UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXII

Nonparametric methods for complex spatial domains: density estimation and hypothesis testing

Coordinatore del Corso: Prof. Massimiliano Caporin

Supervisore: Prof. Laura M. Sangalli

Co-supervisore: Prof. Livio Finos

Dottorando/a: Ferraccioli Federico

September 30, 2019

Abstract

The analysis of not only big, but increasingly complex data represents a thriving branch of statistics. Modern applications ranging from neuroscience, geo-sciences, astronomy and engineering pose stimulating challenges to classical statistics and require the development of novel methodologies. In this thesis we propose nonparametric approaches to density estimation and hypothesis testing over multidimensional domains with complex shapes. The synergy of ideas and techniques from applied mathematics, numerical analysis and statistics allows us to obtain flexible and efficient tools.

The thesis is organized in three main threads. The first considers the problem of density estimation over multidimensional domains with complex shapes. Here we combine a nonparametric likelihood approach with a regularization involving partial differential operators. The second thread examines two sample hypothesis testing. Inspired by the first part, we take advantage of permutation procedures to develop high dimensional multinomial tests for distributions defined over complex domain. The last thread moves toward a parallel direction, that is the study of hypothesis testing procedures for semiparametric spatial regression models. After a careful analysis of their theoretical properties, we propose a nonparametric randomization approach to test the linear components of such models.

Sommario

L'analisi di dati non solo ad alta dimensionalità, ma soprattutto ad alta complessità rappresenta una fervente branca della statistica. Numerose applicazioni recenti derivanti da neuroscienze, geoscienze, astronomia e ingegneria cosituiscono nuove stimolanti sfide alla statistica classica e richiedono lo sviluppo di nuove metodologie. In questa tesi proponiamo degli approcci non-parametrici al problema di stima della densità e di test di ipotesi nel caso di domini multidimensionali aventi forme complesse. La sinergia tra idee e tecniche provenienti dalla matematica applicata, dall'analisi numerica e dalla statistica ci ha permesso di ottenere strumenti flessibili ed efficienti.
La tesi si sviluppa in tre principali direzioni. Nella prima viene considerato il problema di stima della densità nel caso di domini multidimensionali aventi forme complesse. In questo caso proponiamo una soluzione che combina un approccio di verosimiglianza non-parametrico ad una penalizzazione che coinvolge operatori differenziali. La seconda direzione prende in considerazione il problema di test di ipotesi a due campioni. Prendendo spunto dalla prima parte e grazie all'utilizzo di un approccio permutazionale, sviluppiamo dei test per multinomiali ad alta dimensionalità definite su domini complessi. La terza parte si muove in una direzione parrallela a quest'ultima, in particolare viene considerato lo studio di test di ipotesi per modelli semiparametrici di regressione spaziale. Dopo aver sviluppato la teoria asintotica, proponiamo un approccio di randomizzazione per testare la componente lineare di questi modelli.

To the One, true value of the parameter θ

Acknowledgements

I guess the first thanks go to Livio and Laura, the two main academic and economical investors. Thanks to Livio for leaving me the "freedom" to do what I want, for teaching me to follow the flow and for always supporting me since my Bachelor degree. Thanks to Laura for guiding me these last two years and for giving me an enormous amount of opportunities. Thanks also to the kingdom of MOX, who hosted me this last year. Tusen takk, Nils, for the possibility and the patience (and thanks for all the ~~fish~~ whales). And by extension, thanks to all of the FocuStat group and the mostly italian crew for the many pølse med lompe. Thanks also to the department of Statistics of Padova, the professors, the students, the Saviour Patrizia.
A great thanks goes to Aduso for the paychecks spent at Umbe, Pallotta for the sushi boats, Casa for the psychological support and the holiday-conferences during the PhD, Da Pont for the time spent playing LoL and Politi for the smile. Thanks to Chiara for the exchange of countless complaints. And speaking of flatmates, thanks to all the ex inhabitants of Via Altinate 47, aka the "Il vecchio e la bambina" group, Caterina, Francesca, Galia, Giada and the Giorgias. Thanks to Ambra, that still (hopefully) bear me even after more than ten years. A huge hug goes to Angelina, the best random event happened to me and the compass that guided me during this last year. And finally, thanks to my family, for supporting and pushing me forward throughout my whole life.

Contents

List of Figures **xiii**

List of Tables **xvii**

Introduction **3**
Overview . . . 3
Main contributions of the thesis . . . 4

1 Definition of tools and techniques **7**
1.1 Methodology . . . 7
1.1.1 Functional spaces . . . 7
1.1.2 Differential operators . . . 8
1.1.3 Finite element method . . . 9
1.1.4 Delaunay triangulation and Voronoi tessellation . . . 11
1.2 Spatial regression with differential regularization . . . 11
1.2.1 The model . . . 12
1.2.2 Well-posedness of the estimation problem and characterization of the solution . . . 13
1.2.3 Numerical solution . . . 14

2 Nonparametric likelihood density estimation **17**
2.1 Introduction . . . 17
2.2 Density estimation with differential regularization . . . 19
2.2.1 Equivalence to Poisson process intensity estimation . . . 20
2.3 Theoretical properties . . . 21
2.3.1 Well posedness of the estimation problem . . . 21
2.4 Estimation procedure . . . 22
2.4.1 Finite elements . . . 22
2.4.2 Discretization of the infinite dimensional estimation problem . . . 24
2.4.3 Gradient descent algorithm . . . 25
2.4.4 Initialization of the gradient descent algorithm . . . 26
2.4.5 Selection of the smoothing parameter . . . 29
2.5 Simulation studies and applications . . . 30
2.5.1 Simulation 1: mixture of gaussians over squared domain . . . 30
2.5.2 Simulation 2: density over horseshoe domain . . . 32

2.6 Portland crimes . . . 33
2.7 Future research . . . 35

3 Two samples hypothesis testing **39**
3.1 Introduction . . . 39
3.2 Methodology . . . 40
3.2.1 Definition of the test statistic . . . 41
3.2.2 Nonparametric combination . . . 42
3.3 Theoretical properties . . . 44
3.4 Simulations . . . 46
3.4.1 n bigger than K . . . 47
3.4.2 n smaller than K . . . 47
3.4.3 Unbalanced case . . . 47
3.5 Application . . . 48

4 Hypothesis testing for spatial regression models **53**
4.1 Introduction . . . 53
4.2 Model . . . 55
4.2.1 Discrete Estimator . . . 56
4.3 Asymptotic properties . . . 57
4.4 Hypothesis testing . . . 59
4.4.1 Flip-score test . . . 60
4.4.2 Block sign-flip test . . . 62
4.4.3 Random covariates . . . 65
4.5 Simulations and application . . . 66
4.5.1 Application to Switzerland rainfall . . . 68

Conclusions **73**

Appendix **77**
.1 Proof of Theorem 2.1 . . . 77

List of Figures

1 In red, the locations of crimes in the city of Portland, with the boundary of the municipality in black. 4

2.1 The figure displays the municipality of Portland, with the locations of motor vehicle thefts. The city is divided in two parts by the Willamette river. The phenomenon under study appears influenced by the complex shape of the municipality. For instance, in the northern area of the city, a much higher criminality is observed on the East side of the river with respect to the West side. This is also the case for the southern part of the city and for Hayden Island, in the northern part toward Vancouver, where the number of occurrences is much higher than in the inland nearby part of the municipality. 18

2.2 Mesh used for the study of motor vehicle theft in the city of Portland. The mesh represent very well the complex morphology of the domain, cut through by the Willemette river. The mesh is obtained as a constrained Delaunay triangulation using the functions in the `R` package `fdaPDE` (Lila et al., 2016b). 23

2.3 A linear finite element basis function on a triangulation. 23

2.4 Top left: A sample of 200 observations from a mixture of Gaussian distributions with a Voronoi tessellation of the domain. Top right: Voronoi tessellation and the dual Delaunay triangulation. Middle and bottom rows: heat diffusion estimates as the time increases. 27

2.5 Simulation 1: mixture of gaussians over squared domain. Top left: true density. Top right: boxplots of the Mean Integrated Squared Error (MISE), over the 100 simulation repetitions, of the estimates provided by Kernel Density Estimation (KDE) and the proposed Density Estimation with PDE regularization (DE-PDE); for completeness, we also display the MISE of the heat diffusion estimates that we use as starting guesses for the gradient descent algorithm that provides DE-PDE estimates. Bottom left: mean of the estimates yielded by KDE over the 100 simulation repetitions. Bottom right: mean of the estimates yielded by DE-PDE over the 100 simulation repetitions. 31

2.6 Simulation 2: density over horseshoe domain. Panels as in Figure 2.5. KDE cannot accurately deal with data over this non-trivial domain: it returns estimates where the high density values in the upper horseshoe arm are partly smoothed into the low density values in the bottom part of the domain, with isolines very different from those of the true density. The proposed DE-PDE estimates is not affected by this problem and does not display any strange behavior at the boundaries of the domain; instead it is able to very well capture the overall structure of the true density. . . 32
2.7 Estimated density for motor vehicle thefts. 33
2.8 Enlargement of the estimates provide by KDE (left) and the proposed DE-PDE (right) on Hayden Island. The figure shows that KDE estimator is not able to comply with the complex structure of the domain: though a very high number of motor vehicle theft is observed in the island, while almost no theft is observed in the nearby inland part of the municipality, KDE return an estimate that is everywhere low. Instead, the proposed DE-PDE estimator can efficiently deal with the complex morphology of the domain, returning an estimate of the theft density that is much higher over the Hayden Island. 34
2.9 Prostitution in Portland. Top left: locations of prostitution related crimes, in the year 2012. Top right: coarse data-driven mesh. Central and bottom panels: DE-PDE estimated densities, using a fine regular mesh with about 3000 nodes (left) and using the coarse data-driven mesh with about 600 nodes (right) . These images highlight how accurately the proposed method captures the density mass concentrated along Northest 82nd Avenue, that appears as a neat ridge in the three-dimensional visualization. 38

3.1 On the left, the Delaunay triangulation in black and its dual graph, the Voronoi tessellation in red. On the right, the crimes related to human trafficking in the city of Portland, together with the tessellation of the domain. 41
3.2 A comparison between the five tests proposed in four different scenarios. The power of the tests are plotted against the ℓ_1 distance between F and G. The sample sizes are $n_1 = n_2 = 1000$ while the number of cells of the multinomial is $K = 400$. Each point in the graph is an average over 1000 simulations. 48
3.3 A comparison between the five tests proposed in four different scenarios. The power of the tests are plotted against the ℓ_1 distance between F and G. The sample sizes are $n_1 = n_2 = 100$ while the number of cells of the multinomial is $K = 400$. Each point in the graph is an average over 1000 simulations. 49
3.4 A comparison between the five tests proposed in four different scenarios. The power of the tests are plotted against the ℓ_1 distance between F and G. The sample sizes are $n_1 = 4000$ and $n_2 = 400$, while the number of cells of the multinomial is $K = 400$. Each point in the graph is an average over 1000 simulations. 50

3.5 From the top-left, distribution of the crimes concerning human trafficking in the city of Portland from 2012 to 2017. 51
3.6 From the top-left, estimated densities of the distributions of crimes concerning human trafficking in the city of Portland from 2012 to 2017. . . . 51

4.1 Switzerland rainfall data. These include 467 daily rainfall measurements recorded in Switzerland on May 8, 1986. The size and color of point markers represent the value of the rainfall at each location, highlighting a strong spatial anisotropy. 54
4.2 On the left, heatmap of the matrix B_n. On the right, Reordering of B_n with $M = 15$ groups. 62
4.3 On the left, mesh and centroids of the RDD. On the right, the corresponding RDD with $M = 15$ blocks. 63
4.4 In the first column, the mesh used for the regression in the two two simulation scenarios. In the second column, one sample of RDD for each domain. In the third column, the spatial fields f considered in the simulations. 67
4.5 Switzerland rainfall data. These include 467 daily rainfall measurements recorded in Switzerland on May 8, 1986. The size and color of point markers represent the value of the rainfall at each location, highlighting a strong spatial anisotropy. 69
4.6 Comparison of the different test statistics in the case of square domain. The solid line correspond to the flip-score with RDD; the dashed line to the standard flip-score; the dotted line to the Wald test with sandwich estimation of the variance; the dashed-dotted line to the standard Wald test. 70
4.7 Comparison of the different test statistics in the case of horseshoe domain. The solid line correspond to the flip-score with RDD; the dashed line to the standard flip-score; the dotted line to the Wald test with sandwich estimation of the variance; the dashed-dotted line to the standard Wald test. 71

List of Tables

3.1 Number of crimes from 2012 to 2017 50

4.1 Covariance structure of the generated random fields 67
4.2 Covariance structure of the generated random fields 68

Introduction

Overview

The thesis develops nonparametric methods for density estimation and hypothesis testing over multidimensional domains with complex shapes. This is motivated by important applications from varied applicative fields, ranging from life-sciences (and in particular neurosciences), to geo-sciences, astronomy and engineering. Classical statistical methods are often inadequate to handle the complexity of such data, whose analysis calls for the definition of innovative methods. In this work we propose possible solutions that merge ideas and approaches from statistics, applied mathematics and engineering.
Sangalli et al. (2013) introduces a spatial regression model with differential regularization, that enables the estimation of spatial fields over multidimensional domains with a non-trivial geometry. The thesis extends this method in various directions. In the first part, a nonparametric penalized likelihood approach for density estimation is presented. The nonparametric likelihood approach gives great flexibity and allows the estimation over planar domain with complex shapes, characterized by strong concavities or interior holes. For example, figure 1 illustrates the study of criminality in the city of Portland, Oregon, USA. The figure shows the point process of crimes. In this case, the non-trivial conformation of the domain, characterized by the presence of the river, is crucial in the study of the phenomenon. The northern area of the city is affected by much higher criminality on the East side of the river with respect to the West side. Similarly to Sangalli et al. (2013), the proposed method features a regularizing term involving partial differential operators. In particular, we target the smoothness of the estimate by the Laplace operator applied to the logarithm of the density. Standard likelihood approaches are not suitable for the estimation of the proposed model. To solve the problem, we resort to numerical techniques, and in particular we use a finite element approach.
In the same setting of the previous part, we consider nonparametric one and two sample goodness-of-fit tests. As an example, we might be interested in comparing the distribution of crimes in different years. The significance levels are provided via a permutational

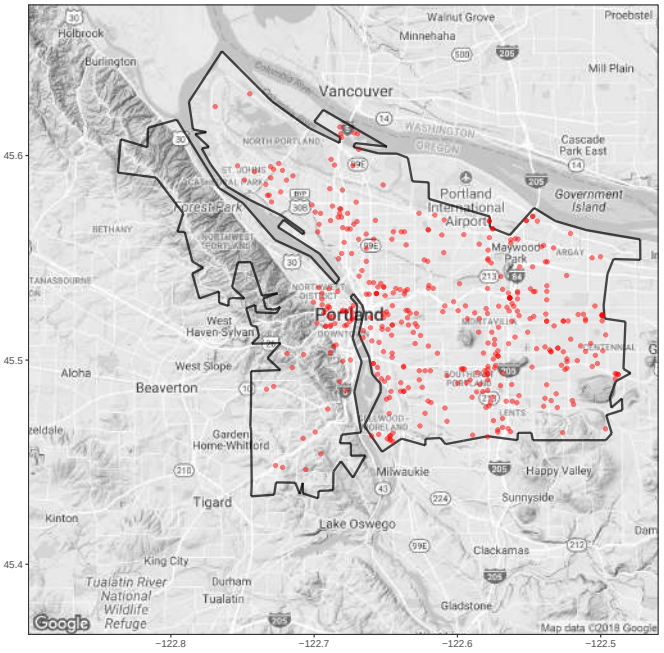

FIGURE 1: In red, the locations of crimes in the city of Portland, with the boundary of the municipality in black.

approach. The complexity of the spatial domain is taken into account using a partition defined by a Voronoi tessellation.

In the third part, we fill an important theoretical and methodological gap for the model introduced by Sangalli et al. (2013). We first study the asymptotic properties of the corresponding estimators. We then develop statistical tests for the linear components of such models. The tests are based on random sign flipping of the score components of the model, inspired by the approach of Hemerik et al. (2019).

Main contributions of the thesis

Density estimation over complex planar domains

The vast majority of density estimators are restricted to univariate domains or to unbounded multivariate domains (Scott, 2015). None of these methods address the case density estimation over domains with complex shapes. In the univariate case, a classical approach is to consider a penalized likelihood formulation. Two main choices for the regularization term has been considered. The first roughness penalty $R(f) = ||(\sqrt{f})^{(1)}||_2^2$ of Good and Gaskins (1980), and the penalization functional $R(f) = ||(\log f)^{(3)}||_2^2$ of Silverman (1982), where the exponent denotes the first and third derivative respectively.

Both proposals naturally impose the positive constraint on f in their formulation. Although both models could be in principle extended to the multivariate setting, the generalization to domain with complex shapes is far for trivial.
We propose a new method for density estimation able to deal with data distributed over complex planar domains. We consider the penalization functional given by the Laplacian of the log-density. This functional is a measure of local curvature and therefore controls the smoothness of the estimates. A key feature of the Laplacian is the invariance with respect to Euclidean transformations of the spatial coordinates. It therefore ensures that the concept of smoothness does not depend on the orientation of the coordinate system. The density corresponding to the null family of the operator is the uniform distribution over the domain. We study the theoretical properties of the method. Moreover, we provide an appropriate discretization of the infinite dimensional estimation problem via the Finite Element Method (FEM). This methodology is often used in engineering applications to solve partial differential equations. A crucial advantage of the use of these numerical techniques consists in the possibility of considering spatial domains with complex shapes (instead of simple tensorized domains as considered by most of the available spatial methods). The strong synergy between statistical and numerical approaches and tools ensures the high flexibility and the computational efficiency of the proposed method. Extensive simulation studies show that the method compare favourably to existing density estimation techniques, both on simple two dimensional domains and, even more, on domain with complex shapes.

One and two samples hypothesis testing for distribution over complex domains

The problem of goodness-of-fit tests has been studied in details in the case of univariate data. Classical well-known nonparametric tests are for example the Kolmogorov-Smirnov test, the number of runs test and the longest run test. Much less work has been devoted to higher dimensional cases, particularly in the cases where no directional alternatives are defined. A recent work by Arias-Castro et al. (2018) develops the theory of histogram tests in arbitrary dimension, extending the work of Ingster (1987).
We propose a nonparametric procedure for two samples hypothesis testing for distributions over complex spatial domains. In particular, we considers tests on smooth distributions by recasting the problem to an appropriate high-dimensional multinomial test. Recent results on these approaches are given in Balakrishnan and Wasserman (2018). The continuous densities are quantized via an appropriate Voronoi tessellation of the spatial domain. Differently from the histogram method in Arias-Castro et al.

(2018), the proposed approach based on Voronoi tessellation has higher flexibility and can handle domains with complex shapes in any dimension. A permutation procedure is used to provide the distribution under the null hypothesis and the corresponding p-values.

Hypothesis testing in spatial regression models

The Spatial Regression models with differential regularization proposed in Sangalli et al. (2013) have been shown to be numerically efficient and capable to handle complex applied problems (see for example Azzimonti et al. (2015) and Lila et al. (2016a)). On the other hand, the theoretical properties are still largely unexplored. Here we study the consistency and the asymptotic normality of such estimators, following and approach proposed by Yu and Ruppert (2002) for penalized spline estimation.
Although the resulting asymptotic distribution of the estimators might be used for hypothesis testing, the corresponding performances are far from acceptable. In the finite sample scenario, the variance is usually overestimated due to the regularization term. To reduce this problem, we proposed nonparametric test procedure based on random sign flipping of the score components of the model, inspired by Hemerik et al. (2019). The proposed method uses Random Domain Decomposition, in the same fashion as Menafoglio et al. (2018), to reduce the effect of the spatial dependence without additional parametric assumption on the form of the correlation structure.

Chapter 1

Definition of tools and techniques

1.1 Methodology

A recurring motif in this thesis is the combination of methods and ideas from statistics, applied mathematics and engineering. This is motivated by the complexity of modern real data applications, where the complexity might come from certain underlying physical process or the geometry of the domains involved. These challenges call for the development of new methods able to handle such complexity. In this first section we introduce some concepts and tools that are not frequently discussed in the statistical literature.

1.1.1 Functional spaces

We first start with some basic definition of functional spaces that we use throughout the thesis. Let $L^p(\Omega)$ the Lebesgue space

$$L^p(\Omega) := \{u : ||u||_{L^p(\Omega)} < \infty\},$$

equipped with the norm

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u|^p \, dx\right)^{1/p},$$

in the case $1 \leq p < \infty$, while for $p = \infty$ we have

$$\|u\|_{L^\infty(\Omega)} := \sup_{x\in\Omega} |u(x)|.$$

Let $W^{k,p}(\Omega)$ the Sobolev space defined as

$$W^{k,p}(\Omega) := \{u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega) \ \forall |\alpha| \leq k\}.$$

The space $W^{k,p}(\Omega)$ is equipped with the norm

$$\|u\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}},$$

in the case $1 \leq p < \infty$, while in the case $p = \infty$

$$\|u\|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| \leqslant k} \|D^\alpha u\|_{L^\infty(\Omega)},$$

where the symbol $D^\alpha u$ denote the weak derivative (see e.g. Adams (1975) for a more detailed treatment of Sobolev spaces). In the rest of the thesis, we denote the Hilbert space $W^{k,2}(\Omega)$ by $H^2(\Omega)$. Sobolev spaces with $p = 2$ are especially important because of their connection with Fourier series and because they form a Hilbert space.
All these concepts are used in the following chapters for the definition of the methods and in some of the proofs.

1.1.2 Differential operators

We now introduce one of the key ingredient of modern mathematics, Partial Differential Equations (PDE). PDEs represents an extremely powerful tool to describe many natural phenomena, such as sound and heat diffusion, fluid dynamics, quantum mechanics and gravitational fields. These equations involve rates of change with respect to continuous variables, i.e. multivariable functions and partial derivatives. In this thesis we consider second order linear PDEs, that have the general form

$$Af_{xx} + 2Bf_{xy} + Cf_{yy} + Df_x + Ef_y + u = 0,$$

In particular we focus on the case where $B^2 - AC < 0$. This lead to the so called elliptic PDEs, equations characterized by smooth solutions within the interior of the region where the equation and solutions are defined.
The most simple and probably most studied elliptic PDE is the Laplace equation, defined in Cartesian coordinates as

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0.$$

The Laplacian operator Δf gives the difference between the average value of a function in the neighboring of a point, and its value at that point. In order to have a unique

solution to this equation, we need to add some boundary conditions, such as

$$f = g \quad \text{on} \quad \Gamma \subset \partial\Omega, \tag{1.1}$$

know as the Dirichlet boundary conditions, or

$$\partial_n f = \nabla f \cdot \mathbf{n} = g \quad \text{on} \quad \partial\Omega \backslash \Gamma,$$

known as Neumann boundary conditions, where $\nabla f \cdot \mathbf{n}$ denotes the normal derivative on the boundary. We use these equations in Chapter 2 and Chapter 4 to introduce regularization functionals in the proposed models.
In the case of Dirichlet boundary conditions, the Laplace equation has a nice interpretation in terms of heat. Imagine we have a square metal plate and fix the temperature of the boundary according to the specified conditions. Then let the heat flow until it reaches a stationary state. The temperature distribution is given by the solution of the corresponding Dirichlet problem. In other words, the Laplace equation can be thought of as the steady-state of another very important equation, the heat equation, defined as

$$\frac{\partial f}{\partial t} = \alpha \Delta f.$$

The equation says that the rate $\frac{\partial f}{\partial t}$ at which a material at a point will heat up (or cool down) is proportional to how much hotter (or cooler) the surrounding material is. The coefficient α in the equation takes into account the thermal conductivity, the specific heat, and the density of the material. This equation has an interesting properties from a statistical points of view. The fundamental solution of the heat equation, corresponding to the initial condition of an initial point source of heat at a known position, is the heat kernel

$$\Phi(x,t) = \frac{1}{\sqrt{4\alpha\pi t}} \exp\left(-\frac{x^2}{4\alpha t}\right).$$

The relationship between the solution of the heat equation and the concept of smoothing has been discussed in various works, such as Chaudhuri and Marron (1999) and Botev et al. (2010). In Chapter 2 we use a similar concept of smoothing in the development of a density estimation method.

1.1.3 Finite element method

Closely related to PDE is the Finite Element Method (FEM), a numerical method widely used in engineering and mathematical physics. Let us briefly introduce the theory

behind it. We start with a domain Ω to which our observations belong. As first step, we need an appropriate partition of this domain. A convenient choice is to consider a subdivision into triangles, also called triangulation. Let us call this triangulation $\mathcal{T}$. The triangles must fulfill a rule, that is, they cannot overlap. If two triangles have some intersection, it is either a common vertex or a common full edge. Having defined a triangulation of the domain, we can now build the second piece of the puzzle. First consider two triangles sharing a common edge, T and T'. We can build a piecewise linear function simply defining its value at the four vertices. Moreover, since the value on the common edge depend only on the values on the two common vertices, the resulting function is continuous.
Doing this for every triangle, we end up with a function that is linear on each triangle and globally continuous. The space of such function is defined as

$$V_{\mathcal{T}} = \left\{ f_{\mathcal{T}} \in \mathcal{C}(\bar{\Omega}) |\ f_{\mathcal{T}}|_T \in \mathbb{P}_1,\, \forall T \in \mathcal{T} \right\},,$$

where $\mathbb{P}_1$ is the space of polynomial function with degree at most one. Fixing the values on the set of vertices of the triangulation gives a unique $f_{\mathcal{T}} \in V_{\mathcal{T}}$ with those values on the vertices. Every element of $V_{\mathcal{T}}$ is therefore uniquely determined by its values on the set of vertices.
Let us now denote $\mathbf{p}_i$ the i-th vertex of the triangulation, with i varying from 1 to the number of vertices K. If we fix one node (vertex) and associate the value one to this node and zero to all others, there exists a unique function $\psi_i \in V_{\mathcal{T}}$ that has this values, that is

$$\psi_i(\mathbf{p}_j) = \delta_{ij} = \begin{cases} 1 & j = i \\ 0 & j \neq i \end{cases}.$$

The function ψ_i vanishes on a triangle T if the triangle has not $\mathbf{p}_i$ as vertex. The support of ψ_i is therefore the union of triangles that has $\mathbf{p}_i$ as vertex.
We now have the building block to define

$$f_{\mathcal{T}} = \sum_{i=1}^{K} f_{\mathcal{T}}(\mathbf{p}_i)\psi_i.$$

It is easy to prove that $\{\psi_i | i = 1, \ldots, K\}$ is a basis of $V_{\mathcal{T}}$ and therefore $\dim V_{\mathcal{T}} = K$. Moreover, every element of $V_{\mathcal{T}}$ can be written as linear combination of the element of the basis, with coefficients the values of the function at the vertices. The FEM is the basis for the development af the methods in Chapter 2 and Chapter 4.

1.1.4 Delaunay triangulation and Voronoi tessellation

As we discussed in the previous section, FEM relies on the definition of a partition of the domain of interest. Since we are dealing with quite general domains Ω, with possibly complex shapes, an appropriate choice is a Delaunay triangulation. Given a set of points $\mathcal{K}$ in some domain $\Omega \in \mathbb{R}^d$, a Delaunay is a triangulation such that no point in $\mathcal{K}$ is inside the circum-hypersphere of any d-simplex of the triangulation. A generalization of this method is the so called constrained Delaunay triangulation. This type of triangulation allows to force certain required segments, for instance boundary segments. We use this method in all the chapters of the thesis as a starting point of our modelization.
In the case of Euclidean space, the dual graph for a Delaunay triangulation corresponds to the Voronoi tessellation for the same set of points. The Voronoi tessellation is a partitioning of a domain into regions based on the distance to a specific set of points. Let $\mathcal{P} = \{D_1, \dots, D_K \subset \Omega\}$ be a partition of the domain into disjoint sub-regions. Each Voronoi cell D_k is defined as

$$D_k = \{x \in X \mid d(x, c_k) \leq d(x, c_j) \text{ for all } j \neq k\},$$

where the points $(c_1, \dots, c_k)$ represent the centers of each cell. Different choices of the distance $d(\cdot, \cdot)$ lead to different types of tessellation. Classical choices are the Euclidean distance $d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}$, or the Manhattan distance $d(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{d} |p_i - q_i|$. We use the Voronoi tessellation in Chapter 3 and Chapter 4 to define an appropriate nonparametric test procedure.

1.2 Spatial regression with differential regularization

In the first section we briefly introduced the reader to some basic concepts of modern applied mathematics and numerical analysis. Now we define what represents the starting point for various chapters of the thesis, the spatial regression with differential regularization method (SR-PDE). The SR-PDE, introduced by Sangalli et al. (2013), comprises a class of models that can handle data displaying complex spatial or spatiotemporal dependencies. In these models, unlike classical spatial data analysis approach (see Cressie (1992) for a review), where the unknown spatial field is assumed to be stochastic and the covariance of the stochastic field is used to model the second order spatial dependence of the phenomenon under study, the unknown spatial field is assumed to be deterministic

and the spatial structure of the phenomenon is modelled via the PDE in the regularizing term. The regularization based on elliptic differential operators represent a rich and flexible tool to model spatial and spatio-temporal variation, and can naturally include various forms of anisotropy and non-stationarity. If problem-specific knowledge about the phenomenon under study is available, this information might be suitably formalized in a PDE that can be used to define the spatial or spatio-temporal structure of the model. We now define the model and describe its numerical solution, that resorts on numerical analysis techniques, such as finite elements analysis or isogeometric analysis based on advanced spline bases (see for example Quarteroni and Quarteroni (2009)).

1.2.1 The model

Suppose we have $\{\mathbf{p}_i, i = 1, \ldots, n\}$ data locations over a bounded regular domain $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega \in \mathcal{C}^2$. We observe n realizations $z_i \in \mathbb{R}$, that correspond to the values of the variable of interest at point $\mathbf{p}_i$. Let also $\mathbf{w}_i \in \mathbb{R}^q$ be a the vector covariates associated to the observation z_i, and f is an unknown deterministic mean field that captures the spatial structure of the problem. With these ingredients, we define a semi-parametric model of the form

$$z_i = \mathbf{w}_i^\top \boldsymbol{\beta} + f(\mathbf{p}_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is the vector of regression coefficients, and ϵ_i are i.i.d. random errors with zero mean and variance σ^2. The estimation problem can be solved using the method proposed in Sangalli et al. (2013), minimizing the regularized least squares

$$\sum_{i=1}^{n} (z_i - \mathbf{w}_i^\top \boldsymbol{\beta} - f(\mathbf{p}_i))^2 + \lambda \int_\Omega (Lf - u)^2 \, \mathrm{d}\mathbf{p}, \tag{1.3}$$

where $\lambda > 0$ is the smoothing parameter and L denotes an elliptic differential operator. To define the form of the operator, we need to introduce a symmetric and positive definite matrix $K = \{K_{ij}\} \in \mathbb{R}^{2\times 2}$, named diffusion tensor, a vector $\mathbf{b} = \{b_j\} \in \mathbb{R}^2$, named transport vector, and a real scalar $c > 0$, named reaction term. The operator L can be defined as

$$Lf = -\mathrm{div}(K\nabla f) + \mathbf{b}\nabla f + cf, \tag{1.4}$$

where the first term in the sum is a second order differential operator defined as

$$\mathrm{div}(K\nabla f) = \frac{\partial}{\partial p_1}\left(K_{11}\frac{\partial f}{\partial p_1} + K_{12}\frac{\partial f}{\partial p_2}\right) + \frac{\partial}{\partial p_2}\left(K_{21}\frac{\partial f}{\partial p_1} + K_{22}\frac{\partial f}{\partial p_2}\right),$$

the second term is a first order differential operator defined as

$$\mathbf{b}\nabla f = b_1 \frac{\partial f}{\partial p_1} + b_2 \frac{\partial f}{\partial p_2},$$

and the last term is a zeroth order operator. The three terms in (1.4) can model various form of anisotropy and non-stationarity. The first element $\text{div}(K\nabla f)$ of (1.4), is a diffusion term that induces a smoothing in all the directions. If the matrix K is a multiple of the identity matrix I, it induces an isotropic smoothing effect. Otherwise it implies an anisotropic smoothing with a preferential direction, that corresponds to the first eigenvector of the diffusion tensor K. The degree of anisotropy induced by the diffusion tensor K is controlled by the ratio between its first and second eigenvalue. The transport term $\mathbf{b}\nabla f$ induces a smoothing only in the direction specified by the transport vector $\mathbf{b}$. Finally, the reaction term cf has instead a shrinkage effect, since penalization of the L_2 norm of f induces a shrinkage of the surface to zero. The term u in (1.3) is called forcing term, and can either be the null function $u = 0$ (so-called homogeneous case), or $u \neq 0$ (non-homogeneous case), even further increasing the flexibility in the modeling of space variation. A special case is the the Laplacian Δf, obtained setting $K = I$, $b = 0$, $c = 0$ and $u = 0$. The Laplacian induces an isotropic and stationary smoothing.

1.2.2 Well-posedness of the estimation problem and characterization of the solution

The well-posedness of the estimation problem (1.3) and the uniqueness of its solutions are discussed in Sangalli et al. (2013) and Azzimonti et al. (2014). In particular, they show that (1.3) is well defined for $\boldsymbol{\beta} \in \mathbb{R}^q$ and $f \in H^2(\Omega)$. The solution is also unique imposing appropriate boundary conditions on f. To state the exact results, we need to introduce some relevant quantities. Let $V(\Omega)$ be the subspace of $H^2(\Omega)$ characterized by the chosen boundary conditions. Set $\mathbf{z} = (z_1, \ldots, z_n)$ the vector of observed data values, and, for any function $\mathbf{v} = (v(\mathbf{p}_1), \ldots, v(\mathbf{p}_n))$ the vector of evaluations of v at the n spatial locations. In presence of covariates, denote by W the $n \times q$ matrix whose i-th row is given by $\mathbf{w}_i^t$, the vector of q covariates associated with observation z_i at $\mathbf{p}_i$, and assume that W has full rank. Moreover, set $Q = I - W(W^\top W)^{-1}W^\top$, the matrix that projects onto the orthogonal complement of $\mathbb{R}^n$ with respect to the subspace of $\mathbb{R}^n$ spanned by the columns of W. Finally, let us denote by $\hat{\mathbf{f}}_n$ the vector of evaluations of f at the n data locations, $\hat{\mathbf{f}}_n = (f(\mathbf{p}_1, \ldots, \mathbf{p}_n))$. The following proposition characterizes the solution to the estimation problem, under homogeneous boundary conditions and

forcing terms $u = 0$. The case for non-homogeneous boundary conditions and forcing terms is discussed in Azzimonti et al. (2014) .

Proposition 1.1. *There exists a unique pair of estimators* $(\hat{\boldsymbol{\beta}} \in \mathbb{R}^q, \hat{f} \in V(\Omega))$ *which minimize* (1.3)*. Moreover*

$$\hat{\boldsymbol{\beta}} = (W^\top W)^{-1} W^\top (\mathbf{z} - \hat{\mathbf{f}}_n),$$

and $\hat{f}$ *satisfies*

$$\mathbf{v}_n^\top Q \hat{\mathbf{f}}_n + \lambda \int_\Omega (Lv)(L\hat{f}) = \mathbf{v}_n^\top Q \mathbf{z}, \quad \forall v \in V(\Omega). \tag{1.5}$$

1.2.3 Numerical solution

The solution to the fourth order differential problem (1.5) cannot be found analytically. It is therefore necessary to use advanced numerical discretization procedures, as shown in Sangalli et al. (2013). One possibility is to use the FEM, described in the previous sections. We now introduce some quantities, related to this method, that are used in the following chapters. First, let $\mathcal{T}$ denote a regular triangulation of the domain Ω. The domain is consequently approximated by the union of all triangles $\Omega_\mathcal{T}$, while the boundary $\partial\Omega$ is approximated by a polygon (or multiple polygons in case of a domain with holes). On $\mathcal{T}$, we define the finite element space $V_\mathcal{T}^r$, with $r = 1, 2, \ldots$, as the space of continuous surfaces over $\Omega_\mathcal{T}$ that are polynomials of degree r when restricted to any triangle in $\mathcal{T}$. Let Ψ be the $n \times N_\mathcal{T}$ matrix evaluating the $N_\mathcal{T}$ basis functions $\psi_1, \ldots, \psi_{N_\mathcal{T}}$ at the n data locations

$$\Psi = \begin{bmatrix} \psi_1(\mathbf{p}_1) & \ldots & \psi_{N_\mathcal{T}}(\mathbf{p}_1) \\ \vdots & \ddots & \vdots \\ \psi_1(\mathbf{p}_n) & \ldots & \psi_{N_\mathcal{T}}(\mathbf{p}_n) \end{bmatrix}.$$

This matrix allows to evaluate the value of each function in $V_\mathcal{T}^r$ as a linear combination of the set of basis. Let also R_0 and R_1 be the $N_\mathcal{T} \times N_\mathcal{T}$ matrices

$$R_0 = \int_{\Omega_\mathcal{T}} \boldsymbol{\psi}\boldsymbol{\psi}^\top \quad \text{and} \quad R_1 = \int_{\Omega_\mathcal{T}} \nabla\boldsymbol{\psi} K \nabla\boldsymbol{\psi}^\top + \mathbf{b}\boldsymbol{\psi}\nabla\boldsymbol{\psi}^\top + c\boldsymbol{\psi}\boldsymbol{\psi}^\top,$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_K)^\top$ is the vector of basis functions. These matrices are called mass and stiffness matrix, and represent the system of linear equations related approximation of the operator in the penalization term. Having defined these quantities, the problem (1.5) can now be recasted in the finite element space.

Proposition 1.2. *There exists a unique pair of estimators* $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}})$ *which solve the discrete counterpart of the estimation problem (Sangalli et al., 2013). Moreover,*

$$\hat{\boldsymbol{\beta}} = (W^\top W)^{-1} W^\top (z - \hat{\mathbf{f}}_n),$$

and $\hat{f} = \Psi\hat{\mathbf{f}}$*, with* $\hat{\mathbf{f}}$ *satisfying*

$$\begin{bmatrix} -\Psi^\top Q\Psi/n & \lambda R_1^\top \\ \lambda R_1 & \lambda R_0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} -\Psi^\top Q\mathbf{z}/n \\ 0 \end{bmatrix}. \tag{1.6}$$

Although the system is typically large, being of order $2N_{\mathcal{T}}$, it is highly sparse because the matrices R_0 and R_1 are highly sparse, since the cross-products of nodal basis functions and of their partial derivatives are mostly zero, due to local support of the bases. For simplicity, we define $P = R_1 R_0^{-1} R_1$, a symmetric positive definite matrix. Moreover, let S be the $n \times n$ matrix

$$S = \Psi(\Psi^\top Q\Psi + \lambda P)^{-1}\Psi^\top Q.$$

With this notation, we have

$$\begin{aligned} \hat{\mathbf{f}}_n &= S\mathbf{z}, \\ \hat{\boldsymbol{\beta}} &= (W^\top W)^{-1} W^\top (I - S)\mathbf{z}. \end{aligned} \tag{1.7}$$

Note that the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ are linear in the observed data values $\mathbf{z}$. In Chapter 4, we will develop the asymptotic properties of this model, in particular focusing on consistency and asymptotic normality.

Chapter 2

Nonparametric likelihood density estimation

2.1 Introduction

Density estimation represents a core tool in statistic. It is essential for the visualization of data structures in exploratory data analysis and represents the starting point for regression and classification problems. Furthermore, it is fundamental in the more recently studied density based clustering and ridge estimation. In this Chapter in particular we consider density estimation over planar domains with non-trivial geometries, including those with complex boundaries, sharp concavities or interior holes. Figure 2.1 illustrates the kind of problem we are considering. The points correspond to crime locations in the municipality of Portland, Oregon. The data come from the Portland Police Bureau, and they comprise a collection of different crime categories in different years. The interest here is to estimate the distribution of crimes, in order to identify critical and dangerous areas in the city. In this case, the complex geographical conformation of the domain, characterized by the presence of the river, is crucial in the study of the phenomenon. For instance, in the northern area of the city, a much higher criminality is observed on the East side of the river with respect to the West side. Standard density estimators, such as the gold standard kernel density estimator (KDE) (Wand and Jones, 1994) or the spline density estimator (Gu and Qiu, 1993), are not appropriate in this case. These methods in fact rely on Euclidean distances, thus leading to inaccurate estimate when the phenomenon under studies is influenced by shape of the spatial domain. The same hold for the other recent proposals to density estimation, such as shape-constrained methods (Carando et al., 2009; Cule et al., 2010; Samworth, 2018) and heat diffusion estimators (Chaudhuri and Marron, 1999; Botev et al., 2010).

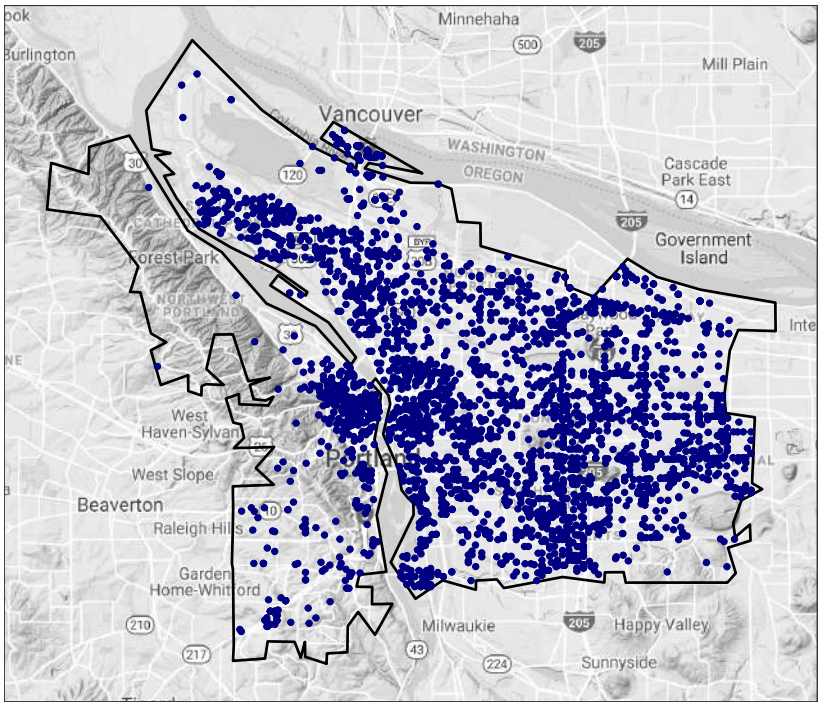

FIGURE 2.1: The figure displays the municipality of Portland, with the locations of motor vehicle thefts. The city is divided in two parts by the Willamette river. The phenomenon under study appears influenced by the complex shape of the municipality. For instance, in the northern area of the city, a much higher criminality is observed on the East side of the river with respect to the West side. This is also the case for the southern part of the city and for Hayden Island, in the northern part toward Vancouver, where the number of occurrences is much higher than in the inland nearby part of the municipality.

The modelling of data distributed over complex planar domains has recently attracted an increasing interest; see, e.g., Ramsay (2002); Lai and Schumaker (2007); Wang and Ranalli (2007); Wood et al. (2008); Sangalli et al. (2013); Scott-Hayward et al. (2014); Menafoglio et al. (2018); Niu et al. (2019). To the best of our knowledge, no one has so far considered the problem of density estimation in such setting.

In this Chapter we develop a flexible density estimation method, based on a nonparametric likelihood approach, with a regularizing term involving a partial differential operator. We study the theoretical properties of the proposed estimator, such as uniqueness and consistency. From a theoretical perspective, an analogous regularized nonparametric likelihood approach has been considered in the context of simple multidimensional domains by Gu and Qiu (1993), and formerly, in the univariate case, by Good and Gaskins (1980) and Silverman (1982); on the other hand, the generalization of the latter estimators to complex domains is not obvious. Another sensible strategy would be to use a log-linear expansion within a sieve likelihood framework, such as in Shen (1997).

Although the sieve maximum likelihood possesses interesting efficiency properties, the choice of an appropriate expansion and the generalization to complex domains is again not straightforward. Here we propose an highly innovative method to tackle the estimation problem. This method leverages on advanced numerical analysis techniques, making use of finite elements. The finite element method (see, e.g., Ciarlet, 2002) is often used in engineering applications to solve partial differential equations. A crucial advantage of such techniques consists in the possibility of considering spatial domains with complex shapes, instead of simple tensorized domains, as considered by Gu and Qiu (1993) and by the other available methods for density estimation. Moreover, the proposed method for density estimation does not impose any shape constraints, allowing for the estimation of fairly complex structures. In particular, thanks to the finite element formulation, the method is able to capture highly localized features, and lower dimensional structures such as ridges. This ability also makes the method particularly well suited in research areas such as density based clustering (Chacón, 2015) and ridge estimation (Genovese et al., 2014). As a byproduct, we also describe an innovative heat diffusion estimator, inspired by the works of Chaudhuri and Marron (1999) and Botev et al. (2010), that is able to handle data distributed over complex domains.
The Chapter is organized as follows. Section 2.2 introduces the proposed nonparametric likelihood density estimator with differential regularization. The same section also outlines the equivalence of the considered problem with Poisson process intensity estimation. In Section 2.3 we study its theoretical properties and prove the consistency of the estimator. In Section 2.4.1 we describe the estimation procedure. Section 5 reports some simulation studies that show the performances of the proposed method with respect to state of the art techniques. Section 6 gives the application to the Portland crime data. Section 2.7 discusses possible directions for future research.

2.2 Density estimation with differential regularization

We consider the problem of estimating a density function f on a spatial domain $\Omega \subset \mathbb{R}^2$. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n independent realizations from f. We use the logarithm transform $g = \log f$, and propose to estimate f by finding the function g that minimizes the negative penalized log likelihood

$$L(g) = -\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i) + \int_{\Omega} \exp(g) \; + \lambda \int_{\Omega} \big(\Delta g\big)^2 \tag{2.1}$$

where $\lambda > 0$. The first term in (2.1) is the negative log-likelihood. The second term is necessary to ensure that the estimate integrates to one. The third term is a regularization, necessary to avoid unbounded likelihoods. In fact, unlike classical parametric density estimation, where the parameter space is finite dimensional and the form of f is assumed known, here we deal with an infinite class of densities. In particular, the regularization term we use involves the Laplacian, a differential operator defined as

$$\Delta g = \frac{\partial^2 g}{\partial x_1^2} + \frac{\partial^2 g}{\partial x_2^2}$$

where $\boldsymbol{x} = (x_1, x_2)$. The Laplacian provides a measure of the local curvature of g. This regularization thus controls the roughness of the estimate. In particular, when the smoothing parameter λ increases, the solution should ideally flatten out and presents less bumps.

Instead of the simple Laplacian, the regularizing term could as well involve more complex partial differential operators, or the misfit of a Partial Differential Equation (PDE). This is particularly interesting when some problem-specific information about the phenomenon is available, that can be formalized in terms of a PDE. This is explored in the context of spatial and spatio-temporal regression methods in Azzimonti et al. (2015) and Arnone et al. (2019).

2.2.1 Equivalence to Poisson process intensity estimation

In thhis section we briefly discuss the relationship of the proposed estimator with the problem of estimating a Poisson intensity and a possible extension of our proposal. The estimation of spatial point processes, especially inhomogeneous, are emerging as fundamental in many applications. Some likelihood approaches for inhomogeneous process have been proposed by Guan and Shen (2010) and Waagepetersen and Guan (2009). Here weighted estimating equations incorporate information on both inhomogeneity and dependence of the process. More recent approaches are studied in Diggle et al. (2013), that focuses on log-gaussian Cox process estimated via MCMC, Coeurjolly and Møller (2014), based on a variational procedure, and in Flaxman et al. (2017), that proposes a nonparametric approach based on Reproducing Kernel Hilbert Spaces. All these models are nonetheless inappropriate and not easily generalizable for data distributed over complex spatial domains.

Let us consider n i.i.d. observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from a Poisson counting process on Ω

with inhomogeneous intensity function γ. The likelihood of the process is

$$\prod_{i=1}^{n} \gamma(\boldsymbol{x}_i) \exp\left(\int_{\Omega} (1 - \gamma(\mathbf{u}))\, \mathrm{d}\mathbf{u}\right).$$

If we set $g(\boldsymbol{x}) = log(\gamma(\boldsymbol{x}))$ and we omits the constant term $\int_{\Omega} 1\, \mathrm{d}\mathbf{u} = |\Omega|$, we obtain the negative log-likelihood

$$-\sum_{i=1}^{n} g(\boldsymbol{x}_i) + n \int_{\Omega} \exp(g(\mathbf{u}))\, \mathrm{d}\mathbf{u}\,.$$

Finally, likewise in the case of density estimation, we can add a regularization, and consider the functional

$$-\sum_{i=1}^{n} g(\boldsymbol{x}_i) + n \int_{\Omega} \exp(g(\mathbf{u}))\, \mathrm{d}\mathbf{u} + \tilde{\lambda} \int_{\Omega} \big(\Delta g\big)^2 \tag{2.2}$$

with a positive smoothing parameter $\tilde{\lambda}$. The minimization of the functional (2.2) is equivalent to the minimization of (2.1), setting $\tilde{\lambda} = n\lambda$. We can thus tackle the minimization of (2.2) along the same lines detailed below for the density estimation problem considered in Section 2.2. Thus, our proposal also defines an innovative method for the study of inhomogeneous Poisson processes, that is able to accurately handle data observed over complex spatial domains.

2.3 Theoretical properties

In this section we formalize the minimization problem introduced in the previous sections, and we demonstrate that this estimation problem is well posed, proving the existence of a unique minimizer in an appropriate functional space.

2.3.1 Well posedness of the estimation problem

Let $L^2(\Omega)$ denote the space of square integrable functions over Ω. The Sobolev space $H^k(\Omega)$ is defined as

$$H^k(\Omega) = \big\{g \in L^2(\Omega) : D^\alpha g \in L^2(\Omega)\ \forall |\alpha| \leq k\big\}$$

and is equipped with the standard norm $\|g\|^2_{H^k(\Omega)} = \sum_{|\alpha|\leq k} \|D^\alpha g\|^2_{L^2(\Omega)}$ where the symbol $D^\alpha g$ denote the weak derivative of order α (see, e.g., Adams, 1975, for a detailed

treatment of Sobolev spaces).
Define the space

$$V = \left\{ g \in H^2(\Omega) : \frac{\partial g}{\partial \nu} = 0 \text{ on } \partial\Omega \right\}$$

where ν is the normal versor to the boundary of the domain $\partial\Omega$, and $\frac{\partial g}{\partial \nu}$ is the derivarive in the normal direction. The space V is the Sobolev space of functions that have two weak derivatives in L^2 and satisfy homogeneous Neuman boundary conditions, i.e., that have null normal derivative at the boundary of the domain. In this space, the density corresponding to the null family of the operator, i.e., when $\lambda \to +\infty$, is the uniform distribution over Ω. In the formulation of the problem of Poisson intensity estimation, outlined in Section 2.2.1, when $\lambda \to \infty$, the obtained estimates tend to an homogeneous Poisson intensity on Ω. The following Theorem states that the minimization problem is well posed in the space V.

Theorem 2.1. *The functional $L(g)$ defined in equation (2.1) has a unique minimizer in V.*

Proof. The proof is deferred to Appendix .1 □

The same result of course holds for the functional in equation (2.2), setting $\tilde{\lambda} = n\lambda$, although in the following, for simplicity of exposition, we restrict our attention to the density setting.

2.4 Estimation procedure

The minimization of the functional (2.1) is an infinite dimensional problem and its solution is not analytically available. Here we consider a discretization of such infinite dimensional problem based on finite elements (see, e.g., the textbook Ciarlet, 2002, for an introduction to finite elements). In particular, we consider a linear approximation of the function g and correspondingly of the functional (2.1). This leads to a tractable estimation procedure. Moreover, the proposed technique permits to efficiently handle data scattered over domains with complicated shapes.
The implementation of the method is based on the `R` package `fdaPDE` (Lila et al., 2016b).

2.4.1 Finite elements

First, we consider a discretization of the domain Ω using a constrained Delaunay triangulation; this is a generalization of the Delaunay triangulation (see for example Hjelle and Dæhlen, 2006) that enables the definition of the boundary of the domain,

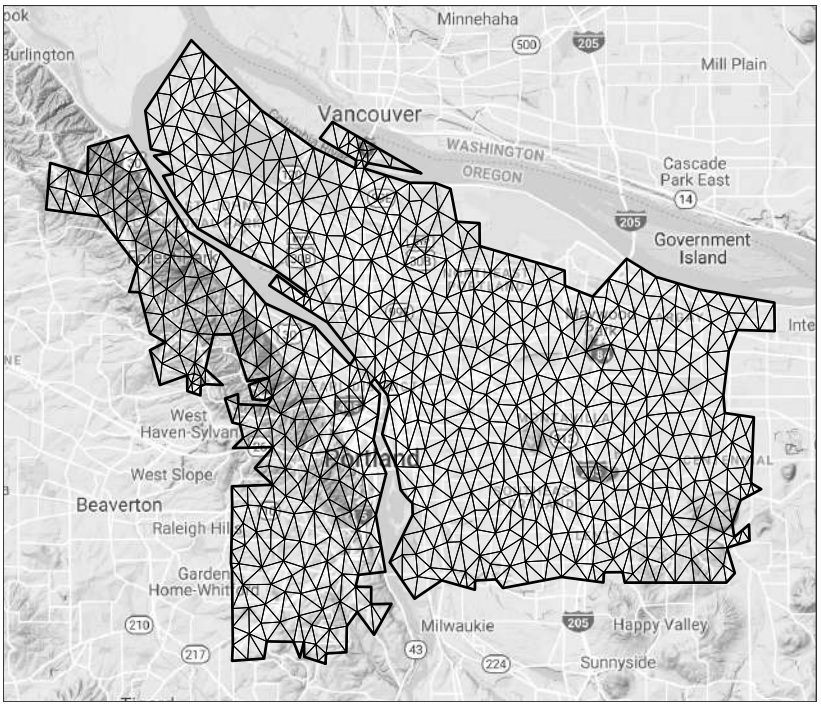

FIGURE 2.2: Mesh used for the study of motor vehicle theft in the city of Portland. The mesh represent very well the complex morphology of the domain, cut through by the Willemette river. The mesh is obtained as a constrained Delaunay triangulation using the functions in the `R` package `fdaPDE` (Lila et al., 2016b).

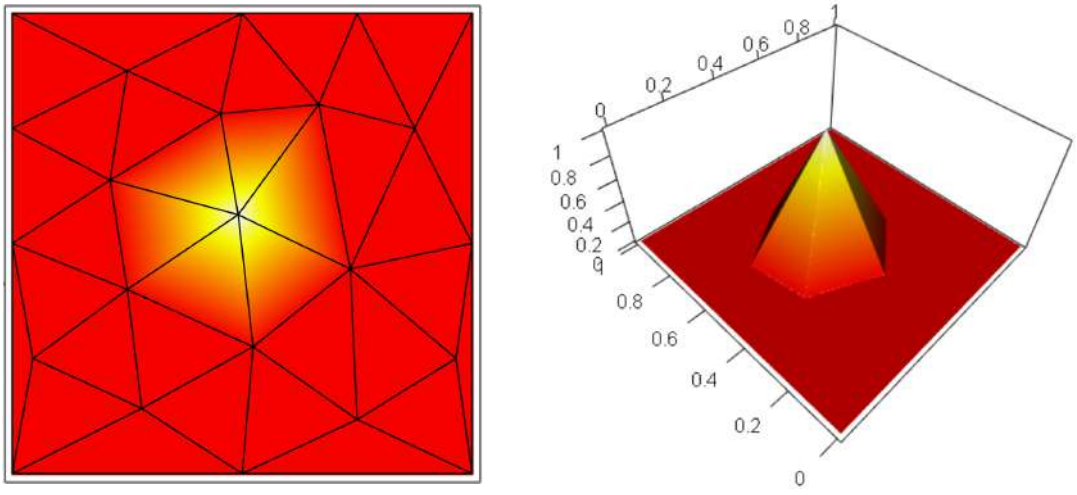

FIGURE 2.3: A linear finite element basis function on a triangulation.

forcing the required segments into the triangulation. The resulting domain is denoted by $\Omega_{\mathcal{T}}$, where $\mathcal{T}$ is the set of all the triangles. In the simulation studies and application presented in the following sections, the triangulation is constructed starting from the boundary; the triangulation is then refined according to criteria concerning maximal allowed triangle area and minimal allowed triangle angle. The `R` package `fdaPDE` (Lila

et al., 2016b) provides the functions to construct the mesh and refine it.
Figure 2.2 shows the mesh that we use for the estimation of the distribution of crimes in Portland. The mesh is able to represent very well this complex domain, accurately rendering the Willamette river, that cuts through the city, and other detailed features of the domain. It is also possible to consider data-driven meshes, that are constructed starting from constrained Delaunay triangulations of the data, or a subsample of them, and then refined according to criteria of minimal allowed triangle angle and maximal allowed triangle edge. Such data-driven meshes permit to capture strongly localized features of the density while being parsimonious, (i.e., using a limited number of mesh nodes). An example is this sense is shown in Section 2.6.
We now define the piecewise polynomial functions over $\Omega_{\mathcal{T}}$. For simplicity of exposition we present the linear case, but higher order polynomials can be used as well. To this aim, we define a system of bases. Let us denote by $\boldsymbol{\xi}_k$, $k = 1, \ldots, K$, the nodes of the mesh. In the case of linear finite elements, these nodes coincide with the vertices of the triangles. Specifically, for each node $\boldsymbol{\xi}_k$ we consider the finite element basis ψ_k, defined as the piecewise linear function that has value 1 at node $\boldsymbol{\xi}_k$ and value 0 at any other node $\boldsymbol{\xi}_\ell$, with $\ell \neq k$. Figure 2.3 shows one such finite element function.
Any function g, that is globally continuous on $\Omega_{\mathcal{T}}$ and is linear when restricted to any triangle of $\mathcal{T}$, can be obtained as an expansion of the K bases $\psi_1, \ldots, \psi_K$, i.e., $g(\cdot) = \mathbf{g}^T \boldsymbol{\psi}(\cdot)$, where $\mathbf{g} = (g_1, \ldots, g_K)^\top$ is the K-vector of coefficients of the basis expansion, and and $\boldsymbol{\psi} := (\psi_1, \ldots, \psi_K)^\top$ is the vector that packages the K finite element basis. Moreover, it turns out that the vector g of coefficients of the basis expansion coincides with the vector of evaluations of the function at the K nodes of the mesh, i.e., $\mathbf{g} = (g(\boldsymbol{\xi}_1), \ldots, g(\boldsymbol{\xi}_K))^\top$. In fact, since $\psi_k(\boldsymbol{\xi}_j) = \delta_{jk}$, where δ_{jk} is the Kronecker delta ($\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 1$ otherwise), we have that

$$g(\boldsymbol{\xi}_j) = \sum_{k=1}^{K} g_k \psi_k(\boldsymbol{\xi}_j) = \sum_{k=1}^{K} g_k \delta_{jk} = g_j.$$

The finite element space of functions has thus been constructed so that any function in such space is completely defined by the values its assumes at the K nodes.

2.4.2 Discretization of the infinite dimensional estimation problem

We now discretize the infinite dimensional estimation problem, associated with the minimization of the functional (2.1), using the finite elements introduced in Section

2.4.1.

Let Ψ be the $n \times K$ matrix having as entries the evaluations of the K finite element basis functions $\psi_1, \dots, \psi_K$ at the n data points $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, i.e.,

$$\Psi := \begin{bmatrix} \psi_1(\boldsymbol{x}_1) & \dots & \psi_K(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \psi_1(\boldsymbol{x}_n) & \dots & \psi_K(\boldsymbol{x}_n) \end{bmatrix}.$$

Moreover, let $\mathbf{1}$ denote the K-vector with entries all equal to 1. With this notation, using the piecewise linear function $g = \mathbf{g}^\top \boldsymbol{\psi}$, we can discretize the negative penalized log-likelihood that constitutes the first term in (2.1) by $-\mathbf{1}^\top \Psi \mathbf{g}$.

To discretize the second term in (2.1), i.e., $\int_\Omega \exp(g)$, we need an appropriate quadrature rule (Quarteroni et al., 2010); here in particular we use a Gaussian quadrature rule. In general, for each triangle $\tau \in \mathcal{T}$, we consider q quadrature nodes and an associated vector of quadrature weights $\mathbf{w} \in \mathbb{R}^q$, and we denote by Ψ_τ the $q \times K$ matrix having as entries the evaluations of the K basis functions $\psi_1, \dots, \psi_K$ at the q quadrature nodes in the triangle τ. The second term in (2.1) can hence be discretized as $\sum_{\tau \in \mathcal{T}} \mathbf{w}^\top \exp(\Psi_\tau \mathbf{g})$.

Finally, to approximate the third term in (2.1), i.e., the roughness penalty, we need to introduce the vectors $\boldsymbol{\psi}_{x_1} := (\partial\psi_1/\partial x_1, \dots, \partial\psi_K/\partial x_1)^\top$ and $\boldsymbol{\psi}_{x_2} := (\partial\psi_1/\partial x_2, \dots, \partial\psi_K/\partial x_2)^\top$, and $K \times K$ matrices

$$R_0 := \int_{\Omega_\mathcal{T}} (\boldsymbol{\psi}\boldsymbol{\psi}^\top) \qquad \text{and} \qquad R_1 := \int_{\Omega_\mathcal{T}} (\boldsymbol{\psi}_{x_1}\boldsymbol{\psi}_{x_1}^\top + \boldsymbol{\psi}_{x_2}\boldsymbol{\psi}_{x_2}^\top).$$

Following Ramsay (2002) and Sangalli et al. (2013), the regularization can hence be discretized by $\lambda \mathbf{g}^\top R_1 R_0^{-1} R_1 \mathbf{g}$. Such approximation only involves the first derivatives of the function $g = \mathbf{g}^\top \boldsymbol{\psi}$.

Summarizing, the negative penalized log-likelihood functional (2.1) can be discretized as

$$L_\mathcal{T}(\mathbf{g}) = -\mathbf{1}^\top \Psi \mathbf{g} + \sum_{\tau \in \mathcal{T}} \mathbf{w}^\top \exp(\Psi_\tau \mathbf{g}) + \lambda \mathbf{g}^\top R_1 R_0^{-1} R_1 \mathbf{g} \tag{2.3}$$

The minimization of the functional (2.3) can be performed using classical steepest descent approaches, such as the gradient descent algorithm, as described in Section 2.4.3. In Section 2.4.4 we propose a possible initialization for such algorithms.

2.4.3 Gradient descent algorithm

A classical steepest descent algorithm is the gradient descent, a first-order iterative optimization method. The algorithm starts with an initial guess $\mathbf{g}^0$ and takes steps

proportional to the negative of the gradient of the function at the current point. In particular, the values at step $m+1$ are computed as

$$\mathbf{g}^{m+1} \leftarrow \mathbf{g}^m - \alpha L_{\mathcal{T}}^{'}(\mathbf{g}^m)\,,$$

where α is the step of the gradient descent algorithm, and $L_{\mathcal{T}}^{'}$ is the derivative of $L_{\mathcal{T}}$ with respect to $\mathbf{g}$.
Here we consider the simplest formulation of the gradient descent method, but other algorithms, such as Nesterov accelerated gradient (Nesterov, 2018), can be implemented with simple modification of the updates. The selection of the optimal step α is a classical problem in the numerical analysis literature; we therefore refer to a classic textbook such as Lange (2013) for a thorough discussion.
The gradient descent method is proved to converge when the functional to be minimized is strictly convex, as it is in our case. However, the number of iterations needed to converge clearly depends on the goodness of the initial guess $\mathbf{g}^0$. A standard choice for the initial condition would be $\mathbf{g}^0 = 0$, that corresponds to a uniform distribution over Ω. In next session we propose a better initial value $\mathbf{g}^0$ which improves the performance in term of computational time.

2.4.4 Initialization of the gradient descent algorithm

We initialize the vector of parameters by means of a heat diffusion estimator, inspired by the work of Chaudhuri and Marron (1999). In particular, Chaudhuri and Marron (1999) propose an approach to curve estimation based on a heat diffusion process, exploiting the close relationship between heat diffusion processes and Gaussian kernels. The approach is motivated by the "scale-space" models from computer visions and the idea is to explore the whole space of solutions for increasing levels of smoothness. Botev et al. (2010) uses the same idea to define a density estimator and studies the properties of the method. This approach to density estimation, based on the heat diffusion process, gives elegant solutions in the case of univariate domains or multivariate domains with simple shapes. On the other hand, the method discussed in Botev et al. (2010) cannot account for domains with complex shapes.

To overcome this problem, differently from Chaudhuri and Marron (1999) and Botev et al. (2010), we consider a discretization of the heat diffusion process, that enables us to deal with domains with complex shapes. We stress the fact that we use such method only to compute an initial guess for the gradient descent algorithm described in Section 2.4.3.

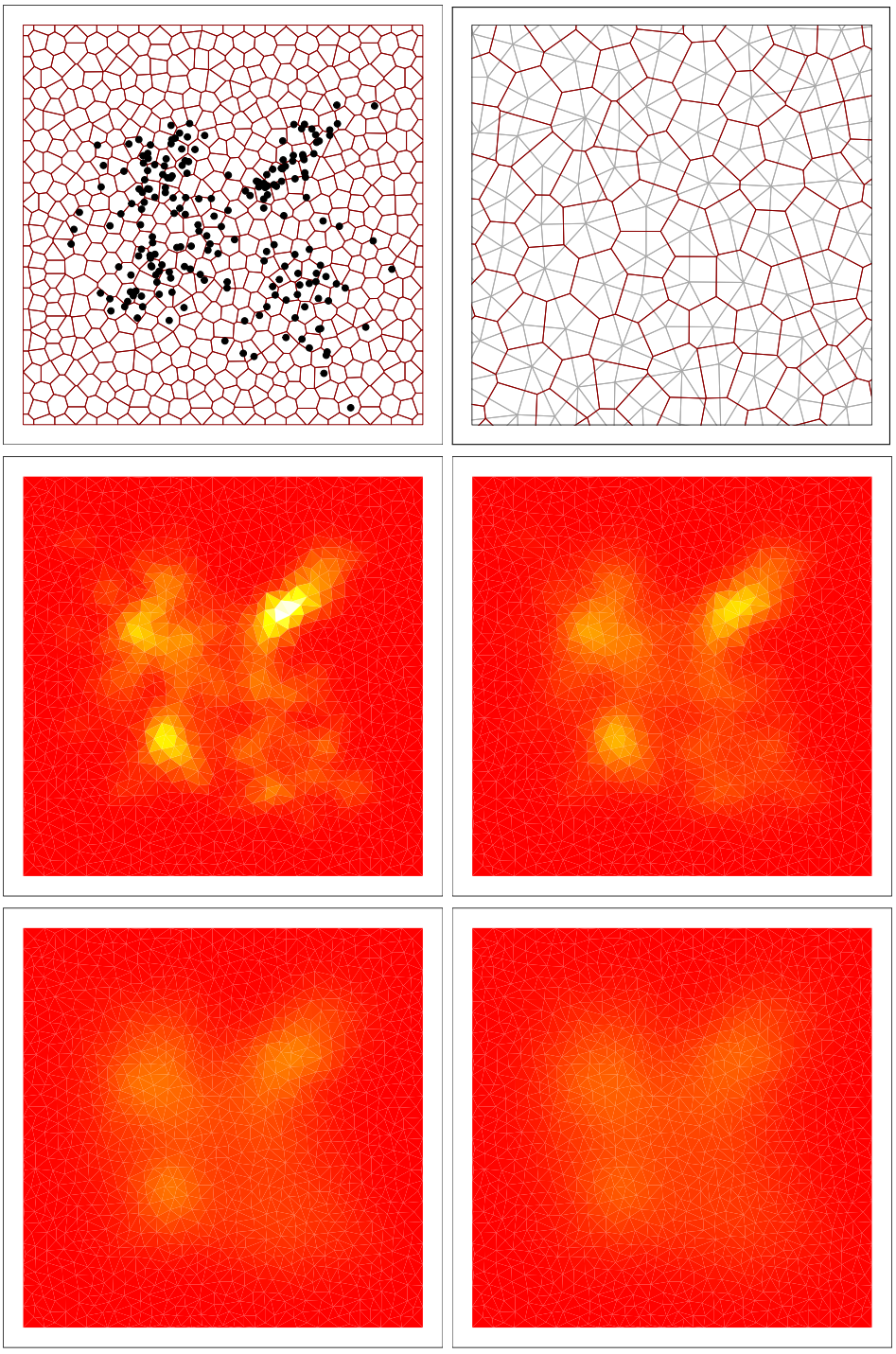

FIGURE 2.4: Top left: A sample of 200 observations from a mixture of Gaussian distributions with a Voronoi tessellation of the domain. Top right: Voronoi tessellation and the dual Delaunay triangulation. Middle and bottom rows: heat diffusion estimates as the time increases.

Let $\delta(\cdot)$ denote the Dirac measure. Given n realizations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, consider the heat

equation

$$\begin{cases} \dfrac{\partial}{\partial t}\tilde{f}(\boldsymbol{x};t) = \dfrac{1}{2}\Delta\tilde{f}(\boldsymbol{x};t) & \boldsymbol{x}\in\Omega, t>0 \\ \dfrac{\partial\tilde{f}}{\partial\nu}(\boldsymbol{x}) = 0 & \boldsymbol{x}\in\partial\Omega \\ \tilde{f}(\boldsymbol{x};0) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i) & \end{cases} \tag{2.4}$$

The initial condition of the equation, $\tilde{f}(\boldsymbol{x};0)$, is the empirical density of the data. The use of homogeneous Neumann boundary conditions (second equation of the system) ensures that, for every $t \geq 0$, the solution $\tilde{f}$ integrates to one over the domain Ω, thus being a proper density (see Botev et al., 2010). While the initial condition, the empirical density, constitutes an extremely rough solution, as t increases, the solution $\tilde{f}(\boldsymbol{x};t)$ becomes progressively more smooth, converging to a uniform density over Ω when $t \to \infty$. The main idea is that, for a certain time t, $\tilde{f}(\boldsymbol{x};t)$ provides a good initial guess for the true density f, that we can use in the gradient descent algorithm. In particular, differently from Chaudhuri and Marron (1999) and Botev et al. (2010), we solve the heat-diffusion problem (2.4) numerically, using a forward Euler integration scheme (see for example Butcher, 2016); moreover, we consider an appropriate finite element formulation. Specifically, let us first of all consider the Voronoi tesselation of the spatial domain of interest, associated with the triangulation of the domain discussed in Section 2.4.1. The triangulation and the Voronoi tessellation constitutes two dual partitions of the domain of interest. Figure 2.4 illustrates the relationship between the triangulation and the Voronoi tesselation: in the top right panel, on a squared domain, we show in gray a triangulation of the domain and in red the corresponding Voronoi tesselation. For $k = 1, \ldots, K$, we denote by R_k the k-th Voronoi tile: this is the set of all points in $\Omega_{\mathcal{T}}$ that are closer to node $\boldsymbol{\xi}_k$ of the triangulation than to any other node $\boldsymbol{\xi}_j$, with $j \neq k$, i.e.: $R_k = \{\boldsymbol{x} \in \Omega \mid d(\boldsymbol{x},\boldsymbol{\xi}_k) \leq d(\boldsymbol{x},\boldsymbol{\xi}_j) \text{ for all } j \neq k\}$, where $d(\cdot,\cdot)$ denotes the Ecuclidean distance, computed within the domain of interest (i.e., without crossing the boundaries of the domain). We hence approximate the empirical density of the data by the finite element function $\tilde{f}^0 = \tilde{\mathbf{f}}^{0\top}\boldsymbol{\psi}$ that takes the following values at the nodes:

$$\tilde{f}^0_k = \tilde{f}^0(\boldsymbol{\xi}_k) = \frac{1}{n}\sum_{i=1}^{n}\frac{|R_k|}{|\Omega|}\mathbb{I}(\boldsymbol{x}_i \in R_k) \qquad \text{for } k = 1, \ldots, K \tag{2.5}$$

where $\mathbb{I}$ is the indicator function, $|R_k|$ denotes the area of the $k-$th tile and $|\Omega|$ the area of the spatial domain Ω. The value of this function at the $k-$th node hence corresponds to the proportion of data that fall within the $k-$th tile, weighted by the relative area of the tile. With a sufficiently fine triangulation, and correspondingly small tiles, such

function provides a good approximation of the empirical density. We thus use this function to approximate the initial condition of the heat diffusion problem (2.4). We hence discretize (2.4) by finite elements in space and a forward Euler scheme in time, setting the temporal step size to Δt. This means that, starting from the initialization in equation (2.5), we compute an approximation of $\tilde{f}(\boldsymbol{x};t)$, at times $t = m\Delta t$, where $m = 1, 2, \ldots$ is the iteration index, by the finite element function $\tilde{f}^m = \tilde{\mathbf{f}}^{m\top}\boldsymbol{\psi}$, setting the following values of the functions at the nodes

$$\tilde{f}_k^{m+1} = \tilde{f}_k^m + \Delta t \frac{1}{\#(\mathcal{N}_k)} \sum_{j \in \mathcal{N}_k} (\tilde{f}_j^m - \tilde{f}_k^m), \quad k = 1, \ldots, K \ldots$$

where $\mathcal{N}_k$ is the set of nodes that are closest neighbours of $\boldsymbol{\xi}_k$ and $\#(\mathcal{N}_k)$ is its cardinality. Looking at the solutions for different time instants (i.e. for different m) we obtain a set of functions that ranges from the extremely rough sum of spikes at the observations ($m = 0$) to the uniform distribution over Ω ($m \to \infty$). Figure 2.4 illustrates this process. The top left panel shows a sample of 200 data from the mixture of Gaussian distributions using in the first simulation case, and detailed in Section 2.5.1; the same figure also displays the Voronoi tessellation of the domain. The top right panel report a zoom of the Voronoi tessellation, with the associated triangulation. The central left panel shows the corresponding approximation of the empirical density, $\tilde{f}^0$. The remaining panels shows progressively smother solutions $\tilde{f}^m$.
Among the various solutions $\tilde{f}^m$, we then use as a starting guess for the gradient descent algorithm the solution $\tilde{f}^m$, such that $g^m = \log(\tilde{f}^m)$ minimizes the functional (2.1). The maximum m is chosen large enough to have a complete diffusion of the heat flow.

2.4.5 Selection of the smoothing parameter

The selection of the smoothing parameter λ is crucial for an accurate estimation and to ensure a right balance between the bias and the variance of the estimator. We consider here a k-fold cross-validation procedure based on the L_2 norm. This norm is frequently used in literature and leads to a particularly tractable selection algorithm (Marron, 1987). The value of λ can be chosen minimizing

$$\hat{R}(\lambda) = \int (\hat{f}_\lambda^{-[k]}(\boldsymbol{x}))^2 - \frac{2}{\#(\boldsymbol{x}^{[k]})} \sum_{i \in [k]} \hat{f}_\lambda^{-[k]}(\boldsymbol{x}^{[k]}) \tag{2.6}$$

where k is the fold index, $\hat{f}_\lambda^{-[k]}(\boldsymbol{x})$ is the density estimated without the k-th fold, $\boldsymbol{x}^{[k]}$ is the subset of observations of the k-th fold and $\#(\boldsymbol{x}^{[k]})$ its cardinality. The first term

in (2.6) can be easily computed thanks to the finite element formulation. The latter consists of the evaluation of the estimated density in the left out points and its a simple linear combinations of the vertices of the mesh.

2.5 Simulation studies and applications

In this section we present some simulation studies where we compare the performance of the proposed nonparametric Density Estimator with Partial Differential regularization (DE-PDE) with Kernel Density Estimation (KDE). KDE is implemented using the `R` package `ks` (Duong, 2018), that consider anisotropic Gaussian kernels, automatically selecting the bandwidth matrix H.
We consider two simulation studies, with density functions defined over two different domains. The first domain is a simple rectangular domain, and the density we consider on this domain is not influenced by the shape of the domain. The second domain has the shape of an horseshoe, and the density we consider over this domain displays different values on the two sides of the horseshoe. In both cases, we simulate 100 samples of 200 observations each. We select the bandwidth matrix H for KDE and the smoothing parameter λ for DE-PDE using 5-fold cross-validation.

2.5.1 Simulation 1: mixture of gaussians over squared domain

We first simulate from a mixture of four Gaussian with means

$$\mu_1 = \begin{pmatrix} -2 \\ -1.5 \end{pmatrix}, \mu_2 = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \mu_3 = \begin{pmatrix} -2 \\ 1.5 \end{pmatrix}, \mu_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix},$$

and variances

$$\Sigma_1 = \begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.8 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}, \Sigma_4 = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix},$$

and mixing weights $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$. The results are shown in Figure 2.5. The top left panel represent the contour lines of the true density. The bottom panels display the means of the estimates obtained by KDE and by DE-PDE over the 100 simulation repetitions. Both estimators are able to capture the shape of the true density. The top right panel shows the boxplots of the Mean Integrated Squared Error (MISE) of the two estimator, over the 100 simulation repetition. For completeness, we also show the MISE of the solutions obtained by the heat-diffusion estimator (Heat) described in

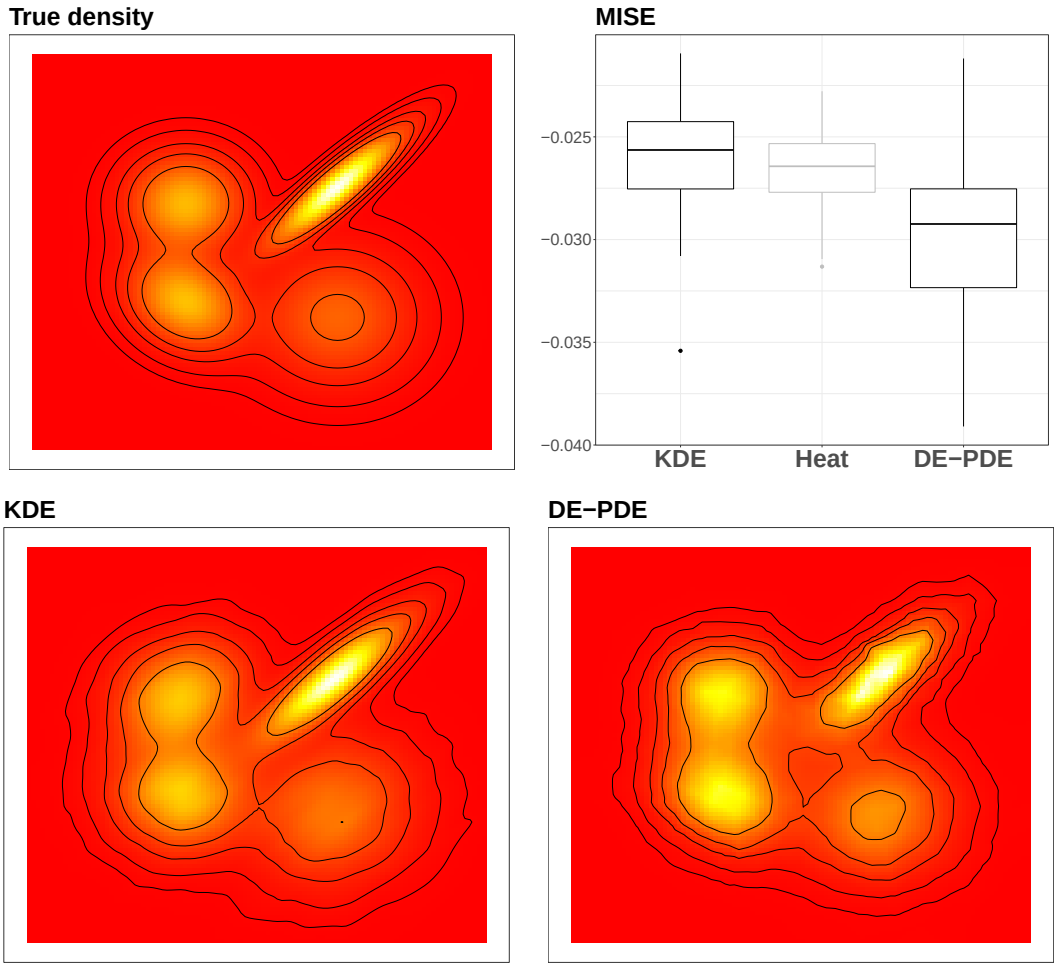

FIGURE 2.5: Simulation 1: mixture of gaussians over squared domain. Top left: true density. Top right: boxplots of the Mean Integrated Squared Error (MISE), over the 100 simulation repetitions, of the estimates provided by Kernel Density Estimation (KDE) and the proposed Density Estimation with PDE regularization (DE-PDE); for completeness, we also display the MISE of the heat diffusion estimates that we use as starting guesses for the gradient descent algorithm that provides DE-PDE estimates. Bottom left: mean of the estimates yielded by KDE over the 100 simulation repetitions. Bottom right: mean of the estimates yielded by DE-PDE over the 100 simulation repetitions.

Section 2.4.4, and that we use as starting values for the gradient descent algorithm that leads to DE-PDE estimates. The boxplots highlight a clear advantage of the proposed DE-PDE over KDE. This is probably due to the different shapes of the components of the mixtures, that make more difficult the choice of an appropriate bandwidth matrix for the KDE.

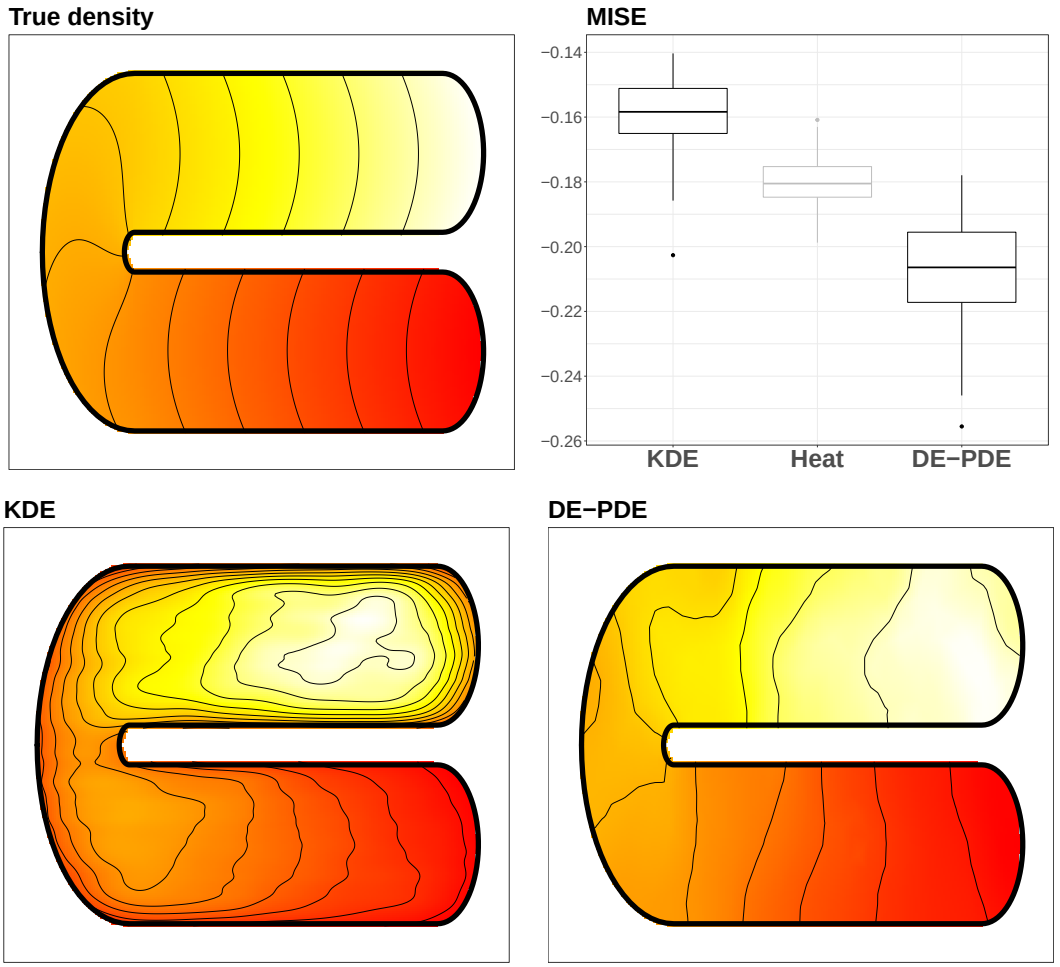

FIGURE 2.6: Simulation 2: density over horseshoe domain. Panels as in Figure 2.5. KDE cannot accurately deal with data over this non-trivial domain: it returns estimates where the high density values in the upper horseshoe arm are partly smoothed into the low density values in the bottom part of the domain, with isolines very different from those of the true density. The proposed DE-PDE estimates is not affected by this problem and does not display any strange behavior at the boundaries of the domain; instead it is able to very well capture the overall structure of the true density.

2.5.2 Simulation 2: density over horseshoe domain

We consider the horseshoe domain defined in Ramsay (2002) and considered in a number of subsequent papers. The density we use in this test is defined starting from the function defined in Wood et al. (2008) (Section 5.1), translating the function by 5 and diving by its integral in order to be a proper density. The top left panel of Figure 2.6 shows the true density. The function follows the shape of the domain, with higher values on the upper branch and lower values on the lower branch. This simulation setting presents similar difficulties as the analysis of crimes in Portland, outlined in the Introduction. In both cases, the domain is characterized by a strong concavity, that almost separates two parts of the domains, with one part displaying much higher density

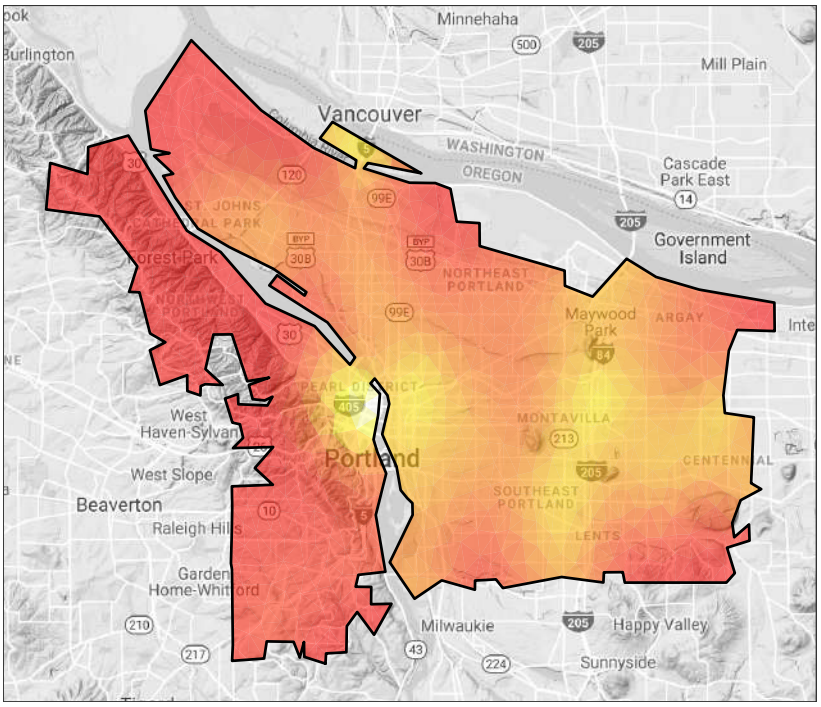

FIGURE 2.7: Estimated density for motor vehicle thefts.

values that the other part. Methods that fully rely on the Euclidean distance, without accounting for the shape of the domain, return in this case inaccurate estimates. This is evidenced by the bottom left panel of Figure 2.6, that report the means of the estimate obtained by KDE over the 100 simulation repetitions. KDE is clearly unable to identify the true structure of the density, and pours the higher density values of the top horseshoe arm into the lower density values of the bottom horseshoe arm, returning estimates that are particularly poor near the boundaries. The proposed DE-PDE instead does not suffer from this problem and is able to accurate handle the data returning estimates that capture the structure of the true density and do not display any particular problem near the boundary of the domain. This is further highlighted by the boxplots of the MISE over the 100 simulation replicates: the boxplots show the neat superiority of DE-PDE over KDE.

2.6 Portland crimes

We consider the problem of estimating the crimes distribution in the city of Montreal. The data come from NIJ "Real-Time Crime Forecasting Challenge" and consists of calls-for-service positions from the Portland Police Bureau. Wilhelm and Sangalli (2016) also offer a study of crime data over the city of Portland, but they aggregate crimes

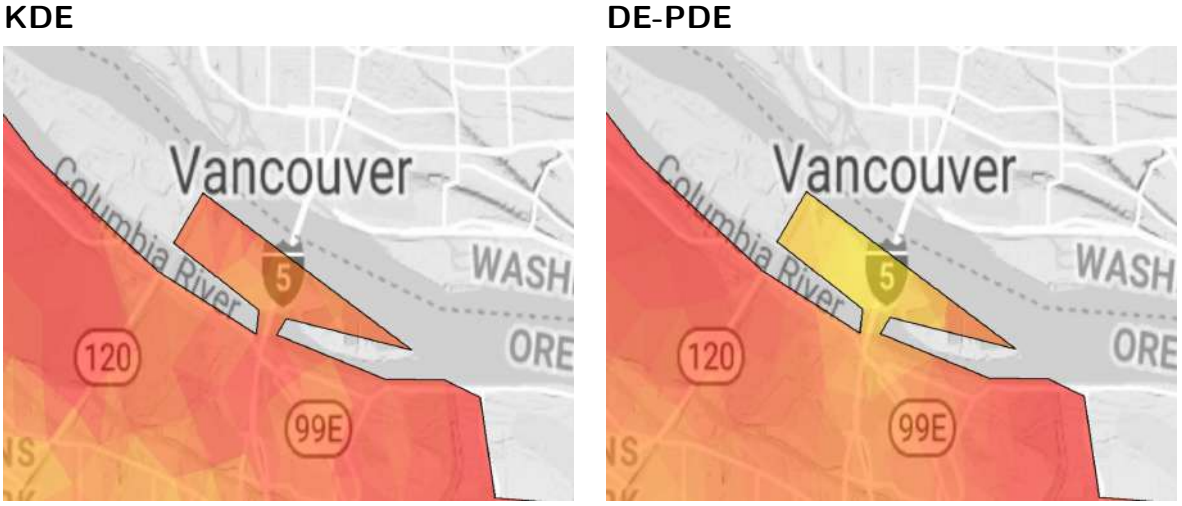

FIGURE 2.8: Enlargement of the estimates provide by KDE (left) and the proposed DE-PDE (right) on Hayden Island. The figure shows that KDE estimator is not able to comply with the complex structure of the domain: though a very high number of motor vehicle theft is observed in the island, while almost no theft is observed in the nearby inland part of the municipality, KDE return an estimate that is everywhere low. Instead, the proposed DE-PDE estimator can efficiently deal with the complex morphology of the domain, returning an estimate of the theft density that is much higher over the Hayden Island.

per district, and consider a generalized linear model to analyze crime counts over the various municipality districts. Figure 2.1 shows the location of motor vehicle theft over the municipality, in the year 2012. Figure 2.3 shows the municipality of Portland, along with a Delaunay triangulation based on 788 nodes. Note that two areas are not part of the domain of interest: the airport, in the northern part of the city, and the western part of Hayden Island, in the north, toward Vancouver. As already commented in the Introduction, the frequency of occurrences of motor vehicle thefts varies significantly over the various part of the municipality; moreover, the complex morphology of the city clearly influences the phenomenon under study. For instance, rather different theft numbers are observed on the two sides of the river. In northern part of the city, a much higher occurrence of vehicle thefts is observed on the east side of the river; the same can be said in the southern part of the city. In the city center instead, more occurrences are present on the west side of the river. A similar situation applies for the Hayden Island, in the north toward Vancouver, where there are clearly more vehicle thefts that in the inland nearby part of the municipality. In general, the phenomenon is not smooth across the river, that acts as a physical barrier. Figure 2.7 shows the estimate of the vehicle theft density obtained by the proposed DE-PDE method. The estimate accurately complies with the shape of the domain and is able to capture localized features. The two main distribution masses are concentrated in the city center and in the Lloyd district, a

primarily commercial neighborhood in the North and Northeast sections of the city. In is also interesting to note the high density region on the eastern part of the city, along the War Veterans Memorial freeway, a main highway that serves the Portland-Vancouver metropolitan area and passes near three of the largest shopping centers of the city. All these areas have huge amounts of parking lots, with cars parked for long periods of time. It is interesting to note the high concentration area in the Eastern part of Hayden Island. This part of the island, named Jantzen Beach, has highly developed retail areas near the freeway, hotels, offices, manufactured home communities, and condominium complexes. Despite the complexity of the domain, our estimator is able to identify the high density region on the island, without interfering with the estimation on the opposite side of the river, where almost no observation are present. See also Figure 2.8 that shows two enlargements of this picture over Hayden Island.

Figure 2.9 instead considers crimes related to prostitution (data corresponding to year 2012). The top left panel of the figure shows the crime locations, that are concentrated along the Northest 82nd Avenue. This is a major arterial on the Eastside, that has long had a reputation as a hub for prostitution and other aspects of Portland's sex industry. The central right and bottom right panels of the same figure show the DE-PDE estimate of the crime density obtained considering a regular fine triangulation with around 3000 nodes and selecting the smoothing parameter by cross validation. These figures highlight how accurately the proposed method captures the density mass concentrated around a segment that corresponds to the Northest 82nd Avenue. The proposed estimator is flexible enough to detect a low dimensional structure of the underlying density, a ridge, without oversmoothing it. The top right panel of the same figure displays a coarse data-drive triangulation, with 612 nodes: this is a constrained Delaunay triangulation of a random subsample of the crime locations, refined setting a minimum value for the triangle angles. The central and bottom right panels shows the corresponding DE-PDE estimate, obtained setting the smoothing parameter by cross validation. This estimate, while being more parsimonious and requiring the estimation of a smaller number of parameters, is nevertheless able to accurately represent the highly anisotropic data density. The KDE, on the contrary, returns a very rough estimate, with many spikes, being unable to select a proper bandwidth matrix that could represent the ridge.

2.7 Future research

The proposed density estimation method can be extended in various directions.
A first fascinating direction goes toward higher dimensional and non-euclidean domains.

These include two-dimensional curved domains with non-trivial geometries, and three-dimensional domains with complex boundaries. Data observed over these complex domains are common in modern applications (see, e.g., (Niu et al., 2019)). In the neurosciences, for instance, data are measured over a domain characterized by a formidably complicated morphology, the brain. Brain imaging studies can either focus on the cerebral cortex, a two dimensional curved domain with an highly convoluted nature (Lila et al., 2016a; Chung et al., 2016), or consider the whole organ, a three-dimensional domain with highly complex internal and external boundaries. In the geo-sciences, data are commonly acquired over two-dimensional curved domains or three-dimensional domains with complex orography. In astronomy, such as in cosmic web reconstruction (Chen et al., 2015), flexible density estimators are needed to identify and characterize intrinsic lower dimensional structures, such as smooth manifolds. Density estimation over complicated multidimensional domains requires flexible methods able to overcome the classical concept of Euclidean distance. Some proposals generalize the kernel density estimation to Riemaniann manifolds, using the concept of exponential map to solve the problem (see, e.g., Kim and Park, 2013; Berry and Sauer, 2017). In our setting, the flexible formulation of DE-PDE in terms of finite elements enables the extension to curved two-dimensional domains and to complex three-dimensional domains. In particular, we can here resort respectively to surface finite elements, likewise in (Lila et al., 2016a), and to volumetric finite elements.

Another interesting direction of research concerns the development of time-dependent density estimation methods. The modeling of densities over time permits the understanding of the evolution of underlying processes generating the data. However, very little attention has so far been devoted to the development of such models, especially in more than one dimension (see, e.g., Gervini (2019) and references therein for some first proposal in this regard). The proposed DE-PDE method could be generalized to space-time point data. This can be done by considering two regularizations, one in time and one in space, or alternatively a unique regularization involving a time-dependent differential operator, in analogy to the spatio-temporal regression methods presented in Bernardi et al. (2017) and Arnone et al. (2019).

An intriguing alternative to the described framework would consist in exploring the proposed approach from a bayesian perspective. The considered penalization could in fact be interpreted as a Gaussian prior over a graph, the triangulation. This may lead to interesting considerations in terms of random processes, especially in the case of Poisson intensity estimation.

Finally, as commented in Section 2.2, we could consider regularizing terms involving

more complex differential operators and partial differential equations, similarly to what done in Azzimonti et al. (2014) and Arnone et al. (2019) in the context of spatial regression. This possibility would enable the inclusion in the density estimator of problem-specific information concerning the physics of the process generating the data.

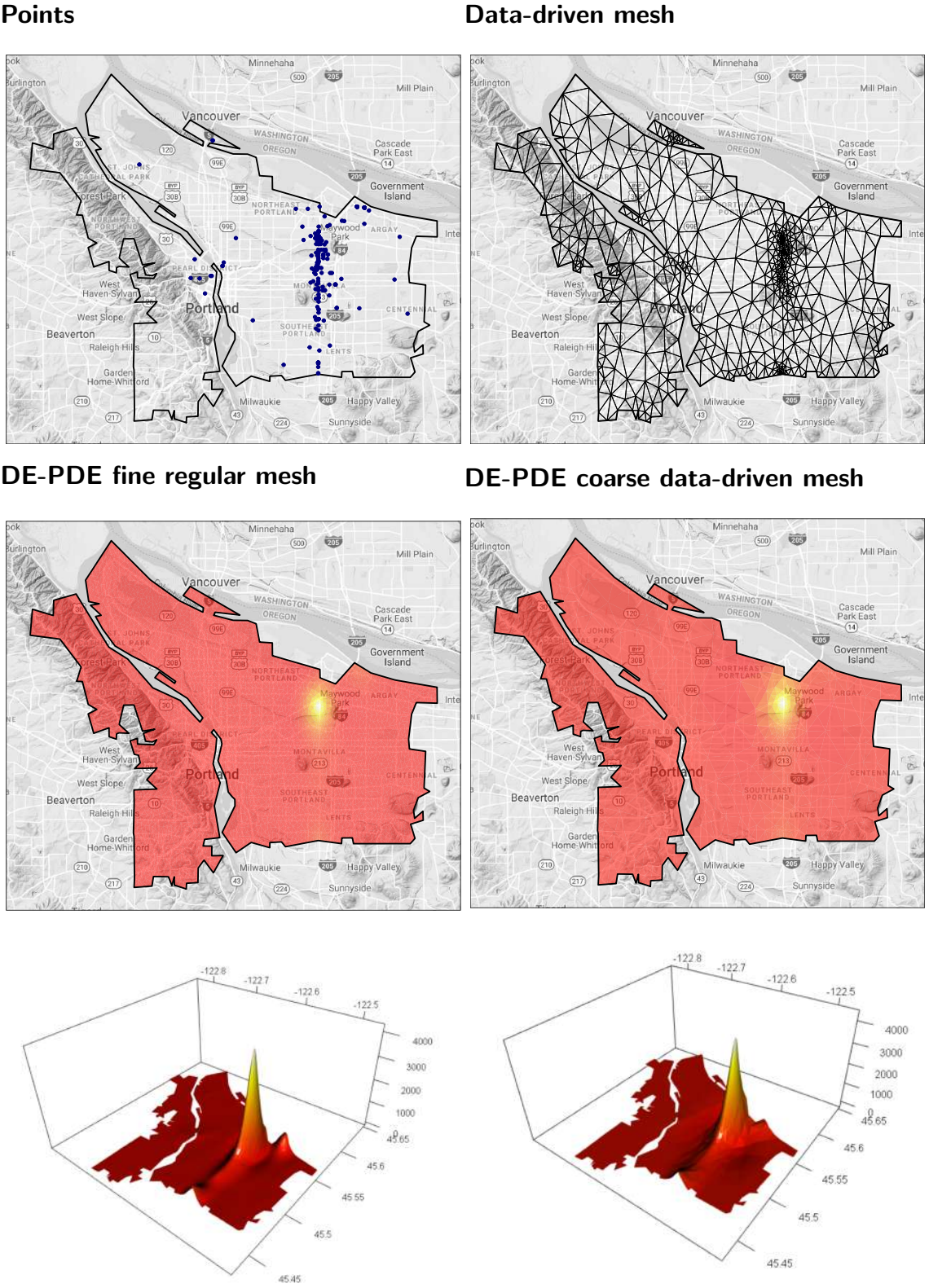

FIGURE 2.9: Prostitution in Portland. Top left: locations of prostitution related crimes, in the year 2012. Top right: coarse data-driven mesh. Central and bottom panels: DE-PDE estimated densities, using a fine regular mesh with about 3000 nodes (left) and using the coarse data-driven mesh with about 600 nodes (right) . These images highlight how accurately the proposed method captures the density mass concentrated along Northest 82nd Avenue, that appears as a neat ridge in the three-dimensional visualization.

Chapter 3

Two samples hypothesis testing

3.1 Introduction

In this second work, we study nonparametric hypothesis two samples testing procedures. In the same setting of the previous chapter, we focus on densities defined over complex domains. As an example, we use data about the criminality in the city of Portland. Our interest is to study if the distribution of a particular crime has changed over the years. We study the problem from a permutational perspective, defining different statistics using nonparametric combinations of partial tests on appropriate partitions of the domain. This allows us to redefine the problem in terms of a high dimensional multinomial test. We then study the theoretical properties of the proposed tests, in particular focusing on unbiasedness and consistency.

The problem of two samples hypothesis testing can be considered a fairly old subject. Some examples among the many nonparametric tests are the Kolmogorov-Smirnoff test (Kolmogorov (1933); Smirnov (1939)) and the number-of-runs test (Wald and Wolfowitz, 1940). On the multivariate setting, a classic example is the works Hotelling (1951), that describes the t-test for multivariate Gaussian distribution. The subject has nonetheless prompted a recent attraction in the statistical literature, specially related to the study of high dimensional multinomials. One of the first works that consider multinomial data in the high-dimensional case is by Fienberg and Holland (1973). Here the authors study the case where the number of cells is large while the number of observations per cell is moderate. A more recent work by Arias-Castro et al. (2018) shows the connection between high dimensional multinomials and the two sample testing on continuous distributions. Generalizing a work of Ingster (1987), they study the minimax properties of the histogram test, under suitable smoothness assumptions, for both goodness of fit and two samples tests. A recent review on this topic is provided in Balakrishnan and

Wasserman (2018), where they compare various testing procedures based on different metrics.

Here we move in the same direction, defining the test as an high dimensional multinomial test. The main difficulty is that in a typical histogram testing framework we consider counts on a regular grid. The problem we are facing here is slightly more complex, because we are dealing with densities defined over a general domain Ω, possibly with a complex shape. It is therefore necessary to define an appropriate partition of the domain of under study. This opens the way to permutation procedures such as the nonparametric combination (NPC), defined in Pesarin and Salmaso (2010). We also evaluate various combining methods that have been proposed for the two samples problem, and identify those that provide the best Type I error control and power across a range of situations.

The chapter develops as follows: in Section 2 we introduce the proposed methodology; in Section 3 we study the theoretical properties; in Section 4 we present some simulation studies and in Section 5 an application to the Portland crime data.

3.2 Methodology

Suppose we have observations $X_1, \dots, X_{n_1}$ from some distribution F and $Y_1, \dots, Y_{n_2}$ from some distribution G. Let f and g be Lebesgue measurable densities corresponding to the distributions F and G, respectively. Our interest is to test whether these two groups of observations have been drawn from the same distribution or not. The formalization of this hypothesis takes the form

$$H_0 : F = G \quad \text{versus} \quad H_1 : \delta(F, G) \geq \epsilon, \tag{3.1}$$

where $d(\cdot, \cdot)$ is some specified distance. In the case of continuous distributions, various distances have been used in the literature. Some examples are the L_1 and the L_2 distances, the Hellinger and the Kullback-Leibler distances. The choice of the distance is far from universal and can lead to more or less interpretable tests. Here we focus on the total variation distance (TV), defined as

$$\delta(F, G) = \sup_A |F(A) - G(A)| .$$

This distance is closely related to the ℓ_1 norm when the set is countable, in particular

$$TV(F, G) = \frac{1}{2}\|f - g\|_1.$$

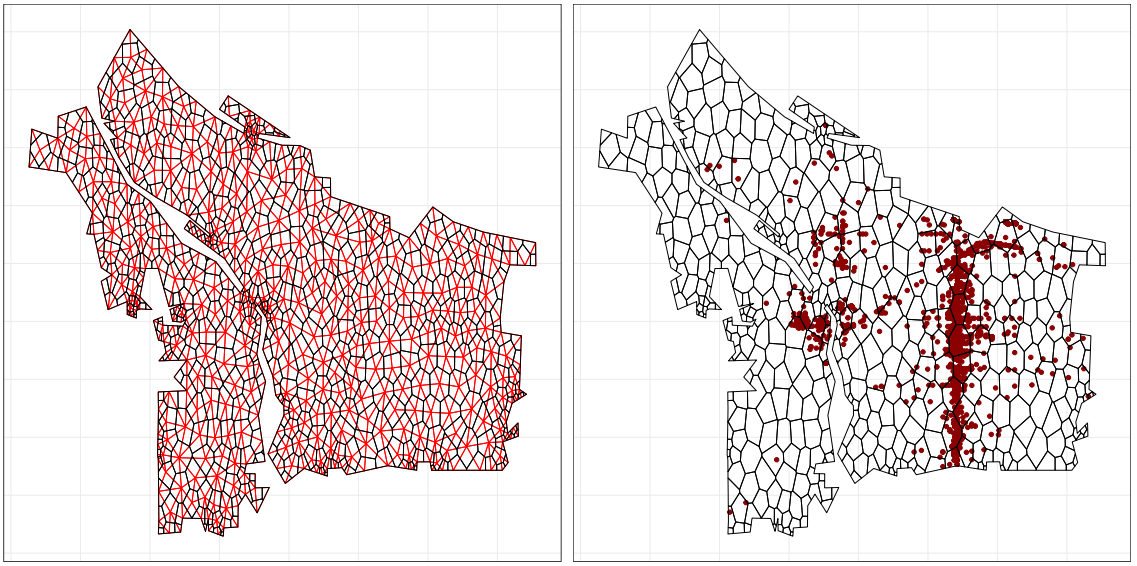

FIGURE 3.1: On the left, the Delaunay triangulation in black and its dual graph, the Voronoi tessellation in red. On the right, the crimes related to human trafficking in the city of Portland, together with the tessellation of the domain.

In the context of probability distribution theory, the TV distance has a clear interpretation. It represent the largest possible difference between the probabilities that the two probability distributions can assign to the same event. This distance is also invariant to natural transformation (Devroye and Györfi, 1985).

3.2.1 Definition of the test statistic

The test in (3.1) is defined in a fully nonparametric fashion, with no specific directional alternatives. In these cases, a possible solution is to consider test statistics based on appropriate partition of the domain under study. The most common example being the widely used histogram test (see for example Ingster (1987) and Arias-Castro et al. (2018)). This type of test is particularly appealing in a variety of applied problem. It is fairly easy to define and simple to use, and it has many interesting theoretical properties. In this work we consider a similar approach. In particular, we redefine (3.1) as a high dimensional multinomial test.

Here we focus our attention to densities defined over a complex domain $\Omega \in \mathbb{R}^2$, possibly with boundaries. Given the non regular shape of the domain, it is reasonable to use a Voronoi tessellation as a possible partition. Let $\mathcal{P} = \{D_1, \dots, D_K \subset \Omega\}$ be a partition of the domain into disjoint sub-regions. Each Voronoi cell D_k is defined as

$$D_k = \{x \in X \mid d(x, c_k) \leq d(x, c_j) \text{ for all } j \neq k\}$$

where the points $(c_1, \dots, c_k)$ represent the centers of each cell. In this case we consider as distance $d(\cdot, \cdot)$ the classic Euclidean distance. Note that the cell satisfies

$$\Omega = \bigcup_{k=1}^{K} D_k \quad \text{and} \quad D_i \cap D_j = \emptyset, \quad \forall\, i, j = 1, \dots, K,\ i \neq j.$$

In figure 3.1 we see the Voronoi tessellation, in black, of the city of Portland. The Voronoi graph is the dual graph of the Delaunay triangulation, in red, defined in the previous chapter for the estimation of the density.
Having defined a sensible partition, we define the quantized versions of f and g as $p = (p_1, \dots, p_K)$ and $q = (q_1, \dots, q_K)$, with

$$p_k = \frac{|D_k|}{|\Omega|} \int_{D_k} f(x)\mathrm{d}x \quad q_k = \frac{|D_k|}{|\Omega|} \int_{D_k} g(x)\mathrm{d}x.$$

The two vectors p and q can be seen as probability vectors indexed by K, the number of tiles in the defined tessellation. In particular, the two vectors defines multinomial probability distributions on the partitioned domain. Note that the methodology we develop is generalizible to higher dimension by considering different partitions.
Given the two vectors of observations $X = (X_1, \dots, X_{n_1})$ and $Y = (Y_1, \dots, Y_{n_2})$, we define on each partition the simple statistics

$$S_k(X) = \frac{|D_k|}{n_1} \sum_{i=1}^{n_1} \mathbb{I}(X_i \in D_k) \quad \text{and} \quad S_k(Y) = \frac{|D_k|}{n_2} \sum_{i=1}^{n_2} \mathbb{I}(Y_i \in D_k).$$

Unlike the classical histogram test, here the tiles does not have the same area. The term $|D_k| = \int_{D_k} \mathbf{1}$ is therefore needed to weight for the dimension of each tile. In the case of Portland, this correspond to count the number of observations belonging to each tiles, as in figure 3.1.

3.2.2 Nonparametric combination

We now define various test statistics using the concept of nonparametric combination (NPC), defined in Pesarin and Salmaso (2010). We specify the hypothesis (3.1) as

$$H_0 : \bigcap_{k=1}^{K} p_k = q_k \quad \text{vs} \quad H_1 : \bigcup_{k=1}^{K} p_k \neq q_k.$$

The NPC is combined with a permutation approach, that gives many advantages. The global properties of the test are easy to derive the global properties using the properties

of the partial tests. Moreover, it enables the researcher to obtain not only a global p-value, like in traditional tests, but also a p-value for each of the defined aspects or domains.

In the following we indicate by T_k^0 and T_k^* the observed and permutation values of T_k. The assumptions necessary for the nonparametric combination of the set of partial tests T_k are:

(a) All permutation partial tests T_k^* are marginally unbiased and significant for large values, so that they are larger in H_1 that in H_0.

(b) All permutation partial tests T_k^* are consistent, that is,

$$\lim_{n\to\infty} \mathbb{P}(T_k^* \geq T_{k\alpha}^0 | X, Y, H_{1k}) \to 1, \quad \forall \alpha > 0, k = 1, \ldots, K,$$

where $T_{k\alpha} < \infty$ is the critical value of T_k at level α. In order to obtain global traditional consistency it suffices that at least one partial test is consistent Pesarin and Salmaso (2010).

Here we only define the test statistics, leaving to the next section the study of the unbiasedness and the consistency. We use different combination of $S(\cdot)$ to define tests based on different metrics:

1. ℓ_1:
$$T_{\ell_1} = \sum_{i=1}^{K} |S_k(X) - S_k(Y)|,$$

2. ℓ_2:
$$T_{\ell_2} = \sum_{i=1}^{K} (S_k(X) - S_k(Y))^2,$$

3. Max:
$$T_\infty = \max_k |S_k(X) - S_k(Y)|,$$

4. Log-Ratio:
$$T_{\text{LR}} = \sum_{i=1}^{K} \log\left(\frac{S_k(X)}{S_k(Y)}\right),$$

5. Centered χ^2:
$$T_{\chi^2} = \sum_{i=1}^{K} \frac{(\tilde{n}S_k(X) - \tilde{n}S_k(Y))^2 - (\tilde{n}^2 S_k(X) + \tilde{n}^2 S_k(Y))}{\tilde{n}^2 S_k(X) + \tilde{n}^2 S_k(Y)},$$

where $\tilde{n} = n_1 n_2$.

These tests are among the most widely used in the literature and may lead to optimality in different scenarios. In particular, the ℓ_1 and the χ^2 seem to be quite promising. The work of Arias-Castro et al. (2018) shows the minimaxity of the χ^2 histogram test under fairly general smoothness conditions. This result generalizes the work of Ingster (1987) in the univariate case. A more recent work by Chan et al. (2014) proposes the centered version of the χ^2 and studies its optimality. On the other hand, the ℓ_1 test appears to be a good contender in the case of unbalanced sample sizes (Balakrishnan and Wasserman, 2018). In the next section we study the theoretical properties of these tests and then propose some simulation studies to compare their performances.

3.3 Theoretical properties

Let us first introduce the partial test

$$T_k(X,Y) = S_k(X) - S_k(Y) = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{I}(x_i \in D_k) - \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbb{I}(y_i \in D_k). \tag{3.2}$$

The partial test T_k has to verify the assumptions (a) and (b), necessary for the nonparametric combination. We first consider the unbiasedness of the considered test statistic with respect to permutations.

Proposition 3.1. *Suppose that the data X and Y are exchangeable under the null hypothesis. Then test T_k defined in* (3.2) *is conditionally and unconditionally unbiased under permutation. The T_k is also an exact test.*

Proof. The statistic S is defined as the count of observations that belong to that cell. The function S is therefore symmetric, that is, invariant with respect to rearrangements of the observations. It is also monotonic non-increasing, that is, $S(X+\delta) \geq S(X)$ for any observation X and any $\delta \geq 0$. This last property ensures that large values of T give evidence against H_0. It follows from Theorem 1 and Theorem 2 of Pesarin and Salmaso (2010) that T_k is conditionally and unconditionally unbiased, and also exact. A similar argument can be used to prove the unbiasedness for $\log S$, used in the log-ratio test. □

We now consider the consistency of the partial tests T_k.

Proposition 3.2. *Suppose that the data $X = (X_1, \ldots, X_{n_1})$ and $Y = (Y_1, \ldots, Y_{n_2})$ are such that $n_1/n_2 \to \lambda$ as $n_1, n_2 \to \infty$. Then test T_k defined in* (3.2) *is consistent under permutation.*

Proof. The study of the consistency of each partial test depends on the asymptotic distribution of the test statistic involved. To simplify the notation, lat us denote $T_k = T$ and define

$$Z = (Z_1, \ldots, Z_n) = (\mathbb{I}(x_1 \in D_k), \ldots, \mathbb{I}(x_{n_1} \in D_k), \mathbb{I}(y_1 \in D_k), \ldots, \mathbb{I}(y_{n_1} \in D_k)),$$

where $n = n_1 + n_2$. Let also $(\pi(1), \ldots, \pi(n))$ and $(\pi^{'}(1), \ldots, \pi^{'}(n))$ be independent random permutations of $1, \ldots, n$ and independent from Z. We now define the vector

$$(T_n^*, T_n^{*'}) = \left(\sqrt{n_1} \sum_{i=1}^{n} Z_i W_i, \sqrt{n_1} \sum_{i=1}^{n} Z_i W_i^{'} \right),$$

where

$$W_i = \begin{cases} 1 & \pi(i) \leq n_1 \\ -\frac{n_1}{n_2} & \pi(i) > n_1 \end{cases},$$

and $W_i^{'}$ is defined in the same way with $\pi^{'}$. The subscript n underline the dependence on the sample size. It is easy to see that $\mathbb{E}(W_i) = \mathbb{E}(W_i Z_i) = 0$. We also have

$$\mathrm{Var}(T_n^*) = \frac{n_1}{n_2} p_k(1 - p_k) + q_k(1 - q_k),$$

and

$$\mathrm{Cov}(T_n^*, T_n^{*'}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}(W_i W_j^{'} Z_i Z_j) = 0,$$

using the independence of W_i and $W_i^{'}$. We therefore have the convergence

$$(T_n^*, T_n^{'*}) \xrightarrow{d} (T^*, T^{'*}),$$

where $(T^*, T^{'*})$ is a bivariate normal with same marginal distribution, having mean 0 and variance $\lambda p_k(1-p_k)+(1-q_k)$. The convergence for the case of the log-ratio test can be obtained in the same way, using the difference of the logarithmic transformations.
In the case of T_k^0, we have the convergence

$$(T_n^0, T_n^{'0}) \xrightarrow{d} (T^0, T^{'0}),$$

where $(T, T^{'})$ is a bivariate normal with same marginal distribution, having mean 0 and variance $p_k(1 - p_k) + \lambda q_k(1 - q_k)$.
Consider the situations where $Z = Z^{n()}$ is an increasing sequence of i.i.d observations. Let us denote by $T_\alpha^0(Z^{(n)})$ the critical value at level α of the statistic T^0.We now have

to verify assumption (b), that is

$$\lim_{n\to\infty} \mathbb{P}(T^*(Z^{(n)}) \geq T^0_\alpha(Z^{(n)})|Z^{(n)}, H_1) \to 1, \quad \forall \alpha > 0, k = 1, \ldots, K.$$

As $n \to \infty$, the value $T^0_\alpha(Z^{(n)})$ converges to the quantile T^0_α of a normal distribution $\mathcal{N}(0, p_k(1-p_k) + \lambda q_k(1-q_k))$. We therefore obtain

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n_1} \left(\sum_{i=1}^{N} Z_i W_i \right) \geq T^0_\alpha \middle| Z^{(n)}, H_1 \right) \to 1.$$

Note that, although the distribution of T^0 and T^* have slightly different variances in the case of unbalanced samples sizes, we still have the asymptotic consistency. In the case on balanced sample sizes, we may be able to obtain stronger properties, such as a test that is locally asymptotically uniformly most powerful (Lehmann and Romano, 2006). □

We now consider the consistency of global tests.

Proposition 3.3. *The global tests (1) to (5) are conditionally and unconditionally unbiased and consistent under permutations.*

Proof. The combining functions (1) to (5) are admissible since they define convex acceptance regions (see Section 4.2.6 of Pesarin and Salmaso (2010) for a characterization of admissible combining functions). This property conbined by the two previous results gives the unbiasedness and the consistency of the globals tests. □

3.4 Simulations

We consider different simulation scenarios in order to compare the performances of the proposed tests. In the first case we consider a distribution with center of mass that shift away under the alternative. In the second case, we consider a distribution with center of mass that shrink toward its center. In the third case we consider a uniform distribution over a increasingly smaller support. In the fourth and last case we consider a uniform distribution against a spike distribution with increasingly smaller support. In each case we performed 1000 simulations and the critical value is provided using a permutation procedure with 500 random permutations. For each scenario we consider three different settings, with different ratios between n, the number of observations, and K, the number of cells of the multinomial.

3.4.1 n bigger than K

In the first setting $n > K$, that is the number of observation is larger that the number of cells of the multinomial. In particular we consider the case where $n_1 = n_2 = 1000$ while $K = 400$. Remind that $n = n_1 + n_2$, the sum of the number of samples in each group. This can be considered as a standard asymptotic setting. In figure 3.2 are presented the power curves in the case where the two samples are unbalanced. In each scenario the centered Chi-square has better performances with respect to all the other tests. The ℓ_1 test is almost as good as the Chi-square, while the ℓ_2 has slightly less power. The log-ratio test performs poorly in the last two cases, where the distributions are supported over a smaller region of the domain and a high number of cells have low probability mass.

3.4.2 n smaller than K

In the second setting, we consider a reverse situation, where $n < K$. In this setting the curse of dimensionality may affect the power of some of . In figure 3.3 are presented the power curves for each test in each case, where the values are an average over the 1000 simulations. In particular we consider the case where $n_1 = n_2 = 100$ while $K = 400$. In this scenario, the centered Chi-square and the ℓ_1 still remain the more powerful tests overall. The ℓ_2 and the log-ratio have slightly worse performances. The max test shows low power when the distributions are more diffuse and when a shift in the center of mass is present. Nonetheless, it regain power in presence of more spiky distributions.

3.4.3 Unbalanced case

In the last setting, we consider an unbalanced situation, where the sample sizes are $n_1 = 4000$ and $n_2 = 400$. This seems to represents a more delicate situation. The Chi-square test, although minimax optimal in the balanced case, has somehow worse performances. This result is in line with the work of Balakrishnan and Wasserman (2018). The ℓ_1 and ℓ_2 have generally the best behavior, with an ℓ_1 predominance in the first two cases where the support of the distributions is more concentrated and most of the multinomial cells have zero counts. The log-ratio has acceptable power is some cases but it seems problematic when the distribution have small support. The max test remains generally the worst of all the proposals.

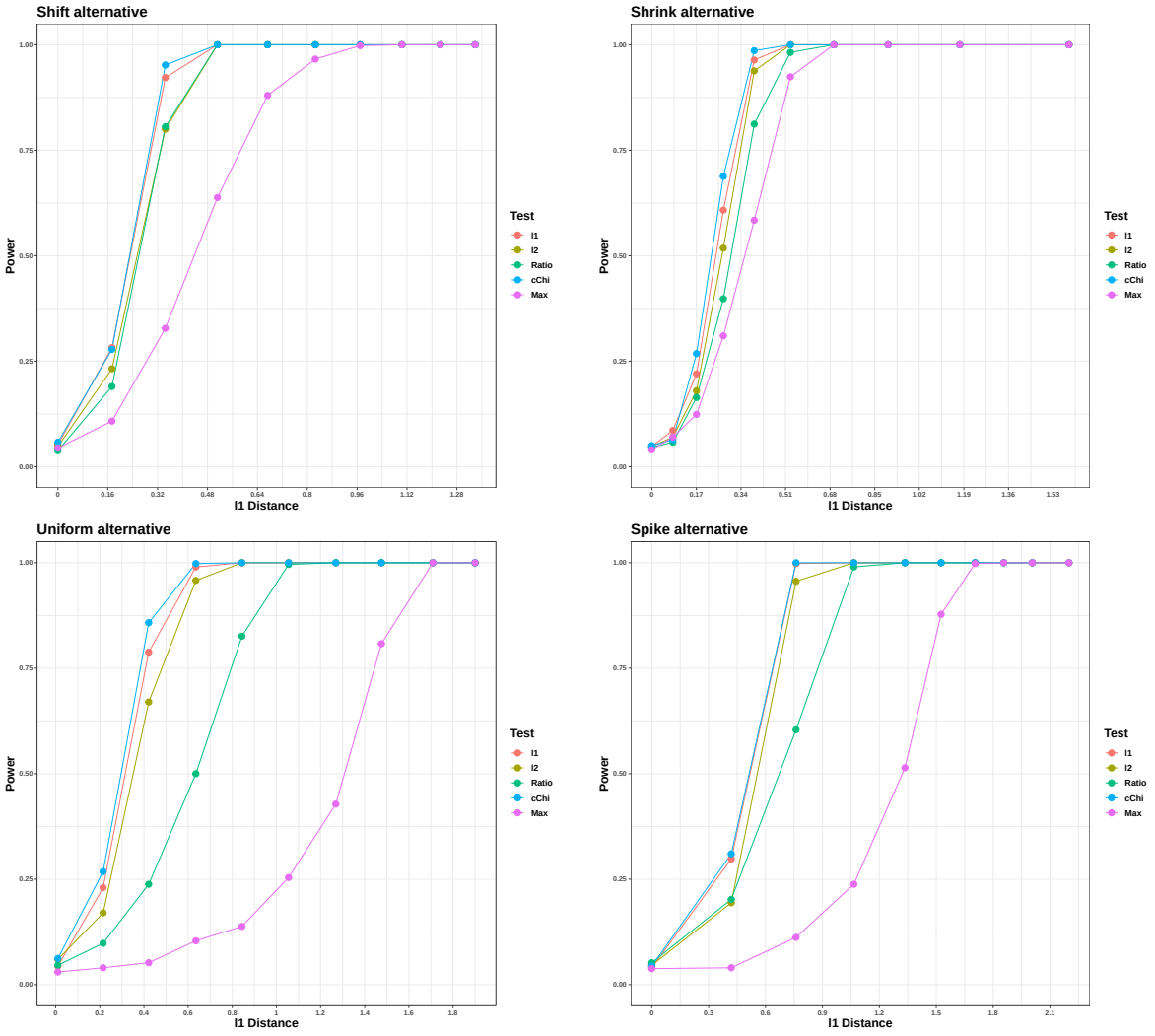

FIGURE 3.2: A comparison between the five tests proposed in four different scenarios. The power of the tests are plotted against the ℓ_1 distance between F and G. The sample sizes are $n_1 = n_2 = 1000$ while the number of cells of the multinomial is $K = 400$. Each point in the graph is an average over 1000 simulations.

3.5 Application

In the previous chapter, we applied the proposed density estimation method to the crime data in the city of Portland, Oregon. It is clear from 2.9 that the East-side of the city seems to be a fairly critical area. The problem of human trafficking was particularly evident in the early 2000's. Since then, the Police department and the State of Oregon started various campaigns to reduce or at least control this problem.

> In 2013 East Precinct began an innovative partnership with the Multnomah County District Attorney's Office to respond to the increase of crime, particularly violent crimes related to gang activity and drug trafficking, in east

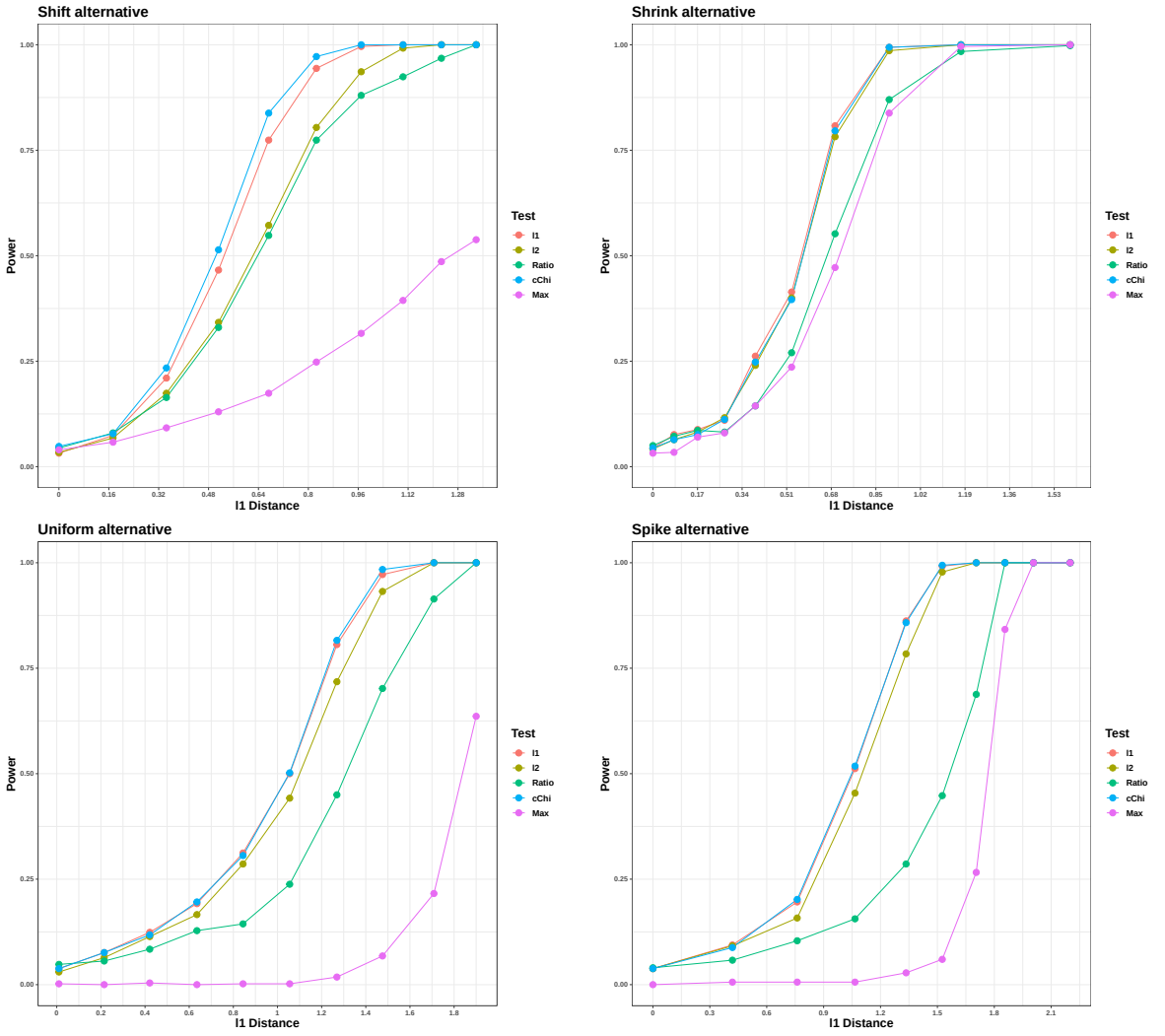

FIGURE 3.3: A comparison between the five tests proposed in four different scenarios. The power of the tests are plotted against the ℓ_1 distance between F and G. The sample sizes are $n_1 = n_2 = 100$ while the number of cells of the multinomial is $K = 400$. Each point in the graph is an average over 1000 simulations.

> Multnomah County. [...] The Street Crimes Unit and the Prostitution Coordination Team were both brought back to full strength. Each team now has four officers reporting to a single sergeant. By bringing these two units under one supervisor the ability to impact the problem has increased as these units grow in understanding of how sex trafficking and gang violence overlap. Hales and Reese (2013)

As a result, we would like to test if the distribution of crimes, in particular related to human trafficking, has changed over the years. We apply the centered χ^2 to test the distribution of each year against the previous one. The differences result all significant except for the last two years. The improvement of prosecution policies and the increase control and prevention of the Police Department seem to have positively influenced

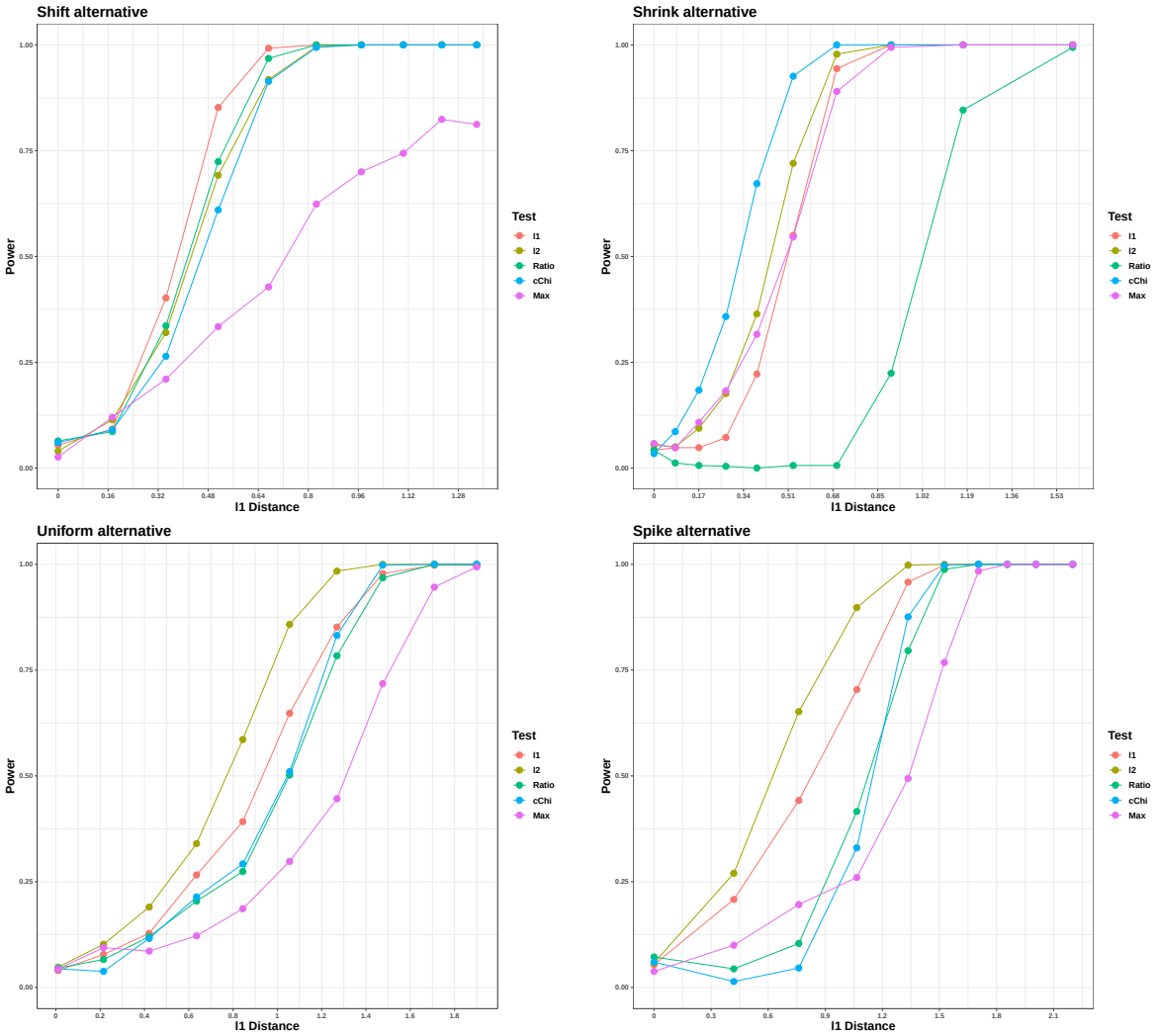

FIGURE 3.4: A comparison between the five tests proposed in four different scenarios. The power of the tests are plotted against the ℓ_1 distance between F and G. The sample sizes are $n_1 = 4000$ and $n_2 = 400$, while the number of cells of the multinomial is $K = 400$. Each point in the graph is an average over 1000 simulations.

the situation. There is a clear reduction in the critical area around the Northeast 82nd Avenue (figure 3.5). It is worth noting that the significant differences in the distributions are not the result of a decrease in number of cases (table 3.1).

TABLE 3.1: Number of crimes from 2012 to 2017

Year	2012	2013	2014	2015	2016	2017
Count	229	188	143	109	198	159

This is more evident if we estimate the distributions with the method proposed in chapter 1, as in figure 3.6. Here we see that the mode of the distribution shifts from the Northeast 82nd avenue to the northern part of the city, near the airport. The modes in

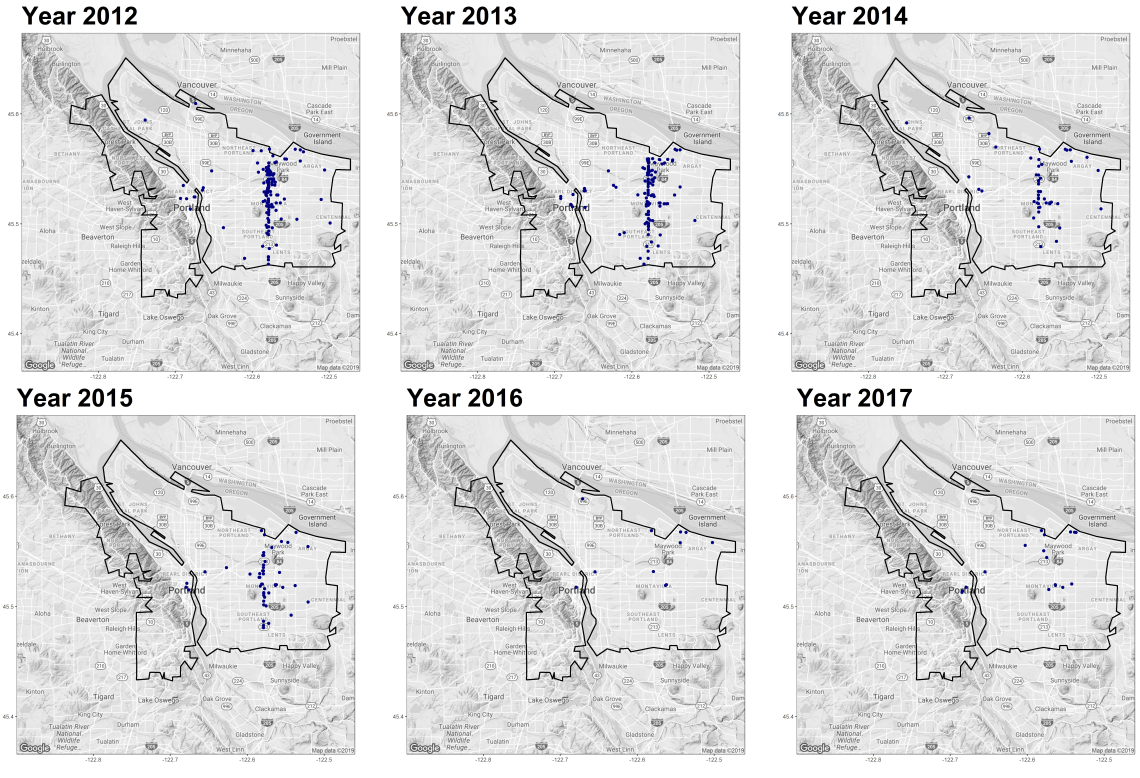

FIGURE 3.5: From the top-left, distribution of the crimes concerning human trafficking in the city of Portland from 2012 to 2017.

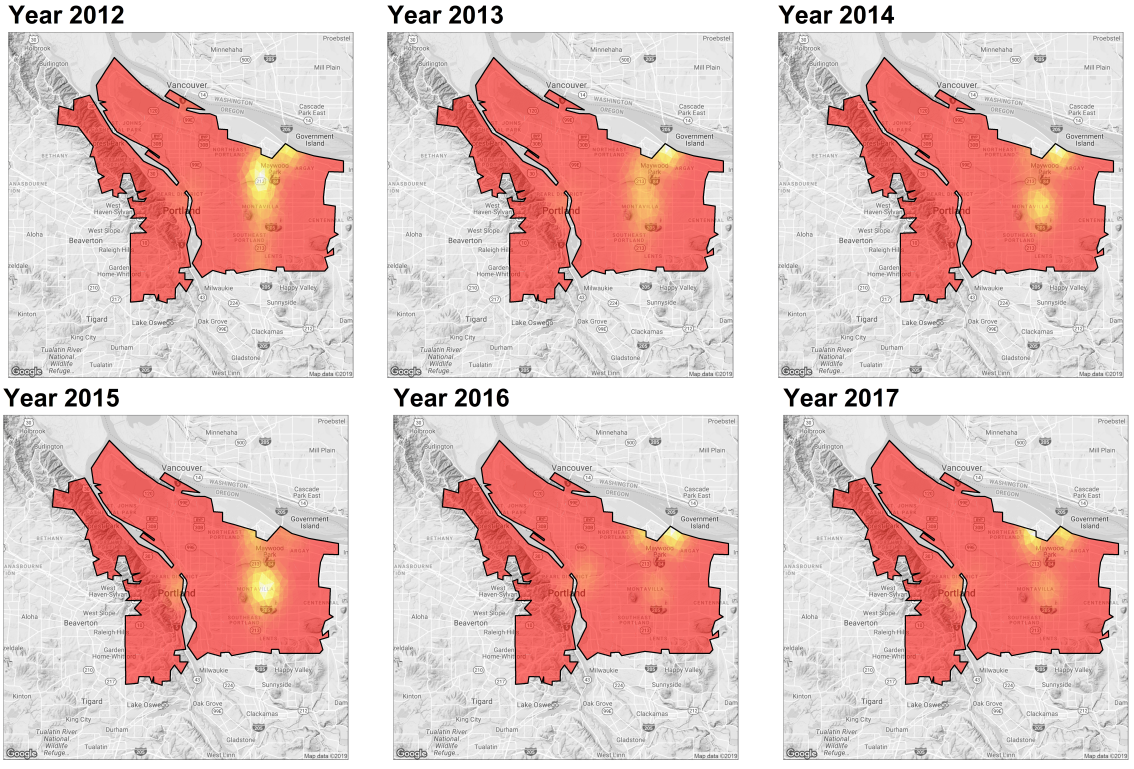

FIGURE 3.6: From the top-left, estimated densities of the distributions of crimes concerning human trafficking in the city of Portland from 2012 to 2017.

the last two years correspond probably to the position of illegal brothels. The human trafficking problem is therefore reduced, especially in the streets, but more efforts are needed for a complete eradication.

Chapter 4

Hypothesis testing for spatial regression models

4.1 Introduction

In this final chapter we move back to the SR-PDE introduced in Sangalli et al. (2013). Here we study the asymptotic properties of such estimators, focusing on the consistency and the asymptotic normality. We also develop two nonparametric procedure to test the linear components of such models.

The SR-PDE represents a set of numerically efficient regression models, capable to handle complex applied problems (see for example Azzimonti et al. (2015) and Lila et al. (2016a)). These models use a regularized estimation scheme, similar to other semiparametric models such as penalized splines regression (Yu and Ruppert, 2002), thin plate splines (Wood, 2003) and soap film smoothing (Wood et al., 2008). The main difference resides in the regularization functional, based on partial differential operator. These operators allows to include various types of anisotropy and non-stationarity, and also to consider boundary conditions. These peculiarities pose new challenges for the study of the theoretical properties of these models, that are still largely unexplored. The work of Sangalli et al. (2013) presents the well posedness of the estimation problem and the characterization of its solution. The work of Azzimonti et al. (2014) shows that the infinite-dimensional estimator $\hat{f}$ in (1.5) and the discrete estimator $\hat{f}$ in (1.6) are asymptotically unbiased. Although promising, the results are nonetheless restricted to the case where no covariates are present. The results that we present here consider focuses on the discrete estimator, with the inclusion of covariates. Although the resulting asymptotic distribution of the estimators might be used for hypothesis testing, the corresponding

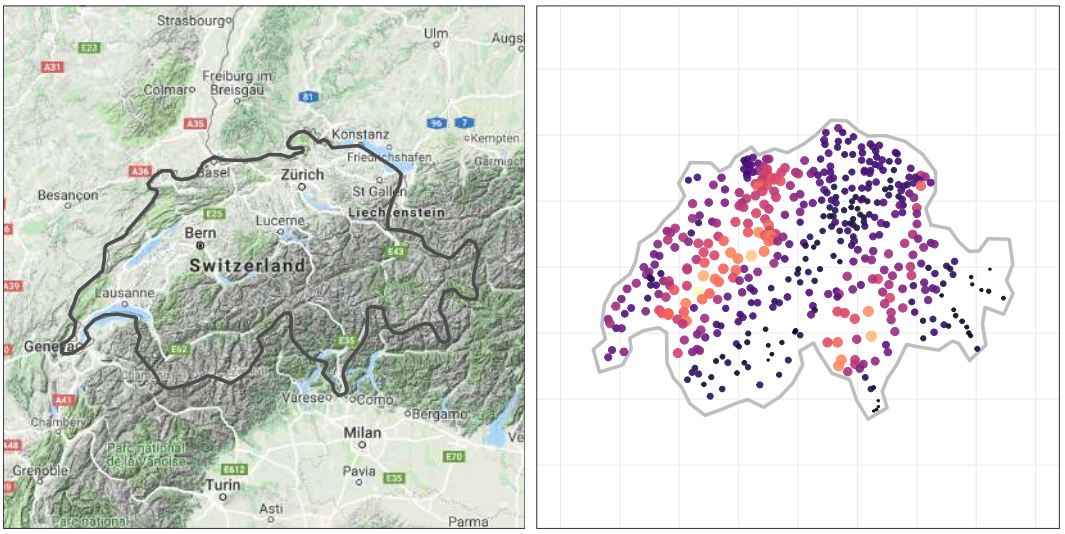

FIGURE 4.1: Switzerland rainfall data. These include 467 daily rainfall measurements recorded in Switzerland on May 8, 1986. The size and color of point markers represent the value of the rainfall at each location, highlighting a strong spatial anisotropy.

performances are far from acceptable. In the finite sample scenario, the variance is usually overestimated due to the regularization term. To reduce this problem, we propose a nonparametric test procedure based on random sign flipping of the score components of the model, inspired by the work of Hemerik et al. (2019). Nonetheless, the presence of possible spatial dependence structures make the testing procedure less accurate. As an example, Figure 4.1 shows rainfall measurements recorded in Switzerland on May 8, 1986. This dataset was used for the Spatial Interpolation Comparison 97 (Dubois et al., 2003). The size and color of point markers represent the value of the rainfall at each location, highlighting a strong spatial anisotropy, with higher rainfall values alternating with lower rainfall values along elongated regions oriented in the northeast-southwest direction. In this case, where a strong spatial dependence is present, we propose a generalization based on a Random Domain Decomposition (RDD). This partitioning approach helps to reduce the effect of the spatial dependence, without the necessity to introduce additional parametric assumptions on the form of the correlation structure. Here we focus on the SR-PDE, but the same tests can be extended to other generalized semiparametric models, such as penalized splines regression (Yu and Ruppert, 2002), thin plate splines (Wood, 2003) and soap film smoothing (Wood et al., 2008).
The chapter develops as follows: in Section 2 we briefly introduce the SR-PDE framework; in Section 3 we study the asymptotic properties of the estimators; in Section 4 we present the hypothesis testing procedure and in Section 5 some simulation studies and an application to rainfall data in Switzerland.

4.2 Model

Let $\{\mathbf{p}_i, i = 1, \ldots, n\}$ be a set of n data locations over a bounded regular domain $\Omega \subset \mathbb{R}^2$, with boundary $\partial\Omega \in \mathcal{C}^2$. Let $z_i \in \mathbb{R}$ be the value of the variable of interest observed at point $\mathbf{p}_i$, and let $\mathbf{w}_i \in \mathbb{R}^q$ be a the vector covariates associated to the observation z_i. We assume a semi-parametric model of the form

$$z_i = \mathbf{w}_i^\top \boldsymbol{\beta}_0 + f_0(\mathbf{p}_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^q$ is the vector of true parameters, f_0 is an unknown deterministic mean field that captures the spatial structure of the problem and ϵ_i are i.i.d. random errors with zero mean and variance σ^2. The estimation problem can be solved minimizing the regularized least squares

$$\sum_{i=1}^{n} (z_i - \mathbf{w}_i^\top \boldsymbol{\beta} - f(\mathbf{p}_i))^2 + \lambda \int_\Omega (\Delta f)^2 \, \mathrm{d}\mathbf{p}, \tag{4.2}$$

where $\lambda > 0$ is the smoothing parameter and Δ denotes the Laplace operator, defined as

$$\Delta f(\mathbf{p}) = \frac{\partial^2}{\partial p_1} f(\mathbf{p}) + \frac{\partial^2}{\partial p_2} f(\mathbf{p}).$$

where $\mathbf{p} = (p_1, p_2)$. The functional (4.2) is well defined for $\boldsymbol{\beta} \in \mathbb{R}^q$ and $f \in \mathcal{H}^2(\Omega)$ and the estimation problem has unique solution imposing appropriate boundary conditions (see Sangalli et al. (2013), Azzimonti et al. (2014) for details). If problem-specific information is available, it is possible to consider more complex regularization term instead of the simple Laplacian. In particular, if this information can be formalized in terms of a PDE $Lf = u$, it makes sense to estimate the unknown $\boldsymbol{\beta}$ and f by minimizing a functional that trades-off a data fidelity criterion, the least square term, and a model-fidelity criterion, the PDE. This can be done by introducing a regularization of the form

$$\int_\Omega (Lf - u)^2 \, \mathrm{d}\mathbf{p}.$$

This regularizing term enables a very rich modeling of space variation, including anisotropic and directional smoothing. We point to the works of Azzimonti et al. (2014) and Azzimonti et al. (2015) for more details.

4.2.1 Discrete Estimator

Here we briefly remind the reader of the results presented in Chapter 1 about the SR-PDE method. In particular, the reduction of the infinite dimensional problem (4.2) to a linear system using the FEM.
We first introduce an appropriate partition of the domain Ω, for instance using a Delaunay triangulation. This triangulation $\mathcal{T}$ defines an approximated domain $\Omega_{\mathcal{T}}$, represented by a mesh with $N_{\mathcal{T}}$ nodes. Given the triangulation, we then define a set of basis functions $\psi_1, \dots, \psi_{N_{\mathcal{T}}}$, and a corresponding $n \times N_{\mathcal{T}}$ matrix $\Psi = \{\psi_j(\mathbf{p}_i)\}_{ij}$. This matrix consists of the evaluations of the $N_{\mathcal{T}}$ basis functions $\psi_1, \dots, \psi_{N_{\mathcal{T}}}$ at the n data locations. Now let $\mathbf{z} = (z_1, \dots, z_n)$ be the vector of observed data values. Let also W be the $n \times q$ matrix whose i-th row is given by $\mathbf{w}_i^t$, the vector of q covariates associated with observation z_i at $\mathbf{p}_i$, and assume that W has full rank. Moreover, set $Q = I - W(W^\top W)^{-1}W^\top$, the matrix that projects into the orthogonal complement of $\mathbf{R}^n$ with respect to the subspace of $\mathbf{R}^n$ spanned by the columns of W. Finally, let us denote by $\hat{\mathbf{f}}_n$ the vector of evaluations of f at the n data locations, $\hat{\mathbf{f}}_n = (f(\mathbf{p}_1, \dots, \mathbf{p}_n))$. As shown in (see Sangalli et al. (2013), there exists a unique pair of estimators $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}})$ which solve the discrete counterpart of the estimation problem (4.2), expressed as

$$\hat{\boldsymbol{\beta}} = (W^\top W)^{-1}W^\top(\mathbf{z} - \hat{\mathbf{f}}_n),$$

with $\hat{\mathbf{f}}$ satisfying

$$\begin{bmatrix} -\Psi^\top Q\Psi/n & \lambda R_1^\top \\ \lambda R_1 & \lambda R_0 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{f}} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} -\Psi^\top Q\mathbf{z}/n \\ 0 \end{bmatrix}. \tag{4.3}$$

The matrices R_0 and R_1 are called mass and stiffness matrix, respectively, and represent the system of linear equations related to theapproximation of the operator in the regularization term. For simplicity, let us denote $P = R_1^\top R_0^{-1} R_1$, the positive semi-definite matrix that represent the discretization of the regularization term in (4.2). Let also S be the $n \times n$ matrix

$$S = \Psi(\Psi^\top Q\Psi + \lambda P)^{-1}\Psi^\top Q.$$

Using this notation, we have

$$\begin{aligned} \hat{\mathbf{f}}_n &= S\mathbf{z}, \\ \hat{\boldsymbol{\beta}} &= (W^\top W)^{-1}W^\top(I - S)\mathbf{z}. \end{aligned} \tag{4.4}$$

Note that the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ are linear in the observed data values $\mathbf{z}$.

4.3 Asymptotic properties

We present some results on the asymptotic properties of the discrete estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$. Here we assume the number of basis $N_{\mathcal{T}}$ to be fixed in advance, and such that the corresponding triangulation results fine enough to capture all the aspects of the problem. Some results on the asymptotic properties of these estimators are proposed in Arnone (2018). These results are nonetheless limited to rates of convergence. Here we provide more practical results, that are the consistency and the convergence to a known normal distribution.
First, we study the asymptotic behavior of the nonparametric component and its consistency. These are necessary to develop the results for the linear part.

Proposition 4.1. *Let $(\hat{\mathbf{f}}_{n,\lambda_n})$ be a sequence of sr-PDE estimators. Assume that a non-singular limit $A = \lim_n A_n$ exists. If $\lambda_n = o(n^{1/2})$ and $n - q \geq N_{\mathcal{T}}$, then $\hat{\mathbf{f}}_{n,\lambda_n}$ is a consistent estimator for f_0, with asymptotic distribution*

$$\sqrt{n}(\hat{\mathbf{f}} - f_0)|W \overset{n}{\sim} \mathcal{N}_{N_{\mathcal{T}}}(0, \sigma^2 A).$$

Proof. Recall from (4.3) that $\hat{\mathbf{f}}$ is the solution of the linear system

$$\begin{cases} -\Psi^\top Q\Psi\hat{\mathbf{f}}/n + \lambda R_1^\top \mathbf{g} = -\Psi^\top Q\mathbf{z}/n \\ \lambda R_1\hat{\mathbf{f}} + \lambda R_0\mathbf{g} = 0 \end{cases}.$$

Using (4.1), the first equation can be rewritten as

$$-\Psi^\top Q\Psi\hat{\mathbf{f}}/n + \lambda R_1^\top \mathbf{g} = -\Psi^\top Q(W\beta_0 + \Psi\mathbf{f}_0 + \epsilon)/n,$$

where $\mathbf{f}_0$ denotes the evaluation of the true function f_0 on the nodes, and $\boldsymbol{\beta}_0$ the vector of parameters of the linear component. Substituting the expression for $\mathbf{g}$, we obtain

$$(\Psi^\top Q\Psi/n + \lambda P)(\hat{\mathbf{f}} - \mathbf{f}_0) + \lambda P\mathbf{f}_0 = \Psi^\top Q\epsilon/n. \tag{4.5}$$

Remember that QW and has all entries equal to 0 by construction. To simplify the notation, let us define $A_n = \left(\Psi^\top Q\Psi/n\right)^{-1}$, where the subscript n underline the dependence on the sample size. The right side of the equation (4.5) can be used as a pivot to derive the asymptotic distribution of our estimator. First, we need to study the bias term

$$-\lambda(\Psi^\top Q\Psi/n + \lambda P)^{-1}P\mathbf{f}_0. \tag{4.6}$$

Starting from the analytic expression for $\hat{\mathbf{f}}$ in (4.4), we can compute the bias expansion

$$b_n(\lambda) = -\lambda A_n P \mathbf{f}_0 + \lambda^2 A_n P A_n P \mathbf{f}_0 + o(\lambda^2).$$

Note that the first order term of the bias expansion corresponds to (4.6). The variance of the estimator has the expansion

$$\mathrm{Var}_n(\lambda) = \frac{\sigma^2}{n}\left(A_n - 2\lambda A_n P A_n + 3\lambda A_n P A_n P A_n + O(\lambda^3)\right).$$

Finally, the mean square error for the sr-PDE estimator can be computed as

$$\begin{aligned} M_n(\lambda) &= \mathrm{Var}_n(\lambda) + b_n(\lambda) b_n(\lambda)^\top \\ &= \frac{\sigma^2}{n}\left(A_n - 2\lambda A_n P A_n + 3\lambda A_n P A_n P A_n + O(\lambda^3)\right) + \lambda^2 A_n P \mathbf{f}_0 \mathbf{f}_0^\top P A_n + o(\lambda^2). \end{aligned} \tag{4.7}$$

The bias of the estimator for the nonparametric component is then of the order $O(\lambda)$. If $\lambda = o(n^{-1/2})$, the asymptotic bias of $\sqrt{n}(\hat{\mathbf{f}} - f_0)$ will vanish. Otherwise, if $\lambda \to 0$ slower than $n^{-1/2}$, the bias term dominates. The work of Yu and Ruppert (2002) develops a similar theory in the case of penalized regression splines. Nonetheless, in our case the higher flexibility of the estimation procedure, provided by the FEM, results in a more delicate handling of the asymptotic quantities. In particular, we need to be careful about the behaviour of the term A_n in the asymptotic regime. Recall that Q is a projection matrix of order n and $\mathrm{rank}(A_n) = n - q$, while Ψ is a $n \times N_{\mathcal{T}}$ matrix with $\mathrm{rank}(\Psi) = \min(n, N_{\mathcal{T}})$. To check if A_n is well defined, let us rewrite

$$\Psi^\top Q \Psi = \Psi^\top V \Lambda V^\top \Psi = (\Psi^\top V)\Lambda^{1/2}\Lambda^{1/2}(V^\top \Psi),$$

where V is the matrix of eigenvectors Q, and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = \lambda_i$. Without loss of generality, let us define Λ so that the last q eigenvalues are equal to zero. These q eigenvalues annihilate the last q rows of $B = \Lambda^{1/2}(V^\top \Psi)$. We than have

$$\Psi^\top Q \Psi = B^\top B,$$

that is a $N_{\mathcal{T}} \times N_{\mathcal{T}}$ square matrix with $\mathrm{rank}(\Psi^\top Q \Psi) = \min(n - q, N_{\mathcal{T}})$. In order to have a nonsingular matrix, we need $n - q \geq N_{\mathcal{T}}$.
To obtain the consistency, we need $\mathbb{E}(\hat{\mathbf{f}} - f_0)^2 \to 0$. Using (4.7), the condition is satisfied

with $\lambda_n = o(n^{1/2})$. Thus $\hat{\mathbf{f}}_{n,\lambda_n}$ converges to f_0 in probability.
The special case where $\lambda = cn^{-1/2}$, with $c > 0$, is more subtle. Here we would obtain

$$\sqrt{n}(\hat{\mathbf{f}} - f_0)|W \overset{n}{\sim} \mathcal{N}_{N_\mathcal{T}}(-c\Sigma^{-1}Pf_0, \sigma^2 A).$$

We can therefore still achieve the $\sqrt{n}$ consistency, but with a more involved asymptotic mean. □

Given (4.1), we now consider a similar result for the linear component of the model.

Proposition 4.2. *Let $(\hat{\boldsymbol{\beta}}_n)$ be a sequence of SR-PDE estimators. Let Θ be a compact parameter space, with β_0 as interior point. Assume $\Sigma = \lim_n \Sigma_n$ exists and is nonsingular. Then given a consistent estimate $\hat{\mathbf{f}}$, the estimator $\hat{\boldsymbol{\beta}}_n$ is consistent for β_0, with asymptotic distribution*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \beta_0)|W \overset{n}{\sim} \mathcal{N}_q(0, \sigma^2(\Sigma^{-1} + \Sigma^{-1}W^\top \Psi A \Psi^\top W \Sigma^{-1})).$$

Proof. Given a consistent estimator $\hat{\mathbf{f}}$, the vector $\hat{\boldsymbol{\beta}}$ is the solution of the score equation

$$\frac{1}{n} W^\top (\mathbf{z} - W\hat{\boldsymbol{\beta}} - \Psi\hat{\mathbf{f}}) = 0,$$

Using (4.1), we get

$$\frac{1}{n} W^\top W(\hat{\boldsymbol{\beta}} - \beta_0) + \frac{1}{n} W^\top \Psi(\hat{\mathbf{f}} - \mathbf{f}_0) = \frac{1}{n} W^\top, \epsilon.$$

Let us define $\hat{\Sigma}_n = W^\top W/n$. We obtain

$$\hat{\Sigma}_n(\hat{\boldsymbol{\beta}} - \beta_0) = \frac{1}{n} W^\top \epsilon - \frac{1}{n} W^\top \Psi(\hat{\mathbf{f}} - \mathbf{f}_0).$$

The right side can be used as a pivoting quantity to obtain the result.
The consistency follows from Lemma 4.1 and the central limit theorem. □

Note that the estimation of the nonparametric component affects the variance of the linear estimator. This result is consistent with classic on smoothing methods for regression.

4.4 Hypothesis testing

The development of the asymptotic theory in the previous section leads to the question on how to define appropriate testing procedure. In the case of semiparametric

regression, we might be interest to test whether the linear component has an effect on the variable of interest. In the example of Figure 4.1, we would like to test if altitude influences the amount of rain. If we consider the definition of (4.1), we specify the system of hypotheses

$$\mathrm{H}_0 : \boldsymbol{\beta}_0 = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \boldsymbol{\beta}_0 \neq 0.$$

A possible solution is to rely on the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and use a classic Wald type test. A more robust alternative is to use the same test together with a sandwich estimate of the variance, as in Yu and Ruppert (2002). Both these parametric alternative may nonetheless be unreliable in a finite sample scenario. The bias may be non negligible and there is an high chance to overestimates the variance.
Here we consider a possible alternative inspired by the work of Hemerik et al. (2019). In particular, we define a randomization test based on random sign flipping of the score component. Although robust to various types of misspecification Hemerik et al. (2019), the test may lose power if a strong spatial correlation in the score component is present. If that is the case, we propose a generalization of such test, able to reduce the effect of the spatial correlation.

4.4.1 Flip-score test

We first introduce the simplest version of the flip-score test. Let $\pi = (\pi_1, \dots, \pi_n)$ be a random vector uniformly distributed in $\{-1, 1\}^n \backslash (1, \dots, 1)$. Let also Π be a diagonal matrix with entries $\Pi_{ii} = \pi_i$. We denote by T the classic score statistic

$$T = W^\top (\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f}).$$

Given a matrix Π, we define in what follows the flip-score statistic T^* as

$$T^* = W^\top \Pi (\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f}).$$

As a first step, we consider the behavior of both the expected value and the variance of T and T^* under the null hypothesis. To simplify the notation, let us define $B_n = (\Psi^\top \Psi/n + \lambda P)^{-1}$. The randomization procedure should leave unchanged these two quantities. For the expected value, it is easy to see that

$$\mathbb{E}(T) = \mathbb{E}(W^\top (\mathbf{z} - \Psi\hat{\mathbf{f}})) = W^\top \mathbb{E}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon) = \lambda_n W^\top \Psi B_n P f_0 \to 0,$$

Algorithm 1 Simple sign-flipping

1: Compute the score components under H_0
2: Compute the the observed test statistic T^{obs}
3: **for** i $\in 1, \ldots, B$ **do**
4: Generate a sign flipping matrix Π^i
5: Compute the test statistic

$$T^i = W^\top \Pi^i (\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f})$$

6: **end for**
7: Use the $T^1, \ldots, T^B$ to obtained the p-value

and

$$\mathbb{E}(T^*) = \mathbb{E}(W^\top \Pi(\mathbf{z} - \Psi\hat{\mathbf{f}})) = W^\top \Pi \mathbb{E}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon) = \lambda_n W^\top \Pi \Psi B_n P f_0 \to 0,$$

where we used the expression (4.5). Both T and T^* are asymptotically unbiased under H_0. However, it is worth noting that in the finite sample case we might have to pay a small price due to the regularization on the nonparametric component. We discuss this problem in the next sections.

We now consider variance of T and T^* under the null. In the case of T, we obtain

$$\begin{aligned}
\text{Var}(T) &= \text{Var}(W^\top(\mathbf{z} - \Psi\hat{\mathbf{f}})) \\
&= W^\top \text{Var}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon) W \\
&= W^\top \Psi \text{Var}(\lambda B_n P f_0 + B_n \Psi^\top \epsilon) \Psi^\top W + W \text{Var}(\epsilon) W \\
&= \sigma^2 W^\top \Psi B_n \Psi^\top \Psi B_n \Psi^\top W + \sigma^2 W^\top W.
\end{aligned}$$

In the case of T^*, we have

$$\begin{aligned}
\text{Var}(T^*) &= \text{Var}(W^\top \Pi(\mathbf{z} - \Psi\hat{\mathbf{f}})) \\
&= W^\top \Pi \text{Var}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon) \Pi W \\
&= W^\top \Pi \Psi \text{Var}(\lambda B_n P f_0 + B_n \Psi^\top \epsilon) \Psi^\top \Pi W + W \text{Var}(\epsilon) W \\
&= \sigma^2 W^\top \Pi \Psi B_n \Psi^\top \Psi B_n \Psi^\top \Pi W + \sigma^2 W^\top W.
\end{aligned}$$

In order to have a permutation invariant test, we should have $\Psi B_n \Psi^\top = I$, so that it can commute with Π. Although this can be achieved in the asymptotic case, in the finite sample case this can be an hypothesis way too restrictive.

To better understand the problem, let us focus on the structure of the matrix B_n, shown in figure 4.2. The matrix has a similar structure to a weighted adjacency matrix of an

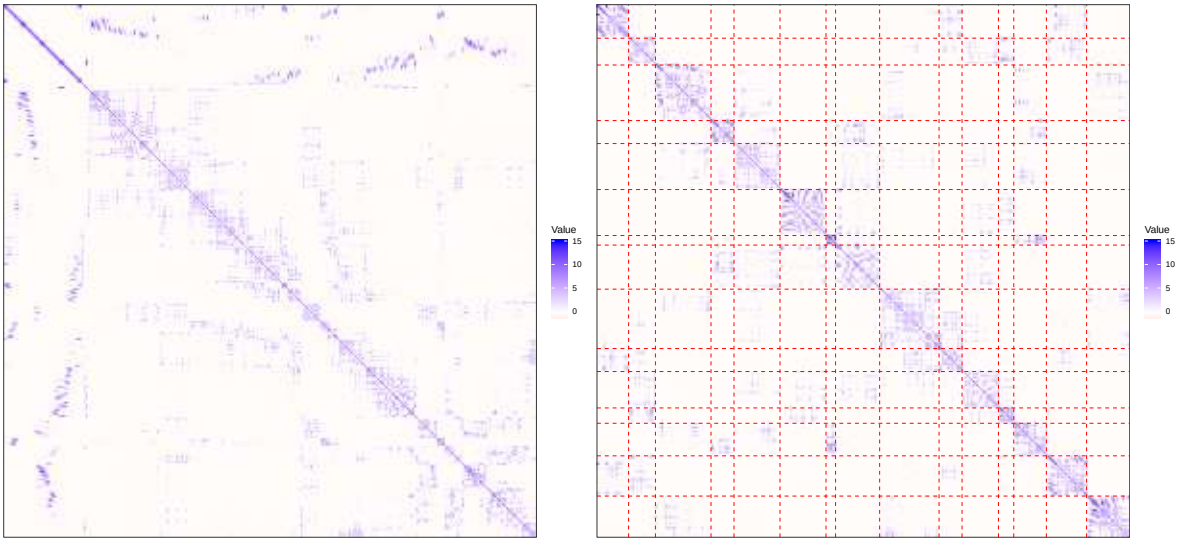

FIGURE 4.2: On the left, heatmap of the matrix B_n. On the right, Reordering of B_n with $M = 15$ groups.

undirected graph, where the graph in this case is given by the mesh. There is a clear distinction of two main blocks. The first block correspond to the nodes forming the boundaries, while the second block correspond to the internal nodes. The second block has itself the structure similar to a block diagonal matrix. This is more clear if we look at the expansion of $\Psi B_n \Psi^\top$, given by

$$\Psi B_n \Psi^\top = (\mathbb{I} + \lambda(P_1 - P_2 P_4^{-1} P_3))^{-1} = \mathbb{I} - \lambda(P_1 - P_2 P_4^{-1} P_3) + \mathcal{O}(\lambda^2),$$

where P_i are the block of the matrix P. Here we used the fact that $\Psi = [\mathbb{I}_n | \mathbb{O}]$ is a $n \times N_\mathcal{T}$ block matrix, with an order n identity matrix as first block and an $n \times N_\mathcal{T} - n$ matrix of zeros as second block. The structure of $\Psi B_n \Psi^\top$ is close to an identity matrix, but it has nonzero element outside the diagonal in correspondence to the connected nodes of the mesh. This induces a local correlation in the residuals that is not preserved by the simple sign flipping.

4.4.2 Block sign-flip test

In order to reduce this problem, we propose a blockwise sign-flip procedure. The idea is to try to define a sign-flipping matrix the mimics the structure of $\Psi B_n \Psi^\top$. Since the blocks are induced by the connections of the nodes of the mesh, it seems reasonable to define a partition of the domain that groups nodes that are close together.
Let $\mathcal{P} = \{D_1, \dots, D_M \subset \Omega\}$ be a partition of the domain into disjoint sub-regions such

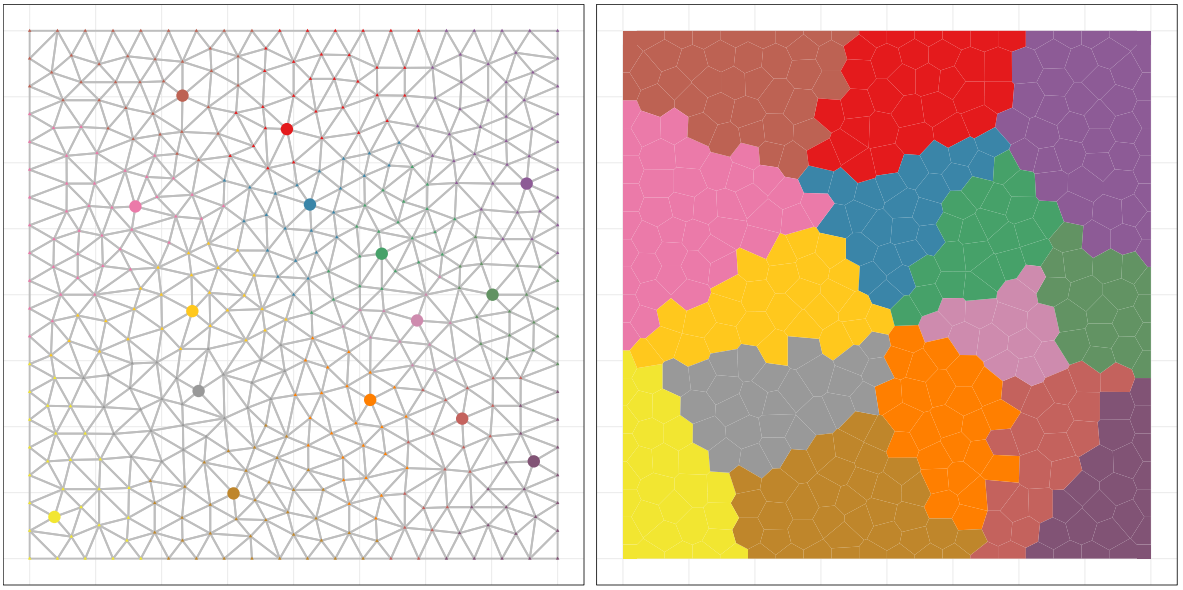

FIGURE 4.3: On the left, mesh and centroids of the RDD. On the right, the corresponding RDD with $M = 15$ blocks.

that

$$\Omega = \bigcup_{m=1}^{M} D_m \quad \text{and} \quad D_i \cap D_j = \emptyset, \quad \forall\, i, j = 1, \dots, M,\ i \neq j.$$

The first step is to use the Delaunay triangulation of the domain in the estimation step as a graph. Let us denote the nodes of such triangulation as $(\xi_1, \dots, \xi_{N_\mathcal{T}})$. We now randomly select a sub-sample of size M of the nodes. These will define the centroid $(c_1, \dots, c_m)$ of each sub-domain $(D_1, \dots, D_M)$. The Delaunay triangulation has a corresponding dual graph, called Voronoi tessellation. Let us call $(t_1, \dots, t_{N_\mathcal{T}})$ the tiles of such tesselletion. We define each block D_m of the RDD as the union of the tiles t_i that are closer to the centroid. As a distance between two nodes we consider the (Euclidean) length of the shortest path on the graph connecting the two sites: this is computed by the Dijkstra's algorithm (Dijkstra, 1959). More generally, the distance between a nodes $\xi_i \in \mathcal{T}$ and a node ξ_j belonging to the graph, is computed by measuring the length of the shortest path connecting ξ_i and ξ_j (see figure 4.3). A similar partition of the domain is used in Menafoglio et al. (2018) for spatial kriging.

Given a random partition, we define a block sign-flipping matrix Π_B, where diagonal elements belonging to the same block shares the same sign. The partition induces a block structure in B_n that mimics the one of Π_B. For simplicity, let us denote $\tilde{B}_n = \Psi B_n \Psi^\top$.

Algorithm 2 Block sign-flipping

1: Compute the score components under H_0
2: Compute the the observed test statistic T^{obs}
3: **for** $l \in 1, \dots, L$ **do**
4: Generate $\mathcal{P}^1, \dots, \mathcal{P}^L$ different partitions of the domain of interest
5: **for** $i \in 1, \dots, B$ **do**
6: Generate a block sign flipping matrix Π_l^i
7: Compute the test statistic

$$T^{m,i} = W^\top \Pi_l^i (\mathbf{z} - W\boldsymbol{\beta} - \Psi\mathbf{f})$$

8: **end for**
9: **end for**
10: Use the $L \cdot B$ obtained statistics to compute the p-value

This gives

$$\begin{aligned}
\mathrm{Var}(T^*) &= \mathrm{Var}(W^\top \Pi_B(\mathbf{z} - \Psi\hat{\mathbf{f}})) \\
&= W^\top \Pi_B \mathrm{Var}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon)\Pi_B W \\
&= W^\top \Pi_B \Psi \mathrm{Var}(\lambda B_n P f_0 + B_n \Psi^\top \epsilon)\Psi^\top \Pi_B W + W\mathrm{Var}(\epsilon)W \\
&= \sigma^2 W^\top \Pi_B \tilde{B}_n \tilde{B}_n \Pi_B W + \sigma^2 W^\top W \\
&= \sigma^2 W^\top \tilde{B}_n^L \tilde{B}_n^R W + \sigma^2 W^\top W,
\end{aligned}$$

where $\tilde{B}_n^L \tilde{B}_n^R$ has the same blocks and the same diagonal elements of $\tilde{B}_n \tilde{B}_n$. The two matrices differs only in the signs of the off the off-diagonal blocks. This gives us a test that is almost exact and it is a reasonable approximation in a spatial regression setting. There is nonetheless a parameter that has to be chosen, that is the number of blocks in each RDD. We suggest the following criterion based on the bandwidth of the matrix. As first step, we use the Reverse Cuthill-McKee algorithm on the matrix $\tilde{B}_n$ (see (Cuthill and McKee, 1969)). This algorithm is widely used in linear algebra to find a permutation of a sparse matrix that has a symmetric sparsity pattern into a band matrix form with a small bandwidth. A band matrix is simply a sparse matrix whose non-zero entries are confined to a diagonal band. For each row i of $\tilde{B}_n$, we compute $\kappa_i = \#\{b_{ij} = 0, \forall j > i\}$. The suggested number of blocks will be given by

$$M = \frac{n}{\sum_{i=1}^n \kappa_i / n} = \frac{n^2}{\sum_{i=1}^n \kappa_i}.$$

The idea behind the criterion is to find a block structure that covers most of the non-zero element of the matrix.

4.4.3 Random covariates

In the previous section we made the somehow usual assumption of fixed covariates. It is nonetheless interesting to consider the case where the covariates W are themselves random variables. This case might be of interest for spatial regression or more general semiparametric regression models, if the goal is to construct a model for prediction purposes. An intriguing perspective on random designs in linear regression is given in Buja et al. (2015). This scenario is also closely connected to the theory of estimating equations Tsiatis (2007) and the model-robust inference with sandwich estimators and bootstrap.

Let W be a random vector with $\mathbb{E}(W) = 0$ and $\mathrm{Var}(W) = \Sigma_W$. For simplicity here we assume $q = 1$. If we consider the test T defined above, we obtain

$$\begin{aligned}
\mathbb{E}(T) &= \mathbb{E}(\mathbb{E}(W^\top(\mathbf{z} - \Psi\hat{\mathbf{f}})|W)) \\
&= \mathbb{E}(W^\top\mathbb{E}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon|W)) \\
&= \mathbb{E}(W^\top\Psi\mathbb{E}(\lambda B_n P f_0 + B_n\Psi^\top\epsilon|W) + \mathbb{E}(\epsilon|W)) \\
&= \lambda\mathbb{E}(W)^\top\Psi B_n P f_0 = 0.
\end{aligned}$$

With fairly general assumptions on the expectation of W, we obtain the unbiasedness even in the finite sample case. Following the same reasoning for T^*, we have

$$\begin{aligned}
\mathbb{E}(T^*) &= \mathbb{E}(\mathbb{E}(W^\top\Pi(\mathbf{z} - \Psi\hat{\mathbf{f}})|W)) \\
&= \mathbb{E}(W^\top\Pi\mathbb{E}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon|W)) \\
&= \mathbb{E}(W^\top\Pi\Psi\mathbb{E}(\lambda B_n P f_0 + B_n\Psi^\top\epsilon|W) + \Pi\mathbb{E}(\epsilon|W)) \\
&= \lambda\mathbb{E}(W)^\top\Pi\Psi B_n P f_0 = 0.
\end{aligned}$$

We now consider the variance of both T and T^*. The former lead to

$$\begin{aligned}
\mathrm{Var}(T) &= \mathbb{E}(\mathrm{Var}(W^\top(\mathbf{z}-\Psi\hat{\mathbf{f}})|W)) + \mathrm{Var}(\mathbb{E}(W^\top(\mathbf{z}-\Psi\hat{\mathbf{f}})|W)) \\
&= \mathbb{E}(W^\top \mathrm{Var}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon|W)W) \\
&= \mathbb{E}(W^\top\Psi\mathrm{Var}(\lambda B_n P f_0 + B_n\Psi^\top\epsilon|W)\Psi^\top W + W\mathrm{Var}(\epsilon|W)W) \\
&= \sigma^2\mathbb{E}(W^\top(\tilde{B}_n\tilde{B}_n + I)W) \\
&= \sigma^2\mathbb{E}(W^\top\tilde{B}_n\tilde{B}_n W) + \sigma^2\mathbb{E}(W^\top W) \\
&= \sigma^2\mathbb{E}(\mathrm{tr}(W^\top\tilde{B}_n\tilde{B}_n W)) + \sigma^2\mathbb{E}(W^\top W) \\
&= \sigma^2\mathbb{E}(\mathrm{tr}(\tilde{B}_n W W^\top\tilde{B}_n)) + \sigma^2\mathbb{E}(W^\top W) \\
&= \sigma^2\mathrm{tr}(\mathbb{E}(\tilde{B}_n W W^\top\tilde{B}_n)) + \sigma^2\mathbb{E}(W^\top W) \\
&= \sigma^2\mathrm{tr}(\tilde{B}_n\mathbb{E}(WW^\top)\tilde{B}_n) + \sigma^2\mathrm{tr}(\mathbb{E}(WW^\top)) \\
&= \sigma^2\mathrm{tr}(\tilde{B}_n\Sigma_W\tilde{B}_n) + \sigma^2\mathrm{tr}(\Sigma_W),
\end{aligned}$$

where we used the fact that the trace of a product is invariant under cyclical permutations of the factors, and that the trace is a linear operator, so it commutes with expectation. With similar calculation we obtain

$$\begin{aligned}
\mathrm{Var}(T^\pi) &= \mathbb{E}(\mathrm{Var}(W^\top\Pi(\mathbf{z}-\Psi\hat{\mathbf{f}})|W)) + \mathrm{Var}(\mathbb{E}(W^\top\Pi(\mathbf{z}-\Psi\hat{\mathbf{f}})|W)) \\
&= \mathbb{E}(W^\top\Pi\mathrm{Var}(\Psi f_0 - \Psi\hat{\mathbf{f}} + \epsilon|W)\Pi W) \\
&= \mathbb{E}(W^\top\Pi\Psi\mathrm{Var}(\lambda B_n P f_0 + B_n\Psi^\top\epsilon|W)\Psi^\top\Pi W + W\Pi\mathrm{Var}(\epsilon|W)\Pi W) \\
&= \sigma^2\mathbb{E}(W^\top\Pi(\tilde{B}_n\tilde{B}_n + I)\Pi W) \\
&= \sigma^2\mathrm{tr}(\tilde{B}_n\Pi\Sigma_W\Pi\tilde{B}_n) + \sigma^2\mathrm{tr}(\Sigma_W).
\end{aligned}$$

Note that in this case a central role is played by the structure of the covariance matrix Σ_W. If the covariates have no spatial dependence, that is the covariance structure $\Sigma_W = I$, then Π commutes with Σ_W and we obtain an exact test. On the other hand, if spatial dependence is present, we may need to consider other strategies, based for instance on the eigenvalue decomposition of Σ_W, in order to obtain more powerful tests.

4.5 Simulations and application

In this section we present some simulation studies to investigate the finite sample performance of the proposed tests. We compare our tests to a classical Wald type test based on the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, and to the possibly more robust sandwich version of the same test. A detailed explanation of sandwich estimator in the case of

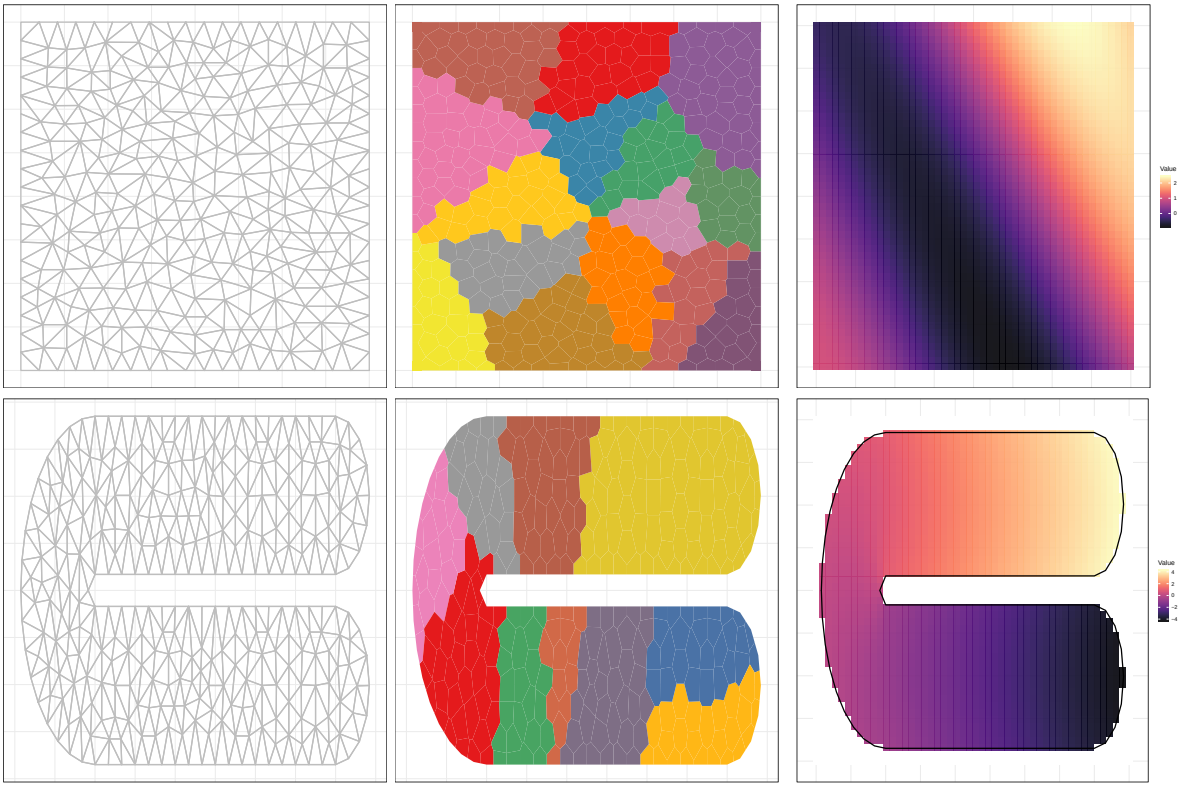

FIGURE 4.4: In the first column, the mesh used for the regression in the two two simulation scenarios. In the second column, one sample of RDD for each domain. In the third column, the spatial fields f considered in the simulations.

penalized semiparametric regression models can be found in Yu and Ruppert (2002). In the first case we consider a square domain with spatial field f_0 defined as $f(x, y) = \cos\left(\frac{2x+y}{4}\right) + \left(\frac{x+y}{15}\right)^2$ (see figure 4.4). The covariates are generated from random fields with different mean and covariance structures using the function `RFsimulate` of the `R` package `RandomFields` (Schlather et al., 2015), but remain fixed over the simulations. In particular, we consider four covariates generated as follows:

TABLE 4.1: Covariance structure of the generated random fields

	Covariance structure	Smoothness, ν	Variance, v	Scale, s
S1	Gaussian			0.3
S2	Matern	1	8	0.5
S3	Exponential		1	2
S4	Matern	5	2	1

In the second case we consider a horseshoe domain (Ramsay, 2002) with spatial field defined using the `fs.test` function of the `R` package `mgcv` (Wood, 2003), shown in figure 4.4 . Similarly to the first simulation, we genereted four covariates as follows:

TABLE 4.2: Covariance structure of the generated random fields

	Covariance structure	Smoothness, ν	Variance, v	Scale, s
S1	Gaussian			0.1
S2	Matern	2	8	0.3
S3	Exponential		1	1
S4	Matern	5	1	0.3

We compared four different test procedures. In red, the block sign-flipping with Random Domain Decomposition, in green, the standard sign-flipping, in cyan, the Wald test based on sandwich estimator, and in purple the classical Wald test. For each of the covariates, we generated 500 samples from with values of the coefficient $\beta = (0, 0.05, 0.1, 0.5)$. The value $\beta = 0$ correspond to the null hypothesis and is used to check the Type-I error control. The other values are used to check the power of the test with increasing values of the parameter.
In the case of square domain, in figure 4.6, the proposed nonparametric tests outperform both the parametric variants. As the scale and the smoothness of the covariate increase, the nonparametric test shows a slightly conservative behavior. This may be due to the problem of exchangeability discussed in the previous section. In all the cases, the power is nonetheless fairly high. The performances of the parametric variants are far worse, due to the overestimation of the variance.
In the case of horseshoe domain, in figure 4.7, the difference in performances are more distinct. The proposed nonparametric tests still outperform both the parametric variants. As in the previous case, when scale and the smoothness of the covariate increase, the nonparametric test shows a slightly less control of the Type-I error. The problem is more pronounces in the last case, where the covariate is particularly smooth. Again, this is due to the problem of exchangeability discussed in the previous section, enhanced by the more complex shape of the domain. The power is nonetheless acceptable in all the cases. The parametric variants have still far worse performances.

4.5.1 Application to Switzerland rainfall

We apply the proposed method to the analysis of the dataset of 467 daily rainfall measurements recorded in Switzerland on May 8, 1986; this dataset was used for the Spatial Interpolation Comparison 97 (Dubois et al., 2003). The data are shown in Figure 4.5. The data include the elevation at the 467 locations, that we use here as a covariate since intuition suggests that the orography of the region may play an important role in

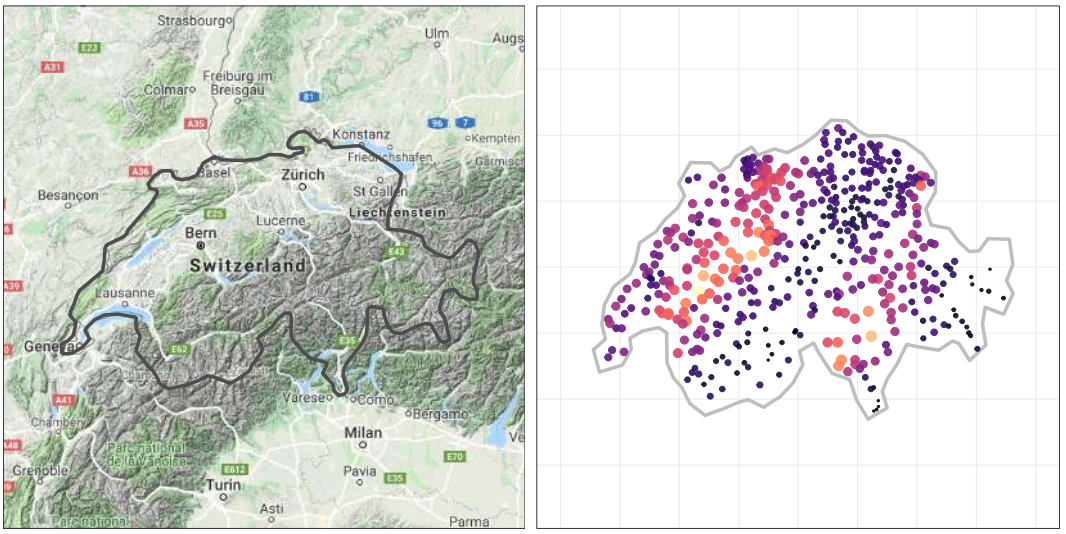

FIGURE 4.5: Switzerland rainfall data. These include 467 daily rainfall measurements recorded in Switzerland on May 8, 1986. The size and color of point markers represent the value of the rainfall at each location, highlighting a strong spatial anisotropy.

the rainfall phenomenon; see figure 4.5 that shows the elevation over Switzerland. We estimated the model under the null using the anisotropic version of the SR-PDE model. Then we used the block sign-flipping test with 15 block to test the coefficient for the altitude. With a p-value of 0.52, the elevation seems not to have a significant impact on the rainfall. This is probably due to the fact that the effect of elevation on rainfall is not linear; the distribution of rainfall is the result of more complex phenomena, such as the interaction between the geomorphology and the atmospheric circulation.

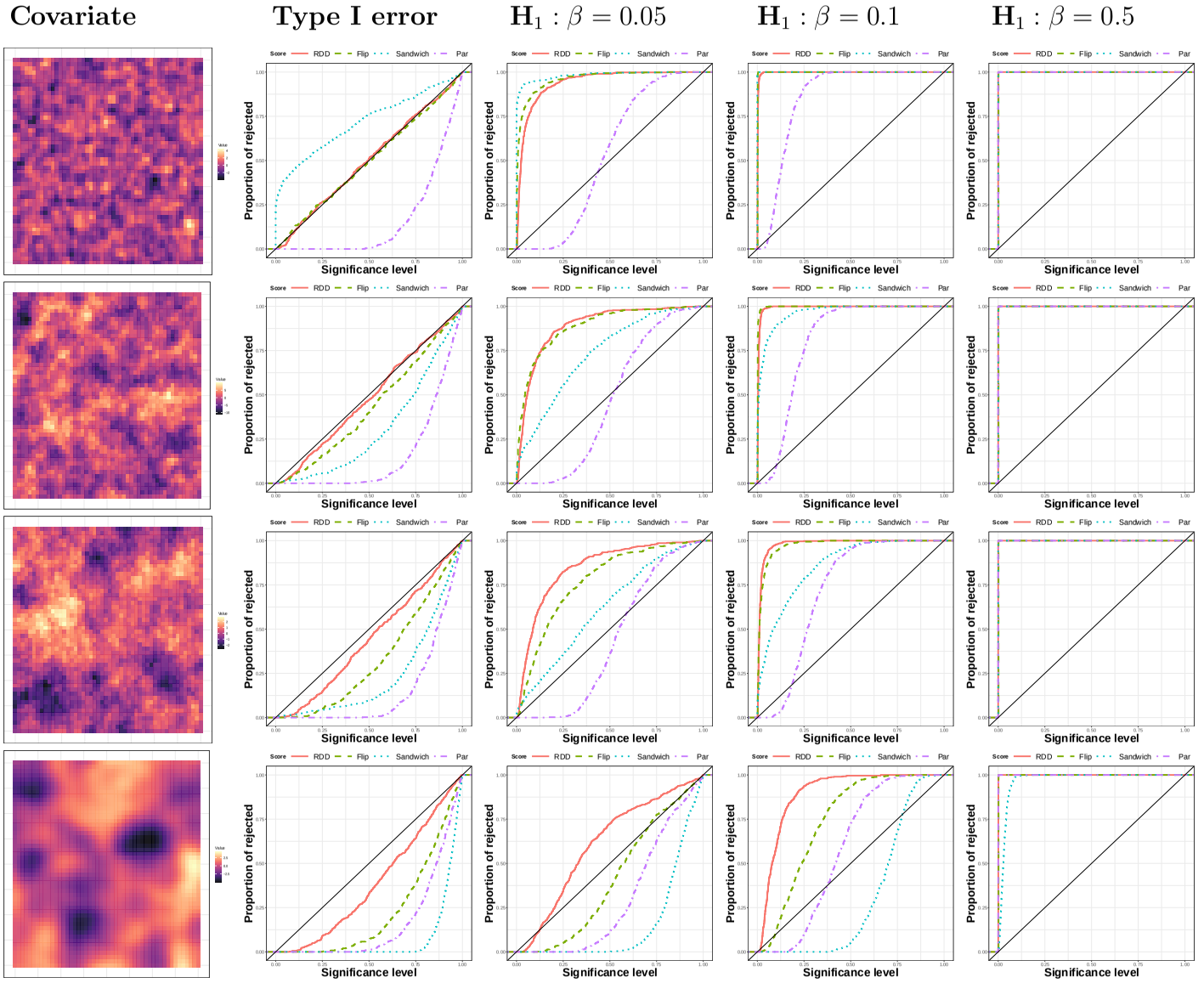

FIGURE 4.6: Comparison of the different test statistics in the case of square domain. The solid line correspond to the flip-score with RDD; the dashed line to the standard flip-score; the dotted line to the Wald test with sandwich estimation of the variance; the dashed-dotted line to the standard Wald test.

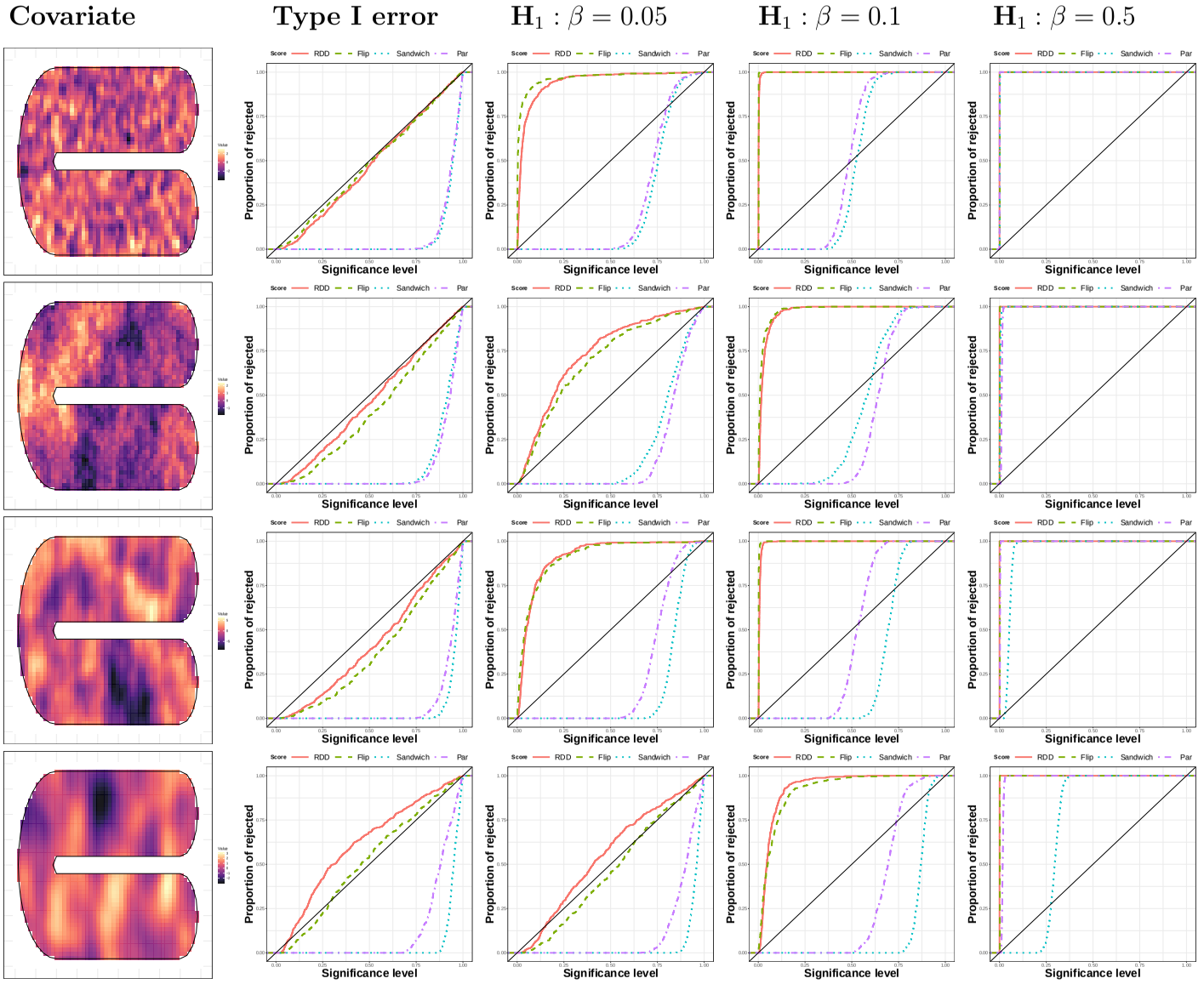

FIGURE 4.7: Comparison of the different test statistics in the case of horseshoe domain. The solid line correspond to the flip-score with RDD; the dashed line to the standard flip-score; the dotted line to the Wald test with sandwich estimation of the variance; the dashed-dotted line to the standard Wald test.

Conclusions

Discussion

The analysis of complex data structures represents an exciting area of research. It poses new stimulating challenges and fuels one of the fastest growing fields of statistics. Taking inspiration from different methodologies from different fields - numerical analysis, engineering, functional analysis - we developed novel procedures for data displaying complex spatial dependencies. The complexity in the structure of spatial variation may be due to different reasons. In some cases, the complexity originates from the complex physics of the phenomenon under study. In other cases, the complex spatial variation is the consequence of the non-trivial conformation of the domain where the data are observed.

In Chapter 2, we developed a nonparametric penalized likelihood approach for density estimation. Our method can be applied to densities defined over planar domains with complex shapes, characterized by strong concavities or interior holes. This type of complexity is handled using a nonparametric likelihood approach combined with a regularization involving partial differential operators. This method gives great flexibity and allows the estimation over domains that influence the behavior of the distribution. To solve the estimation problem, we resorted on numerical techniques, and in particular we used a finite element method. This approach makes the optimization computationally tractable even in high dimensional scenarios. The proposed method compared favorably with state of the art density estimators.

In Chapter 3, we considered various nonparametric two sample tests. In the same setting of the previous chapter, we focused on densities defined over complex domains. Using a permutational approach, we defined the tests as nonparametric combinations based on high dimensional multinomials. These procedures showed promising results and led to satisfactory asymptotic properties. With appropriate partitions of the domain, the proposed methodology can be used for hypothesis testing in any dimension.

In Chapter 4, we started with the development of the asymptotic properties of the

SR-PDE class. We obtained consistency and asymptotic normality of the discrete estimators under general conditions. We then developed two nonparametric tests for the linear components of such models. The randomization procedure used for the tests showed encouraging results, even in cases where strong spatial correlation is present. This peculiarity is extremely useful in applied spatial regression models with covariates.

Future directions of research

The procedures developed in this thesis take an initial step towards addressing important questions for complex spatial data, but several exciting areas remain still open for future researches. In the context of density estimation, a first interesting direction is to consider higher dimensional and non-euclidean domains. These include the cases of curved surfaces with non-trivial geometries, and three-dimensional domains with complex boundaries. These features are important in modern applications in fields such as neuroscience, where the data are observed over a domain characterized by the formidably complicated morphology, the brain (Lila et al., 2016a). Geo-sciences, where the spatio-temporal dependence and the presence of complex boundaries have to be considered (Bernardi et al., 2017). Astronomy, such as cosmic web reconstruction (Chen et al., 2015), where flexibile estimators are needed to identify and characterize intrinsic lower dimensional structures, such as smooth manifolds. Density estimation over complicated multidimensional domains requires flexible methods able to overcome the classical concept of Euclidean distance. Some proposals generalize the kernel density estimation to Riemaniann manifolds, using the concept of exponential map to solve the problem (Kim and Park (2013); Berry and Sauer (2017)). In our setting, the flexibility of FEM allows to define discretizations of curved surfaces and also higher dimensional structures. The proposed method can therefore be generalized to these types of domains. In these cases, more advanced theoretical tools may be required to study the inferential properties of the estimators, driving more challenging and more exciting researches. Another appealing direction concerns the development of time-dependent density estimation methods. The modeling of densities over time permits the understanding of the evolution of underlying processes generating the data. However, very little attention has been devoted to the development of such models, especially in more than one dimension. Finally, a fascinating alternative is to tell the whole story from a Bayesian perspective. The penalization can indeed be associated with prior over a graph, the triangulation. In particular, a Gaussian prior with a covariance structure induced by the specified form

of the partial differential operator. The estimation procedure will follow an equivalent approach, namely a maximum a posteriori solution. This may lead to interesting considerations in terms of random processes, especially in the case of Poisson intensity estimation.

The possible directions for two samples hypothesis testing are pretty much aligned with the ones for density estimation. The development of two samples testing procedures for higher dimensional domains, and in particular for non-euclidean surfaces seems very interesting. Closely related to the development of time-dependent density estimators is the study of sequential two testing procedures. Modern applications often include a time dimension that has to be properly considered. In neuroscience for example, it is of interest to study possible changes in brain activity over times or during different tasks. With regard to this topics, a huge number of new researches is devoted to change points analysis (see for example Cunen et al. (2018); Chen (2019); Gao et al. (2019)). In the case of spatial processes that develop in time, another possibility is to define local tests, i.e. tests that focuses on subset of the domain. This possibility is related to the study of multi-scale tessellations, characterized by a finer discretization where the mass of the distribution is concentrated. This may involve the definition of appropriate two stage procedures.

Moving to hypothesis testing in the case of regression models, a first problem to address is a more thorough study of the asymptotic properties. This may involve group theory methods, used for example in Hoeffding (1952), and more recently by Chung and Romano (2013) and Hemerik and Goeman (2018). Moreover, it is of interest to extend the proposed methodology to the generalized SR-PDE (Wilhelm and Sangalli, 2016). Again, the first step is to study the asymptotic properties of the corresponding estimators. Similar results in the case of other semiparametric regression models are given, for example, in Wood (2012) and Marra and Wood (2012). The second step is to develop nonparametric hypothesis testing, possibly considering the sign-flipping procedure proposed. The work of Hemerik et al. (2019), that develops the theory in the case of classic GLM, may constitute the starting point for a generalization to penalized models. Encouraging results may also drive the extension of the hypothesis testing methodology to more general semiparametric models (for example Wood et al. (2008) and Wood (2003)).

Appendix

.1 Proof of Theorem 2.1

The proof of Theorem 2.1 relies on the following two lemmas.

Lemma .1. *The functional* $J(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int_\Omega \exp(g)$ *is continuous and strictly convex in* V*.*

Proof. The continuity of J is obvious since the first term is linear and both the exponential and the integral are continuous operators. Let now $g_1,\ g_2 \in V$, $\gamma \in [0,1]$ and $g = \gamma g_1 + (1-\gamma)g_2$. We have to show that $J(g) \le \gamma J(g_1) + (1-\gamma)J(g_2)$ and that the equality holds only if $g_1 = g_2$. We have:

$$\begin{aligned} J(g) &= J(\gamma g_1 + (1-\gamma)g_2) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\{\gamma g_1(X_i) + (1-\gamma)g_2(X_i)\} + \int_\Omega \exp(\gamma g_1 + (1-\gamma)g_2) \\ &= \gamma\left\{-\frac{1}{n}\sum_{i=1}^{n} g_1(X_i)\right\} + (1-\gamma)\left\{-\frac{1}{n}\sum_{i=1}^{n} g_2(X_i)\right\} \\ &\qquad + \int_\Omega \exp(\gamma g_1)\exp((1-\gamma)g_2)). \end{aligned}$$

Using Holder's inequality with $p = 1/\gamma$ and $q = 1/(1-\gamma)$ we have

$$\int_\Omega \exp(\gamma g_1)\exp((1-\gamma)g_2)) \le \left\{\int_\Omega \exp(g_1)\right\}^{\gamma}\left\{\int_\Omega \exp(g_2)\right\}^{1-\gamma}.$$

Moreover, using Young's inequality with the same p and q we have

$$\left\{\int_\Omega \exp(g_1)\right\}^{\gamma}\left\{\int_\Omega \exp(g_2)\right\}^{1-\gamma} \le \gamma\left\{\int_\Omega \exp(g_1)\right\} + (1-\gamma)\left\{\int_\Omega \exp(g_2)\right\}.$$

This leads to $J(g) \le \gamma J(g_1) + (1-\gamma)J(g_2)$.

It remains to show that the equality holds if and only if $g_1 = g_2$. In Holder's inequality, the equality holds only if there exists $a, b \neq 0$ such that

$$a \exp(g_1) = b \exp(g_2) \qquad \Leftrightarrow \qquad g_1 = g_2 + \log(b/a).$$

Moreover, in Young's inequality, the equality holds only when

$$\int_\Omega \exp(g_1) = \int_\Omega \exp(g_2).$$

Substituting $g_1 = g_2 + \log(b/a)$ in the equation above, we get $a = b$; this in turn implies $g_1 = g_2$. Thus J is strictly convex in V. □

Let now V_0 denote the null space of the Laplacian in V, i.e., $V_0 = \{g \in V : \|\Delta g\|_{L^2} = 0\}$. Let V_Δ denote the complementary space of V_0 in V, i.e., $V = V_0 \oplus V_\Delta$, where $\oplus$ denotes the direct sum.

Lemma .2. *V_0 is of finite dimension. Moreover $\|\Delta \cdot\|_{L^2}$ is a norm in the space V_Δ, equivalent to the H^2 norm.*

Proof. Jet $g_0 \in V_0$. The g_0 is a solution of the differential equation

$$\begin{cases} \Delta g = 0 & \text{in } \Omega \\ \dfrac{\partial g}{\partial \nu} = 0 & \text{on } \partial\Omega \end{cases}$$

This implies that g_0 is a constant function over Ω, that is, $V_0 = \{g : \Omega \to \mathbb{R} : g = c,\, c \in \mathbb{R}\}$. Thus V_0 is a finite dimensional space.

It remains to prove that $\|\Delta \cdot\|_{L^2}$ and $\|\cdot\|_{H^2}$ are equivalent in V_Δ. By definition of the H^2 norm, we have that, for all $g \in H^2(\Omega)$,

$$\|\Delta g\|_{L^2}^2 \leq \|g\|_{H_2}^2 .$$

In addition, for all $g \in V$,

$$\|g\|_{H_2} \leq C\{\|g\|_{L^2} + \|\Delta g\|_{L^2}\}.$$

Since we can always write $g = c + \tilde{g}$, with $c \in \mathbb{R}$ and $\|\tilde{g}\|_{L^2} = 0$, then, for each $\tilde{g} \in V_\Delta$, we have

$$\|\tilde{g}\|_{H_2} \leq C \|\Delta \tilde{g}\|_{L^2} .$$

□

Thanks to Lemma .1 and Lemma .2, we can leverage on Theorem 4.1 of Gu and Qiu (1993). Thanks to this theorem we have that functional $L(g)$ in (2.1) has a unique minimizer in V if and only if $-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int_\Omega \exp(g)$ has a minimizer in V_0. The latter condition is verified since V_0 is the space of constant functions. This concludes the proof that the functional $L(g)$ in (2.1) has a unique minimizer in V.

Bibliography

Adams, R. A. (1975) Sobolev spaces. 1975.

Arias-Castro, E., Pelletier, B. and Saligrama, V. (2018) Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. Journal of Nonparametric Statistics **30**(2), 448–471.

Arnone, E. (2018) Regression with pde penalization for modelling functional data with spatial and spatio-temporal dependence. PhD thesis .

Arnone, E., Azzimonti, L., Nobile, F. and Sangalli, L. M. (2019) Modeling spatially dependent functional data via regression with differential regularization. Journal of Multivariate Analysis **170**, 275–295.

Azzimonti, L., Nobile, F., Sangalli, L. M. and Secchi, P. (2014) Mixed finite elements for spatial regression with pde penalization. SIAM/ASA Journal on Uncertainty Quantification **2**(1), 305–335.

Azzimonti, L., Sangalli, L. M., Secchi, P., Domanin, M. and Nobile, F. (2015) Blood flow velocity field estimation via spatial regression with pde penalization. Journal of the American Statistical Association **110**(511), 1057–1071.

Balakrishnan, S. and Wasserman, L. (2018) Hypothesis testing for high-dimensional multinomials: A selective review. The Annals of Applied Statistics **12**(2), 727–749.

Bernardi, M. S., Sangalli, L. M., Mazza, G. and Ramsay, J. O. (2017) A penalized regression model for spatial functional data with application to the analysis of the production of waste in venice province. Stochastic Environmental Research and Risk Assessment **31**(1), 23–38.

Berry, T. and Sauer, T. (2017) Density estimation on manifolds with boundary. Computational Statistics & Data Analysis **107**, 1–17.

Botev, Z. I., Grotowski, J. F. and Kroese, D. P. (2010) Kernel density estimation via diffusion. The annals of Statistics **38**(5), 2916–2957.

Buja, A., Berk, R. A., Brown, L. D., George, E. I., Pitkin, E., Traskin, M., Zhao, L. and Zhang, K. (2015) Models as approximations-a conspiracy of random regressors and model deviations against classical inference in regression. Statistical Science p. 1.

Butcher, J. C. (2016) Numerical methods for ordinary differential equations. John Wiley & Sons.

Carando, D., Fraiman, R. and Groisman, P. (2009) Nonparametric likelihood based estimation for a multivariate lipschitz density. Journal of Multivariate Analysis **100**(5), 981–992.

Chacón, J. E. (2015) A population background for nonparametric density-based clustering. Statistical Science **30**(4), 518–532.

Chan, S.-O., Diakonikolas, I., Valiant, P. and Valiant, G. (2014) Optimal algorithms for testing closeness of discrete distributions. In Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms, pp. 1193–1203.

Chaudhuri, P. and Marron, J. S. (1999) Sizer for exploration of structures in curves. Journal of the American Statistical Association **94**(447), 807–823.

Chen, H. (2019) Sequential change-point detection based on nearest neighbors. The Annals of Statistics **47**(3), 1381–1407.

Chen, Y.-C., Ho, S., Freeman, P. E., Genovese, C. R. and Wasserman, L. (2015) Cosmic web reconstruction through density ridges: method and algorithm. Monthly Notices of the Royal Astronomical Society **454**(1), 1140–1156.

Chung, E. and Romano, J. P. (2013) Exact and asymptotically robust permutation tests. The Annals of Statistics **41**(2), 484–507.

Chung, M. K., Hanson, J. L. and Pollak, S. D. (2016) Statistical analysis on brain surfaces. Handbook of Neuroimaging Data Analysis p. 233.

Ciarlet, P. G. (2002) The finite element method for elliptic problems. Volume 40. Siam.

Coeurjolly, J.-F. and Møller, J. (2014) Variational approach for spatial point process intensity estimation. Bernoulli **20**(3), 1097–1125.

Cressie, N. (1992) Statistics for spatial data. Terra Nova **4**(5), 613–617.

Cule, M., Samworth, R. and Stewart, M. (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(5), 545–607.

Cunen, C., Hermansen, G. and Hjort, N. L. (2018) Confidence distributions for change-points and regime shifts. Journal of Statistical Planning and Inference **195**, 14–34.

Cuthill, E. and McKee, J. (1969) Reducing the bandwidth of sparse symmetric matrices. In Proceedings of the 1969 24th national conference, pp. 157–172.

Devroye, L. and Györfi, L. (1985) Density estimation: The l1 view.

Diggle, P. J., Moraga, P., Rowlingson, B. and Taylor, B. M. (2013) Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. Statistical Science pp. 542–563.

Dijkstra, E. W. (1959) A note on two problems in connexion with graphs. Numerische mathematik **1**(1), 269–271.

Dubois, G., Malczewski, J. and De Cort, M. (2003) Mapping radioactivity in the environment: spatial interpolation comparison 97. Office for Official Publications of the European Communities.

Duong, T. (2018) ks: Kernel Smoothing. R package version 1.11.3.

Fienberg, S. E. and Holland, P. W. (1973) Simultaneous estimation of multinomial cell probabilities. Journal of the American Statistical Association **68**(343), 683–691.

Flaxman, S., Teh, Y. W. and Sejdinovic, D. (2017) Poisson intensity estimation with reproducing kernels. Electronic Journal of Statistics **11**(2), 5081–5104.

Gao, Z., Shang, Z., Du, P. and Robertson, J. L. (2019) Variance change point detection under a smoothly-changing mean trend with application to liver procurement. Journal of the American Statistical Association **114**(526), 773–781.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2014) Nonparametric ridge estimation. The Annals of Statistics **42**(4), 1511–1545.

Gervini, D. (2019) Doubly stochastic models for replicated spatio-temporal point processes. arXiv preprint arXiv:1903.09253 .

Good, I. and Gaskins, R. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. Journal of the American Statistical Association **75**(369), 42–56.

Gu, C. and Qiu, C. (1993) Smoothing spline density estimation: Theory. The Annals of Statistics pp. 217–234.

Guan, Y. and Shen, Y. (2010) A weighted estimating equation approach for inhomogeneous spatial point processes. Biometrika **97**(4), 867–880.

Hales, C. and Reese, M. (2013) Portland police bureau statistical report.

Hemerik, J. and Goeman, J. (2018) Exact testing with random permutations. TEST **27**(4), 811–825.

Hemerik, J., Goeman, J. and Finos, L. (2019) Robust testing in generalized linear models by sign-flipping score contributions. arXiv:1909.03796 .

Hjelle, Ø. and Dæhlen, M. (2006) Triangulations and applications. Springer Science & Business Media.

Hoeffding, W. (1952) The large-sample power of tests based on permutations of observations. The Annals of Mathematical Statistics **23**(2), 169–192.

Hotelling, H. (1951) A generalized t test and measure of multivariate dispersion. In Proceedings of the second Berkeley symposium on mathematical statistics and probability.

Ingster, Y. I. (1987) Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. Theory of Probability & Its Applications **31**(2), 333–337.

Kim, Y. T. and Park, H. S. (2013) Geometric structures arising from kernel density estimation on riemannian manifolds. Journal of Multivariate Analysis **114**, 112–126.

Kolmogorov, A. (1933) Sulla determinazione empirica di una lgge di distribuzione. Inst. Ital. Attuari, Giorn. **4**, 83–91.

Lai, M.-J. and Schumaker, L. (2007) Spline functions on triangulations. Volume 110. Cambridge University Press.

Lange, K. (2013) Optimization. Springer Texts in Statistics. Springer New York. ISBN 9781461458388.

Lehmann, E. L. and Romano, J. P. (2006) Testing statistical hypotheses. Springer Science & Business Media.

Lila, E., Aston, J. A. and Sangalli, L. M. (2016a) Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. The Annals of Applied Statistics **10**(4), 1854–1879.

Lila, E., Sangalli, L. M., Ramsay, J. and Formaggia, L. (2016b) fdaPDE: Functional Data Analysis and Partial Differential Equations; Statistical Analysis of Functional and Spatial Data, Based on Regression with Partial Differential Regularizations. R package version 0.1-4.

Marra, G. and Wood, S. N. (2012) Coverage properties of confidence intervals for generalized additive model components. Scandinavian Journal of Statistics **39**(1), 53–74.

Marron, J. (1987) A comparison of cross-validation techniques in density estimation. The Annals of Statistics **15**(1), 152–162.

Menafoglio, A., Gaetani, G. and Secchi, P. (2018) Random domain decompositions for object-oriented kriging over complex domains. Stochastic Environmental Research and Risk Assessment **32**(12), 3421–3437.

Nesterov, Y. (2018) Lectures on convex optimization. Volume 137. Springer.

Niu, M., Cheung, P., Lin, L., Dai, Z., Lawrence, N. and Dunson, D. (2019) Intrinsic gaussian processes on complex constrained domains. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **81**(3), 603–627.

Pesarin, F. and Salmaso, L. (2010) Permutation tests for complex data: theory, applications and software. John Wiley & Sons.

Quarteroni, A. and Quarteroni, S. (2009) Numerical models for differential problems. Volume 2. Springer.

Quarteroni, A., Sacco, R. and Saleri, F. (2010) Numerical mathematics. Volume 37. Springer Science & Business Media.

Ramsay, T. (2002) Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(2), 307–319.

Samworth, R. J. (2018) Recent progress in log-concave density estimation. Statistical Science **33**(4), 493–509.

Sangalli, L. M., Ramsay, J. O. and Ramsay, T. O. (2013) Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75**(4), 681–703.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M. and Strokorb, K. (2015) Analysis, simulation and prediction of multivariate random fields with package RandomFields. Journal of Statistical Software **63**(8), 1–25.

Scott, D. W. (2015) Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons.

Scott-Hayward, L., MacKenzie, M., Donovan, C., Walker, C. and Ashe, E. (2014) Complex region spatial smoother (cress). Journal of Computational and Graphical Statistics **23**(2), 340–360.

Shen, X. (1997) On methods of sieves and penalization. The Annals of Statistics **25**(6), 2555–2591.

Silverman, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. The Annals of Statistics pp. 795–810.

Smirnov, N. V. (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bull. Math. Univ. Moscou **2**(2), 3–14.

Tsiatis, A. (2007) Semiparametric theory and missing data. Springer Science & Business Media.

Waagepetersen, R. and Guan, Y. (2009) Two-step estimation for inhomogeneous spatial point processes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **71**(3), 685–702.

Wald, A. and Wolfowitz, J. (1940) On a test whether two samples are from the same population. The Annals of Mathematical Statistics **11**(2), 147–162.

Wand, M. P. and Jones, M. C. (1994) Kernel smoothing. Crc Press.

Wang, H. and Ranalli, M. (2007) Low-rank smoothing splines on complicated domains. Biometrics **63**(1), 209–217.

Wilhelm, M. and Sangalli, L. M. (2016) Generalized spatial regression with differential regularization. Journal of Statistical Computation and Simulation **86**(13), 2497–2518.

Wood, S. N. (2003) Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65**(1), 95–114.

Wood, S. N. (2012) On p-values for smooth components of an extended generalized additive model. Biometrika **100**(1), 221–228.

Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008) Soap film smoothing. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(5), 931–955.

Yu, Y. and Ruppert, D. (2002) Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association **97**(460), 1042–1054.

Federico Ferraccioli

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.
e-mail: ferraccioli@stat.unipd.it

Current Position

November 2019 – October 2021
Postdoctoral research fellow, University of Padova.
Supervisor: Prof. Giovanna Menardi

Research interests

- *Density estimation*
- *Nonparametric methods*
- *Clustering*
- *High dimensional models*
- *Penalized models*
- *Hypothesis testing*
- *Object Oriented Data Analysis*

Education

October 2016 – September 2019
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Nonparametric methods for complex spatial domains: density estimation and hypothesis testing
Supervisor: Prof. Laura M. Sangalli
Co-supervisor: Prof. Livio Finos.

October 2014 – September 2016
Master (*laurea specialistica/magistrale***) degree in Statistical Sciences**
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Topic modeling, behind the scenes: directed and undirected graphical models."
Supervisor: Prof. Livio Finos
Final mark: 110/110 cum laude

October 2011 – July 2014
Bachelor degree (*laurea triennale***) in Statistics, Economics and Finance**.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Topic model workout: un approccio per l'analisi di microblogging, mass media e dintorni."
Supervisor: Prof. Livio Finos
Final mark: 107/110.

Visiting periods

March 2018 – May 2019
University of Oslo,
Oslo, Norway.
Supervisor: Prof. Nils Lid Hjort

October 2018 – September 2019
Politecnico di Milano,
Milano, Italy.
Supervisor: Prof. Laura M. Sangalli

Further education

July 2018 – August 2018
Numerical Analysis
Scuola Matematica Interuniversitaria 2018 - Perugia
Instructor: Arieh Iserles (University of Cambridge)

July 2018 – August 2018
Commutative Algebra and Geometry
Scuola Matematica Interuniversitaria 2018 - Perugia
Instructor: Juan Migliore (University of Notre Dame)

Work experience

June 2009 – July 2009
Internship.
MyS S.r.l, business intelligence and consulting company.

Computer skills

- *R*
- *Python*
- *C/C++*
- *LaTeX*

Language skills

Italian: native; English: fluent.

Publications

Articles in journals and proceedings

Ferraccioli, F., Sangalli, L.M., Finos, L. (2018) Bounded Domain Density Estimation. Book of Short Papers SIS2019, Proceedings of the 50th Scientific meeting of the Italian Statistical Society, Milano.

Finos, L., Hemerik, J., Ferraccioli, F., Sangalli, L.M., Goeman, J.J. (2018) Robust testing of generalized spatial regression models by conditional score test. Book of Abstracts, ISNPS 2018 4th Conference of the International Society for Nonparametric Statistics, Salerno.

Ferraccioli, F., Sangalli, L.M., Finos, L. Density estimation via nonparametric penalized likelihood. Book of Abstracts, ISNPS 2018 4th Conference of the International Society for Nonparametric Statistics, Salerno.

Matozzo, V., Fabrello, J., Masiero, L., Ferraccioli, F., Finos, L., Pastore, P., Bogialli, S. (2018). Ecotoxicological risk assessment for the herbicide glyphosate to non-target aquatic species: A case study with the mussel Mytilus galloprovincialis. Environmental Pollution, 233, 623-632.

Ferraccioli, F. , Sangalli, L. M., Finos, L. (2018). Nonparametric penalized likelihood for density estimation, Book of Short Papers SIS2018, Proceedings of the 49th Scientific meeting of the Italian Statistical Society, Palermo.

Ferraccioli, F. , Finos, L. (2017). Exponential family graphical models and penalizations, Book of Short Papers SIS2017, Proceedings of the 48th Scientific meeting of the Italian Statistical Society, Firenze.

Chapters in books

Cappozzo, A., Ferraccioli, F., Stefanucci, M., & Secchi, P. (2017). "An Object Oriented Approach to Multimodal Imaging Data in Neuroscience." START UP RESEARCH. Springer, Cham, 2017.

Working papers

Ferraccioli, F. , Sangalli, L. M., Finos, L. Nonparametric density estimation with differential regularization. (submitted)

Ferraccioli, F. , Sangalli, L. M., Finos, L. Testing the effect of covariates in semiparametric regression models.

Ferraccioli, F. , Sangalli, L. M., Finos, L. Two sample hypothesis testing for complex regions.

Conference presentations

Nonparametric likelihood density estimation on bounded domains. EMS 2019, University of Palermo. July 2019

Analysis of Data Over Complex Regions (invited). GRASPA 2019, University of Pescara. July 2019.

Bounded domain density estimation SIS 2019 Università Cattolica del Sacro Cuore, Milano. June 2019.

Density estimation via nonparametric penalized likelihood. ISNPS 2018, University of Salerno. June 2018.

Nonparametric penalized likelihood for density estimation. SIS 2018, University of Palermo. June 2018.

Nonparametric density estimation with differential regularization (invited). ISA Workshop, "Frontiers in Functional Data Analysis", University College Dublin. March 2018.

Exponential family graphical models. SIS 2017. University of Firenze. July 2017.

Teaching experience

March 2019
Statistica Iterazione, Master degree.
Specialized lecture, "Classification models for text mining".
Università degli Studi di Padova.
Instructor: Prof. Guido Masarotto.

Other Interests

I studied at "F. Venezze" Conservatory (Rovigo) for 7 years, classical double bass and music theory. I spend my free time reading and listening to podcasts about different topics, ranging from physical sciences, philosophy and technology. I love music and visual arts, with a particular predilection for architecture (early career dream).

References

Prof. Laura M. Sangalli
Laboratory for Modeling and Scientific Computing MOX, Dipartimento di Matematica
Politecnico di Milano
Piazza L. da Vinci, 32
20133 Milano (Italy)
phone: +39 02 23994554
email: laura.sangalli(at)polimi.it

Prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione
Università degli Studi di Padova
Via Venezia, 8
35131 Padova (Italy)
Phone: +39 049 8276519
e-mail: livio.finos(at)unipd.it