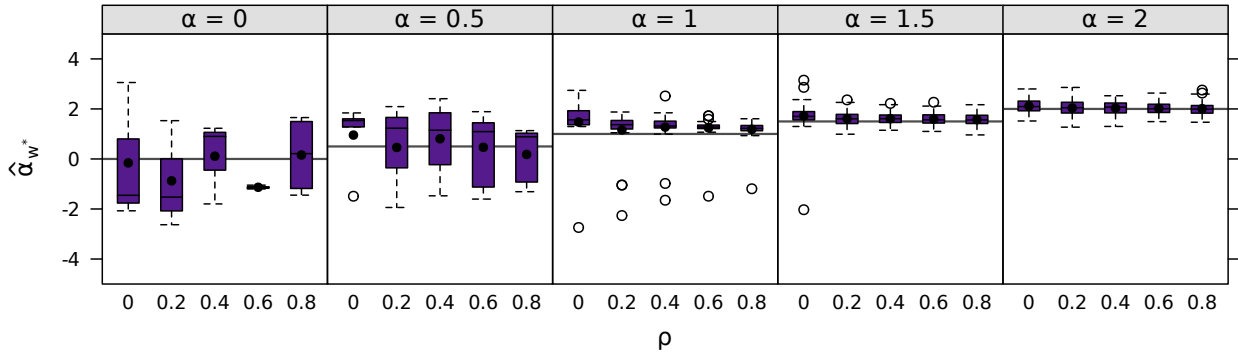
Figure C.2: Simulation study: the selected ρ^* as a function of the parameters ρ and α (panel (a)) and $\hat{\alpha}_{w^*}$ obtained with INLA when we selected $\mathcal{H}_1$ using the Bayes factor criterion (panel (b)) with a threshold of 100. The main horizontal lines correspond to the true value of the parameters ρ and α in panels (a) and (b) respectively, and the black points represent the mean estimates obtained.

4   Supplementary Materials of the application

Table C.1 describes the confounding factors for the detection of clusters of abnormal survival times after the initiation of dialysis in people aged 70 and over in the Nord-Pas-de-Calais region of northern France.

Figure C.3 shows the estimated φ_k for each model: the i.i.d. model, the Leroux CAR model and the ICAR model.

(a)

(b)

(c)

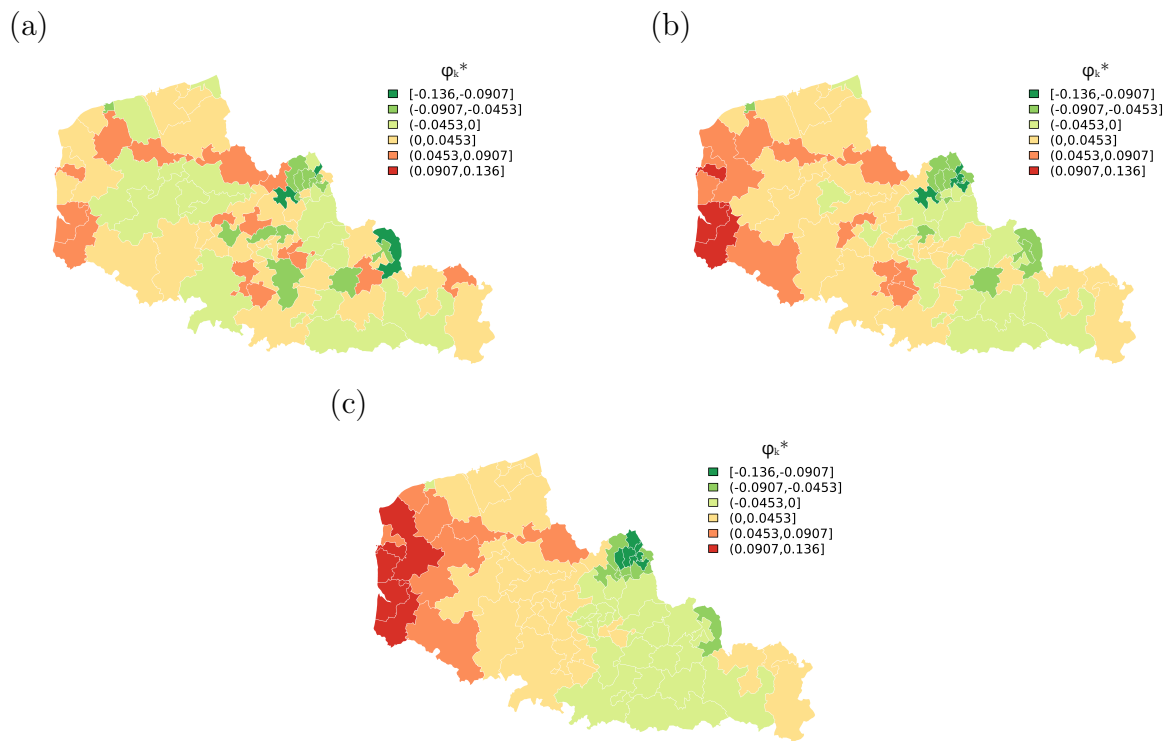Figure C.3: Estimated frailties φ_k^* with the i.i.d. (panel (a)), CAR (panel (b)) and ICAR (panel (c)) models

Table C.1: Description of the confounding factors for the detection of clusters of abnormal survival times in elderly patients with ESRD in northern France, 2004-2020.

Variable	**Overall**, $N = 6071^{1}$
Age (in years)	79.0 (74.8;83.5)
Body mass index (in kg/m^2)	26.2 (23.0;30.1)
Sex (Female)	2582/6071 (42.5%)
Type of nephropathy	
Polycystic kidney disease	104/5342 (1.9%)
Primitive glomerulonephritis	427/5342 (8.0%)
Hypertension or vascular	2048/5342 (38.3%)
Diabetic nephropathy	1683/5342 (31.5%)
Pyelonephritis	283/5342 (5.3%)
Other	797/5342 (14.9%)
Number of cardiovascular comorbidities	
None	962/5380 (17.9%)
One	1566/5380 (29.1%)
Two or more	2852/5380 (53.0%)
Diabetes (Yes)	3059/5960 (51.3%)
Chronic respiratory disease (Yes)	1062/5793 (18.3%)
Respiratory assistance (Yes)	380/5770 (6.6%)
Cirrhosis (Yes)	148/5804 (2.5%)
Severe behavioral disorder (Yes)	257/5532 (4.6%)
Mobility	
Independent walking	3262/4930 (66.2%)
Need for help from a third party	1205/4930 (24.4%)
Total disability	463/4930 (9.4%)
Hemoglobin level (< 11g/dL)	3760/5479 (68.6%)
Serum albumin level (< 35g/dL)	2525/4786 (52.8%)
Dialysis method	
Hemodialysis	5330/6071 (87.8%)
Peritoneal dialysis	741/6071 (12.2%)
Emergency initiation (Yes)	1776/5470 (32.5%)
Active malignant cancer (Yes)	586/5819 (10.1%)
Glomerular filtration rate	
< 7mL/min/1.73m^2	882/5285 (16.7%)
$7 - 10$mL/min/1.73m^2	1535/5285 (29.0%)
> 10mL/min/1.73m^2	2868/5285 (54.3%)
Period of treatment initiation	
2004-2009	1928/6071 (31.8%)
2010-2015	2258/6071 (37.2%)
2016-2020	1885/6071 (31.0%)

[1] Number of observed/ Total number of observed (%) for qualitative variables
  Median (Interquartile range) for quantitative variables

Appendix D

Useful concepts for a good understanding of the perspectives

1   Tensors

Tensors can be seen as a generalization of matrices for high dimensional data. They are increasingly used nowadays with the increase of storage and computing capacities. For example, tensors can represent an image over time (i.e. a video). In this case the tensor will be three-dimensional (two dimensions for the image and one dimension for the time): it is said to be of order 3, or to be a third-order tensor.

Notations and operations. A tensor's mode is a dimension of the tensor and its order is its number of modes. In the following we will give the concepts in the general case of M-order tensors $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ although they will be illustrated in the particular case $M = 3$. Note that when $M = 1$ and $M = 2$, the tensors reduce to vectors and matrices respectively.

A tensor can be transformed into a matrix or a vector. These operations are called unfolding and vectorization respectively.

We call fibers of a tensor the vectors obtained by fixing all indices except one of the tensor. Figure D.1 shows an example of mode-1, 2 and 3 fibers on a third-order tensor.

Then we call mode-m unfolding, the matrix $\boldsymbol{\mathcal{X}}_{(m)}$ whose columns are the mode-m fibers (Figure D.2).

Once the mode-1 unfolding $\boldsymbol{\mathcal{X}}_{(1)}$ of $\boldsymbol{\chi}$ is computed, the vectorization of $\boldsymbol{\mathcal{X}}$, noted $\mathrm{vec}(\boldsymbol{\chi})$ is the vectorization of $\boldsymbol{\mathcal{X}}_{(1)}$:

$$\mathrm{vec}(\boldsymbol{\chi}) = (\boldsymbol{\mathcal{X}}_{(1):1}^{\top}, \ldots, \boldsymbol{\mathcal{X}}_{(1):(d_2 d_3 \ldots d_M)}^{\top})^{\top}$$

where $\boldsymbol{\mathcal{X}}_{(1):i}$ denotes the ith column of $\boldsymbol{\mathcal{X}}_{(1)}$.

An inner product on tensors can be defined as:

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle = \sum_{i_1, \ldots, i_M} \boldsymbol{\mathcal{X}}_{i_1, \ldots, i_M} \boldsymbol{\mathcal{Y}}_{i_1, \ldots, i_M}$$

and thus the corresponding norm of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ is

$$||\boldsymbol{\mathcal{X}}||_2 = \sqrt{\sum_{i_1, \ldots, i_M} \boldsymbol{\mathcal{X}}_{i_1, \ldots, i_M}^2}.$$

We also define a useful product when manipulating tensors: the m-mode matrix product of the tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ by a matrix $A \in \mathbb{R}^{J \times d_m}$ in mode m is defined by $\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{X}} \times_m A$ where
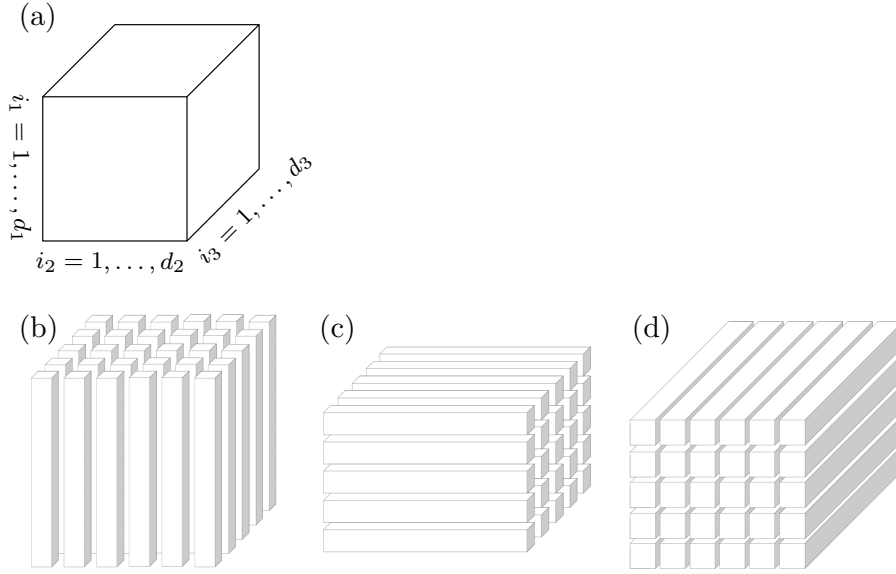
Figure D.1: Example of mode-1 (b),2 (c) and 3 (d) fibers on a third-order tensor (a).
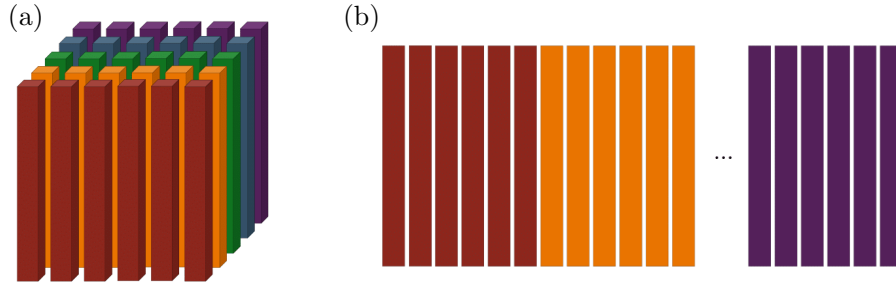

Figure D.2: Example of mode-1 fibers (a) and mode-1 unfolding (b) on a third-order tensor.

$\mathcal{U} \in \mathbb{R}^{d_1 \times \cdots \times d_{m-1} \times J \times d_{m+1} \times \cdots \times d_M}$ and

$$\mathcal{U}_{i_1,\ldots,i_{m-1},j,i_{m+1},\ldots,i_M} = \sum_{i_m=1}^{d_m} \mathcal{X}_{i_1,\ldots,i_M} A_{j,i_m}.$$

Then $\mathcal{U}_{(m)} = A\mathcal{X}_{(m)}$.

Tensor decomposition. There exist many methods of tensor decomposition. The most known are the CANDECOMP/PARAFAC (CP) decomposition (Hitchcock, 1927) and Tucker decomposition (Tucker, 1966). The illustrations of this section are inspired from Kolda and Bader (2009).

CP decomposition, or tensor rank decomposition or canonical polyadic decomposition (CPD) was proposed by Hitchcock (1927).

The principle is to decompose $\mathcal{X}$ as a sum of R rank-1 tensors:

$$\mathcal{X} \approx \sum_{r=1}^{R} p_r^{(1)} \circ p_r^{(2)} \circ \cdots \circ p_r^{(M)}, \tag{D.1}$$

where $\circ$ denotes the outer product (Figure D.3).

The rank of $\mathcal{X}$ is then the smallest R such that D.1 is true.

Under some conditions (Kruskal, 1977, 1989), this decomposition is unique (up to a rescaling

or a modification of the order of the $p_r^{(1)} \circ \cdots \circ p_r^{(M)}$). In particular a sufficient condition is

$$\sum_{m=1}^{M} K^{(m)} \geq 2R + (M-1)$$

where $K^{(m)}$ is the Kruskal rank, defined as the maximum value k such that any k columns of $(p_1^{(m)}, \ldots, p_R^{(m)}) \in \mathbb{R}^{d_m \times R}$ are linearly independent.

Note that the CP decomposition can be performed with the alternating least squares algorithm (CP-ALS, Minster et al. (2021)) for example.
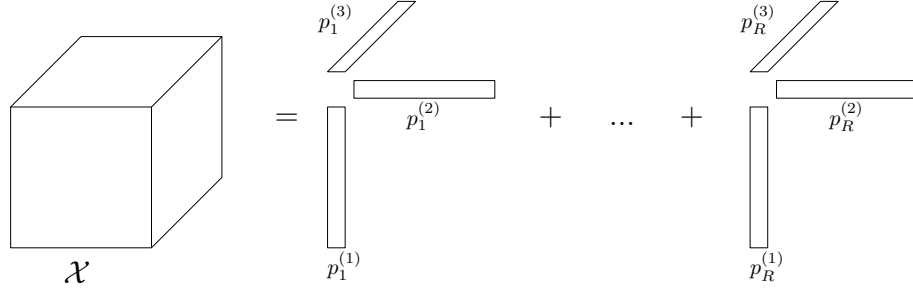

Figure D.3: Example of CP-decomposition for a third-order tensor

Another widely used decomposition of tensors is the Tucker decomposition which consists in writing $\boldsymbol{\mathcal{X}}$ as the multiplication of a smaller tensor $\boldsymbol{\mathcal{C}}$ by matrices $P^{(m)}$ along each mode m (Figure D.4):

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{C}} \times_1 P^{(1)} \times_2 \cdots \times_M P^{(M)}.$$

Note that this decomposition is not unique (Rabanser et al., 2017).
The higher-order singular value decomposition (HOSVD) (Tucker, 1966) and the higher-order orthogonal iteration (HOOI) algorithms (De Lathauwer et al., 2000) are two widely known methods to compute the Tucker decomposition of a tensor.
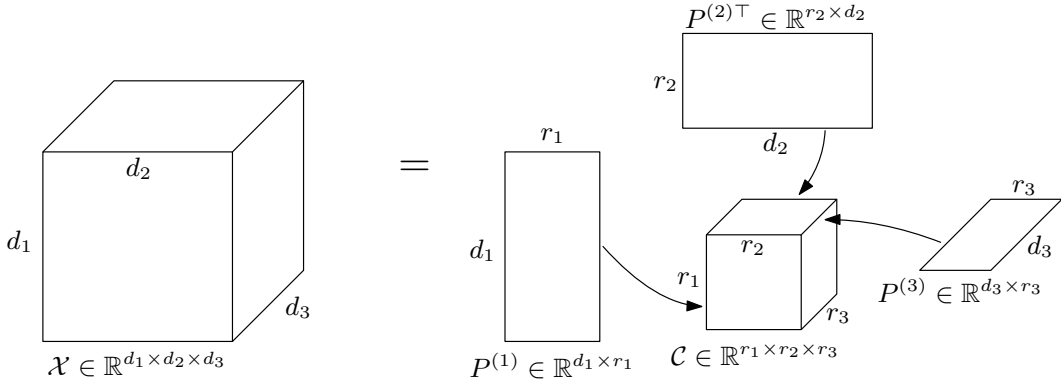

Figure D.4: Example of Tucker decomposition for a third-order tensor

Tensor normal distribution. Here we introduce a useful distribution in the context of tensor analysis: the tensor normal distribution.

Let $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{d_1 \times \cdots \times d_M}$ be a random tensor. We assume that its covariance is separable, that is

$$\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{X}})] = \Sigma_M \otimes \cdots \otimes \Sigma_1$$

where $\otimes$ denotes the Kronecker product, $\Sigma_m \in \mathbb{R}^{d_m \times d_m}$ and $(\Sigma_m)_{i,j}$ is the covariance between the ith and j^{th} elements of the mode-m fibers of $\boldsymbol{\mathcal{X}}$.

We say that $\boldsymbol{\mathcal{X}}$ follows a tensor normal distribution $\mathcal{N}_T(\boldsymbol{\mathcal{U}}, \underline{\Sigma})$ with $\underline{\Sigma} = \{\Sigma_1, \dots, \Sigma_M\}$ if its probability distribution function is

$$f(\boldsymbol{\mathcal{X}}) = (2\pi)^{-\frac{d}{2}} \left[\prod_{m=1}^{M} |\Sigma_m|^{-\frac{d}{2d_m}} \right] \exp\left[-\frac{1}{2} \left\| (\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{U}}) \times \underline{\Sigma}^{-\frac{1}{2}} \right\|_2^2 \right]$$

where $d = \prod_{m=1}^{M} d_m$ and $(\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{U}}) \times \underline{\Sigma}^{-\frac{1}{2}} = (\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{U}}) \times_1 \Sigma_1^{-\frac{1}{2}} \times_2 \cdots \times_M \Sigma_M^{-\frac{1}{2}}$.

Note that this is equivalent to $\text{vec}(\boldsymbol{\mathcal{X}}) \sim \mathcal{N}(\text{vec}(\boldsymbol{\mathcal{U}}), \Sigma_M \otimes \cdots \otimes \Sigma_1)$.

The reader may refer to Kolda and Bader (2009) for a more detailed review on tensor algebra.

Tensor analysis. The emergence of tensors has led to the development of new methods for these data such as missing value processing (Long et al., 2019; Liu and Moitra, 2020; Xia et al., 2021), regression (Tan et al., 2012; Zhou et al., 2013; Guhaniyogi et al., 2017; Zhou et al., 2021) or clustering (Tait et al., 2020; Mai et al., 2021). These approaches are widely used for image analysis, such as facial recognition (Cao et al., 2014; Lin et al., 2019), or large brain related data (Ma et al., 2016; Cai et al., 2021).

In the case where tensors are observed spatially, there is to our knowledge no clustering method allowing to take into account both the proximity of the observations and the proximity of their spatial locations. In this context, it would be interesting to develop a spatial clustering method for high dimensional data modeled by tensors and measured spatially. More precisely, we propose in the perspectives a Gaussian mixture model taking into account the spatial structure of the data and allowing to manage the large dimension of the data by assuming a CP decomposition and by making sparsity assumptions on the tensor parameters.

2 Signatures

The notion of signature was defined by Chen (1957, 1977) for smooth paths and rediscovered in the context of rough path theory (Lyons, 1998; Friz and Victoir, 2010). In recent years, signatures were widely used in many domains such as character recognition (Yang et al., 2015; Liu et al., 2017), finance (Gyurkó et al., 2013; Arribas, 2018), recurrent neural networks (Lai et al., 2017) and medicine (Kormilitzin et al., 2016; Morrill et al., 2019). In the field of functional data, Fermanian (2022) recently proposed a new functional linear regression model using signatures. Thus, the purpose of this section is to introduce the concept of signature.

Let

$$\begin{aligned} X : [0,1] &\rightarrow \mathbb{R}^p \\ t &\mapsto (X_t^{(1)}, \dots, X_t^{(p)})^\top \end{aligned}$$

be a continuous path of bounded variation that is

$$\sup_{(t_0, \dots, t_k) \in I} \sum_{i=1}^{k} ||X_{t_i} - X_{t_{i-1}}|| < \infty$$

where $I = \{(t_0, \dots, t_k)/k \geq 0, 0 = t_0 < \cdots < t_k = 1\}$. We note $X \in BV(\mathbb{R}^p)$.

Then the signature of X is $Sig(X) = (1, \boldsymbol{X}^1, \boldsymbol{X}^2, \dots)$ where

$$\boldsymbol{X}^d = \int \cdots \int_{0 \leq t_1 < t_2 < \cdots < t_d \leq 1} \mathrm{d}X_{t_1} \circ \mathrm{d}X_{t_2} \dots \circ \mathrm{d}X_{t_d},$$

and the truncated signature at order D is then $Sig^D(X) = (1, \boldsymbol{X}^1, \boldsymbol{X}^2, \dots, \boldsymbol{X}^D)$.

Example. Let

$$X_1 : [0,1] \rightarrow [0,1]^2$$
$$t \mapsto (t, t^2)^\top.$$

Then

$$\boldsymbol{X_1^1} = \int_{0<t_1<1} dX_{1,t_1} = \left(\int_{0<t_1<1} dt_1, \int_{0<t_1<1} 2t_1 dt_1 \right)^\top = (1,1)^\top \; ;$$

$$\boldsymbol{X_1^2} = \int_{0<t_2<1} \int_{0<t_1<t_2} dX_{1,t_1} \circ dX_{1,t_2} = \int_{0<t_2<1} \int_{0<t_1<t_2} \begin{pmatrix} 1 & 2t_2 \\ 2t_1 & 4t_1 t_2 \end{pmatrix} dt_1 dt_2 = \begin{pmatrix} 1/2 & 2/3 \\ 1/3 & 1/2 \end{pmatrix}$$

and so the truncated signature at order 2 is $Sig^2(X_1) = \left(1, (1,1)^\top, \begin{pmatrix} 1/2 & 2/3 \\ 1/3 & 1/2 \end{pmatrix} \right).$

Note that if $X : [a,b] \rightarrow \mathbb{R}^p$ with $0 \le a < b \le 1$ then the previous definitions are still valid by replacing the interval $[0,1]$ by $[a,b]$. In this case the signature of X is written $Sig_{[a,b]}(X)$.

Properties. Two well-known properties of signatures are invariance under time reparametrization and invariance by translation.

The first one states that for any non-decreasing surjective function $\beta : [0,1] \rightarrow [0,1]$ and any continuous path of bounded variation $X : [\beta(a), \beta(b)] \rightarrow \mathbb{R}^p$, $0 \le a < b \le 1$,

$$S_{[\beta(a),\beta(b)]}(X) = S_{[a,b]}(\tilde{X})$$

where $\tilde{X}_t = X_{\beta(t)} \; \forall t \in [a,b]$.

Example. Let

$$X_1 : [0,1] \rightarrow [0,1]^2$$
$$t \mapsto (t, t^2)^\top.$$

Then we have seen previously that the truncated signature at order 2 is

$$Sig^2(X_1) = \left(1, (1,1)^\top, \begin{pmatrix} 1/2 & 2/3 \\ 1/3 & 1/2 \end{pmatrix} \right).$$

Let

$$X_2 : [0,1] \rightarrow [0,1]^2$$
$$t \mapsto (t^2, t^4)^\top \; (\beta(t) = t^2).$$

Then the truncated signature at order 2 of X_2 is

$$Sig^2(X_2) = Sig^2(X_1) = \left(1, (1,1)^\top, \begin{pmatrix} 1/2 & 2/3 \\ 1/3 & 1/2 \end{pmatrix} \right) :$$

$$\boldsymbol{X_2^1} = \int_{0<t_1<1} dX_{2,t_1} = \left(\int_{0<t_1<1} 2t_1 \, dt_1, \int_{0<t_1<1} 4t_1^3 \, dt_1 \right)^\top = (1,1)^\top \; ;$$

$$\boldsymbol{X_2^2} = \int_{0<t_2<1} \int_{0<t_1<t_2} dX_{2,t_1} \circ dX_{2,t_2} = \int_{0<t_2<1} \int_{0<t_1<t_2} \begin{pmatrix} 4t_1 t_2 & 8t_1 t_2^3 \\ 8t_1^3 t_2 & 16t_1^3 t_2^3 \end{pmatrix} dt_1 dt_2 = \begin{pmatrix} 1/2 & 2/3 \\ 1/3 & 1/2 \end{pmatrix}.$$

The invariance by translation came directly from the fact that if $\tilde{X}_t = x + X_t, x \in \mathbb{R}^p$ then $d\tilde{X}_t = dX_t$.

Sometimes these invariances are not desirable. To avoid the invariance by time reparametrization a very simple solution is to consider the path $\tilde{X}_t = (X_t^\top, t)^\top$, called time-augmented path, and it is enough to add a point $X_0 = 0$ at the beginning of the path to circumvent the invariance by translation.

Another well-known property is Chen's identity which allows to compute easily the signature of a concatenation of paths: if $X : [a, b] \to \mathbb{R}^p$ and $Y : [b, c] \to \mathbb{R}^p$, $0 \le a < b < c \le 1$ are two continuous paths of bounded variation then the concatenation of X and Y is

$$
\begin{aligned}
X * Y : [a, c] &\to \mathbb{R}^p \\
t &\mapsto \begin{cases} X_t & \text{if } t \in [a, b] \\ X_b + Y_t - Y_b & \text{if } t \in [b, c] \end{cases}
\end{aligned}
$$

and $Sig_{[a,c]}(X * Y) = Sig_{[a,b]}(X) \otimes Sig_{[b,c]}(Y)$.

Signature coefficients. We call signature coefficient along $(i_1, \ldots, i_d) \subset [\![1, p]\!]^d$ the quantity

$$
\mathcal{S}_{(i_1, \ldots, i_d)}(X) = \int \cdots \int_{0 \le t_1 < \cdots < t_d \le 1} \mathrm{d}X_{t_1}^{(i_1)} \ldots \mathrm{d}X_{t_d}^{(i_d)}.
$$

Example. Let

$$
\begin{aligned}
X_1 : [0, 1] &\to [0, 1]^2 \\
t &\mapsto (t, t^2)^\top.
\end{aligned}
$$

We have seen previously that the truncated signature at order 2 is

$$
Sig^2(X_1) = \left(1, (1, 1)^\top, \begin{pmatrix} 1/2 & 2/3 \\ 1/3 & 1/2 \end{pmatrix} \right).
$$

Moreover we can note that

$$
\mathcal{S}_{(1)}(X_1) = \int_{0<t_1<1} \mathrm{d}X_{1,t_1}^{(1)} = \int_{0<t_1<1} \mathrm{d}t_1 = 1
$$

$$
\mathcal{S}_{(2)}(X_1) = \int_{0<t_1<1} \mathrm{d}X_{1,t_1}^{(2)} = \int_{0<t_1<1} 2t_1 \mathrm{d}t_1 = 1
$$

$$
\mathcal{S}_{(1,1)}(X_1) = \int_{0<t_2<1} \int_{0<t_1<t_2} \mathrm{d}X_{1,t_1}^{(1)} \mathrm{d}X_{1,t_2}^{(1)} = \int_{0<t_2<1} \int_{0<t_1<t_2} \mathrm{d}t_1 \mathrm{d}t_2 = \frac{1}{2}
$$

$$
\mathcal{S}_{(1,2)}(X_1) = \int_{0<t_2<1} \int_{0<t_1<t_2} \mathrm{d}X_{1,t_1}^{(1)} \mathrm{d}X_{1,t_2}^{(2)} = \int_{0<t_2<1} \int_{0<t_1<t_2} 2t_2 \, \mathrm{d}t_1 \mathrm{d}t_2 = \frac{2}{3}
$$

$$
\mathcal{S}_{(2,1)}(X_1) = \int_{0<t_2<1} \int_{0<t_1<t_2} \mathrm{d}X_{1,t_1}^{(2)} \mathrm{d}X_{1,t_2}^{(1)} = \int_{0<t_2<1} \int_{0<t_1<t_2} 2t_1 \, \mathrm{d}t_1 \mathrm{d}t_2 = \frac{1}{3}
$$

$$
\mathcal{S}_{(2,2)}(X_1) = \int_{0<t_2<1} \int_{0<t_1<t_2} \mathrm{d}X_{1,t_1}^{(2)} \mathrm{d}X_{1,t_2}^{(2)} = \int_{0<t_2<1} \int_{0<t_1<t_2} 4t_1t_2 \, \mathrm{d}t_1 \mathrm{d}t_2 = \frac{1}{2}.
$$

Thus the vectorization of the truncated signature at order 2 is nothing else than the vector $(1, \mathcal{S}_{(1)}(X_1), \mathcal{S}_{(2)}(X_1), \mathcal{S}_{(1,1)}(X_1), \mathcal{S}_{(2,1)}(X_1), \mathcal{S}_{(1,2)}(X_1), \mathcal{S}_{(2,2)}(X_1))^\top$.

These coefficients can be interpreted geometrically as shown in Figure D.5 inspired from the work of Fermanian (2021): the signature coefficients of order 1 correspond to increments and signature coefficients of order 2 correspond to areas.
We will refer to coefficient-signature for the quantity

$$
\mathcal{S}(X) = (1, \mathcal{S}_{(1)}(X), \ldots, \mathcal{S}_{(p)}(X), \mathcal{S}_{(1,1)}(X), \ldots, \mathcal{S}_{(1,p)}(X), \mathcal{S}_{(2,1)}(X), \ldots)
$$

and to truncated coefficient-signature at order D for the quantity

$$
\mathcal{S}^D(X) = (1, \mathcal{S}_{(1)}(X), \ldots, \mathcal{S}_{(p)}(X), \mathcal{S}_{(1,1)}(X), \ldots, \mathcal{S}_{(1,p)}(X), \ldots, \underbrace{\mathcal{S}_{(p,\ldots,p)}(X)}_{D \text{ terms}}).
$$

Here we have presented the essential concepts of signatures. Nevertheless, the interested reader can refer to Friz and Victoir (2010); Lyons (2014); Fermanian (2021) for more detailed concepts.
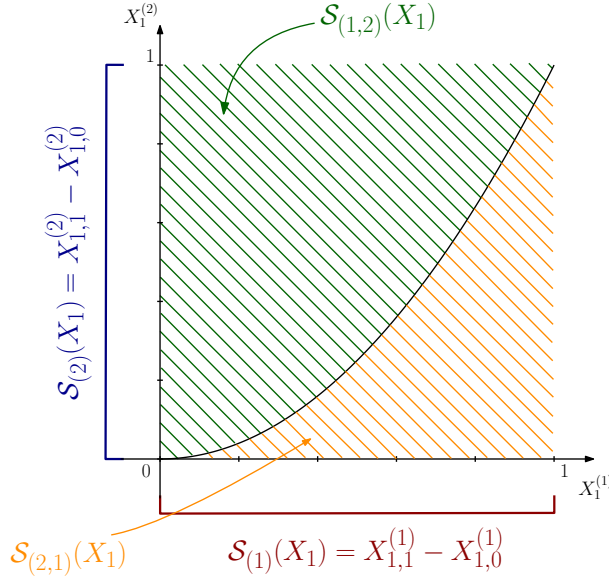
Figure D.5: Illustration of the geometrical interpretation of the signature coefficients with the path $X_{1,t} = (t, t^2)^\top \; \forall t \in [\![0,1]\!]$.

3 Parametric and nonparametric modeling of spatial data

Here we introduce classical approaches to model spatial data. These approaches can be parametric or nonparametric and allow to estimate the variable of interest in unobserved spatial locations. For example, if one is interested in the air quality in a given geographical area, it is almost impossible in practice to measure the air pollution at every geographic location in the study area. These models allow to predict the pollution measurements in the unobserved spatial locations and thus to obtain an estimate of the air quality in the whole study area. Here we focus on parametric estimation using spatial autoregressive (SAR) models and on nonparametric estimation using kernel approaches.

3.1 SAR model for spatial data modeling

Let Y be the real-valued random variable of interest and $X = (X^{(1)}, \ldots, X^{(p)})^\top$ the vector of p real-valued explanatory variables. We note $Y_1, \ldots, Y_n, X_1, \ldots, X_n$ the observations in n spatial locations $s_1, \ldots, s_n$. Spatial models usually include a spatial weight matrix in order to take into account the spatial dependence between the observations. We denote V_n this $n \times n$ matrix composed of elements $v_{i,j}$ usually defined as inversely proportional to the distance between the spatial locations s_i and s_j. In practice the weights $v_{i,j}$ can be defined in several manners among which (i) the inverse distance weights ($v_{i,j} = ||s_i - s_j||^{-\alpha}, \; \alpha > 0$), (ii) the weights based on the k-nearest neighbors ($v_{i,j} = 1$ if s_j belongs to the set of k-nearest neighbors of s_i, 0 otherwise) and (iii) the weights based on the notion of adjacency (i.e., $v_{i,j} = 1$ if s_i and s_j share a common boundary, 0 otherwise).

In addition, several types of interaction can be considered. Here we focus on the case where the variable of interest is spatially correlated (endogenous interaction effect), i.e., the value of Y at the spatial location s_i (Y_i) depends on the values of $\{Y_j\}_{j \neq i}$. This is the interaction considered by the SAR model (Cliff and Ord, 1973):

$$Y_i = \rho_0 \sum_{j \neq i} v_{i,j} Y_j + \boldsymbol{\beta}^\top X_i + \varepsilon_i \tag{D.2}$$

where the ε_i are independent and identically distributed random variables such that $\mathbb{E}[\varepsilon_i] = 0$,

$\mathbb{E}[\varepsilon_i^2] = \sigma_0^2$.

Then Equation D.2 can be rewritten as

$$\boldsymbol{Y} = \rho_0 V_n \boldsymbol{Y} + \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$(I - \rho_0 V_n)\boldsymbol{Y} = \boldsymbol{X}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top, \boldsymbol{X} = (X_1 | \ldots | X_n)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. The parameter $\rho_0 \in [-1, 1]$ controls the intensity of the spatial correlation of Y. Several methods have been proposed to estimate these models (Smirnov and Anselin, 2001; Lee, 2007). However, in the case where the errors do not follow a normal distribution, the quasi-maximum likelihood method is appropriate (Ahmed et al., 2021a):

By defining, $S_0 = I - \rho_0 V_n$, $S_n(\rho) = I - \rho V_n$, the quasi-log-likelihood of the model is

$$\ell(\rho, \beta, \sigma^2) = -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \sigma^2 + \ln |S_n(\rho)| - \frac{1}{2\sigma^2} [S_n(\rho)\boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{\beta}]^\top [S_n(\rho)\boldsymbol{Y} - \boldsymbol{X}^\top \boldsymbol{\beta}].$$

Then the quasi-likelihood estimators of $\rho, \boldsymbol{\beta}, \sigma^2$ are derived from the maximization of $\ell(\rho, \beta, \sigma^2)$. Following Lee (2004), the estimators of $\boldsymbol{\beta}$ and σ^2 are usually derived for a given ρ:

$$\hat{\boldsymbol{\beta}}(\rho) = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}S_n(\rho)\boldsymbol{Y}$$
$$\widehat{\sigma^2}(\rho) = \frac{1}{n}[S_n(\rho)\boldsymbol{Y} - \boldsymbol{X}^\top \hat{\boldsymbol{\beta}}(\rho)]^\top [S_n(\rho)\boldsymbol{Y} - \boldsymbol{X}^\top \hat{\boldsymbol{\beta}}(\rho)].$$

Then the concentrated quasi-log-likelihood is defined as

$$\ell(\rho) = -\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \widehat{\sigma^2}(\rho) + \ln |S_n(\rho)| - \frac{1}{2\widehat{\sigma^2}(\rho)}[S_n(\rho)\boldsymbol{Y} - \boldsymbol{X}^\top \hat{\boldsymbol{\beta}}(\rho)]^\top [S_n(\rho)\boldsymbol{Y} - \boldsymbol{X}^\top \hat{\boldsymbol{\beta}}(\rho)]$$

and $\hat{\rho}$ is found by maximizing $\ell(\rho)$. Finally the estimators of $\boldsymbol{\beta}$ and σ^2 are $\hat{\boldsymbol{\beta}}(\hat{\rho})$ and $\widehat{\sigma^2}(\hat{\rho})$ respectively.

The interested reader can refer to Anselin (1988) for other spatial models.

3.2   Kernel-based nonparametric modeling for spatial data

Here we present a nonparametric spatial regression approach based on kernel estimators. Let consider the spatial set S rectangular: $S = S_{\boldsymbol{n}} = \{i = (i_1, \ldots, i_N), 1 \leq i_d \leq n_d, d = 1, \ldots, N\}$. We note $\boldsymbol{n} = (n_1, \ldots, n_N)$ and $\hat{\boldsymbol{n}} = n_1 \times \cdots \times n_N$. The asymptotic properties of the estimators can be obtained when $\boldsymbol{n} \to \infty$. Two cases can be distinguished i.e., the isotropic case ($S_{\boldsymbol{n}}$ extends to infinity at the same speed in all directions): we write $\boldsymbol{n} \to \infty$ when $\min_{d \in [\![1,N]\!]} n_d \to \infty$ and $\frac{n_{d'}}{n_{d''}} \leq C \,\forall 1 \leq d', d'' \leq N$, and the anisotropic case ($S_{\boldsymbol{n}}$ does not extend to infinity at the same speed in all directions): we write $\boldsymbol{n} \to \infty$ when $\min_{d \in [\![1,N]\!]} n_d \to \infty$.

We consider $\hat{\boldsymbol{n}}$ observations $Z_i = (Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$ of a spatial process Z and we aim at estimating nonparametrically the regression function r of Y on X: $r(x) = \mathbb{E}(Y|X = x)$.

The estimation of regression functions by the kernel method was introduced by Nadaraya (1964) and Watson (1964) in the context of n observations (Y_i, X_i). r is then estimated by a weighted average of the Y_i such that

$$\hat{r}(x) = \frac{\frac{1}{h_n} \sum_{i=1}^{n} Y_i K \left(\frac{x - X_i}{h_n}\right)}{\frac{1}{h_n} \sum_{i=1}^{n} K \left(\frac{x - X_i}{h_n}\right)},$$

where K is a kernel function such that K is a non-negative real-valued integrable function and $\int_{-\infty}^{+\infty} K(u)\, \mathrm{d}u = 1$, and $h_n > 0$ is a bandwidth parameter (such that $\lim_{n\to\infty} h_n = 0$) that can be chosen by cross-validation for example.

In the spatial context Lu and Chen (2002) and Lu and Chen (2004) proposed a nonparametric estimator of the regression function and gave conditions to ensure the weak convergence of this estimator, respectively in the case of an anisotropic spatial process (the spatial dependence depends on the direction) and an isotropic spatial process (the spatial dependence between two locations depends only on their distance and not on the direction) by extending the Nadaraya-Watson estimator:

$$\hat{r}_{\boldsymbol{n}}(x) = \frac{\dfrac{1}{h_{\boldsymbol{n}}^p} \displaystyle\sum_{i \in S_{\boldsymbol{n}}} Y_i K\left(\dfrac{x - X_i}{h_{\boldsymbol{n}}}\right)}{\dfrac{1}{h_{\boldsymbol{n}}^p} \displaystyle\sum_{i \in S_{\boldsymbol{n}}} K\left(\dfrac{x - X_i}{h_{\boldsymbol{n}}}\right)},$$

where $h_{\boldsymbol{n}} \to 0$ when $\boldsymbol{n} \to \infty$.

Note that other nonparametric methods for estimating the regression function have been proposed (Li and Tran, 2009; Robinson, 2011; Dabo-Niang et al., 2016).