CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESE DE DOUTORAMENTO

Nonparametric inference on point processes with covariates

María Isabel Borrajo García

ESCOLA DE DOUTORAMENTO EN CIENCIAS E TECNOLOXÍAS DA USC

PROGRAMA DE DOUTORAMENTO EN ESTATÍSTICA E I.O.

SANTIAGO DE COMPOSTELA

2018

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

DECLARACIÓN DO AUTOR DA TESE

"Nonparametric inference on point processes with covariates"

Dna. María Isabel Borrajo García

presento a miña tese, seguindo o procedemento adecuado ao Regulamento, e declaro que:

1. A tese abarca os resultados da elaboración do meu traballo.
2. No seu caso, na tese faise referencia as colaboracións que tivo este traballo.
3. A tese é a versión definitiva presentada para a súa defensa e coincide ca versión enviada en formato electrónico.
4. Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.

En Santiago de Compostela, a 11 de decembro de 2017

Asdo. María Isabel Borrajo García

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

AUTORIZACIÓN DOS DIRECTORES DA TESE

"Nonparametric inference on point processes with covariates"

D. Wenceslao González Manteiga e Dna. María Dolores Martínez Miranda

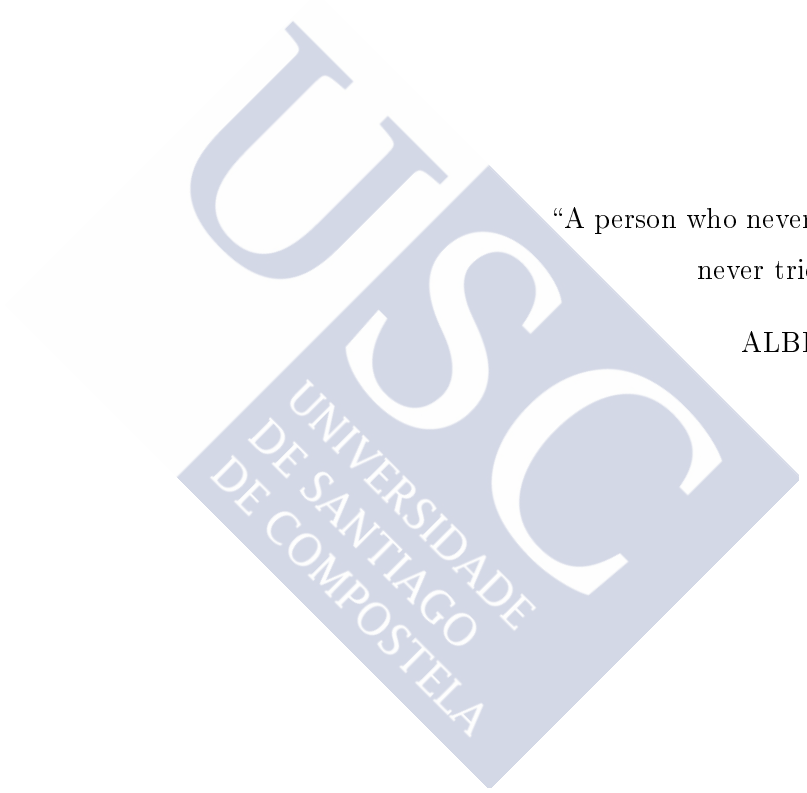
INFORMAN:

que a presente tese correspóndese co traballo realizado por Dna. María Isabel Borrajo García baixo a nosa dirección, que autorizamos a súa presentación considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC e que como directores desta, non incorre nas causas de abstención establecidas na Lei 40/2015.

En Santiago de Compostela, a 11 de decembro de 2017

Asdo. Wenceslao González Manteiga Asdo. María Dolores Martínez Miranda

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA



"A person who never made a mistake,
never tried anything new"

ALBERT EINSTEIN

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Agradecimientos

Quiero comenzar agradeciendo a mis directores de tesis, Wenceslao González Manteiga y María Dolores Martínez Miranda el apoyo recibido y el trabajo realizado estos años. Wences, gracias por haber sabido ver en mí un potencial para hacer investigación, la confianza depositada cuando era una recién licenciada con mi primer contrato de trabajo, y el apoyo y tesón a lo largo de esta etapa. Gracias Lola por haberte incorporado a esta aventura, por tu apoyo, por haberme escuchado y haber intentado entenderme, por los valiosos consejos y por los cálidos recibimientos siempre que nos hemos encontrado.

Gracias a los profesores Rosa Crujeiras y Alberto Rodríguez Casal, vuestro impulso al comienzo de esta etapa fue imprescindible para que hoy yo pueda estar escribiendo estas líneas. Gracias también a los profesores César Sánchez Sellero y José Manuel Prada Sánchez por acompañarme en mi iniciación a la labor docente, sabiendo aceptar y lidiar con los fallos del que comienza en una nueva actividad. Gracias Julia por hacer menos complicada la burocracia que siempre nos envuelve. Agradecer a María los mil arreglos que nos hace en el departamento. Bea, gracias por ese montón de dudas y cuestiones que derivaron en conversaciones muy agradables, gracias por los cafés, los paseos y las buenas noticias que, aún a veces en la distancia, las nuevas tecnologías nos han permitido compartir. Gracias Isa por estar siempre dispuesta a comentar mis mil y una dudas. Gracias a todos los compañeros del departamento de Estadística e Investigación Operativa en particular, y de la Facultad de Matemáticas en general por los momentos compartidos estos años, y gracias especialmente emotivas y cariñosas a mis queridos compañeros de despacho; sin los debates, el papeleo, las "frases de pizarra", las dudas, las risas, algún vídeo de felicitación y la ayuda desinteresada, estos años de doctorado no tendrían ahora un resumen tan positivo.

I would like to thank Prof. Serge Guillas, who gave me the opprotunity to do my research in a different environment, knowing another way of doing the job I love and learning many new things. I also acknowledge the fantastic welcome and support from my workmates at UCL, who made my stay in London far more easy, positive and memorable in the best way possible; special thanks to Alexia, Marta, Anna, Ares and Theo. Thank you also for the nice welcomings every time I have been back to London.

Gracias mamá, papá, padrino, madrina y Ana por todo lo que habéis hecho siempre, porque en gran medida por ello soy lo que soy ahora. Gracias por aguantar estoicamente los peores momentos y por compartir conmigo las más grandes alegrías, incluso cuando no siempre las entendíais. Gracias por vuestro cariño y confianza, por el apoyo incondicional, por dejarme tomar mis propias decisiones, animándome siempre y arropándome cuando no conseguía mis objetivos.

A mis amigos, esa familia que escogemos, tanto los que os habéis quedado con el paso de los años como los que en algún momento han formado parte de mi realidad y se han marchado, gracias por contribuir a mi formación como persona. Gracias un poco más especiales, y sobre todo muy merecidas a Inés, Yésica, Ana, Silvia, Mercedes, Lety, Carol, Lu, Carmen, Isiña y Vero. Todas y cada una de vosotras me habéis regalado un pedacito de lo mejor que tenéis, y os habéis agenciado ya un rincón especial en mi vida; gracias por no haber salido corriendo en los momentos complicados, por haberme tirado de las orejas cuando era necesario, por ayudarme a comprender aquello que yo no era capaz de digerir y por escucharme cuando lo que decía no eran más que tonterías producto de un mal rato.

En definitiva, gracias a todos por haber sabido estar durante esta etapa. Y confesaros que este trabajo que resume de alguna forma unos cuantos años de mi vida, nunca habría sido posible sin vosotros, y por ello, aunque sea yo quien lo firme, también os pertenece.

Santiago de Compostela, diciembre de 2017
Maribel

Preface

Spatial statistics is a big field in statistics, concerned with analysing any random process with a spatial component. This analysis of the spatial structure includes not only the studying of the locations associated to a process using their topological, geometrical or geographical properties, but also the determination of the second order or dependency structures.

Historically, the first mention to something related to what is nowadays called spatial statistics was in Halley (1686) where the author tried to find the cause for some winds and monsoons around the tropics by drawing their "locations" and direction onto a map. We have to come a couple of centuries forward, up to 1854, to find a successful but still rudimentary application of spatial statistics: by that moment there was a cholera outbreak with more than a hundred deaths in the Soho (London, UK) which origin couldn't be found. Dr. Jonh Snow, by speaking to local residents and mapping the cholera cases was able to deduce that the source of contamination was a public water pump on Broad street. Later, in the early 20^{th} century, Student (1907) was interested on the distribution of particles inside a fluid, and he decided to aggregate data in counts and compute the number of particles per unit volume, which we will see later is a herald of the concept of intensity. Fisher (1935) was probably the first one aware about spatial dependence in an agriculture study declaring "after choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart."

In spatial statistics we basically observe of a random value at a certain location; formally we have a random field or random process

$$\{Q(x) : x \in D \subset \mathbb{R}^d\}, \tag{P.1}$$

where D is a set of indexes, $Q(x)$ is any potential datum at a spatial location x and $d \in \mathbb{N}$.

Depending on the properties of both, location, s, and measure, Q, we have different type of scenarios that yield to three distinguished subfields: geostatistics, lattice data and point processes.

- **Geostatistics:** emerged in the early 1980s and studies processes that vary continuous over space, but are measured only at discrete locations; hence D in (P.1) is predetermined and continuous. Common examples are meteorological and air pollution data. In this context the main goals are to determine a spatial pattern, modelling correlation/covariance, make predictions and testing whether there exists a spatial structure or not.

- **Lattice data:** also known as areal data, studies measures (continuous or discrete) associated with areas (that may be regular such as grids or irregular such as geopolitical divisions); hence D in (P.1) is fixed and discrete. In this case there is no possibility of measure "between" data locations and the fact of aggregating the data reduces the available information.

- **Point processes:** are characterized for being random values measured at random locations; hence D in (P.1) is a random set of indexes. We are interested in determining their spatial pattern (if it exists) as well as other important characteristics such as second order dependence, the inclusion of a temporal component in the spatial process or the influence of some external information in the form of marks and/or covariates.

Remark that since the behaviour of the three main types of spatial processes is not the same, then the techniques and methodology used in each of them is also different. Even though, the general aim of spatial statistics in its three branches is to model the process underlying the observed data in order to be able to fully understand its behaviour.

The applications of spatial statistics cover a wide range of areas. Geostatistics are for example applied in hydrogeology, environmental sciences (analysing pollution data measured at fixed locations), climatology (variables like temperature, rain or wind measured at stations) and geology surveys, see Cressie (1993) and Isaaks and Srivastava (1989). Examples of lattice data where spatial techniques has been used are attributes collected by ZIP code, census tract or remotely sensed data reported by pixels; lattice data also appear in public statistics surveys where they used to study economic and social factors by regions, see Schabenberger and Gotway (2017). As we will see all along this manuscript the fields of application of point processes are extensive: forestry, epidemiology, environmental sciences, crime analysis, geology, zoology, astronomy... see P. J. Diggle (2013) for a general approach.

Moreover, we should take into account that in many of these applications, particularly the ones related to the field of point processes, it may exist much more useful and interesting information than only the locations of the occurrences. This information may arise in different forms as we will see in the first chapter of this manuscript, and it can significantly contribute to the analysis of the process. Hence, it is of interest to develop tools, techniques and methodology that allow us to include

this extra information on our analyses.

The main purpose of this dissertation is to provide with a consistent and well established theoretical framework in the field of point processes with covariates, offering innovative statistical methods in estimation and testing for spatial point processes without including the study of the time component beyond some briefly detailed extensions. The manuscript is organised as follows:

Chapter 1: Introduction. In this chapter a brief overview of the theory of spatial point processes is presented. We start by defining some basic concepts in the general point process theory, followed by the main first and second order properties of a point process. Furthermore we introduce spatio-temporal point process and the idea of separability that has been widely studied in the last years. Finally we devote a section to explain different ways of collecting some extra information: marks and covariates, that may have an impact on the process.

Chapter 2: Density estimation with length-biased data. Density estimation is a classical and well-known problem that has attracted lot of interest in the literature for decades. The use of nonparametric techniques in this field started with the kernel density estimator of Parzen (1962) and Rosenblatt (1956). However, in the context of length-biased or weighted data less has been done; the contributions were basically reduced to the kernel estimator in Jones (1991) and a cross-validation bandwidth selector by Guillamón et al. (1998). In this chapter we have filled an existing gap detailing all the asymptotic developments for Jones' estimator, proposing innovative consistent bootstrap resampling methods and defining several new data-driven bandwidth selectors with good performance, that is shown in the finite sample study we have carried out.

Note that this is the only chapter in this manuscript not involving point processes, the reason of having included it is the observable structural similarity between Jones' kernel estimator and our proposal for the intensity estimation in point processes with covariates, this similarity will remain clear in Chapter 3.

This chapter is based on Borrajo et al. (2017a).

Chapter 3: Kernel intensity estimation in point processes with covariates. The first-order intensity function is one of the most important characteristic elements in a point process. Its estimation has been of interest since the scientific community has started to study spatial data and in particular point processes, because it measures the mean number of events per unit length/area/volume. In this chapter we proposed a new estimation procedure for the intensity function in the context of point processes with covariates, based on the ideas of Guan (2008) and Baddeley et al. (2012), whose contributions were mostly limited to the establishment of the model and some practical applications, paying less attention to the theoretical developments. We obtain the

expressions for the pointwise error, its integrated and asymptotic counterparts, and we also propose a new bootstrap method that we use in the definition of the first data-driven bandwidth selectors in this context. The performance of all the proposals are shown through an exhaustive simulation study in which we also compare with the existing competitors. Finally we apply this methods to a real set of data consisting of wildfire locations in Canada during June 2015 including meteorological covariates.

This chapter is based on Borrajo et al. (2017b).

Remark that the model in this chapter is constructed for the intensity function while the one in Chapter 2 is related to the density function. Hence, for convenience and considering that there is no possible confusion between them, we have considered this two chapters independent enough to reuse some of the common notation in statistical methods, i.e, the reader will find the same letter in Chapter 2 and Chapter 3 to denote a common concept, even though in each chapter this concept may have different explicit expressions.

Chapter 4: Testing covariate significance in point processes first-order intensity. In the part of this dissertation devoted to point processes, we have embraced a framework initially defined in Baddeley et al. (2012) where the first-order intensity depends on a spatial covariate through an unknown function that needs to be estimated. This model has been assumed so far without any statistical certainty. In this chapter we propose an L^2-test statistic based on kernel estimators to determine if the dependence on a certain covariate is or not significant. These estimators are the general kernel intensity estimator without covariates proposed by P. Diggle (1985) and the kernel estimator developed in the previous chapter. We have proved the asymptotic normal distribution of the test statistic, however, the poor performance of this approximation (the convergence rate is slow) requires the adaptation of the bootstrap procedure introduced in Chapter 3 in order to better calibrate the test. All the methodology is built based on the real dataset of Murchison geological survey (gold deposits locations and distance to geological faults information) in Western Australia that had already been used in this context, and also on the wildfire dataset in Canada (wildfire locations and temperature information). The first one is finally proved to be covariate dependent, while in the second context temperature seems not to be enough to explain the wildfire distribution in the country. Moreover a simulation study with models based on real situations is carried out to show the good performance of the proposal.

This chapter is based on Borrajo et al. (2017c).

Chapter 5: Nonparametric comparison of first-order intensities in point processes with covariates. Previously in this dissertation we have presented an intensity model in spatial point processes using covariate information. Under such model we have derived an estimator of the intensity in Chapter 3, and addressed the goodness-of-fit of the model in Chapter 4. Now, under the same model, we consider the

classical two sample problem. Hence, using the theoretical framework already presented and following a similar methodology than the one developed in Chapter 4, we define a new L^2-test statistic comparing kernel estimators coming from the two samples. We prove the asymptotic normality of the statistic based on the results of Duong (2013) for multivariate densities. As for the test in Chapter 4, this asymptotic distribution performs poorly, so we use again a bootstrap procedure to improve the calibration, specially for small expected sample sizes. We carry out a simulation study to better understand the performance of this new proposal. To this goal we build two models based on the Murchison dataset and the Canadian wildfires data, previously introduced in Chapter 3 and Chapter 4. We show that our test statistic provides with good approximations of the level, and has high power values in both models.

We finally enclose a summary in Spanish and a notation index.

This work has been supported by FPU grant FPU2013/00473 from the Spanish Ministry of Education. It is acknowledged the support of the Spanish Ministry of Economy and Competitiveness through grants MTM2013-41383P and MTM2016-76969P, which include the support of the European Regional Development Fund (ERDF). IAP network P7/06 StUDyS of the Belgian Government (Belgian Science Policy) is also acknowledge for its support.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

Chapter 1

Introduction

Contents

Point process theory has its roots in the introduction, in 1837 by Siméon-Denis Poisson, of the Poisson distribution, which was derived as a passage to the limit from the binomial distribution, and it allows to predict the pattern in which random events of low probability occurs in the course of a large number of trials. However, the origins of the theory itself are not clear, and a couple of possibilities have been suggested as being on the grounds of spatial point processes: life tables, renewal theory or counting problems on particle physics. The main advances in the field of point processes have been done during the last decades of the 20th century, see Daley and Vere-Jones (1988), Moller and Waagepetersen (2003), Illian et al. (2008), P. J. Diggle (2013) and Baddeley et al. (2015) for an overview of the theoretical background and applications in this field.

Along this chapter, an introduction to point processes is developed. In Section 1.1 basic definitions, concepts and notation are introduced. Section 1.2 is devoted to some of the most important features that characterise a point process. We introduce in Section 1.3 spatio-temporal point process with its basic concepts extended from the spatial case. Finally, in Section 1.4 our attention is focused on different ways of retrieving extra useful information for the process, how to deal with each one and the main differences between them.

1.1 Basic concepts in point process theory

There are different perspectives from which address the study of point processes; we are going to approach the theory from the statistical one rather than from the point of view of measure theory, see Daley and Vere-Jones (1988) for more detail on this.

Definition 1.1. A ***point process***, X, is a random mechanism that generates locations irregularly distributed within a region of space.

Let X be a point process defined in a region $W \subset \mathbb{R}^d$, with d a nonnegative integer that is usually set as one (point process on the real line), two (spatial point process) or three (3D and spatio-temporal point process), we introduce the following concepts:

Definition 1.2. Let $\mathcal{P}(W)$ denote the parts of W, i e., the family of all possible subsets of W. Then $N : \mathcal{P}(W) \to \mathbb{Z}^+$, where $N(A)$ denote the number of occurrences of the process in every set $A \in \mathcal{P}(W)$ is defined as the ***counting measure*** associated with the process X.

Definition 1.3. Let X be a point process in $W \subset \mathbb{R}^d$ and N its associated counting measure, then a realisation (or a sample) of the process, $X_1, \ldots, X_N$, is called a ***point pattern***, and each of the points is called ***event***.

To be precise, the last subindex N should be $N(W)$ but we will for now on follow

the literature and denote it as N to ease the notation.

1.1.1 Complete spatial randomness (CSR)

Figure 1.1 provides examples of different realisations of point processes. These graphics gather the existing possibilities for a pattern in terms of "clusterisation": the one on the left is the most clustered, i.e, we can clearly distinguish groups of points that lie together; the one on the right is regular and looks like the points tend to avoid each other, and the one in the middle, where the points seem to be randomly placed and do not follow any pattern. It may occur, as in our example, that visual inspection gives a correct idea about the nature of the process, however this is not always the case and this is why we require a statistical analysis of point processes.

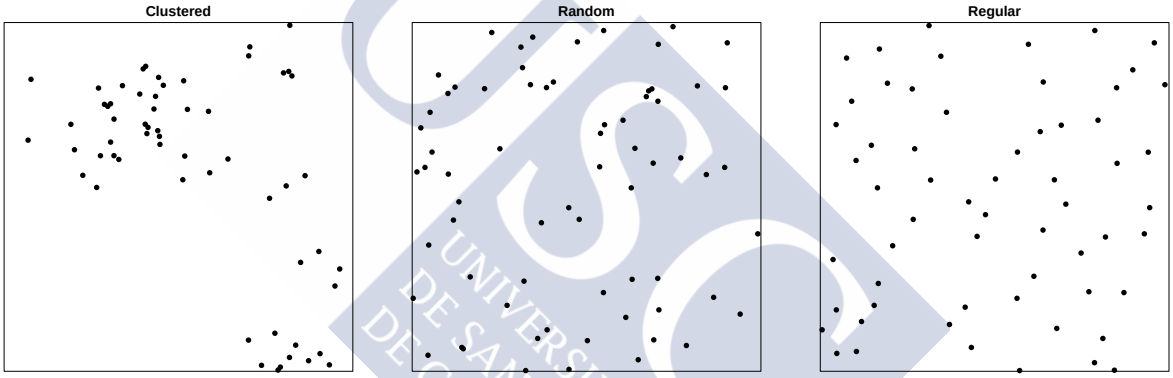


Figure 1.1: Examples of point patterns.

Let focus on the random case; in this situation the points have no preference for any spatial location (homogeneity) and the information about the outcome in one region of space has no influence on the outcome in other regions (independence). This situation is known as complete spatial randomness (CSR) and is formally characterized by the following properties:

CSR1 The number of events in a region $A \subset W$ follows a Poisson distribution with mean $\lambda|A|$, with $\lambda \in \mathbb{R}^+$, where $|A|$ denotes the measure (length, area or volume) of A.

CSR2 Conditional to $N = n$ events in W, the random variables, $X_1, \ldots, X_n$, are an independent identically distributed (i.i.d.) sample from the uniform distribution in W.

CSR represents an ideal scenario and it is of interest to check whether it holds in any practical situation.

1.1.2 Testing CSR

When a point process satisfies the CSR hypothesis, the interpretation of the pattern becomes easier. Hence, testing CSR must be one of the first steps in analysing point patterns in practice. There are, as we will see below, many different procedures to test this hypothesis, and they are all generally used as exploratory tools that provide information to formulate an appropriate model.

There exists several possible alternatives when CSR does not hold, depending on the type of the process, we will briefly describe below some of the most common ones.

- ***Nonhomogeneous Poisson processes:*** we will go into further detail in the next section, but this type of processes is used to model situations where the points are independent among them.
- ***Cox processes:*** this family of processes is used to model point patterns where there is a tendency to form clusters.
- ***Inhibition processes:*** this type is the opposite to the previous one and model situations where the points tends to repulse each other.
- ***Markov processes:*** this family serves to model both, point patterns where there is a tendency to clusterisation as well as others where points tend to avoid each other, with the particularity that the dependence among events affects only to the points lying on the neighbourhood.

Independently on the alternative or the techniques applied, most of the CSR test are constructed as follows: a suitable summary statistic is estimated from the data and compared to the corresponding one under the homogeneous Poisson assumption; if the difference is big, we confirm that there is evidence against the null hypothesis and the alternative is accepted, whereas if the difference is small we remain at the null hypothesis, i.e, CSR.

Tests based on quadrats

These tests are based on the idea of the observation region being divided in k disjoint subregions (quadrats), $A_1, \ldots, A_k$. Denote $N_1, \ldots, N_k$ the random variables representing the number of events in each quadrat. Under CSR assumption, this variables are independent with $N_i \sim Pois(\lambda|A_i|)$, hence the CSR hypothesis can be tested using a χ^2 independence test. Note that when performing this test for any point pattern we are loosing information by aggregating its points, for this reason this type of tests has been replaced by better adapted methods. An extended explanation on quadrat based tests can be consulted in Illian et al. (2008).

Tests based on distances
This family of tests relies on the distance between pairs of events or between points of the observation region and events. Some of the techniques used are the mean distance to the nearest neighbour, the distribution function of the distance to the nearest neighbour and the distribution function of the distance point-event. The idea is that, if these numerical values/functions differ from the expected ones under homogeneous Poisson hypothesis, we reject the CSR assumption, if the distances are larger than expected we will think of a regular pattern, and if they are smaller on a clustered one. See P. J. Diggle (2013) for more details.

Tests based on Voronoi's diagram and Delaunay's triangulation
These statistics are based on dividing the observation region in Voronoi cells (Delaunay triangulation is the one linking the vertex where Voronoi cells are neighbours). Under CSR assumption, we can compute some characteristics of the Voronoi cells and the Delaunay triangulation, and then compare them with the empirical versions obtained from the sample. Chiu (2003) presents different procedures with this idea.

K-test and L-test
These tests are based on this two characteristic functions that we will defined in Section 1.2.3. The idea is that this two functions have, under homogeneous Poisson model, a particular simple shape, $K(t) = \pi t^2$ and $L(t) = t$ for $t > 0$. The tests perform Monte Carlo simulations and construct confidence intervals under the null hypothesis, and then decide whether the empirical versions of K and L based on the data lie or not inside this band. For more detail on this see P. J. Diggle (2013) and Illian et al. (2008).

Test based on spacings
This theory has been mainly developed in Cucala (2006), and the idea is to construct functions of the spacings between x-coordinates and the spacings between y-coordinates, and generalise the Greenwood and Sherman statistics. The main advantage of this approach is the lower computational cost compared to other classical tests.

1.1.3 Poisson point processes

The Poisson point process is one of the most used and studied point process models due to its particularly convenient properties. There are two different types of Poisson processes: the homogeneous and the nonhomogeneous. The first one is the simplest model in point processes theory and it corresponds exactly to CSR, previously defined in the section above, and it is usually denoted by $Pois(\lambda)$ where λ is known as rate or intensity value.

The homogeneous Poisson point process satisfies some interesting properties, such as stationarity and isotropy, that we will discuss later on from a general point of

view, but also has some particularities; for example, the so called ***finite-dimensional distributions*** derived from CSR1 and CSR2; this means that, if we have $A_1, \ldots, A_k$ bounded disjoint subsets in W, then $N(A_1), \ldots, N(A_k)$ are independent Poisson random variables with means $\lambda|A_1|, \ldots, \lambda|A_k|$, respectively. Based on this property, the joint probabilities $\mathbb{P}(N(B_1) = n_1, \ldots, N(B_1) = n_1)$, with $B_k \subset W \; \forall k$, can be evaluated for general (possibly overlapped) subsets.

The applications of homogeneous Poisson point processes are very different and arise in many fields. This process is applied for example in physical sciences such as a model developed to detect alpha particles, see Stoyan et al. (1995). Recently, it has also been used by Andrews et al. (2010) in wireless networks.

The nonhomogeneous Poisson point process is an extension of the simplest one, where the rate value λ is now a function of the location, so $\lambda(x)$ with $x \in W$. It is used in many practical situations, because it allows much more freedom on the model while it still remains mathematically simple and understandable. Moreover it fits well in a lot of different contexts and situations as we will see along this manuscript, and it is also shown in different real examples in Illian et al. (2008), P. J. Diggle (2013) and Baddeley et al. (2015).

Formally, the nonhomogeneous Poisson point process is defined as follows:

Definition 1.4. Let X be a point process in $W \subset \mathbb{R}^d$ with N its associated counting measure and a spatially varying function $\lambda(x)$. Then we say that X is an nonhomogeneous Poisson point process with intensity $\lambda(x)$ if it verifies the following properties:

NHP1 $N(A)$ has a Poisson distribution with mean $\int_A \lambda(x)dx$ for any $A \subset W$.

NHP2 Conditional to $N = n$, the n events in W form an independent random sample from the distribution in W with density proportional to $\lambda(x)$.

This type of Poisson point processes have many more applications than the homogeneous one due to the freedom of the spatially varying intensity. They are used in stochastic geometry, Stoyan et al. (1995); in studies of salmon and sea life in the oceans, Krkošek et al. (2005); forestry as we have already pointed out in the introduction, and naval search problems, Lewis and Shedler (1979).

Examples of realisations of homogeneous and nonhomogeneous ($\lambda(u, v) = 190e^{-3u}$) Poisson point processes are shown in Figure 1.2; both of them are drawn in the square unit region and their mean intensity value is 60.

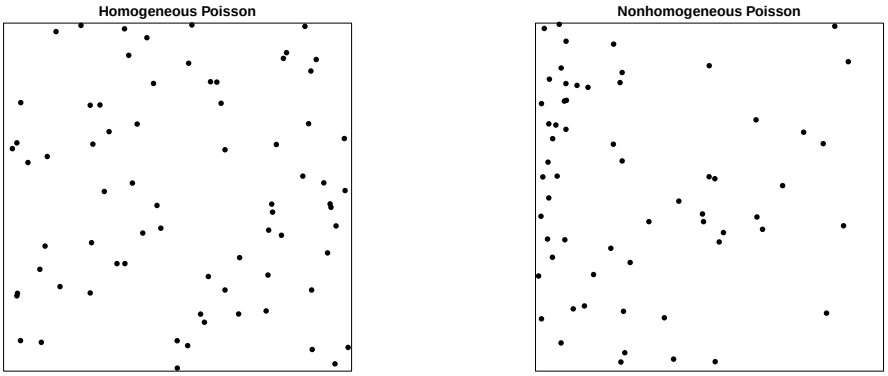


Figure 1.2: Realisations of homogeneous (left) and nonhomogeneous Poisson point processes in the square unit with mean intensity value of 60.

An important point that we have not gone into yet is the simulation of a point process, particularly the simulation of a Poisson point process in a region W. The homogeneous case is almost trivial: to simulate a pattern with intensity λ, we firstly generate a random number, n, from a Poisson distribution with mean $\lambda|W|$. Then the n points are placed independently following a uniform distribution in W.

To simulate a nonhomogeneous Poisson point process it is useful a simulation technique called ***independent thinning***, that consists of using some predefined rules to remove points of a process and form a new one. Suppose that X is a Poisson process with intensity function $\lambda(x)$, and that each point of X is either deleted or retained, independently of other points. If the retention probability is $p(x)$, then the resulting process of retained points is Poisson with intensity $\lambda(x)p(x)$. Then, to simulate a nonhomogeneous Poisson point process with intensity $\lambda(x)$ is enough to start by simulating a homogeneous one with intensity $\lambda = \max_{x \in W} \lambda(x)$ and perform a thinning with retention probabilities $p(x) = \frac{\lambda(x)}{\lambda}$. Along this manuscript all the simulations involving Poisson point processes have been performed with R Core Team (2016) using the package `spatstat`, developed by Baddeley et al. (2015).

1.2 Properties of spatial point processes

When studying a point process or when analysing a point pattern we seek for some properties that allow us to select an appropriate model. There are many possibilities attending to different aspects of the process, the most relevant ones are detailed below.

- ***Stationary:*** this means that the distribution of the point process, i.e., the distribution of any of its realisations, is invariant under translations. Formally for any regions $A_1, \ldots, A_k$, the joint distribution of $N(A_1), \ldots, N(A_k)$ is invariant

under translation of the regions by any value.

- ***Isotropic:*** in this case the distribution of the point process is invariant under rotations around the origin. Hence, given again any regions $A_1, \ldots, A_k$, the joint distribution of $N(A_1), \ldots, N(A_k)$ is invariant under rotations around the origin.

- ***Motion-invariant:*** this property gathers the two previous ones so, the point process is motion-invariant when it is simultaneously stationary and isotropic. The distribution of these processes is then invariant with respect to rotations around any arbitrary point (compositions of translations and rotations around the origin). Moreover a motion-invariant process has the same distribution as every transformed process obtained from the initial one by all rigid motions.

- ***Orderly:*** this means that the process does not contain duplicated points, i.e, $\lim_{|dx|\to 0} \frac{\mathbb{P}(N(dx)>1)}{|dx|} = 0$. This property is also referred in the literature as ***simplicity***.

- ***Second-order orderly:*** a point process is second-orderly if for each pair of points x and y satisfies $\lim_{|dx|\to 0, |dy|\to 0} \frac{\mathbb{P}(N(dx)>1)P(N(dy)>1)}{|dx||dy|} = 0$.

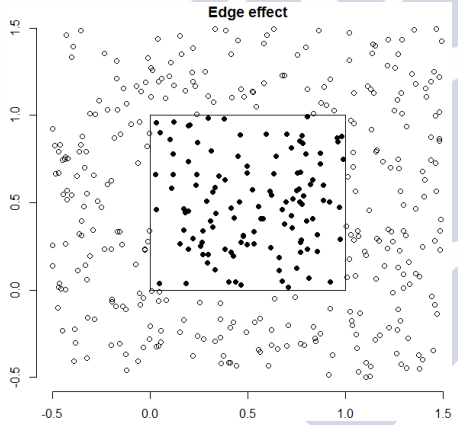


Figure 1.3: Illustrative example of the problematic related to edge effect.

There is a relevant characteristic that usually arises in point process theory, the ***edge effect***. Point processes are stochastic mechanisms that are generally observed in a bounded region, even though they may be defined in a bigger one.

Figure 1.3 illustrates this problem of the edge effect in a very simple but clarifying situation, where we see a point process existing in the square region limited between the abscissae and ordinates -0.5 and 1.5, but only observed in the square unit. Hence, when performing inference near to the boundary inside the observation region we are not taking into account the events lying in the neighbourhood but outside, and this generates an inconsistency that need to be solved.

Hence, when studying point processes we have to take care about the edge effect and try to avoid or correct it. We will see later in this manuscript some assumptions that allow us to directly avoid this effect, and when this is not possible, how edge correction terms may be introduced in the methodology.

Another characteristic describing point processes are the ***void probabilities***, defined as $\nu(A) = \mathbb{P}(N(A) = 0)$ for every $A \subset W$. It was Renyi in 1967 who proved that a point process is uniquely determined by these void probabilities.

1.2.1 First-order intensity function

The first-order intensity function, intensity from now on, is probably one of the most important characteristics of a point process. The intensity is a basic descriptive characteristic that measures the mean number of events per unit length, area or volume. Let define it from a more formal perspective:

Definition 1.5. Following the notation previously introduced in this manuscript, the ***first-order intensity function*** is defined as:

$$\lambda(x) = \lim_{|dx| \to 0} \frac{E[N(dx)]}{|dx|}, \quad x \in W,$$

where E denotes the expectation of a random variable.

A first exploratory and simple approach to intensity is to count the number of points lying in a region, and divide this value by the length, area or volume of the region. Obviously this only gives a value that is not much informative unless the process is homogeneous.

Some calculations related to the intensity may be simplified by the use of ***Campbell's theorem***:

Theorem 1.1. *Let q be a measurable function defined at least in W and $X_1, \ldots, X_N$ a point pattern lying also in W, then*

$$E\left(\sum_{i=1}^{N} q(X_i)\right) = \int_W q(x)\lambda(x)dx.$$

In the particular case of the Poisson point process, the following result is also satisfied:

Theorem 1.2. *Let q be a measurable function defined at least in W and $X_1, \ldots, X_N$ a point pattern lying also in W, then*

$$Var\left(\sum_{i=1}^{N} q(X_i)\right) = \int_W q^2(x)\lambda(x)dx.$$

1.2.2 Relationship between density and intensity function

The general idea we all have in mind is that the density is a nonnegative integrable function whose integral over its domain must be the unit. In the previous section we have given the intuitive and formal definitions of the intensity function that necessarily imply the intensity being nonnegative. Of course, as it computes the mean number of events per unit area, its integral value, which would be the number of events in the integration region, does not have to be equal to one.

The fact of intensity being nonnegative makes it almost trivial to construct a, let say, "artificial" bivariate density function just dividing it by the value of its integral:

$$f(u) = \frac{\lambda(u)}{m}, \quad u \in W,$$

where recall that $m = E[N] = \int_W \lambda(u)du$.

A similar idea of using the close relationship between density and intensity has already been used in P. Diggle and Marron (1988) and Berman and Diggle (1989) to construct, respectively, a bandwidth selector for the classical kernel intensity estimator (we will go deeper into it in Chapter 3) and for a second-order intensity estimator, respectively. Moreover, taking into account the methodology developed in density theory has gone further than the theoretical developments in point processes, particularly in the intensity domain, we can think of adapting those techniques, initially defined for the density, to the intensity case. We will make use of this idea in several occasions along this manuscript.

1.2.3 Second-order characteristics

In comparison with first-order characteristics, which are concerned about the individual behaviour of points, the second-order ones describe the relationship between a pair of points, i.e., try to describe the variation and correlation in the process. We define below the most important ones.

Definition 1.6. The ***second-order intensity function*** is a measure of the dependency of the events and it is defined, for any two different spatial locations x and y in W, as:

$$\lambda_2(x,y) = \lim_{|dx|,|dy|\to 0} \frac{E[N(dx)N(dy)]}{|dx||dy|}. \tag{1.1}$$

In the particular case of a stationary point process, the second-order intensity verifies that $\lambda_2(x,y) \equiv \lambda_2(x-y)$, i.e., it only depends on the distance between the points and not their locations. If in addition the process is also isotropic, the second-order intensity is reduced to $\lambda_2(x,y) \equiv \lambda_2(||x-y||)$, where $||\cdot||$ denotes the L^2-norm in the euclidean space. Due to the independence among its points, for an nonhomogeneous Poisson point process with intensity function $\lambda(x)$, we have $\lambda_2(x,y) = \lambda(x)\lambda(y)$. Moreover we will say that a point process is ***second-order stationary*** if its first and second-order properties are invariant under translations. Intuitively, the value $\lambda_2(x,y)|dx||dy|$ can be interpreted as the probability of observing exactly one event in dx and one event in dy.

In this general context, we can also define the pair correlation function as follows.

Definition 1.7. Let x and y be two different spatial locations in W, the ***pair correlation function*** is

$$g_2(x,y) = \frac{\lambda_2(x,y)}{\lambda(x)\lambda(y)}. \tag{1.2}$$

This function does not have any practical value unless we set some further assumption, because it is not possible to estimate it in such a general case. Baddeley et al. (2000) dealed with this problem and proposed the ***second-order intensity reweighted stationary*** which implies that $g_2(x,y) \equiv g_2(x-y)$.

Given the first and second-order intensities, we can define the ***conditional intensity function*** $\lambda_c(x|y) = \lambda_2(x,y)/\lambda(y)$, which measures the intensity at point x given that there is an event in y.

The second-order intensity function does not have any physical interpretation, this is the reason why Ripley (1977) offers a new second-order measure for homogeneous processes that fulfils this gap, the K-function.

Definition 1.8. ***Ripley's K-function*** or simply ***K-function*** is defined as

$$K(r) = \lambda^{-1} E[N_0(r)],$$

where $N_0(r)$ is the number of events within distance r of an arbitrary event.

Cressie (1993) also refers to a more general K-function as the ***reduced second-order measure***.

The K-function has some interesting properties not shared by the second-order intensity function, such as physical interpretation, invariance through random thinning and simple estimation, see Cressie (1993) and Baddeley et al. (2015) for details. The counterpart is that it is not unique, i.e, two different point process may have identical K-function. However, it is unique and well defined under CSR, in that situation $K(r) = \pi r^2$ and this can be used to test CSR, by comparing the theoretical one with the empirical K-function as we detailed in Section 1.1.2; see also Cressie (1993) and P. J. Diggle (2013).

A variation of the K-function is the L-function:

$$L(r) = \sqrt{\frac{K(r)}{\pi}},$$

which under CSR is $L(r) = r$, and due to this simplicity is also used to perform CSR tests, see Section 1.1.2.

In order to stablish a link between K and λ_2 we need to assume that the point process is orderly, so there are no duplicated points, and second-order orderly, (see Section 1.2). Under these assumptions, the expected number of events within distance r of an arbitrary event, $E[N_0(r)]$, can be computed by integrating the conditional intensity in the disk centred at the origin and radius r:

$$E[N_0(r)] = \lambda K(r) = \int_0^{2\pi} \int_0^r \lambda_c(x|o) x dx d\theta,$$

where o denotes the origin, and taking into account that $\lambda_c(x|o) = \lambda_2(r)/\lambda$ we obtain that

$$K(r) = 2\pi\lambda^{-1} \int_0^r \lambda_2(x) x dx \text{ or conversely } \lambda_2(r) = \lambda(2\pi r)^{-1} K'(r).$$

1.3 Spatio-temporal point processes

A spatial pattern observed within some bounded region W is usually a realisation of a continuous-time process observed at a fixed time. Sometimes point patterns are observed at discrete times for a continuous-time process. Motivating examples in these situations are described in P. J. Diggle (2013), gastro-intestinal illness in Hampshire (UK) from 1$^{\text{st}}$ January 2001 until 31$^{\text{st}}$ December 2003, and in Ogata (1998), earthquake occurrences in Japan from 1926 until 1995, among others.

Spatio-temporal point processes can also be considered as a hybrid of spatial and temporal components. Extending the definition of a random field in (P.1) to include time, we obtain

$$\{Q(x,t) : x \in D(t) \subset \mathbb{R}^2\},$$

where $D(t)$ is the spatial index at time t.

In this context we have to extend also Definition 1.3, here a point pattern is of the form $\{(X_1, t_1), \ldots, (X_N, t_n)\} \in W \times T \subset \mathbb{R}^2 \times \mathbb{R}^+$, where remark that every event has two spatial coordinates and one temporal component.

Following Cressie (1993) we may distinguish two types of spatio-temporal point processes based on the duration of the events over time:

- **Spatio-temporal shock point processes**: events occur simultaneously over space and time, such as for example earthquakes.

- **Spatio-temporal survival point processes**: events are born at a random location and time, and then live for a random length of time. This arises for example in ecology, epidemiology and geography.

It exists also a generalisation of CSR in this context, ***Complete spatio-temporal randomness (CSTR)*** which means that there is a lack of structure in space as well as in time, so a process verifying CSTR is homogeneous Poisson in $W \times \mathbb{R}^+$.

We also need to extend the definitions of the first and second-order characteristics. The ***first-order intensity function for spatio-temporal processes*** is defined in P. J. Diggle (2013) as

$$\lambda(x,t) = \lim_{|dx \times dt| \to 0} \frac{E[N(dx, dt)]}{|dx \times dt|},$$

where dx is an infinitesimal disc containing the location x, dt is an infinitesimal interval containing t and $N(dx, dt)$ represents the number of events in $dx \times dt$.

The general ***conditional intensity function*** of a spatio-temporal process, $\lambda_c(x, t|H_t)$, is defined as the expected rate of points that occur around the spatio-temporal location (x, t) conditionally on the time history H_t which is the set of locations and times of all events of the process occurring before time t:

$$\lambda_c(x,t|H_t) = \lim_{|dx \times dt| \to 0} \frac{E[N(dx, dt|H_t)]}{|dx \times dt|}.$$

We can also define in $W \times T$ the spatial and temporal marginal intensities of the process:

$$\lambda_s(x) = \int_T \lambda(x,t)dt \quad \text{and} \quad \lambda_t(t) = \int_W \lambda(x,t)dx.$$

Furthermore, if $\lambda(x, t)$ is equal to either or both of the marginal intensities, then the process is said to be ***first-order stationary*** in space, time or both.

Another problem affecting spatial point process which is also present in the spatio-temporal context is the ***edge effect***. The first and easier approach is found in P. J. Diggle et al. (1995) where they propose to correct the edge effect of time and space separately, so that the edge correction term in the spatio-temporal context is a product of the other two. Gabriel (2014) extends the classical spatial edge correction term and compares the performance of some estimators of second-order characteristics. There are different types of edge corrections, see González et al. (2016) for a brief overview of the most important ones.

The ***second-order intensity function for spatio-temporal processes*** is a trivial extension of its analogous in the spatial case given in (1.1):

$$\lambda_2((x,t),(y,s)) = \lim_{|dx\times dt|,|dy\times ds|\to 0} \frac{E[N(dx,dt)N(dy,ds)]}{|dx\times dt||dy\times ds|}, \tag{1.3}$$

that can be for sure interpreted as the probability of observing one single event in $dx \times dt$ and one single event in $dy \times ds$. Similarly, the ***pair correlation function*** is the natural extension of (1.2) in the same sense, that can be interpreted as the standardised probability density that an event occurs in each of the two small volumes $dx \times dt$ and $dy \times ds$.

We have seen before in Section 1.2 the definition of second-order stationarity; Baddeley et al. (2000) defined a weaker form of stationary for spatial point processes called ***second-order intensity reweighted stationary*** that requires the intensity to be bounded away from zero and the pair correlation function $g_2(x,y)$ to depend only on $||x-y||$ but does not need the intensity to be constant on the observation region. Then, let X be a second-order intensity reweighted stationary process, the ***nonhomogeneous K-function*** is $K(r) = 2\pi \int_0^r g_2(t)tdt$.

Gabriel and Diggle (2009) extended these concepts to spatio-temporal processes; so they define a ***second-order intensity reweighted stationary and isotropic*** spatio-temporal point process if its intensity function is bounded away from zero and the pair correlation function depends only on $||x-y||$ and $|t-s|$, being (x,t) and (y,s) the spatio-temporal coordinates. For this type of processes the ***space-time nonhomogeneous K-function*** is defined as $K_{\text{ST}}(u,r) = 2\pi \int_0^u \int_0^r g_2(u,r)ududr$, with $x,y \in W$, $s,t \in \mathbb{R}^+$, $u = ||x-y||$, $r = |t-s|$ and g_2 is the spatio-temporal pair correlation function associated with the process.

Similarly to what has been done for spatial point process, testing for CSTR can be performed using the K-function, as well as other inferential procedures that are beyond the scope of this manuscript, see for example Cressie (1993).

1.3.1 Separability

Separability is a desirable property in any spatio-temporal point process. There are different, but equivalent, definitions of this concept. We will use here the one involving the intensity function.

Definition 1.9. Let X be a spatio-temporal point process in $W \times T \subset \mathbb{R}^2 \times \mathbb{R}^+$ with intensity function $\lambda(x,t)$. We will say that X is ***separable*** if it satisfies

$$\lambda(x,t) = \lambda_s(x)\lambda_t(t), \quad (x,t) \in W \times T,$$

where λ_s and λ_t are the marginal intensities defined in (1.3).

Hence, spatio-temporal separability allows to analyse the spatial component disregarding the temporal one, or analyses the temporal component leaving out the spatial one. In such case the data can be studied in terms of the marginal intensities.

Few works have addressed a rigorous analysis of separability. Ogata (1988) assumes the separability in the parametric ETAS models for earthquake occurrences; and recently a couple of nonparametric test have been developed in this field, see Schoenberg (2004) and Díaz-Avalos et al. (2013).

1.4 Gathering extra information

In a point process we basically focus on two sources of information: the number of points given by a random variable and their locations, also random. However, it is often possible to have more information either about the process itself or about the observation region in general. On the one hand ***marks*** are closely related to the process, and there must have been an event to have a mark. On the other hand we found the ***covariates***, where the information given is about the whole observation region and not only on the events themselves. As it has been said in Baddeley et al. (2015) the main differences between marks and covariates in terms of classical statistics are that marks are associated with data points, and they are part of the "response" (the point pattern), while covariates are "explanatory".

Let clarify this with an example. Imagine we have some data about wildfires in a certain region, this is a realization of a spatial point process (indeed there will be a random number of fires in random spatial locations). Moreover, in this situation, we can have for instance the burnt area in each of the fires, which is a mark; it is clear that it should have been a fire to obtain the information about the burnt area (in any other point of the region where there has not been a fire, this information does not exist). But we can also have for example some information about the different types of vegetation

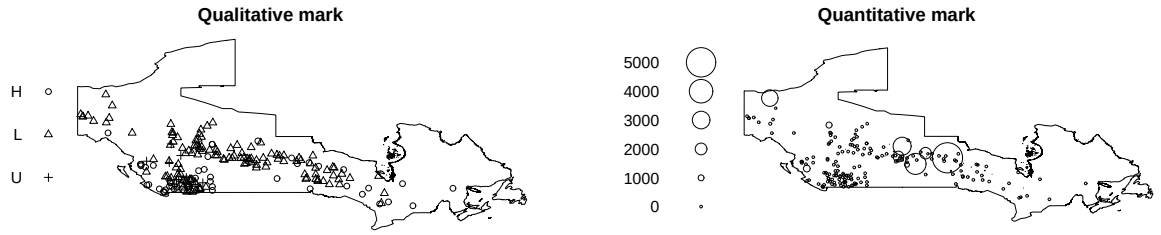


Figure 1.4: Examples of marked point patterns: subset of wildfires in Canada during june 2015 with qualitative mark, cause of the fire H, L, U (left) and quantitative mark, size of the fire in km^2 (right).

scattered in that region, and now we do not need a fire to have this information, we can get it from any point, so this is a covariate.

1.4.1 Marks

Data coming from marked point processes consist of observations of variables given at irregularly distributed locations; these variables that somehow describe properties of the objects represented by the points are called marks, and they provide extra information. Formally, a marked point pattern is denoted by $\{(X_1, m_1), \ldots, (X_N, m_N)\}$ with $X_i \in W$ the locations, and $m_i \in \mathcal{M}$ the marks.

This scenario of marked point processes is close to geostatistics, which also consists of both information of locations and associated values. However the difference lies on the fact that the objective of geostatistics is to study a continuous field through discrete measures at points specifically chosen for this aim; while in marked point processes the locations are random, and constitute indeed an important part of the analysis.

Marks can also be understood as an additional coordinate for the point, for example, any spatio-temporal point process, see Section 1.3, can be seen as a spatial marked point process where the marks are the occurrence time of each event.

The marks can be either ***quantitative*** (continuous generally real-valued measures that describe physical properties of the objects in study) or ***qualitative*** (discrete or categorical data, generally integer-valued that classify the points in types/groups). In particular, if the qualitative mark distinguishes only two groups the process is called ***bivariate*** and in other cases is ***multivariate***. In Figure 1.4 we illustrate the same point pattern (subset of wildfires in Canada during june 2015) with qualitative and quantitative marks.

When dealing with marked point processes we need to take into account that it

may exist spatial correlation between the points and the marks. Consider for example a tree point pattern which would include different species, it is common that different species grow more favourable depending on soil conditions (humidity, composition...), so it may happen that if we consider the type of species as a mark, it might be strongly correlated with the position of the tree itself. Hence, any marked point process model may take this correlation into account.

As we have previously done for spatial point processes, we can define in this context the extended versions of ***stationarity***, ***isotropy***, ***first-order properties***, ***second-order properties***... See Illian et al. (2008) and P. J. Diggle (2013) for more details about the exploratory analysis, inference and examples on marked point processes.

1.4.2 Covariates

A covariate is any kind of data collected as an explanatory variable. It may consist of a ***spatial function*** or ***spatial covariate*** (see an example in Figure 1.5) which will be denoted as $Z(x)$ and that needs to be potentially observable at every location $x \in W$. In practice it is enough that the covariate would be known in a sufficient quantity of locations, note that it is not enough to have this information only on the points of the pattern when our aim is to investigate the dependence of the process on the covariate, so it must cover at least the observation region in an appropriate way.

Another type of covariate that may be used is a point pattern or line segment pattern, but even in this case this type of covariates would be used to define a spatial function, as introduces above, that is usually the distance from every point of the observation region to the nearest point or segment. Consider for example a gold deposit dataset where we know the coordinates of the geological faults, in this case we have a line segment pattern from which we construct the spatial covariate as the distance from every point in the observation region to the nearest fault; see Figure 1.5 (centre and right) for illustration.

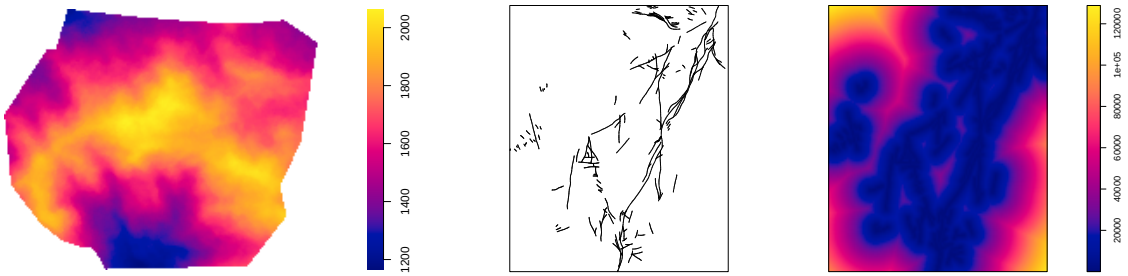


Figure 1.5: Examples of a pure spatial covariate, elevation of the terrain (left); line segment covariate, geological faults (centre) and the spatial covariate constructed based on it (right).

Until now we have focused on continuous covariates, however categorical covariates may also be present. For example, recall the tree point pattern used in the previous subsection; we could have the information of the type of vegetation (classified in a nominal scale) over the observation region and this would be a categorical covariate. For sure the interpretation of both, numerical and categorical covariates is different. In the categorical case, the observation region would be divided according to the covariate in subregions of different type where we should analyse the process, for more detail on this see (Baddeley et al., 2015, Chap. 9). Also numerical covariates can be transformed into categorical ones by discretising the data. We illustrate these two possible situations in Figure 1.6.

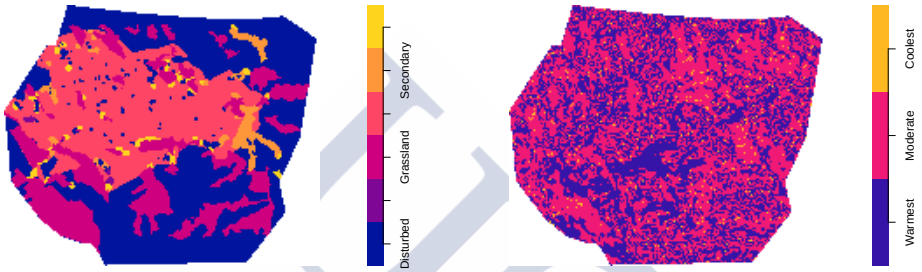


Figure 1.6: Examples of categorical covariates: types of vegetation (left) and a categorical covariate constructed by discretising the Heat Load Index (right).

This extra information is very important and useful because it allows to introduce in the process another possible element influencing the generation of the events. Typically, we would like to know if a covariate has influence on the intensity of the process and, if that occurs, try to quantify it.

Baddeley et al. (2012) proposed a way of including a numerical spatial covariate in the intensity by assuming that $\lambda(x) = \rho(Z(x))$, where ρ is an unknown function that needs to be estimated either parametrically or nonparametrically. We will assume this model in chapters 3, 4 and 5 of this manuscript.

Chapter 2

Density estimation for length-biased data

Contents

We have previously seen in Chapter 1 some of the most important first and second-order characteristic functions in a point process. We have explained that the intensity function has a huge interest in different research fields and we have also detailed the close relationship between the intensity and the density function. As it will remain clear in Chapter 3, there exist evident structural similarities between kernel intensity estimation in point processes with covariates and kernel density estimation for length-biased data, hence we want to make use of this connection, and extend to the spatial field the new developments in the context of density estimation with length-biased data presented in this chapter.

Length-biased data are a particular case of weighted data, which arise in many situations: biomedicine, quality control or epidemiology among others. Jones (1991) proposed a kernel density estimator for this type of data which has the same structure as a kernel intensity estimator proposed in Baddeley et al. (2012) for point processes with covariates (we will go deeper into it in Chapter 3). Along the present chapter we study the theoretical properties of kernel density estimation in the context of length-biased data, proposing two consistent bootstrap methods that we use for bandwidth selection. Apart from the bootstrap bandwidth selectors we also suggest a rule-of-thumb and all these bandwidth proposals are compared with an existing least-squares cross-validation method. A simulation study is accomplished to understand the behaviour of the procedures in finite samples.

Remark that the notation included in this chapter refers to the context of density estimation, not point processes, so it may occur that, to ease the comprehension of the chapter and follow the common notation in this field, we reuse some of the notation previously introduced in the Preface and Introduction of this manuscript to denote slightly different concepts.

2.1 Introduction

In general a sample is supposed to have the same basic characteristics as the population it represents. However, in practice the sample may not be completely representative of the population, and bias is introduced in the sampling scheme, we known them as weighted data. This type of samples is produced when the probability of choosing an observation depends on its value and/or on other covariates of interest. Weighted data arise in many sampling processes, see Patil and Rao (1977), and in a wide variety of fields such as biomedicine, Chakraborty and Rao (2000), epidemiology, Simon (1980), textile fibres, Cox (2005), as well as social sciences, quality control, or economics, Heckman (1990).

Some specific examples are the visibility bias problem that arises when using aerial survey techniques to estimate, for instance, wildlife population density; or a damage

model where an observation may be damaged by a process depending on the variable and then the observed data are clearly biased. Also the textile fibres problem is a classical motivating example.

Let us denote by f the density function of an unobserved random variable X, and assume that the available information refers to a closely related random variable Y with weighted or biased distribution determined by the density function:

$$f_{Y,\omega}(y) = \frac{\omega(y)f(y)}{\mu_\omega} \qquad y > 0,$$

where ω is a known function and $\mu_\omega = \int \omega(x)f(x)dx < \infty$.

A particular case of weighted data is the length-biased data, where the probability of an observation to be sampled is directly proportional to its value in a simple linear way. In this case the weight function that determines the bias is the identity function, i.e., $\omega(y) = y$. This sort of data is quite common in problems related to renewal processes, epidemiological cohort studies or screening programs for the study of chronic diseases, see Zelen and Feinleib (1969).

Cox (2005) proposed an estimator for the mean and another for the distribution function in the context of weighted data. Vardi (1982, 1985) showed that this last estimator was the maximum likelihood estimator of the distribution function under weighted sampling and that the estimation of the mean is $\sqrt{n}-$consistent. Density estimation for this type of data started in the 80's when Bhattacharyya et al. (1988) defined the first density estimator for length-biased data based on the problem of textile fibres, which was continued with theoretical developments in Richardson et al. (1991). Furthermore, Jones (1991) proposed a modification of the common kernel density estimator adapted to length-biased data which is widely used. In the same paper he showed that this proposal has some advantages over the previous one, and better asymptotic properties. Ahmad (1995) extended to the multivariate case these two kernel density estimators. Another extensions using Fourier series have been proposed in Jones and Karunamuni (1997). Later a third nonparametric estimator has been considered in Guillamón et al. (1998).

Density estimation for weighted data has also been studied from other points of view, Barmi and Simonoff (2000) proposed a simple transformation-based approach motivated by the form of the nonparametric maximun likelihood estimator of the density. Efromovich (2004) presented asymptotic results on sharp minimax density estimation. Projection methods are developed in Brunel et al. (2009). Asgharian et al. (2002) and de Uña-Álvarez (2004) studied the problem under the common settings of survival analysis. Also wavelet theory has been used in this context, see Chesneau (2010) which constructed an adaptative estimator based on the Block Shrink algorithm and Ramírez and Vidakovic (2010) who applied dyadic wavelet density estimation. Cutillo et al. (2014) proposed linear and nonlinear wavelet density estimators and recently Comte

and Rebafka (2016) defined the estimation through out the distribution function and using a known link.

The use of nonparametric methods implies to choose a bandwidth parameter, which determines the degree of smoothness to be considered in the estimation. The choice of the bandwidth parameter is crucial and it has motivated several papers in the literature in the recent decades. Marron (1988), Scott (1992) and Silverman (1986) provide a full description of the problem as well as a review of several bandwidth selection methods. Later methods such as plug-in or bootstrap methods, have been defined in Hall and Marron (1987), Sheather and Jones (1991) and Marron (1992). Fourier transforms have also been used in this context, see Chiu (1992). To explore the most relevant bandwidth selection methods in density estimation for complete data see the reviews of Turlach (1993), Cao et al. (1994), Jones et al. (1996) or Heidenreich et al. (2013), and the recent work on local linear density estimation by Mammen et al. (2011, 2014).

This chapter is organised as follows. In Section 2.2 we develop asymptotic theory for the kernel density estimator of Jones (1991) for length-biased data, and we also define two different consistent bootstrap procedures. In Section 2.3 we propose new data-driven bandwidth selection methods: a rule-of-thumb based on the Normal distribution and two bootstrap bandwidth selectors based on the procedures presented in the previous section. These proposals are competitors of a cross-validation method which, to the extent of our knowledge, is the only existent data-driven bandwidth selector in this context. In Section 2.4 we carry out an extensive simulation study to evaluate the performance of the presented bandwidth selectors for finite samples. Some final remarks are given in Section 2.5, as well as a discussion of how the methodology developed in this chapter can be generalised to a widespread weight function. In Section 2.6 we draw our conclusions.

2.2 Theoretical developments

Hereafter in this chapter we will work under the scenario of the length-biased data even though all the results can be generalised to the weighted data case under appropriate assumptions, see final remarks in Section 2.5.

Hence, let us write the density function of the observed variable Y as

$$f_Y(y) = \frac{yf(y)}{\mu}, \qquad y > 0,$$

with $\mu = \int yf(y)dy$.

Let $Y_1, \ldots, Y_n$ be an independent identically distributed (i.i.d.) sample from f_Y, Jones (1991) defined the following kernel density estimator of f, based on the structure

of the one proposed in Parzen (1962) and Rosenblatt (1956):

$$\hat{f}_h(y) = \frac{1}{n}\hat{\mu}\sum_{i=1}^{n}\frac{1}{Y_i}K_h(y - Y_i), \tag{2.1}$$

where $\hat{\mu} = \left(\frac{1}{n}\sum_{i=n}^{n}\frac{1}{Y_i}\right)^{-1}$, see Cox (2005) to find out this estimation, and $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$, with K being a symmetric kernel function.

In the following result we obtain the value of the pointwise mean and variance of $\hat{f}_h$ with the corresponding error rates, as well as its mean squared error (MSE), which is defined as:

$$\text{MSE}(h, y) = E\left[(\hat{f}_h(y) - f(y))^2\right]. \tag{2.2}$$

We need to introduce the following hypotheses:

(A.1) $E\left[\frac{1}{X}\right] < +\infty$, $E\left[\frac{1}{Y^{2\nu}}\right] < +\infty$ where $\nu \in \mathbb{N}$, $\nu \geq 3$,

(A.2) $\int K(u)du = 1$, $\int uK(u)du = 0$ and $\mu_2(K) < +\infty$,

(A.3) $\lim_{n\to\infty} nh = +\infty$,

(A.4) y a continuity point of f,

(A.5) f has two continuous derivates and

(A.6) K is twice differentiable.

Theorem 2.1. *Under conditions (A.1) to (A.4) we have:*

$$E\left[\hat{f}_h(y)\right] = (K_h \circ f)(y) + O\left(\tfrac{1}{n}\right) \text{ and}$$
$$\text{Var}\left[\hat{f}_h(y)\right] = n^{-1}\left[(K_h^2 \circ \gamma)(y) - (K_h \circ f)^2(y)\right] + O\left(\tfrac{1}{n}\right),$$

where $\circ$ denotes the convolution between two functions and $\gamma(y) = \mu f(y)/y$. Moreover, adding condition (A.5), we have:

$$\text{MSE}(h, y) = \frac{1}{4}h^4\left(f''(y)\right)^2\mu_2^2(K) + \frac{\gamma(y)}{nh}R(K) + o\left(h^4 + \frac{1}{nh}\right), \tag{2.3}$$

where $\mu_2(K) = \int u^2K(u)du$ and $R(K) = \int K^2(u)du$.

– Proof –

First of all we rewrite the estimator in (2.1) as follows:

$$\hat{f}_h(y) = \frac{1}{n}\frac{1}{\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i}}\sum_{i=1}^n \frac{1}{Y_i}K_h(y-Y_i) = \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i}K_h(y-Y_i)}{\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i}} = \frac{\phi_n(y)}{\xi_n}. \tag{2.4}$$

We start by calculating the punctual mean of (2.1) for which we need the mean of the numerator and denominator in (2.4), so:

$$\overline{\phi_n(y)} := E[\phi_n(y)] = E\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i}K_h(y-Y_i)\right] = \int \frac{1}{z}K_h(y-z)f_Y(z)dz = \frac{1}{\mu}(K_h \circ f)(y)$$

and

$$\overline{\xi_n} := E[\xi_n] = E\left[\frac{1}{n}\sum_{j=1}^n \frac{1}{Y_j}\right] = \int \frac{1}{z}f_Y(z)dz = \frac{1}{\mu}.$$

We divide this proof in two separated but linked parts, detailing all the results involving mean and variance calculations respectively.

Mean

Applying the linearisation technique used in Collomb (1976) we can write down

$$\begin{aligned}\hat{f}_h(y) &= \frac{\phi_n(y)}{\xi_n} = \frac{\phi_n(y)}{\overline{\xi_n}}\cdot\frac{\overline{\xi_n}}{\xi_n} = \frac{\phi_n(y)}{\overline{\xi_n}}\left[1+\sum_{k=1}^{\nu-1}(-1)^k\left(\frac{\xi_n-\overline{\xi_n}}{\overline{\xi_n}}\right)^k + (-1)^\nu\frac{\overline{\xi_n}}{\xi_n}\left(\frac{\xi_n-\overline{\xi_n}}{\overline{\xi_n}}\right)^\nu\right] \\ &= \frac{\phi_n(y)}{\overline{\xi_n}}\left[1+\sum_{k=1}^{\nu-1}(-1)^k\left(\frac{\xi_n-\overline{\xi_n}}{\overline{\xi_n}}\right)^k\right] + (-1)^\nu\hat{f}_h(y)\left(\frac{\xi_n-\overline{\xi_n}}{\overline{\xi_n}}\right)^\nu.\end{aligned}$$

with $\nu \geq 2$ and taking into account that $\xi_n \neq 0\ \forall\, n$.

Using the notation $S_n^{a,b}(y) := E[\phi_n^a(y)(\xi_n-\overline{\xi_n})^b]$, $s_n^{a,b}(y) := E[(\phi_n(y)-\overline{\phi_n(y)})^a(\xi_n-\overline{\xi_n})^b]$, $\sigma_n^{a,b}(y) = E[\hat{f}_h(y)^a(\xi_n-\overline{\xi_n})^b]$ for $a \in \{0,1,2\}$, $b \in \{1,\ldots,\nu\}$ and knowing that $\overline{\phi_n}S_n^{0,k}(y) + s_n^{1,k}(y) = S_n^{1,k}(y)$ we can write:

$$\begin{aligned}E[\hat{f}_h(y)] &= E\left[\frac{\phi_n(y)}{\overline{\xi_n}} + \frac{\phi_n(y)}{\overline{\xi_n}}\sum_{k=1}^{v-1}(-1)^k\left(\frac{\xi_n-\overline{\xi_n}}{\overline{\xi_n}}\right)^k + (-1)^\nu\hat{f}_h(y)\left(\frac{\xi_n-\overline{\xi_n}}{\overline{\xi_n}}\right)^\nu\right] \\ &= \frac{\overline{\phi_n(y)}}{\overline{\xi_n}} + \sum_{k=1}^{\nu-1}(-1)^k\frac{E[\phi_n(y)(\xi_n-\overline{\xi_n})^k]}{\overline{\xi_n}^{k+1}} + \frac{(-1)^\nu}{\overline{\xi_n}^\nu}E[\hat{f}_h(y)(\xi_n-\overline{\xi_n})^\nu] \\ &= \frac{\overline{\phi_n(y)}}{\overline{\xi_n}} + \sum_{k=1}^{\nu-1}(-1)^k\frac{S_n^{1,k}(y)}{\overline{\xi_n}^{k+1}} + (-1)^\nu\frac{\sigma_n^{1,\nu}(y)}{\overline{\xi_n}^\nu} \\ &= \frac{\overline{\phi_n(y)}}{\overline{\xi_n}} + \frac{\overline{\phi_n(y)}S_n^{0,2}(y)-\overline{\xi_n}s_n^{1,2}(y)-\overline{\xi_n}S_n^{1,1}(y)}{\overline{\xi_n}^3} + \sum_{k=1}^{\nu-1}(-1)^k\frac{S_n^{1,k}(y)}{\overline{\xi_n}^{k+1}} + (-1)^\nu\frac{\sigma_n^{1,\nu}(y)}{\overline{\xi_n}^\nu}.\end{aligned}$$

Then, we have proved that

$$E[\hat{f}_h(y)] = \frac{\overline{\phi_n(y)}}{\overline{\xi_n}} + c_n(y) + c_n^{(\nu)}(y) + \frac{(-1)^\nu\sigma_n^{1,\nu}(y)}{\overline{\xi_n}^\nu},$$

where

$$c_n(y) = \frac{\overline{\phi_n(y)}S_n^{0,2}(y) - \overline{\xi_n}S_n^{1,1}(y)}{\overline{\xi_n}^3} = \frac{\overline{\phi_n(y)}E[(\xi_n - \overline{\xi_n})^2] - \overline{\xi_n}E[\phi_n(y)(\xi_n - \overline{\xi_n})]}{\overline{\xi_n}^3} \text{ and}$$

$$c_n^{(\nu)}(y) = \frac{s_n^{1,2}(y)}{\overline{\xi_n}^3} + \sum_{k=1}^{\nu-1}(-1)^k\frac{S_n^{1,k}(y)}{\overline{\xi_n}^{k+1}} = \frac{E[(\phi_n(y) - \overline{\phi_n(y)})(\xi_n - \overline{\xi_n})^2]}{\xi_n^3} + \sum_{k=1}^{\nu-1}(-1)^k\frac{E[\phi_n(y)(\xi_n - \overline{\xi_n})^k]}{\overline{\xi_n}^{k+1}}.$$

The first addend corresponds to the asymptotic expression of the mean obtained by Jones (1991). We want now to expand each of the other terms and study the rate of convergence. To this aim we use basic statistical properties and we proceed as follows:

$$\begin{aligned} E\left[(\xi_n - \overline{\xi_n})^2\right] &= E\left[\xi_n^2\right] - \overline{\xi_n}^2 = \tfrac{1}{n^2}E\left[\textstyle\sum_{i=1}^n \frac{1}{Y_j^2} + \sum_{i\neq j}\frac{1}{Y_i}\frac{1}{Y_j}\right] - \frac{1}{\mu^2} \\ &= \tfrac{1}{n^2}nE\left[\tfrac{1}{Y_1^2}\right] + \tfrac{(n^2-n)}{n^2}E\left[\tfrac{1}{Y_1}\tfrac{1}{Y_2}\right] - \tfrac{1}{\mu^2} = \tfrac{1}{n}\textstyle\int \frac{1}{z^2}g(z)dz + \frac{n-1}{n}E\left[\frac{1}{Y_1}\right]^2 - \frac{1}{\mu^2} \\ &= \tfrac{1}{n}E\left[\tfrac{1}{X}\right] + \tfrac{n-1}{n\mu^2} - \tfrac{1}{\mu^2} = \tfrac{1}{n\mu}\left(E\left[\tfrac{1}{X}\right] - \tfrac{1}{\mu}\right). \end{aligned}$$

$$\begin{aligned} E\left[\phi_n(y)(\xi_n - \overline{\xi_n})\right] &= E\left[\phi_n(y)\xi_n\right] - \overline{\xi_n}\,\overline{\phi_n(y)} \\ &= E\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{1}{Y_i}K_h(y-Y_i)\right)\left(\frac{1}{n}\sum_{j=1}^n\frac{1}{Y_j}\right)\right] - \tfrac{1}{\mu^2}\left(K_h \circ f\right)(y) \\ &= \tfrac{1}{n^2}E\left[\textstyle\sum_{i=1}^n \frac{1}{Y_i^2}K_h(y-Y_i) + \sum_{i\neq j}\frac{1}{Y_i}\frac{1}{Y_j}K_h(y-Y_i)\right] - \tfrac{1}{\mu^2}\left(K_h \circ f\right)(y) \\ &= \tfrac{n}{n^2}E\left[\tfrac{1}{Y_1^2}K_h(y-Y_1)\right] + \tfrac{n(n-1)}{n^2}E\left[\tfrac{1}{Y_1}\tfrac{1}{Y_2}K_h(y-Y_1)\right] - \tfrac{1}{\mu^2}\left(K_h \circ f\right)(y) \\ &= \tfrac{1}{n\mu^2}\left(K_h \circ \gamma\right)(y) - \tfrac{1}{n\mu^2}\left(K_h \circ f\right)(y). \end{aligned}$$

Hence,

$$\begin{aligned} c_n(y) &= \mu^3\left[\tfrac{1}{\mu}\left(K_h \circ f\right)(y)\tfrac{1}{n\mu}\left(E\left[\tfrac{1}{X}\right] - \tfrac{1}{\mu}\right) - \tfrac{1}{\mu}\left(\tfrac{1}{n\mu^2}\left(K_h \circ \gamma\right)(y) - \tfrac{1}{n\mu^2}\left(K_h \circ f\right)(y)\right)\right] \\ &= \underbrace{\frac{\mu}{n}\left(K_h \circ f\right)(y)\left(E\left[\frac{1}{X}\right] - \frac{1}{\mu}\right)}_{(a)} \underbrace{- \frac{1}{n}\left(K_h \circ \gamma\right)(y)}_{(b)} + \underbrace{\frac{1}{n}\left(K_h \circ f\right)(y)}_{(c)}. \end{aligned}$$

Applying Theorem 2.1 of Cacoullos (1966) with $g(z) = f(z)$ for (a) and (c), and $g(z) = \frac{f(z)}{z}$ for (b), we easily obtain that each of these addends is $O(1/n)$. Therefore, $c_n(y) = O(1/n)$.

To expand the next two terms we use the Hölder's inequality and, taking into account that K is bounded, we only require the finiteness of the second order moment of $\frac{1}{Y}$ to obtain

$$E\left[\phi_n(y)\left(\xi_n - \overline{\xi_n}\right)^2\right] \leq E\left[\phi_n^2(y)\right]^{\frac{1}{2}}E\left[\left(\xi_n - \overline{\xi_n}\right)^4\right]^{\frac{1}{2}} = O(1/n)^{\frac{1}{2}}O(1/n^2)^{\frac{1}{2}} = O(1/n)^{\frac{3}{2}}$$

and $E\left[\left(\xi_n - \overline{\xi_n}\right)^2\right] = O(1/n)$.

Therefore $c_n^{(\nu)}(y) = O(1/n)$. And finally,

$$\sigma_n^{1,\nu}(y) = E\left[\hat{f}_h(y)\left(\xi_n - \overline{\xi_n}\right)^{\nu}\right] \leq E\left[\hat{f}_h^2(y)\right]^{\frac{1}{2}} E\left[\left(\xi_n - \overline{\xi_n}\right)^{2\nu}\right]^{\frac{1}{2}} = O(1/n)^{\frac{\nu}{2}}.$$

Variance

To get the variance, we compute the expected value of the squared estimator. We follow the same techniques as in the previous operations replacing $\hat{f}_h(y)$ by $\hat{f}_h^2(y)$. Applying again the linearisation method of Collomb (1976) with $\nu \geq 2$,

$$\begin{aligned}
\hat{f}_h^2(y) &= \tfrac{\phi_n^2(y)}{\xi_n^2} = \tfrac{\phi_n^2(y)}{\overline{\xi}_n^2} \cdot \tfrac{\overline{\xi}_n^2}{\xi_n^2(y)} = \tfrac{\phi_n^2(y)}{\overline{\xi}_n^2}\left[1 + \textstyle\sum_{k=1}^{\nu-1}(-1)^k \left(\tfrac{\xi_n^2(y) - \overline{\xi}_n^2}{\overline{\xi}_n^2}\right)^k\right] \\
&\quad + (-1)^{\nu} \hat{f}_h^2(y) \left(\tfrac{\xi_n^2(y) - \overline{\xi}_n^2}{\overline{\xi}_n^2}\right)^{\nu} \\
&= \tfrac{\phi_n^2(y)}{\overline{\xi}_n^2}\left[1 + \textstyle\sum_{k=1}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!}{j!(k-j)!} 2^{k-j} \left(\tfrac{\xi_n - \overline{\xi}_n}{\overline{\xi}_n}\right)^{k+j}\right] \\
&\quad + (-1)^{\nu} \hat{f}_h^2(y) \textstyle\sum_{j=0}^{\nu} \tfrac{\nu!}{j!(\nu-j)!} 2^{\nu-j} \left(\tfrac{\xi_n - \overline{\xi}_n}{\overline{\xi}_n}\right)^{\nu+j}.
\end{aligned}$$

To obtain the mean of $\hat{f}_h^2(y)$ we need to work on $\frac{S_n^{2,l}(y)}{\overline{\xi}_n^2} = \frac{\overline{\phi}_n^2 s_n^{0,l}(y) + 2\overline{\phi}_n^2 s_n^{1,l}(y) + s_n^{2,l}(y)}{\overline{\xi}_n^2}$:

$$\begin{aligned}
E\left[\hat{f}_h^2(y)\right] &= \tfrac{1}{\overline{\xi}_n^2} E\left[\phi_n(y)^2\right] + E\left[\tfrac{\phi_n(y)^2(y)}{\overline{\xi}_n^2} \textstyle\sum_{k=1}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!}{j!(k-j)!} 2^{k-j} \left(\tfrac{\xi_n - \overline{\xi}_n}{\overline{\xi}_n}\right)^{k+j}\right] \\
&\quad + (-1)^{\nu} \textstyle\sum_{j=0}^{\nu} \tfrac{\nu! 2^{\nu-j}}{j!(\nu-j)!} \tfrac{E\left[\hat{f}_h^2(y)\left(\xi_n - \overline{\xi}_n\right)^{\nu+j}\right]}{\overline{\xi}_n^{\nu+j}} \\
&= \tfrac{E\left[\phi_n^2(y)\right]}{\overline{\xi}_n^2} + \textstyle\sum_{k=1}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!}{j!(k-j)!} 2^{k-j} \tfrac{S_n^{2,k+j}(y)}{\overline{\xi}_n^{k+j+2}} \\
&\quad + (-1)^{\nu} \textstyle\sum_{j=0}^{k} \tfrac{\nu!}{j!(\nu-j)!} 2^{\nu-j} \tfrac{\sigma_n^{2,\nu+j}(y)}{\overline{\xi}_n^{\nu+j}} \\
&= \tfrac{\overline{\phi}_n^2}{\overline{\xi}_n^2} + \tfrac{s_n^{0,2}(y)}{\overline{\xi}_n^2} + \textstyle\sum_{k=1}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!}{j!(k-j)!} 2^{k-j} \tfrac{\overline{\phi}_n^2 s_n^{0,k+j}(y) + 2\overline{\phi}_n s_n^{1,k+j}(y) + s_n^{2,k+j}(y)}{\overline{\xi}_n^{k+j+2}} \\
&\quad + (-1)^{\nu} \textstyle\sum_{j=0}^{\nu} \tfrac{\nu!}{j!(\nu-j)!} 2^{\nu+j} \tfrac{\sigma_n^{2,\nu+j}(y)}{\overline{\xi}_n^{\nu+j}} \\
&= \tfrac{\overline{\phi}_n^2}{\overline{\xi}_n^2} + \tfrac{s_n^{2,0}(y)}{\overline{\xi}_n^2} - 2\tfrac{\overline{\phi}_n s_n^{1,1}(y) + s_n^{2,1}(y)}{\overline{\xi}_n^3} - \tfrac{\overline{\phi}_n^2 s_n^{0,2}(y) + 2\overline{\phi}_n s_n^{1,2}(y) + s_n^{2,2}(y)}{\overline{\xi}_n^4} \\
&\quad + 4\tfrac{\overline{\phi}_n^2 s_n^{0,2}(y) + 2\overline{\phi}_n s_m^{1,2}(y) + s_n^{2,2}(y)}{\overline{\xi}_n^4} + 4\tfrac{S_n^{2,3}(y)}{\overline{\xi}_n^5} + \tfrac{S_n^{2,4}(y)}{\overline{\xi}_n^6} \\
&\quad + \textstyle\sum_{k=3}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!}{j!(k-j)!} 2^{k-j} \tfrac{S_n^{2,k+j}(y)}{\overline{\xi}_n^{k+j+2}} + (-1)^{\nu} \textstyle\sum_{j=0}^{\nu} \tfrac{\nu!}{j!(\nu-j)!} 2^{\nu+j} \tfrac{\sigma_n^{2,\nu+j}(y)}{\overline{\xi}_n^{\nu+j}} \\
&= \tfrac{\overline{\phi}_n^2}{\overline{\xi}_n^2} + \tfrac{s_n^{0,2}(y)}{\overline{\xi}_n^2} + 3\tfrac{\overline{\phi}_n s_n^{0,2}(y)}{\overline{\xi}_n^4} - 4\tfrac{\overline{\phi}_n s_n^{1,1}(y)}{\overline{\xi}_n^3} - 2\tfrac{s_n^{2,1}(y)}{\overline{\xi}_n^3} + 3\left(\tfrac{s_n^{2,2}(y)}{\overline{\xi}_n^4} + \tfrac{2\overline{\phi}_n s_n^{1,2}(y)}{\overline{\xi}_n^4}\right) \\
&\quad + 4\tfrac{S_n^{2,3}(y)}{\overline{\xi}_n^5} + \tfrac{S_n^{2,4}(y)}{\overline{\xi}_n^6} + \textstyle\sum_{k=3}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!}{j!(k-j)!} 2^{k-j} \tfrac{S_n^{2,k+j}(y)}{\overline{\xi}_n^{k+j+2}} \\
&\quad + (-1)^{\nu} \textstyle\sum_{j=0}^{\nu} \tfrac{\nu!}{j!(\nu-j)!} 2^{\nu+j} \tfrac{\sigma_n^{2,\nu+j}(y)}{\overline{\xi}_n^{\nu+j}}.
\end{aligned}$$

Hence,

$$E\left[\hat{f}_h^2(y)\right] = \frac{\overline{\phi}_n^2(y)}{\overline{\xi}_n^2} + \varphi_n(y) + \Gamma_n^{(\nu)}(y) + (-1)^\nu \Delta^{(\nu)}(y),$$

where

$$\begin{aligned}
\varphi_n(y) &= \tfrac{s_n^{0,2}(y)}{\overline{\xi}_n^2} + 3\tfrac{\overline{\phi}_n s_n^{0,2}(y)}{\overline{\xi}_n^4} - 4\tfrac{\overline{\phi}_n s_n^{1,1}(y)}{\overline{\xi}_n^3},\\
\Gamma_n^{(\nu)}(y) &= -2\tfrac{s_n^{2,1}(y)}{\overline{\xi}_n^3} + 3\left(\tfrac{s_n^{2,2}(y)}{\overline{\xi}_n^4} + \tfrac{2\overline{\phi}_n s_n^{1,2}(y)}{\overline{\xi}_n^4}\right) + 4\tfrac{S_n^{2,3}(y)}{\overline{\xi}_n^5} + \tfrac{S_n^{2,4}(y)}{\overline{\xi}_n^6}\\
&\quad + \textstyle\sum_{k=3}^{\nu-1}(-1)^k \sum_{j=0}^{k} \tfrac{k!2^{k-j}}{j!(k-j)!} \tfrac{S_n^{2,k+j}(y)}{\overline{\xi}_n^{k+j+2}} \text{ and}\\
\Delta^{(\nu)}(y) &= \textstyle\sum_{j=0}^{\nu} \tfrac{\nu!}{j!(\nu-j)!} 2^{\nu+j} \tfrac{\sigma_n^{2,\nu+j}(y)}{\overline{\xi}_n^{\nu+j}}.
\end{aligned}$$

As we have done before for the mean, we must study the order of convergence of these terms:

$$\begin{aligned}
s_n^{2,0}(y) &= E\left[\left(\phi_n(y) - \overline{\phi}_n(y)\right)^2\right] = \mathrm{Var}\left[\phi_n(y)\right] = E\left[\phi_n(y)^2\right] - \overline{\phi}_n(y)^2\\
&= \tfrac{1}{n\mu^2}\left(K_h^2 \circ \gamma\right)(y) + \tfrac{n-1}{n\mu^2}\left(K_h \circ f\right)^2(y) - \tfrac{1}{\mu^2}\left(K_h \circ f\right)^2(y)\\
&= \tfrac{1}{n\mu^2}\left(K_h^2 \circ \gamma\right)(y) - \tfrac{1}{n\mu^2}\left(K_h \circ f\right)^2(y),\\
\overline{\phi}_n^2(y) s_n^{0,2}(y) &= \overline{\phi}_n^2(y) E\left[\left(\xi_n - \overline{\xi}_n\right)^2\right] = \overline{\phi}_n^2(y) \tfrac{1}{n\mu}\left(E\left[\tfrac{1}{X}\right] - \tfrac{1}{\mu}\right)\\
&= \tfrac{1}{n\mu^3}\left(K_h \circ f\right)^2(y)\left(E\left[\tfrac{1}{X}\right] - \tfrac{1}{\mu}\right),\\
\overline{\phi}_n(y) s_n^{1,1}(y) &= \overline{\phi}_n(y) E\left[\left(\phi_n(y) - \overline{\phi}_n(y)\right)\left(\xi_n - \overline{\xi}_n\right)\right]\\
&= \overline{\phi}_n(y) E\left[\phi_n(y)\left(\xi_n - \overline{\xi}_n\right)\right] - \overline{\phi}_n^2(y) E\left[\left(\xi_n - \overline{\xi}_n\right)\right]\\
&= \tfrac{1}{\mu}\left(K_h \circ f\right)(y)\left[\tfrac{1}{n\mu^2}\left(K_h \circ \gamma\right)(y) - \tfrac{1}{n\mu^2}\left(K_h \circ f\right)(y)\right]\\
&= \tfrac{1}{n\mu^3}\left(K_h \circ f\right)(y)\left(K_h \circ \gamma\right)(y) - \tfrac{1}{n\mu^3}\left(K_h \circ f\right)^2(y),\\
s_n^{2,1}(y) &= E\left[\left(\phi_n(y) - \overline{\phi}_n(y)\right)^2\left(\xi_n - \overline{\xi}_n\right)\right] = O(1/n)\\
&\quad -\tfrac{1}{n\mu^3}\left(K_h \circ \gamma\right)(y) - \tfrac{1}{n\mu^3}\left(K_h \circ f\right)^2(y) \text{ and}\\
s_n^{2,2}(y) &= E\left[\left(\phi_n(y) - \overline{\phi}_n(y)\right)^2\left(\xi_n - \overline{\xi}_n\right)^2\right] \le E\left[\left(\phi_n(y) - \overline{\phi}_n(y)\right)^4\right]^{\frac{1}{2}}\\
&\quad E\left[\left(\xi_n - \overline{\xi}_n\right)^2\right]^{\frac{1}{2}} = O(1/n)O(1/n)^{\frac{1}{2}} = O(1/n)^{\frac{3}{2}}.
\end{aligned}$$

In the same way as with this last term and assuming that the l-th order centred moment of the variable $\frac{1}{Y} < +\infty$ with $l = 1, \ldots, 2\nu$, we obtain

$$\overline{\phi}_n(y) s_n^{1,2}(y) = O(1/n)^{\frac{3}{2}}, \quad S_n^{2,3}(y) = O(1/n)^{\frac{5}{2}}, \quad S_n^{2,4}(y) = O(1/n^3),$$

$$S_n^{2,k+j}(y) = O(1/n^{\frac{k+j}{2}+1}) \text{ and } \sigma_n^{2,\nu+j}(y) = O(1/n^{\nu+j}).$$

Finally, gathering all the addends properly, we get

$$\begin{aligned} E\left[\hat{f}_h^2(y)\right] &= \frac{\overline{\phi}_n^2(y)}{\overline{\xi}_n^2} + \varphi_n(y) + \Gamma_n^{(\nu)}(y) + (-1)^\nu \Delta^{(\nu)}(y) = (K_h \circ f)^2(y) + \frac{1}{n}\left(K_h^2 \circ \gamma\right)(y) \\ &\quad -\frac{1}{n}(K_h \circ f)^2(y) + O(1/n). \end{aligned}$$

Then,

$$\begin{aligned} \mathrm{Var}\left[\hat{f}_h(y)\right] &= (K_h \circ f)^2(y) + \frac{1}{n}\left(K_h^2 \circ \gamma\right)(y) - \frac{1}{n}(K_h \circ f)^2(y) - (K_h \circ f)^2(y) + O(1/n) \\ &= \frac{1}{n}\left[\left(K_h^2 \circ \gamma\right)(y) - (K_h \circ f)^2(y)\right] + O(1/n). \end{aligned}$$

To get the MSE it is enough to realise that

$$\mathrm{MSE}\,(h, y) = \mathrm{Bias}^2\left(\hat{f}_h(y)\right) + \mathrm{Var}\left(\hat{f}_h(y)\right),$$

and apply a Taylor expansion as it is done with the kernel density estimator with complete data, then:

$$\mathrm{MSE}\,(h, y) = \frac{1}{4}h^4\left(f''(y)\right)^2 \mu_2^2(K) + \frac{\gamma(y)}{nh}R(K) + o\left(h^4 + \frac{1}{nh}\right).$$

□

Now, defining the mean integrated squared error (MISE) as

$$\mathrm{MISE}(h) = E\int\left(\hat{f}_h(y) - f(y)\right)^2 dy, \tag{2.5}$$

and denoting by AMISE its asymptotic version, the following result is a consequence of Theorem 2.1.

Corollary 2.2. *Under conditions (A.1), (A.2), (A.3) and (A.5),*

$$\mathrm{MISE}\,(h) = \frac{1}{4}h^4\mu_2^2(K)R(f'') + \frac{R(K)\mu c}{nh} + o\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{AMISE}\,(h) = \frac{1}{4}h^4\mu_2^2(K)R(f'') + \frac{R(K)\mu c}{nh}, \ \textit{with} \ \ c = \int \frac{1}{y}f(y)dy.$$

As a consequence, the optimal bandwidth value which minimises AMISE(h) *is:*

$$h_{\mathrm{AMISE}} = \left(\frac{R(K)\mu c}{n\mu_2^2(K)R(f'')}\right)^{1/5}. \tag{2.6}$$

– Proof –

The MISE is computed just by integrating the expression of the MSE obtained in the previous result; and its asymptotic version, AMISE, is inmediately obtained removing the negligible terms.

□

Remark 2.1. Regarding the expressions obtained in Theorem 2.1 and Corollary 2.2, we want to point out the similarities between them and the analogous results in the kernel density estimator for complete data defined by Parzen (1962) and Rosenblatt (1956): the main terms differ only in constants, the convergence rates are preserved, and hence, the expression of the optimal bandwidths are also resemble.

2.2.1 Resampling bootstrap methods

In this section we develop two different bootstrap procedures that can be applied in the context of length-biased data. Both of them are consistent in the way it is shown below and they conform the basis to define different data-driven bandwidth selection methods.

Bootstrapping using Jones' estimator

In this first method we follow the work by Cao (1990); Cao (1993) using the so-called smooth bootstrap to develop a bandwidth selector for the kernel density estimator of Jones (1991), given in (2.1). It is remarkable that one bootstrap bandwidth selector can be implemented in practice without requiring resampling and any Monte Carlo approximation.

Given an i.i.d. sample, $Y_1, \ldots, Y_n$ from f_Y, and $\hat{f}_g$ the density estimator introduced in (2.1) with pilot bandwidth g, the smooth bootstrap samples, $Y_1^*, \ldots, Y_n^*$, are generated by sampling randomly with replacement n times from the estimated density $\hat{f}_{Y,g}(y) = y\hat{f}_g(y)/\hat{\mu}$.

Let Y^* denote the random variable generated by the bootstrap method presented above. From the bootstrap sample let define the bootstrap density estimator of Y^* as

$$\hat{f}_h^*(y) = \frac{1}{n}\hat{\mu}^* \sum_{i=1}^{n} \frac{1}{Y_i^*} L_h\left(y - Y_i^*\right), \tag{2.7}$$

where $\hat{\mu}^* = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i^*}\right)^{-1}$, and $L_h(\cdot) = \frac{1}{h}L(\frac{\cdot}{h})$, with L being a symmetric kernel function fulfilling the same conditions as K but not necessarily being the same function.

The following result provides the expression of the mean, the variance and the mean squared error of $\hat{f}_h^*(y)$ under the bootstrap distribution. We use the notation E^*, Var^* and MSE^* to refer to the bootstrap distribution.

Theorem 2.3. *Under conditions (A.1) to (A.4)*

$$E^*\left[\hat{f}_h^*(y)\right] = \left(L_h \circ \hat{f}_g\right)(y) + O_P\left(\tfrac{1}{n}\right) \text{ and}$$
$$\mathrm{Var}^*\left[\hat{f}_h^*(y)\right] = n^{-1}\left[\left(L_h^2 \circ \hat{\gamma}_g\right)(y) - \left(L_h \circ \hat{f}_g\right)^2(y)\right] + O_P\left(\tfrac{1}{n}\right).$$

Moreover, adding condition (A.6), we obtain

$$\mathrm{MSE}^*(h,y) = \frac{1}{4}h^4\left(\hat{f}_g''(y)\right)^2\mu_2^2(L) + \frac{\hat{\gamma}_g(y)}{nh}R(L) + o_P\left(h^4 + \frac{1}{nh}\right),$$

where $\hat{\gamma}_g(y) = \hat{\mu}\hat{f}_g(y)/y$, $\mu_2(L) = \int u^2 L(u)du$ *and* $R(L) = \int L^2(u)du$.

– Proof –

We now obtain the MSE of the bootstrap estimator under the bootstrap distribution. To this aim we follow similar steps as in the proof of Theorem 2.1. Remind that now, the estimator is given by (2.7), and it can be rewritten as follows:

$$\hat{f}_h^*(y) = \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i^*}L_h(y - Y_i^*)}{\frac{1}{n}\sum_{j=1}^n \frac{1}{Y_j^*}} = \frac{\phi_n^*(y)}{\xi_n^*}.$$

From the expression above we compute the mean of the numerator and the denominator

$$\overline{\phi}_n^*(y) := E^*\left[\phi_n^*(y)\right] = E^*\left[\tfrac{1}{n}\sum_{i=1}^n \tfrac{1}{Y_i^*}L_h(y - Y_i^*)\right] = \int \tfrac{1}{z}L_h(y-z)\hat{f}_{Y,g}(z)dz$$
$$= \tfrac{1}{\hat{\mu}}\left(L_h \circ \hat{f}_g\right)(z)$$
and
$$\overline{\xi}_n^* := E^*[\xi_n^*] = E^*\left[\tfrac{1}{n}\sum_{j=1}^n \tfrac{1}{Y_j^*}\right] = \int \tfrac{1}{z}\hat{f}_{Y,g}(z)dz = \tfrac{1}{\hat{\mu}}.$$

Using the linearisation procedure in Collomb (1976) with $\nu \geq 2$ we have that

$$E^*[\hat{f}_h^*(y)] = \frac{\overline{\phi}_n^*(y)}{\overline{\xi}_n^*} + c_n^*(y) + c_n^{*(\nu)}(y) + \frac{(-1)^\nu \sigma_n^{*1,\nu}(y)}{\overline{\xi}_n^\nu},$$

where

$$
\begin{aligned}
c_n^*(y) &= \frac{\overline{\phi}_n^*(y)S_n^{*0,2}-\overline{\xi}_n^*S_n^{*1,1}}{\overline{\xi}_n^{*3}} = \frac{\overline{\phi}_n^*(y)E^*\left[\left(\xi_n^*-\overline{\xi}_n^*\right)^2\right]-\overline{\xi}_n^*E^*\left[\phi_n^*(y)\left(\xi_n^*-\overline{\xi}_n^*\right)\right]}{\overline{\xi}_n^{*3}},\\
c_n^{*(\nu)}(y) &= \frac{s_n^{*1,2}(y)}{\overline{\xi}_n^{*3}} + \sum_{k=3}^{\nu-1}(-1)^k\frac{S_n^{*1,k}(y)}{\overline{\xi}_n^{*k+1}} = \frac{E^*\left[\left(\phi_n^*(y)-\overline{\phi}_n^*(y)\right)\left(\xi_n^*-\overline{\xi}_n^*\right)^2\right]}{\overline{\xi}_n^{*3}}\\
&\quad + \sum_{k=3}^{\nu-1}(-1)^k\frac{E^*\left[\phi_n^*(y)\left(\xi_n^*-\overline{\xi}_n^*\right)^k\right]}{\overline{\xi}_n^{*k+1}} \quad \text{and}\\
\sigma_n^{*1,\nu}(y) &= E^*\left[\hat{f}_h^*(y)\left(\xi_n^*-\overline{\xi}_n^*\right)^\nu\right].
\end{aligned}
$$

To obtain the variance of the bootstrap estimator we compute

$$
E^*[\hat{f}_h^{*2}(y)] = \frac{\overline{\phi}_n^{*2}(y)}{\overline{\xi}_n^{*2}} + \varphi_n^*(y) + \Gamma_n^{*(\nu)}(y) + (-1)^\nu\Delta^{*(\nu)}(y),
$$

with

$$
\begin{aligned}
\varphi_n^*(y) &= \frac{s_n^{*2,0}(y)}{\overline{\xi}_n^{*2}} + 3\frac{\overline{\phi}_n^{*2}(y)s_n^{*0,2}}{\overline{\xi}_n^{*4}} - 4\frac{\overline{\phi}_n^*(y)s_n^{*1,1}}{\overline{\xi}_n^{*3}},\\
\Gamma_n^{*(\nu)}(y) &= -2\frac{s_n^{*2,1}(y)}{\overline{\xi}_n^{*3}} + 3\left(\frac{s_n^{*2,2}(y)}{\overline{\xi}_n^{*4}} + \frac{2\overline{\phi}_n^*(y)s_n^{*1,2}}{\overline{\xi}_n^{*4}}\right) + 4\frac{S_n^{*2,3}(y)}{\overline{\xi}_n^{*5}} + \frac{S_n^{*2,4}(y)}{\overline{\xi}_n^{*6}} +\\
&\quad + \sum_{k=3}^{\nu-1}(-1)^k\sum_{j=0}^{k}\frac{k!2^{k-j}}{j!(k-j)!}\frac{S_n^{*2,k+j}(y)}{\overline{\xi}_n^{*k+j+2}} \quad \text{and}\\
\Delta^{*(\nu)}(y) &= \sum_{j=0}^{\nu}\frac{\nu!}{j!(\nu-j)!}2^{\nu-j}\frac{\sigma_n^{*2,\nu+j}(y)}{\overline{\xi}_n^{*\nu+j}}.
\end{aligned}
$$

Here the dominant terms are $\frac{\overline{\phi}_n^{*2}(y)}{\overline{\xi}_n^{*2}} + \frac{s_n^{*2,0}(y)}{\overline{\xi}_n^{*2}}$, which are the ones we need to study.

$$
\begin{aligned}
\frac{\overline{\phi}_n^{*2}(y)}{\overline{\xi}_n^{*2}} &= \frac{\left(\frac{1}{\hat{\mu}}(L_h\circ\hat{f}_g)(y)\right)^2}{\left(\frac{1}{\hat{\mu}}\right)^2},\\
s_n^{*2,0}(y) &= E^*\left[\left(\phi_n^*(y)-\overline{\phi}_n^*(y)\right)^2\right] = \mathrm{Var}^*\left[\phi_n^*(y)\right] = E^*\left[\phi_n^{*2}(y)\right] - E^*\left[\phi_n^*(y)\right]^2\\
&= E^*\left[\phi_n^{*2}(y)\right] - \overline{\phi}_n^{*2}(y) = \tfrac{1}{n\hat{\mu}^2}(L_h^2\circ\hat{\gamma}_g)(y) + \tfrac{n-1}{n\hat{\mu}^2}(L_h\circ\hat{f}_g)^2(y)\\
&\quad -\tfrac{1}{\hat{\mu}^2}(L_h\circ\hat{f}_g)^2(y) = \tfrac{1}{n\hat{\mu}^2}(L_h^2\circ\hat{\gamma}_g)(y) - \tfrac{1}{n\hat{\mu}^2}(L_h\circ\hat{f}_g)^2(y) \quad \text{and then}\\
\frac{s_n^{*2,0}(y)}{\overline{\xi}_n^{*2}} &= \hat{\mu}^2 s_n^{*2,0}(y) = \tfrac{1}{n}(L_h^2\circ\hat{\gamma}_g)(y) - \tfrac{1}{n}(L_h\circ\hat{f}_g)^2(y),
\end{aligned}
$$

where we have taken into account that $E^*[\phi_n^{*2}(y)] = \frac{1}{n\hat{\mu}}(L_h^2\circ\hat{\gamma}_g)(y) + \frac{n-1}{n\hat{\mu}^2}(L_h\circ\hat{f}_g)^2(y)$. As the other terms are negligible, we get

$$
\mathrm{Var}^*\left[\hat{f}_h^*(y)\right] = \frac{1}{n}\left[(L_h^2\circ\hat{\gamma}_g)(y) - (L_h\circ\hat{f}_h)^2(y)\right].
$$

Finally, under the regularity conditions previously established we get that

$$\mathrm{MSE}^*(h,y) = \frac{1}{4}h^4\left(\hat{f}_g''(y)\right)^2\mu_2^2(K) + \frac{\hat{\gamma}_g(y)}{nh}R(L) + o_P\left(h^4 + \frac{1}{nh}\right).$$

□

The same way we have done in Corollary 2.2, the integrated versions of the MSE*, MISE* and its asymptotic version are easily deduced from the theorem above.

Corollary 2.4. *Under hypotheses (A.2), (A.3), (A.4) and (A.6)*

$$\mathrm{MISE}^*(h) = \tfrac{1}{4}h^4\mu_2^2(L)R(\hat{f}_g'') + \tfrac{R(L)\hat{\mu}\hat{c}}{nh} + o_P\left(h^4 + \tfrac{1}{nh}\right),$$

$$\mathrm{AMISE}^*(h) = \tfrac{1}{4}h^4\mu_2^2(L)R(\hat{f}_g'') + \tfrac{R(L)\hat{\mu}\hat{c}}{nh}, \qquad \text{with } \hat{c} = \hat{\mu}\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{1}{Y_i^2}.$$

Therefore, the asymptotic expression of the optimal bootstrap bandwidth is:

$$h_{\mathrm{AMISE}^*} = \left(\frac{R(L)\hat{\mu}\hat{c}}{n\mu_2^2(L)R(\hat{f}_g'')}\right)^{1/5},$$

which is a plug-in version of (2.6).

– *Proof* –

Similarly to Corollary 2.2, the proof of this result is an immediate consequence of the result above.

□

The following corollary is a consequence of the previous results.

Corollary 2.5. *Under assumptions (A.1) to (A.6),* MISE* *and* AMISE* *are consistent estimators of* MISE *and* AMISE*, respectively.*

– *Proof* –

Taking into account the expressions obtained in Theorem 2.1, Theorem 2.3, Corollary 2.2 and Corollary 2.4 this result is immediate to deduce.

□

Remark 2.2. As in Remark 2.1, we want to emphasize the similarities between the expressions obtained in Theorem 2.3 and Corollary 2.4 and the analogous ones for the kernel density estimator defined by Parzen (1962) and Rosenblatt (1956) using the smooth bootstrap procedure proposed by Cao (1993).

Bootstrapping using a common kernel density estimator

In this second method, we are also using a smooth bootstrap procedure but considering the common kernel density estimator by Parzen (1962) and Rosenblatt (1956) instead of Jones'. The idea of defining this second procedure is to be able to use the methodology developed in Bose and Dutta (2013) for bandwidth selection in kernel density estimation.

Given an i.i.d. sample, $Y_1, \ldots, Y_n$ from f_Y, and denoting by $\tilde{f}_{K,g}$ the common kernel density estimator with pilot bandwidth g and a kernel function K, the smooth bootstrap samples, $Y_1^*, \ldots, Y_n^*$, are generated by sampling randomly with replacement n times from $\tilde{f}_{K,g}$. Let Y^* denote again the random variable generated by the bootstrap method presented above. From the bootstrap sample let us define the bootstrap density estimator of Y^* as the one presented in (2.7), taking into account that the bootstrap sample is generated differently.

Now we provide the expression for the pointwise mean and variance of $\hat{f}_h^*(y)$ under the bootstrap distribution.

Theorem 2.6. *Under conditions (A.1) to (A.4)*

$$E^*\left[\hat{f}_h^*(y)\right] = \frac{1}{\int \frac{1}{z}\tilde{f}_{K,g}(z)dz}\int \frac{1}{z}L_h(y-z)\tilde{f}_{K,g}(z)dz + O_P\left(\frac{1}{n}\right) \text{ and}$$

$$\begin{aligned}\mathrm{Var}^*\left[\hat{f}_h^*(y)\right] &= \frac{1}{n\left(\int \frac{1}{z}\tilde{f}_{K,g}(z)dz\right)^2}\left[\int \frac{1}{z^2}L_h^2(y-z)\tilde{f}_{K,g}(z)dz - \int \frac{1}{z}L_h(y-z)\tilde{f}_{K,g}(z)dz^2\right] \\ &\quad + O_P\left(\frac{1}{n}\right).\end{aligned}$$

Moreover,

$$\begin{aligned}\mathrm{MSE}^*(h,y) &= \left(\frac{\int \frac{1}{z}L_h(y-z)\tilde{f}_{K,g}(z)dz}{\int \frac{1}{z}\tilde{f}_{K,g}(z)dz} - \hat{f}_h(y)\right)^2 \\ &\quad + \frac{1}{n\left(\int \frac{1}{z}\tilde{f}_{K,g}(z)dz\right)^2}\left[\int \frac{1}{z^2}L_h^2(y-z)\tilde{f}_{K,g}(z)dz - \int \frac{1}{z}L_h(y-z)\tilde{f}_{K,g}(z)dz^2\right] \\ &\quad + O_P\left(\frac{1}{n}\right).\end{aligned}$$

Remark that for this bootstrap method we do not get manageable explicit expressions of the error criteria as we got in the previous one; and the way to obtain MISE$^*(h)$ *is integrating the expression above, but we neither obtain an explicit expression.*

– Proof –

This proof follows similar steps as the one of Theorem 2.3, with the particularity that the generation of the bootstrap sample is made differently.

Let us remind the expression of the bootstrap estimator

$$\hat{f}_h^*(y) = \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i^*} L_h\left(y - Y_i^*\right)}{\frac{1}{n}\sum_{j=1}^n \frac{1}{Y_j^*}} = \frac{\phi_n^*(y)}{\xi_n^*}.$$

Firstly we obtain the mean. Following again the linearisation procedure in Collomb (1976) and the previous proof, we only need the mean of the numerator and the denominator, so:

$$\overline{\phi}_n^*(y) := E^*\left[\phi_n^*(y)\right] = E^*\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{Y_i^*} L_h\left(y - Y_i^*\right)\right] = \int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz$$

and

$$\overline{\xi}_n^* := E^*[\xi_n^*] = E^*\left[\frac{1}{n}\sum_{j=1}^n \frac{1}{Y_j^*}\right] = \int \frac{1}{z}\tilde{f}_{K,g}(z)dz.$$

Recall that in this context, $\tilde{f}_{K,g}$ denote the common kernel density estimator with kernel K and bandwidth g.

Hence,

$$E^*\left[\hat{f}_h^*(y)\right] = \frac{\overline{\phi}_n^*(y)}{\overline{\xi}_n^*} + O_P\left(1/n\right) = \frac{\int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz}{\int \frac{1}{z}\tilde{f}_{K,g}(z)dz} + O_P\left(1/n\right).$$

To compute the variance we follow again the previous proof, and taking only into account the dominant terms we get

$$E^*[\hat{f}_h^{*^2}(y)] = \frac{\overline{\phi}_n^{*^2}(y)}{\overline{\xi}_n^{*\,2}} + \frac{s_n^{*^{2,0}}(y)}{\overline{\xi}_n^{*\,2}} + O_P(1/n),$$

where

$$\begin{aligned}
\frac{\overline{\phi}_n^{*^2}(y)}{\overline{\xi}_n^{*\,2}} &= \frac{\left(\int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz\right)^2}{\left(\int \frac{1}{z}\tilde{f}_{K,g}(z)dz\right)^2} \text{ and} \\
s_n^{*^{2,0}}(y) &= E^*\left[\left(\phi_n^*(y) - \overline{\phi}_n^*(y)\right)^2\right] = \mathrm{Var}^*\left[\phi_n^*(y)\right] = E^*\left[\phi_n^{*^2}(y)\right] - E^*\left[\phi_n^*(y)\right]^2 \\
&= E^*\left[\phi_n^{*^2}(y)\right] - \overline{\phi}_n^{*^2}(y) = \frac{1}{n}\int \frac{1}{z^2} L_h^2(y-z)\tilde{f}_{K,g}(z)dz \\
&\quad -\frac{1}{n}\left(\int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz\right)^2,
\end{aligned}$$

considering that $E^*[\phi_n^{*2}(y)] = \frac{1}{n}\int \frac{1}{z^2} L_h^2(y-z)\tilde{f}_{K,g}(z)dz - \frac{n-1}{n}\left(\int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz\right)^2$.

Then we get

$$\mathrm{Var}^*\left[\hat{f}_h^*(y)\right] = \frac{1}{n}\int \frac{1}{z^2} L_h^2(y-z)\tilde{f}_{K,g}(z)dz - \frac{1}{n}\left(\int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz\right)^2.$$

Finally, just noting that MSE* can be computed as the sum of the squared bias and the variance we obtain the final equation.

□

2.3 Bandwidth selection

In this section we describe bandwidth selection methods for the density estimator defined in (2.1). These methods consist of adaptations of common automatic selectors for kernel density estimation with complete data to the context of length-biased data. We propose a Normal scale rule and two bootstrap selectors derived from the consistent resampling procedures given in the previous section. These proposals are defined as competitors of the cross-validation method proposed in Guillamón et al. (1998).

Two of these new methods are based on estimating the infeasible optimal expression (2.6), in which the unknown elements are $R(f'')$, c and μ. However, we have previously shown that these last two terms can be easily estimated, and then the only term that still needs to be estimated is $R(f'')$. The last bootstrap bandwidth selection procedure is based on the minimisation of the $MISE^*(h)$ and does not require those estimations.

2.3.1 Rule-of-thumb

This method is based on the rule-of-thumb, Silverman (1986), for complete data. The idea is to assume that the underlying distribution is Normal, $N(\mu, \sigma)$, and in this situation

$$R(f'') = \frac{3}{8}\pi^{-1/2}\sigma^{-5}.$$

To get a suitable estimator of σ in the context of length-biased data is not trivial. We suggest to estimate it as follows. Cox (2005) states that $E_Y[X^r] = \frac{\mu_{r+1}}{\mu}$, where μ_{r+1} denotes the (r+1)-th order moment of not observable variable X. So,

$$\mu_2 = \mu E_Y[X] \Rightarrow \hat{\mu}_2 = \hat{\mu}\widehat{E_Y[X]} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\right)^{-1}\left(\frac{1}{n}\sum_{j=1}^{n}Y_j\right),$$

thus,

$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2 = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\right)^{-1}\left[\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i}\right)^{-1}\right],$$

and finally

$$\hat{h}_{\mathrm{RT}} = \left(\frac{R(K)\hat{\mu}\hat{c}8\sqrt{\pi}}{n\mu_2^2(K)3} \right)^{1/5} \hat{\sigma}.$$

Another possible estimator for σ could be obtained using a robust method such as the interquartile range (IQR)

$$\hat{\sigma}_{\mathrm{IQR}} = \frac{\mathrm{IQR}}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)},$$

where Φ is the Normal distribution function.

2.3.2 Cross-validation

The method previously defined is based on minimising estimations of the MISE, more precisely of the AMISE. This procedure relies on the minimisation of the integrated squared error (ISE), the methodology is the same as in Rudemo (1982) and Bowman (1984) applied to (2.1), and it was developed in Guillamón et al. (1998).

Let write:

$$\mathrm{ISE}(h) = \int \left(\hat{f}_h(z) - f(z) \right)^2 dz = \int \hat{f}_h^2(z)dz - 2\int \hat{f}_h(z)f(z)dz + \int f^2(z)dz. \quad (2.8)$$

Note that $\int f^2(z)dz$ does not depend on h, so the minimisation of the ISE is equivalent to minimise the following function:

$$\int \hat{f}_h^2(z)dz - 2\int \hat{f}_h(z)f(z)dz = \int \hat{f}_h^2(z)dz - 2E[\hat{f}_h],$$

which can be estimated by

$$\mathrm{CV}(h) = \int \hat{f}_h^2(z)dz - 2\widehat{E[\hat{f}_h]}.$$

The addends of this estimation may be expressed as follows:

$$\begin{aligned} \int \hat{f}_h^2(z)dz &= \int \left(\tfrac{1}{nh}\hat{\mu} \textstyle\sum_{i=1}^n \tfrac{1}{Y_i} K\left(\tfrac{z-Y_i}{h} \right) \right) \left(\tfrac{1}{nh}\hat{\mu} \textstyle\sum_{j=1}^n \tfrac{1}{Y_j} K\left(\tfrac{z-Y_j}{h} \right) \right) \\ &= n^{-2}h^{-1}\hat{\mu}^2 \textstyle\sum_{i=1}^n \sum_{j=1}^n \tfrac{1}{Y_i}\tfrac{1}{Y_j}(K \circ K)\left(\tfrac{Y_i - Y_j}{h} \right) \text{ and} \end{aligned}$$

$$\widehat{E[\hat{f}_h]} = \hat{\mu}n^{-1} \sum_{i=1}^n \frac{\hat{f}_{-i}(Y_i)}{Y_i} = \hat{\mu}n^{-1} \sum_{i=1}^n Y_i^{-1} \left(\sum_{j \neq i} \frac{1}{Y_j} \right)^{-1} \left(\sum_{j \neq i} \frac{1}{Y_j} K_h(Y_i - Y_j) \right),$$

realising that $E[\hat{f}_h] = \int \hat{f}_h(z)f(z)dz = \int \hat{f}_h(z)\frac{\mu f_Y(z)}{z}dz$, and defining $\hat{f}_{-i}$ as the estimator in (2.1) calculated with all the data points except Y_i.

The cross-validation bandwidth is obtained following the proposal in Dutta (2016), where the CV function is minimized in a compact interval of the form $[c_1\text{IQR}n^{-1/5}, c_2\text{IQR}(\log(n)/n)^{1/5}]$, where IQR is the inter-quartile range and c_1 and c_2 are positive constants (see Dutta (2016) for the choice of these values). We will be denoted hereafter this bandwidth value as $\hat{h}_{\text{CV}}$.

2.3.3 Bootstrap for bandwidth selection

Using Jones' estimator

The asymptotic expression of the optimal bootstrap bandwidth can be considered to derive a consistent bandwidth estimate. Cao (1993) suggested such approach for kernel density estimation with complete data. Since all the quantities involved in the expression are known, the result will be a bandwidth estimate which can be computed in practice without involving any resampling and Monte Carlo approximations. The only issue is to determine the pilot bandwidth g involved in the estimation of $R(f'')$. To this goal we first obtain the asymptotical (infeasible) optimal pilot bandwidth and then we propose two feasible estimations.

We define the optimal pilot bandwidth by optimising the MSE of $R(\hat{f}_g'') = \frac{1}{n}\hat{\mu}\sum_{i=1}^{n}\frac{1}{Y_i}\frac{1}{h^3}L''\left(\frac{y-Y_i}{h}\right)$ as an estimator of $R(f'')$. Let $\hat{f}_g$ be the estimator in (2.1) with L a symmetric kernel function and assume the following conditions:

(A.7) $\int |u|^3 L(u)du < \infty$,

(A.8) L is twice differentiable, with bounded second derivative and verifies that $\lim_{u\to\pm\infty} u^3L(u) = 0$, $\lim_{u\to\pm\infty} u^4L'(u)du = 0$; $\int |u|^4|L''(u)|du < \infty$, $\int L''^2(u)du < \infty$,

(A.9) $\int u^4 L(u)du < \infty$ and

(A.10) f is six times differentiable with $f, f'', f''', f^{(4)} \in L_2(\mathbb{R})$ and verifies the limit condition $\lim_{y\to\pm\infty} f''(y)f'''(y) = 0$.

The result below provides us with the exact value of the optimal pilot bandwidth, in terms of the asymptotic mean squared error (AMSE) of the curvature of Jones' estimator.

Theorem 2.7. *Under hypotheses (A.7) to (A.10) we have that:*

$$\begin{aligned}\mathrm{AMSE}\left(\int(\hat{f}_g''(y))^2dy\right) &= n^{-2}g^{-10}\left(\int L''(y)^2dy\right)c^2\mu^2 + g^4\mu_2^2(L)\left(\int f''''(y)^2dy\right)^2 \\ &\quad +2n^{-1}g^{-3}\mu_2(L)c\mu\int L''(y)^2dy\end{aligned}$$

and

$$g_0 = \arg\min_g \mathrm{AMSE}\left(R(\hat{f}_g''(y))\right) = d_0 n^{-1/7},$$

with

$$d_0 = \left[\frac{5}{2}\mu_2(L)^{-1}c\mu\int(L''(y))^2dy\left(\int(f''''(y))^2dy\right)^{-1}\right]^{1/7}.$$

– Proof –

We calculate the mean and the variance of $R(\hat{f}_g'')$, as an estimator of $R(f'')$, in order to determine its MSE and the expression of the optimal pilot bandwidth g. To this purpose we use the U-statistics theory and its projections, as it has been done for complete data in Cao (1990).

We start by calculating its mean. First of all rewrite

$$\begin{aligned}R(\hat{f}_g'') = \int\left(\hat{f}_g''(y)\right)^2dy &= n^{-1}g^{-5}\int\left(L''(u)\right)^2du\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n\frac{1}{Y_i^2}\right)\hat{\mu}^2 \\ &\quad +n^{-2}g^{-6}\hat{\mu}^2\textstyle\sum_{i\neq j}\frac{1}{Y_i}\frac{1}{Y_j}\int L''\left(\frac{y-Y_i}{g}\right)L''\left(\frac{y-Y_j}{g}\right)dy.\end{aligned}$$

Hence,

$$\begin{aligned}E\left[\int\hat{f}_g''^2(y)dy\right] &= n^{-1}g^{-5}\int\left(L''(u)\right)^2duE\left[\frac{\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{Y_i^2}\right)}{\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{Y_i}\right)^2}\right] \\ &\quad +n^{-2}g^{-6}E\left[\frac{\sum_{i\neq j}\frac{1}{Y_i}\frac{1}{Y_j}\int L''\left(\frac{y-Y_i}{g}\right)L''\left(\frac{y-Y_j}{g}\right)dy}{\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{Y_i}\right)^2}\right] \\ &= n^{-1}g^{-5}\int\left(L''(u)\right)^2duc\mu + n^{-1}(n-1)g^{-6}\mu^2\int\mu_g^2(y)dy + o(n^{-2}g^{-5}),\end{aligned} \tag{2.9}$$

where $\mu_g(y) := E\left[\frac{1}{Y}L''\left(\frac{y-Y}{g}\right)\right]$ and we use that

$$E\left[\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n\frac{1}{Y_i}\right)^2\right] = \frac{1}{\mu^2}+O(n^{-1}) \quad \text{and} \quad E\left[\tfrac{1}{n}\textstyle\sum_{i=1}^n\frac{1}{Y_i^2}\right] = E\left[\frac{1}{Y^2}\right] = \frac{c}{\mu},$$

which have been obtained using Taylor expansions.

Moreover, using again Taylor expansion and the regularity conditions imposed on L and f, we can rewrite

$$\mu_g(y) = \tfrac{1}{\mu}\left(g^3 f''(y) + \tfrac{1}{6}g^5 \int u^4 L''(u) \int (1-t)^3 f^{(iv}(y-gut)dtdu\right),$$

and applying Fubini's theorem and Cauchy-Schwarz inequality, we obtain

$$\int \mu_g^2(y)dy = \tfrac{1}{\mu^2}\left(g^6 \int f''^2(y)dy + g^8\mu_2(L)\int (f'''(y))^2dy + o(g^9)\right).$$

Then, replacing this value in (2.9), we get the mean of the estimator as follows:

$$\begin{aligned} E\left[\int \hat{f}_g''^2(y)dy\right] &= n^{-1}g^{-5}c\mu \int \left(L''(u)\right)^2 du + n^{-1}(n-1)\int (f''(y))^2dy \\ &\quad + n^{-1}(n-1)g^2\mu_2(L)\int (f'''(y))^2dy + o(g^3). \end{aligned}$$

Once having the mean, the bias can be immediately obtained:

$$Bias\left[R\left(\hat{f}_h''\right)\right] = n^{-1}g^{-5}c\mu \int \left(L''(u)\right)^2 du + g^2\mu_2(L)\int (f'''(y))^2dy + o(n^{-1}) + o(g^3). \tag{2.10}$$

The next step is calculate the variance. To this aim we need to rewrite the expression of the estimator $R(\hat{f}_g'')$ in an appropriate way.

$$\begin{aligned} R\left(\hat{f}_g''\right) &= n^{-1}g^{-5}\int (L''(u))^2du\hat{\mu}^2\left(\tfrac{1}{n}\textstyle\sum_{i=1}\tfrac{1}{Y_i^2}\right) \\ &\quad + n^{-2}g^{-6}\hat{\mu}^2\textstyle\sum_{i\neq j}\tfrac{1}{Y_i}\tfrac{1}{Y_j}\int L''\left(\tfrac{y-Y_1}{g}\right)L''\left(\tfrac{y-Y_j}{g}\right)dy \\ &= n^{-1}g^{-5}\int (L''(u))^2du\hat{\mu}^2\left(\tfrac{1}{n}\textstyle\sum_{i=1}\tfrac{1}{Y_i^2}\right) + n^{-2}g^{-6}\hat{\mu}^2\textstyle\sum_{i\neq j}[H_n(Y_i,Y_j) \\ &\quad + \int\left(\tfrac{1}{Y_i}L''\left(\tfrac{y-Y_i}{g}\right) - \mu_g(y)\right)\mu_g(y) + \int\left(\tfrac{1}{Y_j}L''\left(\tfrac{y-Y_j}{g}\right) - \mu_g(y)\right)\mu_g(y) + \int\mu_g^2(y) \\ &= n^{-1}g^{-5}\int (L''(u))^2du\hat{\mu}^2\left(\tfrac{1}{n}\textstyle\sum_{i=1}\tfrac{1}{Y_i^2}\right) + n^{-1}(n-1)g^{-6}\hat{\mu}^2\int\mu_g^2(y)dy \\ &\quad + n^{-2}g^{-6}\hat{\mu}^2\textstyle\sum_{i\neq j}H_n(Y_i,Y_j) + n^{-2}g^{-6}\hat{\mu}^2 2(n-1)\sum_{i=1}^n W_i, \end{aligned}$$

where $H_n(Y_i,Y_j) := \int\left(\tfrac{1}{Y_i}L''\left(\tfrac{y-Y_i}{g}\right) - \mu_g(y)\right)\left(\tfrac{1}{Y_j}L''\left(\tfrac{y-Y_j}{g}\right) - \mu_g(y)\right)$,
$W_i = \int\left(\tfrac{1}{Y_i}L''\left(\tfrac{y-Y_i}{g}\right) - \mu_g(y)\right)\mu_g(y)dy \quad i = 1,\ldots,n$ and we use that

$$\begin{aligned} \int L''\left(\tfrac{y-Y_1}{g}\right)L''\left(\tfrac{y-Y_j}{g}\right)dy &= H_n(Y_i,Y_j) + \int\left(\tfrac{1}{Y_i}L''\left(\tfrac{y-Y_i}{g} - \mu_g(y)\right)\mu_g(y)dy\right) \\ &\quad + \int\left(\tfrac{1}{Y_j}L''\left(\tfrac{y-Y_j}{g} - \mu_g(y)\right)\mu_g(y)dy\right) + \int\mu_g^2(y)dy \quad \text{and} \end{aligned}$$

$$\begin{aligned} \int \tfrac{1}{Y_i}L''\left(\tfrac{y-Y_i}{g}\right)\tfrac{1}{Y_j}L''\left(\tfrac{y-Y_j}{g}\right) &= H_n(Y_i,Y_j) + \int\left(\tfrac{1}{Y_i}L''\left(\tfrac{y-Y_i}{g}\right) - \mu_g(y)\right)\mu_g(y)dy \\ &\quad + \int\left(\tfrac{1}{Y_j}L''\left(\tfrac{y-Y_j}{g}\right) - \mu_g(y)\right)\mu_g(y)dy + \int\mu_g''(y)dy. \end{aligned}$$

Now we calculate the variance of each term. For the first two we use common statistical techniques while for the others we need to do more complex expansions.

$$\begin{aligned}\mathrm{Var}\left[n^{-1}g^{-5}\int (L''(u))^2du\hat{\mu}^2\left(\tfrac{1}{n}\sum_{i=1}\tfrac{1}{Y_i^2}\right)\right] &= n^{-2}g^{-10}\int (L''(u))^2du^2\mathrm{Var}\left[\frac{\frac{1}{n}\sum_{i=1}^n\frac{1}{Y_i^2}}{\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{Y_i}\right)}\right]\\ &= o(n^{-3}g^{-10}) \text{ and}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left[n^{-1}(n-1)g^{-6}\hat{\mu}^2\int\mu_g^2(y)dy\right] &= n^{-2}(n-1)^2g^{-12}\int\mu_g(y)dy\mathrm{Var}\left[\frac{1}{\left(\frac{1}{n}\sum_{i=1}^n\frac{1}{Y_i}\right)^2}\right]\\ &= o(n^{-1}g^{-12}).\end{aligned}$$

Before considering the variance of the third term we need some previous developments related to $H_n(Y_i, Y_j)$.

Firstly, we obtain the expression of $\int\int\mu_g(x)\mu_g(y)\int L''\left(\frac{x-z}{g}\right)L''\left(\frac{y-z}{g}\right)\gamma(z)dzdxdy$ using Taylor expansions on f and γ:

$$\begin{aligned}\int\int\mu_g(x)\mu_g(y)\int L''\left(\tfrac{x-z}{g}\right)L''\left(\tfrac{y-z}{g}\right)\gamma(z)dzdxdy &= \tfrac{g^8}{\mu^2}\int\int\int\int\int_0^1\int_0^1 u_1^2L''(u_1)\\ &u_2^2L''(u_2)(1-t_1)(1-t_2)f''(x-gu_1t_1)f''(x+gw-gu_2t_2)L''(v)L''(v+w)\\ &\gamma(x-gv)dt_1dt_2du_1du_2dvdwdx = I_0-I_1+I_2-I_3+I_4,\end{aligned} \tag{2.11}$$

with

$$\begin{aligned}I_j = \tfrac{1}{\mu^2}\tfrac{g^{8+j}}{j!}\int u_1^2L''(u_1)\int u_2^2L''(u_2)\int_0^1(1-t_1)\int_0^1(1-t_2)\int f''(x-gu_1t_1)\gamma^{(j)}(x)\\ \int f''(x+gw-gu_2t_2)N_j(w)dxdwdt_1dt_2du_1du_2 \quad j=0,\ldots,3 \text{ and}\end{aligned}$$

$$\begin{aligned}I_4 = \tfrac{1}{\mu^2}\tfrac{g^{12}}{6}\int u_1^2L''(u_1)\int u_2^2L''(u_2)\int_0^1(1-t_1)\int_0^1(1-t_2)\int f''(x-gu_1t_1)\\ \int f''(x+gw-gu_2t_2)\int v^4L''(v)L''(v+w)\int_0^1(1-z^3)\gamma^{(iv}(x-gvz)\\ dzdvdwdxdt_2dt_1du_2du_1.\end{aligned}$$

From these expressions, we obtain:

$$\begin{aligned}I_0 = \tfrac{g^{12}}{\mu^2}\int\gamma(y)f''(y)f^{(vi}(y)dy+o(g^{12}), \quad I_1 = \tfrac{-2g^{12}}{\mu^2}\int\gamma'(y)f''(y)f^{(v}(y)dy+o(g^{12}),\\ I_2 = \tfrac{g^{12}}{\mu^2}\int\left(\gamma''(y)\right)^2f^{(iv}(y)dy+o(g^{12}), \quad I_3 = o(g^{12}) \quad \text{and} \quad I_4 = o(g^{12}).\end{aligned}$$

Hence, getting back on (2.11), we have

$$\int\int\mu_g(x)\mu_g(y)\int L''\left(\tfrac{x-z}{g}\right)L''\left(\tfrac{y-z}{g}\right)\gamma(z)dzdxdy = \tfrac{g^{12}}{\mu^2}\int\gamma(y)\left(f^{(iv}(y)\right)^2dy+o(g^{12}).$$

Secondly, we need to remark that

$$\begin{aligned}\int\int\left[\int L''\left(\tfrac{x-z}{g}\right)L''\left(\tfrac{y-z}{g}\right)dz\right]^2 dxdy &= g^3\left(\int\gamma^2(z)dz\right)\left(\int(L''\circ L'')^2(v)dv\right)\\ &\quad +o(g^3).\end{aligned}$$

And now, we are in the position to obtain the expression of $E\left[H_n(Y_i,Y_j)^2\right]$:

$$\begin{aligned}E\left[H_n(Y_i,Y_j)^2\right] &= \int\int\left(\tfrac{1}{\mu^2}\int L''\left(\tfrac{x-z}{g}\right)L''\left(\tfrac{y-z}{g}\right)\gamma(z)dz-\mu_g(x)\mu_g(y)\right)^2 dxdy\\ &= \tfrac{1}{\mu^4}\int\int\left(\int L''\left(\tfrac{x-z}{g}\right)L''\left(\tfrac{y-z}{g}\right)\gamma(z)dz\right)^2 dxdy\\ &\quad -\tfrac{2}{\mu^4}\int\int\mu_g(x)\mu_g(y)\int L''\left(\tfrac{x-z}{g}\right)L''\left(\tfrac{y-z}{g}\right)\gamma(z)dzdxdy\\ &\quad +\tfrac{1}{\mu^4}\left(\int\mu_g^2(y)dy\right)^2\\ &= \tfrac{1}{\mu^4}\left(g^3\left(\int\gamma^2(z)dz\right)\left(\int(L''\circ L'')^2(v)dv\right)\right)-\tfrac{2g^{12}}{\mu^6}\int\gamma(y)(f^{(iv}(y))^2dy\\ &\quad +\tfrac{g^{12}}{\mu^8}\int f(y)(f^{(iv}(y))^2dy-\tfrac{g^{12}}{\mu^8}\left(\int f(y)f^{(iv}(y)dy\right)^2+o(g^{12}).\end{aligned}$$

Recall that we aim to obtain the variance of the third term of (2.11), so

$$\begin{aligned}\mathrm{Var}\left[n^{-2}g^{-6}\hat{\mu}^2\textstyle\sum_{i\neq j}H_n(Y_i,Y_j)\right] &= 2n^{-3}g^{-12}\mu^4E\left[H_n^2(Y_i,Y_j)\right]+o(n^{-5}g^{-12})\\ &= 2n^{-2}g^{-9}\int\gamma^2(z)dz\int(L''\circ L'')^2(v)dv+o(n^{-3}g^{-9}).\end{aligned}$$

The variance of the fourth term of (2.11) is calculated as follows:

$$\begin{aligned}&\mathrm{Var}\left[n^{-2}g^{-6}\hat{\mu}^2 2(n-1)\textstyle\sum_{i=1}^n W_i\right]=4n^{-4}(n-1)^2g^{-12}\left(\mathrm{Var}\left[\textstyle\sum_{i=1}^n W_i\right]\mu^4+o(n^{-1})\right)\\ &=4n^{-3}(n-1)^2g^{-12}\mu^4\left[\tfrac{g^{12}}{\mu^4}\int\gamma(y)\left(f^{(iv}(y)\right)^2dy-\tfrac{g^{-12}}{\mu^4}\left(\left(\int(f'''(y))^2dy\right)^2+o(g^{12})\right)\right],\end{aligned}$$

where we have used that W_i are centred and independent variables, as well as the expression of its second order moment.

Lastly, it can be seen, using Cauchy-Schwarz inequality, that the covariate terms between the addends of (2.11) are negligible.

Then, we finally obtain

$$\begin{aligned}\mathrm{Var}\left[R\left(\hat{f}_f''\right)\right] &= 2n^{-2}g^{-9}\int\gamma^2(y)dy\int(L''\circ L'')^2(v)dv+o(n^{-2}g^{-9})+O(n^{-1})\\ &\quad +O(n^{-3/2}n^{-9/2}).\end{aligned}$$

Gathering (2.10) and (2.12), we have

$$\begin{aligned}\mathrm{MSE}\left(R(\hat{f}_g'')\right) &= 2n^{-2}g^{-9}\int\gamma^2(y)dy\int(L''\circ L'')^2(v)dv\\ &+n^{-2}g^{-10}\left(\int(L''(y))^2dy\right)^2c^2\mu^2+g^4\mu_2^2(L)\left(\int(f''''(y))^2dy\right)^2\\ &+2n^{-1}g^{-3}\mu_2(L)c\mu\int(L''(y))^2dy\int(f''''(y))^2dy\\ &+o(n^{-3}g^{-9})+o(n^{-2}g^{-9})+O(n^{-1})+O(n^{-3/2}g^{-9/2})\\ &= A^2n^{-2}g^{-10}+B^2g^4+2ABn^{-1}g^{-3}+o(n^{-2}g^{-9})+O(n^{-1})\\ &+O(n^{-3/2}g^{-9/2}),\end{aligned}$$

where $A := c\mu\int(L''(u))^2du$ and $B := \mu_2(L)\int(f''''(y))^2dy$.

Then, its asymptotic version is

$$\mathrm{AMSE}\left(R(\hat{f}_h'')\right) = (An^{-1}g^{-5}+Bg^2)^2,$$

and the value of the bandwidth g minimising the quantity above is

$$g_0 = \arg\min_g \mathrm{AMSE} = d_o n^{-1/7},$$

with $d_0 = \left(\frac{5}{2}AB\right)^{1/7}$.

□

From the expression of the optimal pilot bandwidth we can get an estimator, $\hat{g}_0$, just by plugging-in estimates of the unknown quantities. A simpler proposal could be to estimate the pilot by rescaling the rule-of-thumb for bandwidth selection with the corresponding order of the pilot, this is multiplying that value by the factor $\frac{n^{-1/7}}{n^{-1/5}} = n^{2/35}$.

Hence, we define two possibilities for the bootstrap bandwidth estimate:

$$\hat{h}_{\mathrm{Bopt}} = \left(\frac{R(L)\hat{\mu}\hat{c}}{n\mu_2^2(L)R(\hat{f}_{\hat{g}_0}'')}\right)^{1/5} \text{ and}$$

$$\hat{h}_{\mathrm{BRT}} = \left(\frac{R(L)\hat{\mu}\hat{c}}{n\mu_2^2(L)R(\hat{f}_{\hat{g}_1}'')}\right)^{1/5}, \text{ with } \hat{g}_1 = n^{2/35}\hat{h}_{\mathrm{RT}}.$$

Remark 2.3. The asymptotic expression of the MSE in (2.3) given by

$$\mathrm{AMSE}\left(\hat{f}_h(y)\right) = \frac{1}{4}h^4\left(f''(y)\right)^2\mu_2^2(K)+\frac{\gamma(y)}{nh}R(K),$$

and can be used to obtain the expression of an optimal local bandwidth, following similar steps as for the global one, but from the expression:

$$h_{\mathrm{AMSE}}(y) = \left(\frac{\gamma(y) R(K)}{n(f''(y))^2 \mu_2^2(K)} \right)^{1/5}.$$

Then, a similar method as the one described by González-Manteiga et al. (2004) for local linear regression, could be proposed in this context of density estimation with length-biased data.

Using a common kernel density estimator

Bose and Dutta (2013) proposed a new bootstrap bandwidth selector for complete data arguing that they do not need to assume a shape for the unknown density at any stage, and moreover they only require f to be four times differentiable instead of the six times needed in the method presented above.

Following their methodology we propose to obtain a smooth bootstrap bandwidth minimising the $\mathrm{MISE}^*(h)$ in a compact interval I, and assuming that the pilot bandwidth g can be set as $\frac{1}{8} n^{-1/(2p+2s+1)}$, where p and s are the orders of the kernels K and L, respectively. This fixed value for the pilot has been set in Bose and Dutta (2013) after extensive simulation studies using different mixtures of normals on f.

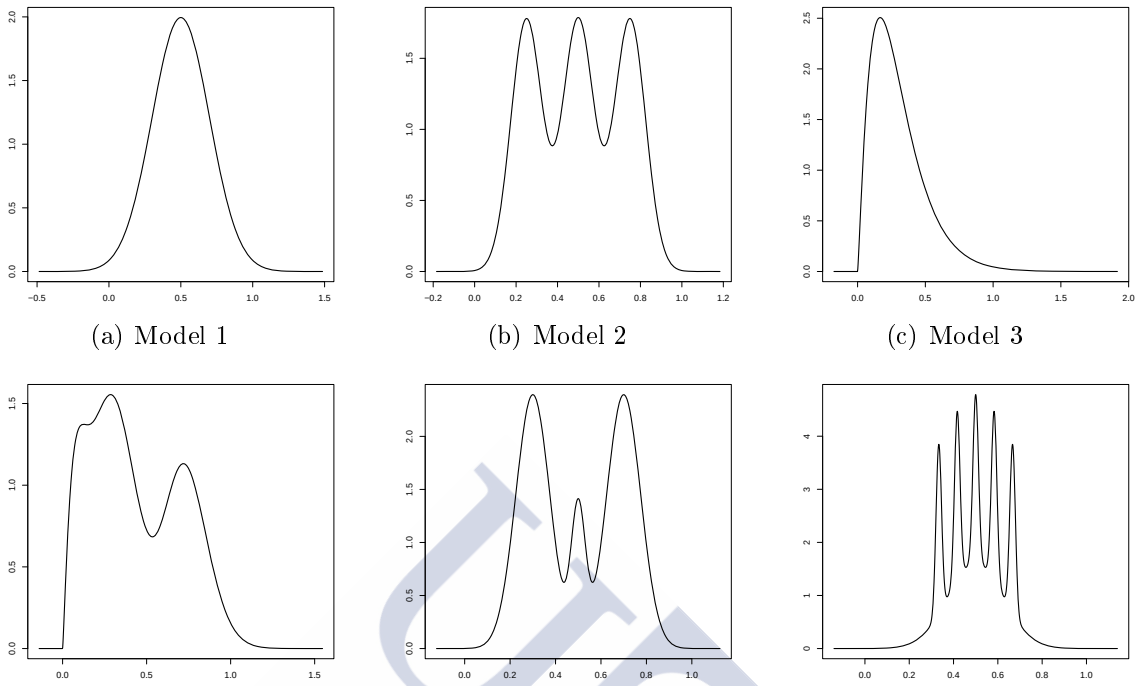
Hence, we define this bootstrap bandwidth as follows:

$$\hat{h}_{\mathrm{B}} = \arg\min_{h \in I} \mathrm{MISE}^*(h).$$

2.4 Finite sample study

In this section we evaluate the performance of the bandwidth selection procedures presented in Section 2.3. To this goal we have carried out a simulation study including rule-of-thumb ($\hat{h}_{\mathrm{RT}}$), cross-validation bandwidth ($\hat{h}_{\mathrm{CV}}$), the bootstrap bandwidths ($\hat{h}_{\mathrm{Bopt}}$) and ($\hat{h}_{\mathrm{B_{RT}}}$) with the two possible pilots and ($\hat{h}_{\mathrm{B}}$). We have considered as benchmarks the infeasible optimal bandwidth values h_{MISE} and h_{ISE} which correspond, respectively, to the optimal bandwidths obtained from MISE and ISE criteria defined in (2.5) and (2.8).

We have simulated six models with densities shown in Figure 2.1, taken from Marron and Wand (1992) and Mammen et al. (2011) but rescaled to the interval $[0, 1]$. We have chosen these models to cover a wide range of densities with different complexity levels, including different number of modes and asymmetry.



(a) Model 1 (b) Model 2 (c) Model 3

(d) Model 4 (e) Model 5 (f) Model 6

Figure 2.1: The six simulated densities in the finite sample study.

These six models are:

- Model 1: a Normal distribution $N(0.5, 0.2^2)$.
- Model 2: a trimodal mixture of three Normal distributions, $N(0.25, 0.075^2)$, $N(0.5, 0.075^2)$ and $N(0.75, 0.075^2)$ with coefficients $\frac{1}{3}$.
- Model 3: a gamma distribution, $Gamma(a, b)$, with $a = b^2$ and $b = 1.5$ applied on $5x$ with $x \in \mathbb{R}^+$.
- Model 4: a mixture of three gamma distributions, $Gamma(a_i, b_i)$, $i = 1, \ldots, 3$ with $a_i = b_i^2$, $b_1 = 1.5$, $b_2 = 3$ and $b_3 = 6$ applied on $8x$ and $x \in \mathbb{R}^+$, with coefficients $\frac{1}{3}$.
- Model 5: a mixture of three Normal distributions, $N(0.3, \frac{3}{40}^2)$, $N(0.7, \frac{3}{40}^2)$ and $N(0.5, \frac{1}{32}^2)$ with coefficients $\frac{9}{20}$, $\frac{9}{20}$ and $\frac{1}{10}$ respectively.
- Model 6: a mixture of six Normal distributions, $N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, 6$ with $\mu_1 = 0.5$, $\mu_2 = \frac{1}{3}$, $\mu_3 = \frac{5}{12}$, $\mu_4 = \frac{1}{2}$, $\mu_5 = \frac{7}{12}$, $\mu_6 = \frac{2}{3}$, $\sigma_1 = \frac{1}{8}$ and $\sigma_i = \frac{1}{80}$, $i = 2, \ldots, 6$ with coefficients $c_1 = \frac{1}{2}$ and $c_i = \frac{1}{10}$, $i = 2, \ldots, 6$.

We have simulated 1000 length-biased Monte Carlo samples from each model considering sample sizes $n = 50$, 100, 200 and 500, using the Epanechnikov kernel function. From these samples we have evaluated the performance of the selectors through the following error criteria:

$$e_1 = \text{mean}(\text{ISE}(\hat{h})), \quad e_2 = \text{std}(\text{ISE}(\hat{h})),$$

$$e_3 = \text{mean}(\hat{h} - h_{\text{ISE}}), \quad e_4 = \text{std}(\hat{h} - h_{\text{ISE}}).$$

The first two measures, e_1 and e_2 are referred to the error of the estimation, so they provide us with information about the overall performance and variability of the different methods. Meanwhile, e_3 and e_4 measure respectively, the bias and variability of the difference between the theoretical benchmark and the value selected by the proposals. This provides information about the way the methods are choosing the bandwidth parameter.

				Model 1			
	h_{ISE}	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{CV}}$	$\hat{h}_{\text{Bopt}}$	$\hat{h}_{\text{B}_{\text{RT}}}$	$\hat{h}_{\text{B}}$	h_{MISE}
				$n = 50$			
e_1	4.61	7.95	7.52	8.72	8.93	9.05	5.20
e_2	5.31	14.39	7.38	8.82	11.23	8.33	8.41
e_3	—	-8.19	3.51	-9.65	-9.86	-9.27	1.15
e_4	—	6.25	10.75	5.83	6.03	5.78	5.32
				$n = 100$			
e_1	3.11	4.92	4.52	5.65	5.79	5.64	3.55
e_2	4.21	10.83	4.83	6.14	8.12	5.06	7.85
e_3	—	-6.62	2.25	-8.63	-8.82	-8.02	0.45
e_4	—	5.48	8.22	5.06	5.23	4.67	4.70
				$n = 200$			
e_1	1.94	2.98	2.73	3.61	3.70	3.29	2.23
e_2	2.98	7.45	3.29	4.08	5.49	3.25	6.15
e_3	—	-5.77	1.28	-8.19	-8.33	-7.16	0.37
e_4	—	4.53	6.87	4.11	4.34	3.79	4.00
				$n = 500$			
e_1	1.00	1.51	1.35	1.95	1.99	1.16	1.53
e_2	1.95	3.99	2.09	2.45	3.02	3.79	2.07
e_3	—	-5.01	0.80	-7.53	-7.60	0.49	-5.56
e_4	—	3.78	5.14	3.32	3.55	3.52	3.14

Table 2.1: Mean and standard deviations of the ISE and of the difference between the benchmark and the bandwidths selectors (criteria e_1 to e_4) for Model 1 multiplied by 10^2.

	Model 2						
	h_{ISE}	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{CV}}$	$\hat{h}_{\text{Bopt}}$	$\hat{h}_{\text{B}_{\text{RT}}}$	$\hat{h}_{\text{B}}$	h_{MISE}
				$n = 50$			
e_1	11.91	15.15	17.99	14.27	14.27	14.28	13.42
e_2	5.24	5.70	6.82	6.33	6.39	7.57	7.21
e_3	—	1.59	14.13	-0.24	-0.31	-3.95	-3.18
e_4	—	10.14	19.332	10.01	10.07	9.02	9.39
				$n = 100$			
e_1	7.89	10.60	12.25	9.01	9.00	8.80	8.23
e_2	3.89	3.17	5.90	3.77	3.77	4.32	4.27
e_3	—	4.83	8.94	2.26	2.22	0.23	-0.54
e_4	—	4.50	14.16	4.41	4.43	4.55	4.15
				$n = 200$			
e_1	4.90	7.30	6.85	5.50	5.51	5.78	5.06
e_2	2.49	2.14	4.55	2.46	2.45	2.48	2.57
e_3	—	5.11	2.99	2.14	2.18	2.23	0.33
e_4	—	1.35	8.68	1.34	1.32	1.98	1.20
				$n = 500$			
e_1	2.47	4.21	3.06	2.74	2.74	4.28	2.55
e_2	1.45	1.45	2.45	1.52	1.55	1.44	1.68
e_3	—	4.52	0.68	1.52	1.50	4.48	0.03
e_4	—	1.25	4.29	0.98	1.07	1.31	1.27

Table 2.2: Mean and standard deviations of the ISE and of the difference between the benchmark and the bandwidths selectors (criteria e_1 to e_4) for Model 2 multiplied by 10^2.

				Model 3			
	h_{ISE}	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{CV}}$	$\hat{h}_{\text{Bopt}}$	$\hat{h}_{\text{B}_{\text{RT}}}$	$\hat{h}_{\text{B}}$	h_{MISE}
				$n = 50$			
e_1	7.66	9.14	25.01	9.51	9.47	15.24	8.75
e_2	5.99	7.65	10.41	8.41	8.34	11.40	6.95
e_3	—	-0.12	24.39	-1.62	-1.55	-6.87	0.53
e_4	—	4.84	11.82	4.83	4.82	9.24	4.27
				$n = 100$			
e_1	5.18	6.00	22.71	6.32	6.29	8.13	5.86
e_2	3.67	4.33	7.37	4.79	4.80	6.25	4.16
e_3	—	0.43	25.88	-1.80	-1.78	-5.08	0.50
e_4	—	3.83	8.52	3.78	3.76	4.39	3.49
				$n = 200$			
e_1	3.52	4.00	21.60	4.27	4.25	4.39	3.94
e_2	2.54	2.81	5.56	3.20	3.19	3.41	2.77
e_3	—	0.75	27.58	-1.90	-1.87	-2.16	0.70
e_4	—	3.09	6.41	3.07	3.05	3.30	2.89
				$n = 500$			
e_1	1.99	2.21	21.24	2.35	2.35	2.39	2.17
e_2	1.35	1.42	3.45	1.59	1.61	1.44	1.47
e_3	—	1.13	30.01	-1.76	-1.75	2.10	0.54
e_4	—	2.15	3.80	2.10	2.11	2.69	2.06

Table 2.3: Mean and standard deviations of the ISE and of the difference between the benchmark and the bandwidths selectors (criteria e_1 to e_4) for Model 3 multiplied by 10^2.

				Model 4			
	h_{ISE}	$\hat{h}_{\mathrm{RT}}$	$\hat{h}_{\mathrm{CV}}$	$\hat{h}_{\mathrm{Bopt}}$	$\hat{h}_{\mathrm{B_{RT}}}$	$\hat{h}_{\mathrm{B}}$	h_{MISE}
				$n = 50$			
e_1	7.75	9.27	17.23	9.61	9.57	17.69	8.92
e_2	4.93	7.20	4.65	7.58	7.59	9.13	6.29
e_3	—	-4.12	37.86	-5.33	-5.44	-15.51	-1.42
e_4	—	13.53	26.01	13.71	13.66	12.12	12.47
				$n = 100$			
e_1	5.47	6.12	15.71	6.39	6.35	10.32	6.03
e_2	3.56	4.07	2.96	4.44	4.44	5.81	3.87
e_3	—	-0.06	43.64	-2.32	-2.41	-10.33	0.49
e_4	—	7.63	18.89	7.72	7.65	7.38	7.22
				$n = 200$			
e_1	3.68	4.13	15.22	4.16	4.16	5.63	4.09
e_2	2.29	2.75	2.48	2.67	2.76	3.46	2.93
e_3	—	1.65	48.79	-1.20	-1.33	-6.50	0.57
e_4	—	4.61	13.04	4.53	4.50	5.08	4.54
				$n = 500$			
e_1	2.10	2.40	14.83	2.30	2.30	2.53	2.30
e_2	1.24	1.27	2.62	1.42	1.43	1.70	1.46
e_3	—	2.56	52.42	-0.62	-0.70	-2.32	0.27
e_4	—	3.06	11.46	2.99	3.00	3.74	3.08

Table 2.4: Mean and standard deviations of the ISE and of the difference between the benchmark and the bandwidths selectors (criteria e_1 to e_4) for Model 4 multiplied by 10^2.

				Model 5			
	h_{ISE}	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{CV}}$	$\hat{h}_{\text{Bopt}}$	$\hat{h}_{\text{B}_{\text{RT}}}$	$\hat{h}_{\text{B}}$	h_{MISE}
				$n = 50$			
e_1	5.05	20.39	25.64	18.38	18.28	18.84	15.89
e_2	8.12	6.82	15.53	7.41	7.44	8.17	8.64
e_3	—	5.86	10.47	4.03	3.92	3.60	-0.36
e_4	—	3.58	18.54	3.53	3.56	4.31	3.21
				$n = 100$			
e_1	9.26	14.44	13.91	11.59	11.51	13.76	9.70
e_2	4.63	4.16	10.80	4.47	4.48	4.79	4.89
e_3	—	6.05	3.39	3.55	3.47	5.12	-0.16
e_4	—	1.67	11.47	1.63	1.64	2.86	1.50
				$n = 200$			
e_1	5.73	10.21	11.27	7.19	7.26	11.52	5.92
e_2	2.69	2.71	13.25	2.73	2.72	3.26	2.74
e_3	—	5.65	5.36	2.83	2.91	6.44	-0.08
e_4	—	1.16	14.83	1.13	1.15	2.13	1.02
				$n = 500$			
e_1	2.9	6.25	3.44	3.67	3.68	11.21	2.99
e_2	1.2	1.44	2.48	1.28	1.29	1.93	1.22
e_3	—	5.03	0.12	2.15	2.17	8.58	0.07
e_4	—	0.87	2.66	0.83	0.84	1.36	0.77

Table 2.5: Mean and standard deviations of the ISE and of the difference between the benchmark and the bandwidths selectors (criteria e_1 to e_4) for Model 5 multiplied by 10^2.

	h_{ISE}	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{CV}}$	$\hat{h}_{\text{Bopt}}$	$\hat{h}_{\text{B}_{\text{RT}}}$	$\hat{h}_{\text{B}}$	h_{MISE}
				Model 6			
				$n = 50$			
e_1	46.43	67.75	63.66	70.76	71.31	68.98	49.66
e_2	13.19	8.09	11.56	8.14	8.25	4.84	18.32
e_3	—	3.84	8.73	2.86	2.56	17.72	-2.58
e_4	—	5.37	8.08	5.34	5.35	6.35	5.09
				$n = 100$			
e_1	30.11	65.97	55.52	64.74	64.50	66.83	30.99
e_2	10.14	4.25	8.73	5.08	5.22	4.35	10.99
e_3	—	5.56	8.69	4.22	4.15	19.51	-0.38
e_4	—	1.51	4.54	1.47	1.47	4.62	1.32
				$n = 200$			
e_1	18.27	62.74	47.91	52.12	51.45	65.21	18.49
e_2	5.58	2.47	14.92	4.51	4.03	4.17	5.58
e_3	—	5.07	7.57	3.51	3.44	19.12	0.14
e_4	—	0.52	4.17	0.46	0.42	4.39	0.30
				$n = 500$			
e_1	9.2	53.41	41.59	32.98	32.71	64.64	9.21
e_2	2.7	3.37	20.43	3.58	3.24	3.17	2.71
e_3	—	4.38	7.12	2.70	2.68	19.81	-0.03
e_4	—	0.36	4.60	0.24	0.22	3.07	0.14

Table 2.6: Mean and standard deviations of the ISE and of the difference between the benchmark and the bandwidths selectors (criteria e_1 to e_4) for Model 6 multiplied by 10^2.

An overview of these numbers indicates that the performance of the methods depends on the complexity of the underlying model. Let classify the models in "easy" (Model 1, Model 3 and Model 4), "intermediate" (Model 2 and Model 5) and "hard" (Model 6) estimation problems.

Regarding to the measure e_1, the rule-of-thumb performs better in smoother densities, such as Model 1, Model 3 and Model 4, however the bootstrap bandwidths are also really competitive for these models, while cross-validation has in general a poorer performance. We have to remark that in Model 4, the bootstrap bandwidth $\hat{h}_{\text{B}}$ needs a large sample size in order to be competitive. Increasing the complexity of the densities, Model 2 and Model 5, the performance of the rule-of-thumb decreases considerably and the bootstrap procedures $\hat{h}_{\text{B}_{\text{RT}}}$ and $\hat{h}_{\text{B}_{\text{NS}}}$ seems to be more accurate; however $\hat{h}_{\text{B}}$ has a

worse performance and the gain with the increasing of the sample size is slower. Note also that depending on the design and the sample size, cross-validation can also produce competitive results. In hard estimation problems as Model 6, bootstrap bandwidths $\hat{h}_{\mathrm{B_{RT}}}$ and $\hat{h}_{\mathrm{B_{NS}}}$ are still valuable competitors.

In terms of variability, which is measured by criteria e_2 and e_4, the cross-validation method exhibits the highest values. The variability of the rule-of-thumb and the bootstrap bandwidths is in general moderate, with the only exception of $\hat{h}_{\mathrm{B}}$ in Model 6 where it exhibits higher values than the other bootstrap rules and the rule-of-thumb.

The bias in bandwidth selection is measured through e_3. Rule-of-thumb and bootstrap bandwidths with pilots generally show bias in the same direction and amount, except for Model 3 where they do not follow this pattern, even though the overall result is good. In smoother models both, tend to oversmooth and the opposite happens with cross-validation. The bias of the other bootstrap bandwidth selector, $\hat{h}_{\mathrm{B}}$, tends to be higher except for very large sample sizes of Model 1.

Moreover, the performance of all the methods proposed may also be compared graphically through the box plots of the errors (ISE's) computed in the 1000 Monte Carlo samples (see Figures 2.2 to 2.7). These plots confirm that the behaviour of the selectors depends on the complexity of the underlying model. The bootstrap proposals seems to have in general a good behaviour outperformed only by cross-validation in Model 6, due to its complexity.

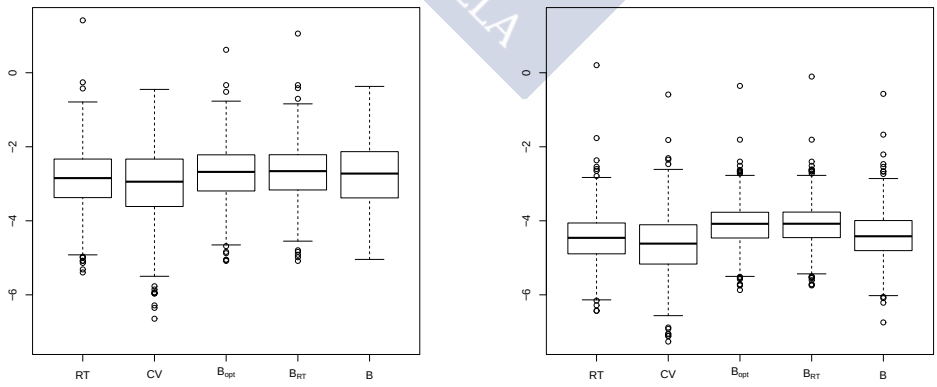


Figure 2.2: The plot of log(ISE) for the density estimation using the different bandwidth selectors for Model 1 and sizes $n = 50, 500$ (from left to right).

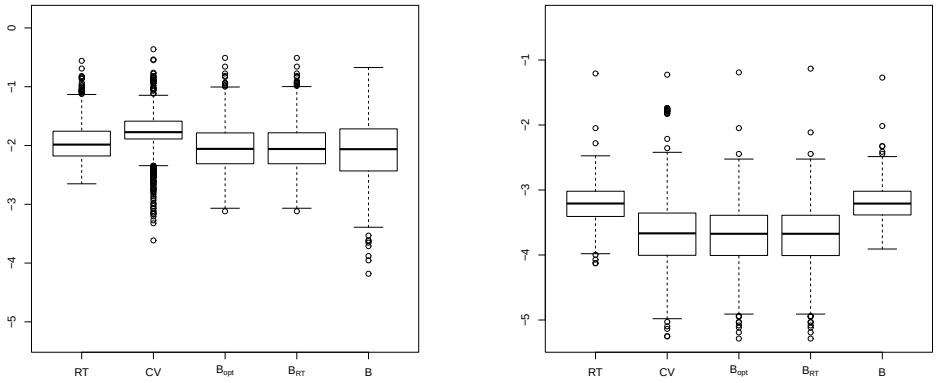


Figure 2.3: The plot of log(ISE) for the density estimation using the different bandwidth selectors for Model 2 and sizes $n = 50, 500$ (from left to right).

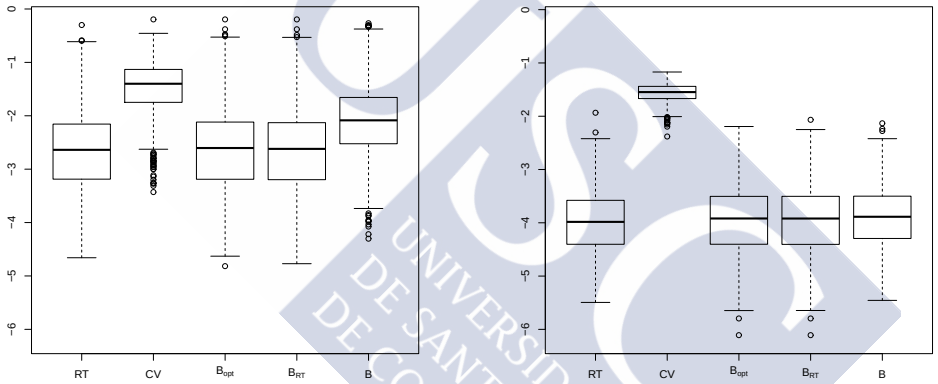


Figure 2.4: The plot of log(ISE) for the density estimation using the different bandwidth selectors for Model 3 and sizes $n = 50, 500$ (from left to right).

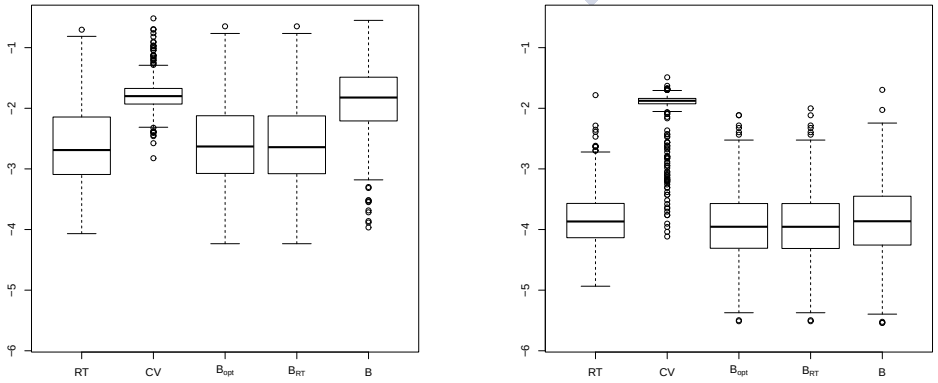


Figure 2.5: The plot of log(ISE) for the density estimation using the different bandwidth selectors for Model 4 and sizes $n = 50, 500$ (from left to right).

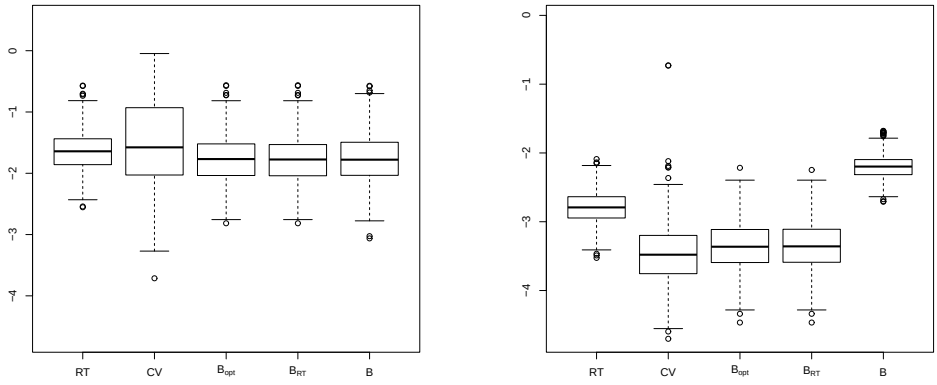


Figure 2.6: The plot of log(ISE) for the density estimation using the different bandwidth selectors for Model 5 and sizes $n = 50, 500$ (from left to right).

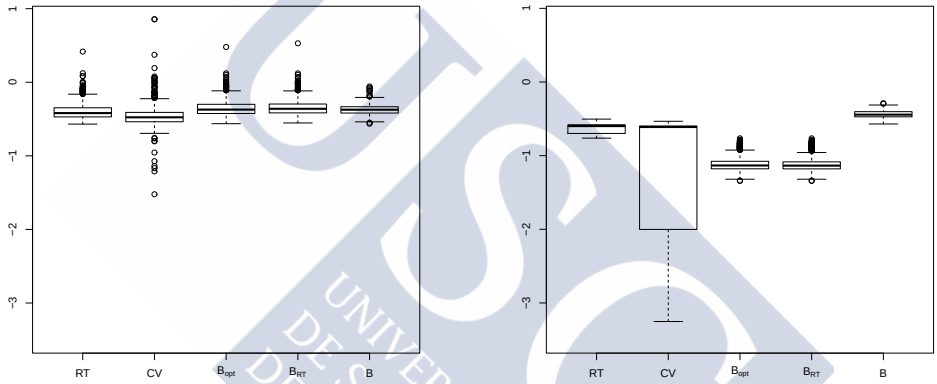


Figure 2.7: The plot of log(ISE) for the density estimation using the different bandwidth selectors for Model 6 and sizes $n = 50, 500$ (from left to right).

2.5 Further extensions

As we have remarked in Section 2.2, the methods previously presented in this chapter can be easily generalised for a general known weight function ω, where the particular case of length-biased data is that of $\omega(y) = y$. First, an appropriate modification of the estimator in (2.1) must be defined, as it has already been presented in Jones (1991):

$$\hat{f}_{h,\omega}(y) = \frac{1}{n}\hat{\mu}_\omega \sum_{i=1}^{n} \omega(Y_i)^{-1} K_h(y - Y_i),$$

with $\hat{\mu}_\omega = \left(\frac{1}{n}\sum_{i=1}^{n} \omega(Y_i)^{-1}\right)^{-1}$.

Theorem 2.1 and Corollary 2.2 can be generalised assuming the following conditions:

(B.1) $E\left[\frac{1}{\omega(X)}\right] < \infty,\ E\left[\frac{1}{\omega(Y)^l}\right] < \infty \quad l = 1, \ldots, 2\nu,$

(B.2) $\int K(u)du = 1,\ \int uK(u)du = 0$ and $\mu_2(K) < +\infty$

(B.3) $\lim_{n\to\infty} nh = +\infty,$

(B.4) y a continuity point of f,

(B.5) f and ω are two times differentiable in y.

We immediately get the error measures as and their optimal bandwidth parameters for the length-biased data:

$$\mathrm{MSE}(\hat{f}_{h,\omega}(y)) = \frac{1}{4}h^4\left(f''(y)\right)^2\mu_2^2(K) + \frac{\gamma_\omega(y)}{nh}R(K) + o\left(h^4 + \frac{1}{nh}\right),$$

with $\gamma_\omega(y) = \dfrac{f(y)\mu_\omega}{\omega(y)}$ and

$$h_{\mathrm{AMSE},\omega}(y) = \left(\frac{\gamma_\omega(y)R(K)}{n(f''(y))^2\mu_2^2(K)}\right)^{1/5}.$$

We also obtain:

$$\mathrm{MISE}(\hat{f}_{h,\omega}) = \frac{1}{4}h^4\mu_2^2(K)R(f'') + \frac{R(K)\mu_\omega c_\omega}{nh} + o\left(h^4 + \frac{1}{nh}\right)$$

where $c_\omega = \displaystyle\int \frac{1}{\omega(y)}f(y)dy$ and then

$$h_{\mathrm{AMISE},\omega} = \left(\frac{R(K)\mu_\omega c_\omega}{n\mu_2^2(K)R(f'')}\right)^{1/5}.$$

The bootstrap methods can be also modified in the same way. Then, the smooth bootstrap samples, $Y_1^*, \ldots, Y_n^*$, can be generated by sampling randomly with replacement n times from the estimated density $\hat{f}_{Y,g,\omega}(y) = \omega(y)\hat{f}_{g,\omega}(y)/\hat{\mu}_\omega$. Here g is again a pilot bandwidth.

For the extension of the bandwidth selectors we need to take into account not only the above modification of the kernel density estimator but also

$$\hat{\sigma}_\omega^2 = \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\omega(Y_i)}\right)^{-1}\left[\left(\frac{1}{n}\sum_{i=1}^n \omega(Y_i)\right) - \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\omega(Y_i)}\right)^{-1}\right].$$

Apart from these considerations the procedures can be obtained in the same way as the length-biased case.

Moreover, we want to remark that the problem of intensity estimation in point processes with covariates, as we have developed it in Chapter 3, may be seen as an extension of the methodology studied in this second chapter, because both estimators (Jones' one and our proposal to estimate the intensity function) have similar structure. This leads to the fact that the theoretical results obtained for both contexts are similar, excluding the fact that in point processes the sample size is a random variable.

2.6 Conclusions

We have considered density estimation in the context of length-biased data, specifically we have focused on the kernel estimator introduced by Jones (1991). We have developed in great detail asymptotic expansions of the MSE, MISE and AMISE of this estimator. Furthermore, we have proposed new bandwidth selection methods and we have studied their behaviour in finite samples through an extensive simulation study. As a general comment, some methods outperforms the others depending on the complexity of the underlying model. Nevertheless, our bandwidth selection proposals have shown to perform quite well and in general, better than the current available cross-validation method. The only exception is the case of very complex densities, with several features and peaks, where cross-validation exhibits the best results, but even in this case our proposals are still competitive.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Kernel intensity estimation in spatial point processes with covariates

Contents

In the spatial point process context, kernel intensity estimation has been mainly restricted to exploratory analysis due to its lack of consistency. However the use of covariates has allowed to design consistent kernel estimators under some restrictive assumptions. In this chapter we focus on defining a theoretical framework to derive a consistent kernel intensity estimator using covariates, as well as a consistent smooth bootstrap procedure. We define two new data-driven bandwidth selectors specifically designed for our estimator: a rule-of-thumb and a plug-in bandwidth based on our consistent bootstrap method. A simulation study is accomplished to understand the behaviour of our proposals in finite samples. Finally, we describe an application to a real dataset consisting of wildfires in Canada during June 2015, using meteorological information as covariates.

Remark that in this chapter, we will go back again to the notation introduced in the Preface and the Introduction of this manuscript related to the point process field, which differs from the one detailed in the previous chapter, even though some of the elements are reused.

3.1 Introduction

Point processes are a branch of spatial statistics whose main aim is to study the geometrical structure of patterns formed by objects (events) that are distributed randomly in number and space. This type of data arise in many different fields such as ecology, Illian et al. (2009) and Law et al. (2009); epidemiology, P. J. Diggle (1990) and Gatrell et al. (1996); astronomy, Babu and Feigelson (1996) and Kerscher (2000); forestry, Stoyan and Penttinen (2000); seismology, Ogata (1988), Ogata (1998), Ogata and Zhuang (2006) and Schoenberg (2011). General theory on point processes as well as some classical applications can be found in Daley and Vere-Jones (1988), Moller and Waagepetersen (2003), Illian et al. (2008), P. J. Diggle (2013) and Baddeley et al. (2015).

Modelling the first-order intensity function is one of the main aims in point process theory. This function computes the mean number of events per (length, area or volume) unit, and it is one of the functions that can completely characterizes a point process. Assuming a parametric model for the intensity function may be a way of estimating it, using for instance a likelihood score such as the Akaike information criteria (AIC), see Van Lieshout (2000), Moller and Waagepetersen (2003) and P. J. Diggle (2013) or pseudolikelihood procedures, see for example Waagepetersen (2007). In the Bayesian context Illian et al. (2012) proposed some models based on log-gaussian Cox processes. However, it is well-known that these techniques can provide inappropriate estimations when the assumed model does not fit the real intensity function. Hence, an alternative through nonparametric methods such as quadrat counts and kernel estimation may apply.

P. Diggle (1985) proposed the first kernel intensity estimator, based on the structure of the common kernel density estimator defined by Parzen (1962) and Rosenblatt (1956), with the inclusion of an edge correction term. As we have already pointed out in the introduction of this manuscript, the main drawback of Diggle's proposal is its lack of consistency, which has almost limited its use to exploratory analysis. To overcome this problem, Cucala (2006) developed asymptotic theory for Diggle's estimator, introducing the concept of "density of events locations" which is based on the idea that the intensity and the density functions differ only in a constant, see Section 1.2.2.

The use of nonparametric methods implies to choose a bandwidth value, which determines the degree of smoothness to be considered in the estimation. The choice of the bandwidth parameter is crucial and it has motivated several papers in the literature in the recent decades, see for example Marron (1988), Scott (1992) and Silverman (1986) for an earlier full description of this problem. There is a lot of theory developed on this issue in areas of statistics such as density estimation and regression, meanwhile in the context of point processes it has received less attention. P. Diggle (1985) proposed a bandwidth selector based on the minimization of the mean squared error (MSE) of his estimator. Later, P. Diggle and Marron (1988) showed the equivalence, for Cox processes in the real line, between that procedure and the standard least-squares cross-validation method used in kernel density estimation. This is an example of the strong connection between this two problems, density and intensity estimation. Brooks and Marron (1991) proved the optimality of the least-squares cross-validation bandwidth for one-dimensional nonhomogeneous Poisson point processes. Fuentes-Santos et al. (2015) develop an extension of Cucala's theory to the two-dimensional case, and propose a bandwidth selection procedure based on a bootstrap method.

Marks and covariates are two different ways of including some extra information in a point process model, as we have explained in the Introduction. Recall that the main difference among them is that marks are directly linked to the events, while covariates include information about the whole observation region and this second scenario is what we consider mainly in this manuscript. Guan (2008) develops a kernel intensity estimator, assuming that the intensity depends on some observed covariates through a continuous unknown function. This estimator turns out to be consistent under some hypotheses over the kernel, the boundary of the region and the pair correlation function. The author also deals with the problem of high dimensional covariates proposing a method to reduce the number of them before applying the kernel techniques. Baddeley et al. (2012) also use information coming from covariates but in a slightly different way. They also assume that the intensity depends on a continuous covariate and propose some intensity estimators based on local likelihood and kernel methodology. In the latest, the bandwidth parameter is chosen using the common rule-of-thumb for density estimation, Silverman (1986), directly applied to the point pattern.

This chapter is organised as follows. In Section 3.2, we make a brief overview on the existing methods in kernel intensity estimation. Section 3.3 is devoted to set up the new framework for kernel intensity estimation with covariates and to develop asymptotic theory for it. In Section 3.4 we propose a new smooth bootstrap method and we prove its consistency. Section 3.5 includes the description of two new data-driven bandwidth selection methods: a rule-of-thumb based on assuming normality and a bootstrap bandwidth selector. In Section 3.6 we carry out an extensive simulation study over all these new procedures and we compare them with the existing competitors. In Section 3.7 we apply all these methods to a real dataset of wildfires in Canada and we draw some conclusions in Section 3.8. Finally, we include some possible extensions of this theory in Section 3.9.

3.2 Kernel intensity estimation

Let X be a point process defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area. Let $X_1, \ldots, X_N$ be a realisation of the process where N is the random variable counting the number of events. Remember that the first-order intensity function, from now on referred as intensity, is defined following P. J. Diggle (2013) as:

$$\lambda(x) = \lim_{|dx| \to 0} \frac{E[N(dx)]}{|dx|},$$

where $|dx|$ denotes the area of an infinitesimal region containing the point $x \in \mathbb{R}^2$.

There is extensive an literature on parametric point processes models and intensity estimation under this assumption, see for instance Schoenberg (2005). However it is well known that we can obtain unreliable estimates when the assumed parametrization deviates from the true intensity. This is the main reason that supports the use of nonparametric techniques. P. Diggle (1985) proposed the first kernel intensity estimator for one dimensional point process, which have been easily extended to the plane:

$$\hat{\lambda}_H^D(x) = \frac{\sum_{i=1}^N K_H(x - X_i)}{p_H(x)}, \quad x \in \mathbb{R}^2,$$

where H is a bandwidth matrix, K denotes a kernel function, $K_H(x) = |H|^{-1/2} K\left(H^{-1/2}x\right)$ and $p_H = \int_W |H|^{-1/2} K(H^{-1/2}(x-y))dy$ is an edge correction term.

This estimator has been widely used during decades for exploratory analysis, but the inference performed with it has been limited due to its lack of consistency. As we have already pointed out, to overcome this problem, Cucala (2006) defines the "density of events locations" as $\lambda_0(x) = \lambda(x)/m$, where $m = \int_{\mathbb{R}} \lambda(x)dx$ is the expected number of events. He proposes a one-dimensional kernel estimator for point processes on the

real line:

$$\hat{\lambda}_{0,h}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)1_{\{N\neq 0\}},\ x \in \mathbb{R}$$

with $1_{\{\cdot\}}$ denoting the indicator function and h a one-dimensional bandwidth parameter. He proves its consistency under an infill structure asymptotic framework, which means that while the observation region remains fixed, the number of events lying on it increase. Fuentes-Santos et al. (2015) extended these ideas to the two-dimensional euclidean space using bandwidth matrices, as it has been done in the context of multivariate density estimation.

Now, let $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ be a spatial continuous covariate that is exactly known in every point of the region of interest W and $Z_1, \ldots, Z_N$ the realisation of the transformed process, i.e, $Z_i = Z(X_i)$. As we have already pointed out in Section 1.4, in practice this covariate will commonly be know in an enough amount of points spread over the region, so the values for the rest of the points can be interpolated and it can be assumed that these values are indeed the real ones.

In some cases it can be assumed that spatial point process intensity depends on the covariate, see for instance Baddeley et al. (2012), so it can be written

$$\lambda(u) = \rho(Z(u)),\ u \in W \subset \mathbb{R}^2, \tag{3.1}$$

where ρ is an unknown function. As Z is known, only ρ needs to be estimated in order to obtain an estimation of λ, which is our target.

For this purpose, it is necessary to deal with the transformed univariate point process, $Z(X)$, and establish the theoretical relationship between this one and the original spatial point process X. If X is a Poisson point process in $W \subset \mathbb{R}^2$ with intensity function $\lambda(u)$ satisfying (3.1), then $Z(X)$ is a Poisson point process in $\mathbb{R}$ with intensity $\rho g^\star(z)$ and with the same expected number of events, where $g^\star$ is the unnormalized version of the derivative of the spatial cumulative distribution function (see Section 3.3 for details).

The proposals in Guan (2008) and Baddeley et al. (2012), following assumption (3.1), are similar kernel intensity estimators. Guan (2008) develop a kernel estimator based on the definition of the distance between two points by the distance through their covariates values:

$$\hat{\lambda}_h^G(u) = \frac{\sum_{i=1}^{N} K_h(||\mathbf{Z}(u) - \mathbf{Z}(X_i)||)}{q_h(u)},$$

with $q_h(u) = \int_W K_h(||\mathbf{Z}(u) - \mathbf{Z}(s)||)ds$ the edge correction term and $\mathbf{Z} = (Z_1, \ldots, Z_p)$ a possibly multivariate covariate where every Z_i fulfils the same conditions as Z.

Considering the increasing domain asymptotic framework and adding also some suitable assumptions the consistency of his proposal is proved. A bandwidth selection

criterion using cross-validation techniques is defined, as well as a dimension reduction tool that allows to handle with high-dimensional covariates.

Baddeley et al. (2012) propose two types of nonparametric intensity estimators founded on two nonparametric density estimators, one based on local likelihood and the other based on kernels. We will focus on the last one, in particular on a weighted kernel intensity estimator for the ρ function with a one-dimensional covariate:

$$\hat{\rho}_W(z) = \sum_{i=1}^{N} \frac{1}{g^\star(Z_i)} K_h(z - Z_i), \tag{3.2}$$

To obtain the bandwidth parameter h in practice, Baddeley et al. (2012) use the common Silverman's rule-of-thumb for density estimation applied to the transformed data.

Our aim in this chapter is not only to define a good kernel estimator for the intensity function under the assumption (3.1), but to be able to set a theoretical framework in which we could prove its consistency and develop optimal bandwidth methods; in short, we want to be able to characterise the intensity estimator in a proper theoretical framework, which is described in the following section.

3.3 A consistent theoretical framework for kernel intensity estimation based on covariates

In this section we work under the transformed space assuming (3.1), and the point process obtained from the original one, X, through the covariate, $Z(X)$, defined in the previous section and that will be detailed below. We need to introduce some definitions and notation.

Definition 3.1. The spatial cumulative distribution function of Z is defined as

$$G(z) = \frac{1}{|W|} \int_W 1_{\{Z(u) \leq z\}} du, \tag{3.3}$$

where $|W|$ denotes the area of the region $W \subset \mathbb{R}^2$.

Let assume that G has a first derivative g, for which we need Z to be differentiable with nonzero gradient, and let denote the unnormalised versions by $g^\star(\cdot) = |W|g(\cdot)$ and $G^\star(\cdot) = |W|G(\cdot)$.

Following Daley and Vere-Jones (1988) and Reiss (2012) we can obtain this result that allow us to establish the model for the transformed point processes in the covariate space:

Theorem 3.1. *Let X be a spatial point process in $W \subset \mathbb{R}^2$ with intensity function of the form (3.1) for some real-value function ρ. Then, $Z(X)$ is a univariate point process in $\mathbb{R}$ with intensity function $\rho g^\star$. Moreover, if the original is a Poisson point process, then the transformed one preserves this property.*

The relationship between X and $Z(X)$ can be extended to the expected number of events through this result adapted from Federer (1969):

Theorem 3.2. *Let $W \subset \mathbb{R}^2$ be a bounded subset, $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ a differentiable function with nonzero gradient in every point of W. Then, for any integrable function $q : W \to \mathbb{R}$, in our particular case $q(u) = \lambda(u)||\nabla Z(u)||^{-1} : W \to \mathbb{R}^+$, it holds:*

$$\int_W \lambda(u)du = \int_W \lambda(u)||\nabla Z(u)||^{-1}||\nabla Z(u)||du = \int_{\mathbb{R}} \int_{Z^{-1}(y)} \lambda(u)||\nabla Z(u)||^{-1} dH(u)dy,$$

where $Z^{-1}(y) = \{u \in W / Z(u) = y\}$ and dH is the one-dimensional Haussdorf measure.

Applying this result to the unnormalised spatial cumulative distribution function, $G^\star$, we have:

$$\begin{aligned} G^\star(z) &= \int_W 1_{\{Z(u)\leq z\}}du = \int_W 1_{\{Z(u)\leq z\}}||\nabla Z(u)||^{-1}||\nabla Z(u)||du \\ &= \int_{\mathbb{R}} \int_{Z^{-1}(y)} 1_{\{Z(u)\leq z\}}||\nabla Z(u)||^{-1} dH(u)du = \int_{-\infty}^{z} \int_{Z^{-1}(y)} ||\nabla Z(u)||^{-1} dH(u)dy. \end{aligned}$$

And deriving with respect to z, we get $g^\star(z) := (G^\star)'(z) = \int_{Z^{-1}(y)} ||\nabla Z(u)||^{-1} dH(u)$.

We can now rewrite the relationship between the original spatial point process intensity and the transformed one through an integral, which also implies that the expected number of events, in the corresponding region, is the same in both processes:

$$\begin{aligned} m &= \int_W \lambda(u)du = \int_{\mathbb{R}} \int_{Z^{-1}(y)} \lambda(u)||\nabla Z(u)||^{-1} dH(u)dy \\ &= \int_{\mathbb{R}} \int_{Z^{-1}(y)} \rho(Z(u))||\nabla Z(u)||^{-1} dH(u)dy = \int_{\mathbb{R}} \rho(y) \int_{Z^{-1}(y)} ||\nabla Z(u)||^{-1} dH(u)dy \\ &= \int_{\mathbb{R}} \rho(y) g^\star(y)dy. \end{aligned}$$

Hence, we can finally have a one-dimensional nonhomogeneous Poisson point process, and moreover the results above reach to the fact that $\rho g^\star$ is the intensity function of the transformed point process.

Now, following the idea of Cucala (2006) we use the close relationship between the intensity and the density function. We can define the following "artificial" density function:

$$f(\cdot) = \frac{\rho(\cdot)g^\star(\cdot)}{m}. \tag{3.4}$$

Basically, what we propose in this section is to take profit of this relationship: firstly estimating the density and then going back to our target problem, that is the intensity estimation, multiplying just by a constant.

Following the pre-established notation, let define the estimator of the relative density as follows:

$$\hat{f}_h(z) = g^\star(z)\frac{1}{N}\sum_{i=1}^{N}\frac{1}{g^\star(Z_i)}K_h\left(z - Z_i\right)1_{\{N\neq 0\}}, \tag{3.5}$$

where K is a univariate kernel function and $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. This is an estimator of f, and once we have it, we can go back to the intensity function by plug-in and letting $\hat{\lambda}(u) = \hat{\rho}_h(Z(u))$, where $\hat{\rho}_h$ can be replaced by (3.2).

In the following statement we obtain the value of the pointwise mean and variance of $\hat{f}_h$ with the corresponding error rates, as well as its mean squared error (MSE), which is defined in the same way as it has been previously done in (2.2).

Hereafter, we establish that our point process X is a nonhomogeneous Poisson point process in $W \subset \mathbb{R}^2$. Although the intensity estimator we propose, as well as the bandwidth selectors, can be applied to non-Poisson processes, the previous assumption is required to prove the consistency of the estimator.

We also need to introduce some regularity conditions:

(B.1) $\int_{\mathbb{R}} K(z)dz = 1$; $\int_{\mathbb{R}} zK(z)dz = 0$ and $\mu_2(K) := \int_{\mathbb{R}} z^2K(z)dz < \infty$,

(B.2) $\lim_{m\to\infty} h = 0$ and $\lim_{m\to\infty}\frac{A(m)}{h} = 0$, where $A(m) = E\left[\frac{1}{N}1_{\{N\neq 0\}}\right]$,

(B.3) G is three times differentiable,

(B.4) z is a continuity point of ρ and

(B.5) ρ is three times differentiable.

Theorem 3.3. *Under conditions (B.1) to (B.4) we have that:*

$$E\left[\hat{f}_h(z)\right] = \frac{g^\star(z)(K_h\circ\rho)(z)}{m}\left(1 - e^{-m}\right) \qquad \textit{and}$$

$$\mathrm{Var}\left[\hat{f}_h(z)\right] = A(m)\frac{(g^\star(z))^2}{n}\left(K_h^2\circ\frac{\rho}{g^\star}\right)(z) - (A(m) + e^{-2m} - e^{-m})(g^\star(z))^2(K_h\circ\rho)^2(z),$$

where $\circ$ denotes the convolution between two functions. Moreover, adding condition (B.5) we have:

$$\begin{aligned} MSE(h,z) \;&= e^{-2m}f^2(z) + (1-e^{-m})^2\frac{h^4}{4}\left(\frac{\rho''(z)g^\star(z)}{m}\right)^2\mu_2^2(K) \\ &\quad -e^{-m}(1-e^{-m})h^2\mu_2(K)\frac{(g^\star(z))^2\rho(z)\rho''(z)}{m^2} + \frac{A(m)}{h}f(z)R(K) \\ &\quad +o(h^2(1-e^{-m})e^{-m}) + o(h^4(1-e^{-m})^2) + o\left(\frac{A(m)}{mh}\right), \end{aligned}$$

with $R(K) = \int_{\mathbb{R}} K^2(z)dz$.

– Proof –

Recall that N is a random variable, so we have a double stochastic scenario, on the one hand the randomness provided by N and on the other the randomness of the point process itself. To deal with this we use the conditional mean and we consider also some tools related to real number series.

Consider first the mean,

$$\begin{aligned} E\left[\hat{f}_h(z)\right] &= E\left[E\left[\hat{f}_h(z)|N=n>0\right]\right] = \textstyle\sum_{n=1}^{\infty} E\left[\hat{f}_h(z)|N=n\right]\mathbb{P}\left(N=n\right) \\ &= \textstyle\sum_{n=1}^{\infty} E\left[\hat{f}_h(z)|N=n\right]\frac{e^{-m}m^n}{n!} = \sum_{n=1}^{\infty}\frac{g^\star(z)}{m}\left(K_h\circ\rho\right)(z)\frac{e^{-m}m^n}{n!} \\ &= \frac{g^\star(z)(K_h\circ\rho)(z)}{m}e^{-m}\textstyle\sum_{n=1}^{\infty}\left(\frac{m^n}{n!}-1\right) = \frac{g^\star(z)(K_h\circ\rho)(z)}{m}\left(1-e^{-m}\right), \end{aligned}$$

where we have used that

$$\begin{aligned} E\left[\hat{f}_h(z)|N=n\right] &= E\left[g^\star(z)\tfrac{1}{n}\textstyle\sum_{n=1}^{\infty}\frac{1}{g^\star(Z_i)}K_h(z-Z_i)\right] = g^\star(z)E\left[\frac{1}{g^\star(Z_1)}K_h(z-Z_1)\right] \\ &= g^\star(z)\textstyle\int_{\mathbb{R}}\frac{1}{g^\star(s)}K_h(z-s)\frac{\rho(s)g^\star(s)}{m}ds = \frac{g^\star(z)}{m}\left(K_h\circ\rho\right)(z). \end{aligned}$$

Next, the variance:

$$\begin{aligned} \mathrm{Var}\left[\hat{f}_h(z)\right] &= E\left[\hat{f}^2(z)\right] - E\left[\hat{f}_h(z)\right]^2 = A(m)\tfrac{(g^\star(z))^2}{n}\left(K_h^2\circ\tfrac{\rho}{g^\star}\right)(z) \\ &- \left(A(m)+e^{-m}-1\right)\tfrac{(g^\star(z))^2}{m^2}\left(K_h\circ\rho\right)^2(z) - \tfrac{(g^\star(z))^2(K_h\circ\rho)^2(z)}{m^2}(1-e^{-m})^2 \\ &= A(m)\tfrac{(g^\star(z))^2}{n}\left(K_h^2\circ\tfrac{\rho}{g^\star}\right)(z) - (A(m)+e^{-2m}-e^{-m})(g^\star(z))^2(K_h\circ\rho)^2(z), \end{aligned}$$

where we have used the notation $A(m) = E\left[\frac{1}{N}1_{\{N\neq 0\}}\right]$, the expectation:

$$\begin{aligned} E\left[\hat{f}_h^2(z)\right] &= E\left[E\left[\hat{f}_h{}^2(z)|N=n>0\right]\right] = \textstyle\sum_{n=1}^{\infty} E\left[\hat{f}_h{}^2(z)|N=n\right]\mathbb{P}\left(N=n\right) \\ &= \textstyle\sum_{n=1}^{\infty} E\left[\hat{f}_h{}^2(z)|N=n\right]\frac{e^{-m}m^n}{n!} = \frac{(g^\star(z))^2}{m}\left(K_h^2\circ\frac{\rho}{g^\star}\right)(z)\sum_{n=1}^{\infty}\frac{e^{-m}m^n}{nn!} \\ &+\tfrac{(g^\star(z))^2}{m^2}\left(K_h\circ\rho\right)^2(z)\textstyle\sum_{n=1}^{\infty}\frac{n-1}{n}\frac{e^{-m}m^n}{n!} = \frac{(g^\star(z))^2}{m}\left(K_h^2\circ\frac{\rho}{g^\star}\right)(z)E\left[\frac{1}{N}1_{\{N\neq 0\}}\right] \\ &+\tfrac{(g^\star(z))^2}{m^2}\left(K_h\circ\rho\right)^2(z)E\left[\tfrac{N-1}{N}1_{\{N\neq 0\}}\right] = A(m)\tfrac{(g^\star(z))^2}{m}\left(K_h^2\circ\tfrac{\rho}{g^\star}\right)(z) \\ &-\left(A(m)+e^{-m}-1\right)\tfrac{(g^\star(z))^2}{m^2}\left(K_h\circ\rho\right)^2(z), \end{aligned}$$

and

$$
\begin{aligned}
&E\left[\hat{f}_h^2(z)|N=n\right] = E\left[(g^\star(z))^2\,\tfrac{1}{n^2}\left(\textstyle\sum_{i=1}^n \frac{1}{g^\star(Z_i)}K_h(z-Z_i)\right)\right]\\
&= \tfrac{(g^\star(z))^2}{n^2}E\left[\left(\textstyle\sum_{i=1}^n \frac{1}{g^\star(Z_i)}K_h(z-Z_i)\right)^2\right]\\
&= \tfrac{(g^\star(z))^2}{n^2}E\left[\textstyle\sum_{i=1}^n \frac{1}{(g^\star(Z_i))^2}K_h^2(z-Z_i) + \sum_{i\neq j}\frac{1}{g^\star(Z_i)}\frac{1}{g^\star(Z_j)}K_h(z-Z_i)K_h(z-Z_j)\right]\\
&= \tfrac{(g^\star(z))^2}{n}E\left[\tfrac{1}{(g^\star(Z_i))^2}K_h^2(z-Z_i)\right] + \tfrac{(g^\star(z))^2}{n^2}n(n-1)\left(E\left[\tfrac{1}{g^\star(Z_1)}K_h(z-Z_1)\right]\right)^2\\
&= \tfrac{(g^\star(z))^2}{n}\textstyle\int_{\mathbb{R}}\frac{1}{(g^\star(s))^2}K_h^2(z-s)\frac{\rho(s)g^\star(s)}{m}ds + \frac{(g^\star(z))^2(n-1)}{n}\left(\int_{\mathbb{R}}\frac{1}{g^\star(s)}K_h(z-s)\frac{\rho(s)g^\star(s)}{m}ds\right)^2\\
&= \tfrac{(g^\star(z))^2}{mn}\left(K_h^2\circ\tfrac{\rho}{g^\star}\right)(z) + \tfrac{(g^\star(z))^2(n-1)}{nm^2}\left(K_h\circ\rho\right)^2(z).
\end{aligned}
$$

Lastly to compute the MSE in terms of bias and variance, we apply a Taylor expansion of second-order to the mean and of first-order to the variance, getting:

$$
\begin{aligned}
E\left[\hat{f}_h(z)\right] &= \tfrac{g^\star(z)}{m}\left(K_h\circ\rho\right)(z)(1-e^{-m}) = \tfrac{g^\star(z)}{m}\left(\rho(z)+\tfrac{1}{2}h^2\mu_2(K)\rho^{''}(z)+o(h^2)\right)\\
&\quad (1-e^{-m})\\
Bias\left[\hat{f}_h(z)\right] &= \tfrac{g^\star(z)}{m}\left(-e^{-m}\rho(z)+(1-e^{-m})\tfrac{1}{2}h^2\rho^{''}(z)\mu_2(K)+o(h^2(1-e^{-m}))\right),\\
\mathrm{Var}\left[\hat{f}_h(z)\right] &= A(m)\tfrac{(g^\star(z))^2}{m}\left(K_h^2\circ\tfrac{\rho}{g^\star}\right)(z) - (A(m)+e^{-2m}-e^{-m})(g^\star(z))^2(K_h\circ\rho)^2(z)\\
&= A(m)\tfrac{(g^\star(z))^2}{m}\tfrac{1}{h}\tfrac{\rho}{g^\star}(z)R(K)+o\left(\tfrac{A(m)}{mh}\right),
\end{aligned}
$$

where $\mu_2(K)=\int_{\mathbb{R}}z^2K(z)dz$ and $R(K)=\int_{\mathbb{R}}K^2(u)du$. Hence,

$$
\begin{aligned}
MSE\,(h,z) &= e^{-2m}f^2(z)+(1-e^{-m})^2\tfrac{h^4}{4}\left(\tfrac{\rho^{''}(z)g^\star(z)}{m}\right)^2\mu_2^2(K)\\
&\quad -e^{-m}(1-e^{-m})h^2\mu_2(K)\tfrac{(g^\star(z))^2\rho(z)\rho^{''}(z)}{m^2}+\tfrac{A(m)}{h}f(z)R(K)\\
&\quad +o(h^2(1-e^{-m})e^{-m})+o(h^4(1-e^{-m})^2)+o\left(\tfrac{A(m)}{mh}\right);
\end{aligned}
$$

from which we can easily derive its asymptotic version

$$
\begin{aligned}
AMSE\,(h,z) &= e^{-2m}f^2(z)+(1-e^{-m})^2\tfrac{h^4}{4}\left(\tfrac{\rho^{''}(z)g^\star(z)}{m}\right)^2\mu_2^2(K)\\
&\quad -e^{-m}(1-e^{-m})h^2\mu_2(K)\tfrac{(g^\star(z))^2\rho(z)\rho^{''}(z)}{m^2}+\tfrac{A(m)}{h}f(z)R(K).
\end{aligned}
$$

□

Now, considering the mean integrated square error (MISE) as defined in (2.5) and denoting by $AMISE$ its asymptotic version, the following result is a consequence of Theorem 3.3:

Corollary 3.4. *Under conditions (B.1) to (B.3) and (B.5),*

$$\begin{aligned} MISE(h) &= e^{-2m}R(f) + (1-e^{-m})^2\tfrac{h^4}{4}R\left(\tfrac{\rho'' g^\star}{m}\right)\mu_2^2(K) \\ &\quad -e^{-m}(1-e^{-m})h^2\mu_2(K)\textstyle\int_{\mathbb{R}}\frac{g^\star(z)\rho''(z)f(z)}{m}dz + \frac{A(m)}{h}R(K) \\ &\quad +o(h^2(1-e^{-m})e^{-m}) + o(h^4(1-e^{-m})^2) + o\left(\tfrac{A(m)}{mh}\right) \qquad \textit{and} \\ AMISE(h) &= (1-e^{-m})^2\tfrac{h^4}{4}R\left(\tfrac{\rho'' g^\star}{m}\right)\mu_2^2(K) + \tfrac{A(m)}{h}R(K). \end{aligned}$$

As a consequence, the optimal bandwidth value which minimises AMISE is:

$$h_{AMISE} = \left(\frac{A(m)R(K)}{\mu_2^2(K)(1-e^{-m})^2 R\left(\frac{\rho'' g^\star}{m}\right)}\right)^{1/5} = \left(\frac{R(K)}{\mu_2^2(K)R(\rho'' g^\star)}\frac{A(m)m^2}{(1-e^{-m})^2}\right)^{1/5}. \tag{3.6}$$

– Proof –

This result is easily deduced by integrating and removing the negligible terms in Theorem 3.3, and then deriving with respect to h to obtain the explicit expression of the optimal bandwidth.

□

Remark 3.1. Note that the expressions obtained in Theorem 3.3 and Corollary 3.4 are close to the ones obtained in Chapter 2 in density estimation for length-biased data, and they are also similar to the analogous for the classical kernel density estimator by Parzen (1962) and Rosenblatt (1956). The difference lies on some extra terms involving the expected sample size that appear in this context of intensity estimation, due to the randomness of the variable N representing the number of events.

3.4 Resampling bootstrap method

Nonparametric bootstrap procedures have been widely used in different contexts to perform inference and calibrate the distribution of statistics in goodness-of-fit tests. The smooth bootstrap procedure for point processes with covariates proposed in this section is inspired in the work of Cao (1993) for kernel density estimation, and Cowling et al. (1996), for Poisson point process.

Let $X_1, \ldots, X_n$ be a realisation of the spatial point process X, construct $Z_1, \ldots, Z_n$ the associated realisation of the transformed univariate process, let $\hat{f}_b$ be the density estimator in (3.5) and $\hat{\rho}_b$ the estimator defined in (3.2), where b is a pilot bandwidth.

Now, conditional on $Z_1, \ldots, Z_n$, let $N^* \sim Poiss\,(\hat{m})$ with $\hat{m} := \int_{\mathbb{R}} \hat{\rho}_b(z) g^\star(z) dz$, generate n^* a realisation of this random variable N^* and then draw $Z_1^*, \ldots, Z_{n^*}^*$ by sampling randomly with replacement n^* times from the distribution with density proportional to $g^\star \hat{\rho}_b$, i.e. $\tilde{f}_b = \frac{\hat{\rho}_b g^\star}{\hat{m}}$.

Denote by Z^* the random variable generated by the bootstrap method presented above. From the bootstrap sample we define the density estimator as:

$$\hat{f}_h^{\,*}(z) = g^\star(z) \frac{1}{N^*} \sum_{i=1}^{N^*} \frac{1}{g^\star(Z_i^*)} K_h\left(z - Z_i^*\right) 1_{\{N^* \neq 0\}}, \tag{3.7}$$

hence, using equation (3.4) we get the associated estimator of ρ:

$$\hat{\rho}_h^*(z) = \sum_{i=1}^{N^*} \frac{1}{g^\star(Z_i^*)} K_h\left(z - Z_i^*\right),$$

and then we plug it in (3.1) to obtain an estimator of λ.

The following result provides the expression of the mean, variance and mean squared error of $\hat{f}_h^*$ under the bootstrap distribution; hereafter we use again, as in Chapter 2, the notation E^*, Var^* and MSE^* to refer to the mean, variance and mean squared error respectively, under the bootstrap distribution.

Theorem 3.5. *Under hypotheses (B.1) to (B.4) we get:*

$$E^*\left[\hat{f}_h^{\,*}(z)\right] = \tfrac{g^\star(z)}{\hat{m}} (K_h \circ \hat{\rho}_b)(z)(1 - e^{-\hat{m}}) \qquad \textit{and}$$

$$\mathrm{Var}^*\left[\hat{f}_h^*(z)\right] = \tfrac{(g^\star(z))^2}{\hat{m}} \left(K_h^2 \circ \tfrac{\hat{\rho}_b}{g^\star}\right)(z) A(\hat{m}) - \tfrac{(g^\star(z))^2}{\hat{m}^2} (K_h \circ \hat{\rho}_b)^2(z)(A(\hat{m}) + e^{-2\hat{m}} - e^{-\hat{m}}),$$

where $A(\hat{m}) := E^*\left[\frac{1}{N^*} 1_{\{N^* \neq 0\}}\right]$. *Moreover, adding condition (B.5) we have*

$$\begin{aligned} MSE^*(h,z) = {} & e^{-2\hat{m}} (\tilde{f}_b(z))^2 + \frac{h^4}{4} (\hat{\rho}_h''(z))^2 \frac{(g^\star(z))^2}{\hat{m}^2} \mu_2^2(L)(1 - e^{-\hat{m}})^2 \\ & - e^{-\hat{m}}(1 - e^{-\hat{m}}) h^2 \tilde{f}_b(z) \frac{\hat{\rho}_b''(z) g^\star(z)}{\hat{m}} \mu_2(L) + \frac{A(\hat{m})}{h} R(L) + o_P(h^4 (1 - e^{-\hat{m}})^2) \\ & + o_P(h^2 (1 - e^{-\hat{m}}) e^{-\hat{m}}) + o_P\left(\frac{A(\hat{m})}{\hat{m} h}\right); \end{aligned} \tag{3.8}$$

recall that we have defined $\tilde{f}_b(\cdot) = \frac{\hat{\rho}_b(\cdot) g^\star(\cdot)}{\hat{m}}$.

– Proof –

To compute the MSE^* of (3.7) we follow the same structure we have established for Theorem 3.3, with the difference that we are now under the bootstrap distribution.

The pointwise mean is:

$$\begin{aligned} E^*\left[\hat{f}_h^*(z)\right] &= E^*\left[E^*\left[\hat{f}_h^*(z)|N^*=n^*>0\right]\right] = \sum_{n^*=1}^{\infty} E^*\left[\hat{f}_h{}^*(z)|n^*=n^*\right]\mathbb{P}^*\left(N^*=n^*\right) \\ &= \sum_{n^*=1}^{\infty} E^*\left[\hat{f}_h{}^*(z)|N^*=n^*\right]\frac{e^{-\hat{m}}\hat{m}^{n^*}}{n^*!} = \sum_{n^*=1}^{\infty}\frac{g^\star(z)}{\hat{m}}(K_h\circ\hat{\rho}_b)(z)\frac{e^{-\hat{m}}\hat{m}^{n^*}}{n^*!} \\ &= \frac{g^\star(z)}{\hat{m}}(K_h\circ\hat{\rho}_b)(z)(1-e^{-\hat{m}}), \end{aligned}$$

where we have used that

$$\begin{aligned} E^*\left[\hat{f}_h^*(z)|N^*=n^*>0\right] &= E^*\left[g^\star(z)\frac{1}{n^*}\sum_{i=1}^{n^*}\frac{1}{g^\star(Z_i^*)}K_h(z-Z_i^*)\right] \\ &= g^\star(z)E^*\left[\frac{1}{g^\star(Z_i^*)}K_h(z-Z_i^*)\right] = \frac{g^\star(z)}{\hat{m}}\int K_h(z-s)\hat{\rho}_b(s)ds \\ &= \frac{g^\star(z)(K_h\circ\hat{\rho}_b)(z)}{\hat{m}}. \end{aligned}$$

We now apply a second-order Taylor expansion to the mean in order to obtain the bias:

$$Bias^*\left[\hat{f}_h^*(z)\right] = -e^{-\hat{m}}\frac{\hat{\rho}_b(z)g^\star(z)}{\hat{m}} + \frac{h^2}{2}\frac{\hat{\rho}_b''(z)g^\star(z)}{\hat{m}}\mu_2(L)(1-e^{-\hat{m}}) + o(h^2(1-e^{-\hat{m}})).$$

To derive the variance we work as follows:

$$\begin{aligned} \mathrm{Var}^*\left[\hat{f}_h{}^*(z)\right] &= E^*\left[\hat{f}_h{}^{*2}(z)\right] - E^*\left[\hat{f}_h{}^*(z)\right]^2 \\ &= \frac{(g^\star(z))^2}{\hat{m}}\left(K_h\circ\frac{\hat{\rho}_b}{g^\star}\right)(z)A(\hat{m}) - \frac{(g^\star(z))^2}{\hat{m}^2}(K_h\circ\hat{\rho}_b)^2(z)(A(\hat{m})+e^{-2\hat{m}}-e^{-\hat{m}}). \end{aligned}$$

Applying now a first-order Taylor expansion we get:

$$\mathrm{Var}^*\left[\hat{f}_h{}^*(z)\right] = \tilde{f}_b(z)R(L)\frac{A(\hat{m})}{h} + o\left(\frac{A(\hat{m})}{\hat{m}h}\right),$$

where we have used:

$$\begin{aligned} E^*\left[(\hat{f}_h{}^*(z))^2\right] &= E^*\left[E^*\left[\hat{f}_h{}^{*2}(z)|N^*=n^*\right]\right] \\ &= \sum_{n^*=1}^{\infty} E^*\left[\hat{f}_h{}^{*2}(z)|n^*=n^*\right]\mathbb{P}^*\left(N^*=n^*\right) \\ &= \sum_{n^*=1}^{\infty} E^*\left[\hat{f}_h{}^{*2}(z)|N^*=n^*\right]\frac{e^{-\hat{m}}\hat{m}^{n^*}}{n^*!} \\ &= \frac{(g^\star(z))^2}{\hat{m}}\left(K_h^2\circ\frac{\hat{\rho}_b}{g^\star}\right)(z)\sum_{n^\star=1}^{\infty}\frac{e^{-\hat{m}}\hat{m}^{n^*}}{n^*n^*!} \\ &+ \frac{(g^\star(z))^2}{\hat{m}^2}(K_h\circ\hat{\rho}_b)^2(z)\sum_{n^*=1}^{\infty}\frac{n^*-1}{n^*}\frac{e^{-\hat{m}}\hat{m}^{n^*}}{n^*!} \\ &= \frac{(g^\star(z))^2}{\hat{m}}\left(K_h^2\circ\frac{\hat{\rho}_b}{g^\star}\right)(z)A(\hat{m}) - \frac{(g^\star(z))^2}{\hat{m}^2}(K_h\circ\hat{\rho}_b)^2(z)(A(\hat{m}+e^{-\hat{m}}-1), \end{aligned}$$

with

$$\begin{aligned}
E^*\left[\hat{f}_h^{*\,2}(z)|N^*=n^*\right] &= E^*\left[\left(g^*(z)\tfrac{1}{n^*}\textstyle\sum_{i=1}^{n^*}\tfrac{1}{g^\star(Z_i^*)}K_h(z-Z_i^*)\right)^2\right]\\
&= \tfrac{(g^\star(z))^2}{n^{*\,2}}E^*\left[\textstyle\sum_{i=1}^{n^*}\tfrac{1}{(g^\star(Z_i^*))^2}L_h^2(z-Z_i^*)\right.\\
&\left.+\textstyle\sum_{i\neq j}\tfrac{1}{g^\star(Z_i^*)}\tfrac{1}{g^\star(Z_j^*)}K_h(z-Z_i^*)K_h(z-Z_j^*)\right]\\
&= \tfrac{(g^\star(z))^2}{n^{*\,2}}\left(n^*E^*\left[\tfrac{1}{(g^\star(Z_1^*))^2}K_h^2(z-Z_1^*)\right]\right.\\
&\left.+n^*(n^*-1)E^*\left[\tfrac{1}{g^\star(Z_1^*)}K_h(z-Z_1^*)\right]^2\right)\\
&= \tfrac{(g^\star(z))^2}{n^*\hat{m}}\left(K_h^2\circ\tfrac{\hat{\rho}_b}{g^\star}\right)(z)+\tfrac{(g^\star(z))^2(n^*-1)}{n^*\hat{m}^2}(K_h\circ\hat{\rho}_b)^2(z).
\end{aligned}$$

Finally, gathering squared bias and variance together we obtain the MSE^* in (3.8). □

In the same way as we have done for Corollary 3.4, the integrated and asymptotic version of the MSE^* can be easily deduced from the previous result in the next Corollary.

Corollary 3.6. *Under conditions (B.1) to (B.3) and (B.5),*

$$\begin{aligned}
MISE^*(h) &= e^{-2\hat{m}}R(\tilde{f}_b)+\tfrac{h^4}{4}R\left(\tfrac{\hat{\rho}_b''g^\star}{\hat{m}}\right)\mu_2^2(L)(1-e^{-\hat{m}})^2\\
&\quad -e^{-\hat{m}}(1-e^{-\hat{m}})h^2\mu_2(L)\textstyle\int\tfrac{\tilde{f}_b(z)\hat{\rho}_b''(z)g^\star(z)}{\hat{m}}dz+\tfrac{A(\hat{m})}{h}R(L)\\
&\quad +o(h^4(1-e^{-\hat{m}})^2)+o_P(h^2(1-e^{-\hat{m}})e^{-\hat{m}})+o_P\left(\tfrac{A(\hat{m})}{\hat{m}h}\right) \qquad \textit{and}\\
AMISE^*(h) &= \tfrac{h^4}{4}R\left(\tfrac{\hat{\rho}_b''g^\star}{\hat{m}}\right)\mu_2^2(L)(1-e^{-\hat{m}})^2+\tfrac{A(\hat{m})}{h}R(L).
\end{aligned}$$

Therefore the asymptotic expression of the optimal bootstrap bandwidth is:

$$h_{AMISE^*}=\left(\frac{A(\hat{m})R(L)}{\mu_2^2(L)(1-e^{-\hat{m}})^2R(\frac{\hat{\rho}_b''g^\star}{\hat{m}})}\right)^{1/5}, \tag{3.9}$$

which is a plug-in version of (3.6).

– Proof –

Similarly to Corollary 3.4, this result is obtained by first integrating and removing negligible terms from the expression in Theorem 3.5, and then differentiating with respect to h.

□

All these results above lead to the following Corollary.

Corollary 3.7. *Under assumptions (B.1) to (B.5) $MISE^*$ and $AMISE^*$ are consistent estimators of MISE and AMISE, respectively.*

– Proof –

The consistency is almost trivial taking into account Theorem 3.3, Theorem 3.5, Corollary 3.4 and Corollary 3.6.

□

Remark 3.2. The theory developed in this section is restricted to the context of spatial point processes, i.e., processes in $\mathbb{R}^2$. A generalisation to the $\mathbb{R}^p$ space can be done, however it does not seem to be of any practical value due to the nature of the data we are using in spatial statistics.

3.5 Data-driven bandwidth selection

In this section we describe two new bandwidth selection methods for the kernel intensity estimator based on (3.5). These methods consist of adaptations of common selectors in the field of density estimation that have not yet been defined nor implemented in the point process framework. We propose a Normal scale rule (rule-of-thumb) and a bootstrap selector derived from the consistent resampling bootstrap procedure detailed in the previous section. All these proposals are based on estimating the infeasible optimal expression h_{AMISE} in (3.6).

3.5.1 Rule-of-thumb for bandwidth selection

This method is inspired in Silverman (1986). We assume that the underlying density (3.4) is Normal, $N(\mu, \sigma)$, with the parameters being estimated from the data, and in this way we replace the unknown values in (3.6).

In the point processes framework the computation is slightly different from the one used in the context of density estimation, because here the density is only a feature to get the intensity. To begin with, we have to remark that in our context, (3.6) has some other unknown elements apart from f, such as m and $A(m)$. The first one is the expected number of points, that in practice can be estimated by the sample size n, and the second one by $1/n$.

The only unknown element left is $\rho^{''}$. However, assuming that $f = \frac{\rho g^\star}{m}$ is Normal, we can derive that

$$\rho''(z) = m\left(\frac{f''(z)}{g^\star(z)} - \frac{2f'(z)(g^\star(z))'}{(g^\star(z))^2} - \frac{f(z)(g^\star(z))''}{(g^\star(z))^2} + \frac{2f(z)(g^\star(z)')^2}{(g^\star(z))^3}\right),$$

and then compute $R\left(\frac{\rho'' g^\star}{m}\right)$ using numerical integration methods. Replacing all those estimations in (3.6) we have the rule-of-thumb bandwidth selector that we will denote by $\hat{h}_{\text{RT}}$.

3.5.2 Bootstrap for bandwidth selection

The asymptotic expression of the optimal bootstrap bandwidth can be considered to derive a consistent bandwidth estimate. Cao (1993) suggested such approach for kernel density estimation with complete data and in Chapter 2 the result for length-biased data is detailed, as well as some remarks to extend it to general weighted data situation.

The expression (3.9) that we use to build this selector, has a few quantities that need to be computed: $\hat{m}$, $A(\hat{m})$ and $R(\frac{\hat{\rho}_b'' g^\star}{\hat{m}})$. The first two can be easily calculated through numerical integration methods such as Simpson's rule, while the last one requires some more development.

The main challenge in the estimation of $R(\frac{\hat{\rho}_b'' g^\star}{\hat{m}})$ is to obtain an appropriate value for the pilot bandwidth b. Regarding Cao (1993) and Theorem 2.7 we can assume that the order of that bandwidth in our context is $m^{-1/7}$, and that the constant has a slight influence on the final result. Hence we propose to use as pilot bandwidth a re-scaled version of the rule-of-thumb previously defined:

$$\hat{b} = m^{2/35}\hat{h}_{RT}.$$

Obviously in practice we do not know the value of m, so we use the best approximation we can have, which is the sample size of the corresponding realization of the point process.

Then, the bootstrap bandwidth we propose is:

$$\hat{h}_{\text{Boot}} = \left(\frac{A(\hat{m})R(L)}{\mu_2^2(L)(1-e^{-\hat{m}})^2 R(\frac{\hat{\rho}_b'' g^\star}{\hat{m}})}\right)^{-1/5}.$$

3.6 Finite sample study

In this section we perform a simulation study to analyse the behaviour of two different bandwidth selectors proposed in the previous section, $\hat{h}_{\mathrm{RT}}$ and $\hat{h}_{\mathrm{Boot}}$. We have also included in this comparison the only bandwidth selector that has been previously used in this context by Baddeley et al. (2012) for their estimator, this is, the common Silverman's rule-of-thumb for density estimation that we denote as $\hat{h}_{\mathrm{Silv}}$.

We have chosen three different models, all satisfying assumption (3.1). These models are nonhomogeneous Poisson point processes in the square unit with intensity function $\lambda(u) = \exp(\beta_0 + \beta_1 Z(u)), \quad u \in W = [0,1] \times [0,1]$, where β_0 and β_1 are given values. To define the three models we have constructed two different covariates that can be seen in Figure 3.1. The first one on the left, Z_1, is a realization of a gaussian random field, with zero mean and exponential covariance structure with parameters $\sigma = 0.1$ and $s = 0.1$, i.e., the covariance function is given by $C(r) = \sigma^2 \exp(-r/s)$, and it is used in Model 1 and Model 2. The second one on the right, dR, is a rescaled version into the square unit of the "distance to letter R" defined in Baddeley et al. (2012) that has been included in Model 3. Moreover, in Model 2 we have added an error term, ε, to perturb the covariate, in order to see how the estimator performs using only partial information on the real covariate generating the process. We have chosen a different realisation of the same gaussian random field we have defined for Z_1, see Figure 3.2.

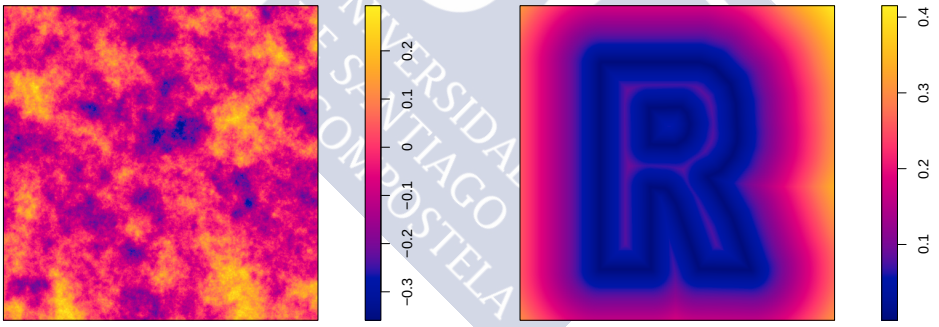


Figure 3.1: Representation of the covariates used in the three models, Z_1 on the left and dR on the right.

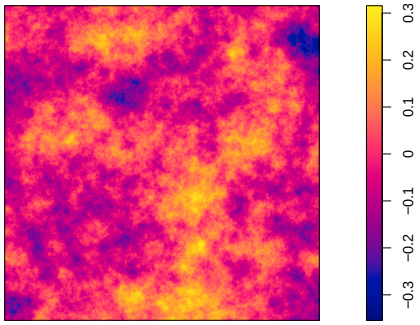


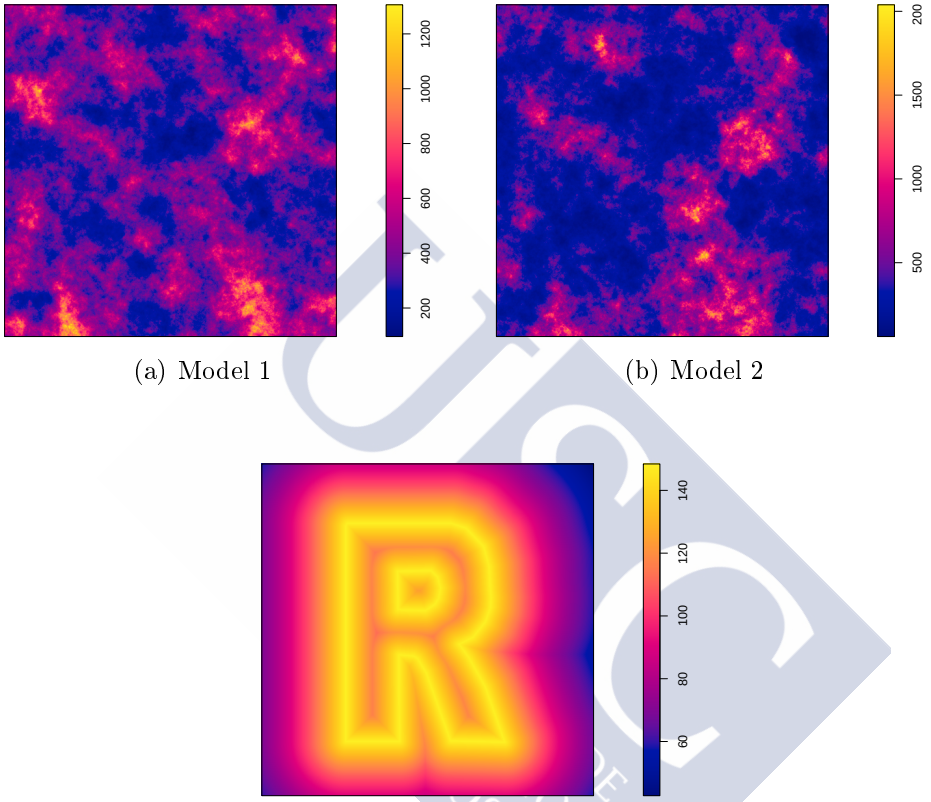
Figure 3.2: Representation of the error term, ε, used to perturb the information given to the estimator by the real covariate.

The three models are determined therefore by the three following intensities:

$$\lambda_1(u) = \exp(6 + 4Z_1(u)), \qquad \lambda_2(u) = \exp(6 + 4(Z_1 + \varepsilon)(u)),$$

$$\lambda_3(u) = \exp(5 - 3\mathrm{dR}(u)),$$

which are represented in Figure 3.3.



(a) Model 1 (b) Model 2

(c) Model 3

Figure 3.3: Intensity functions characterising the three simulated models.

We need to define appropriate error criteria for bivariate intensity estimation; we have set down the following:

$$\mathrm{ISE}_{\mathrm{rel}} = \int_W \left(\frac{\hat{\lambda}(u) - \lambda(u)}{\lambda(u)} \right)^2 du,$$

which is basically a relative integrated squared error. This error criteria is used to obtain the infeasible optimal bandwidth value $\hat{h}_{\mathrm{ISE}}$, that we consider as a benchmark, and which corresponds to the optimal bandwidth obtained from $\mathrm{ISE}_{\mathrm{rel}}$.

We simulate 500 Monte Carlo realizations for each model and different expected sample sizes covering a wide range of possibilities, $m = 50, 100, 250$ and 500. Notice

that the underlying model generating these samples is not exactly the same, we need to rescale the intensity function to guarantee that the mean number of events in the observation region is m.

From the samples, we evaluate the performance of the automatic bandwidth selectors through the following measures:

$$e_1 = \text{mean}\left(\text{ISE}_{\text{rel}}(\hat{h})\right), \qquad e_2 = \text{std}\left(\text{ISE}_{\text{rel}}(\hat{h})\right),$$
$$e_3 = \text{mean}\left((\hat{h} - \hat{h}_{\text{ISE}})/\hat{h}_{\text{ISE}}\right), \qquad e_4 = \text{std}\left((\hat{h} - \hat{h}_{\text{ISE}})/\hat{h}_{\text{ISE}}\right).$$

Similarly to what we have used in Chapter 2, measures e_1 and e_2 refer to the error of the estimation, providing with information about the overall performance and variability of the different methods. Meanwhile, e_3 and e_4 measure respectively, the bias and variability of the difference between the theoretical benchmark $\hat{h}_{\text{ISE}}$, and the value selected by the proposals. This provides information about how well the methods are choosing the bandwidth parameter.

	Model 1			
	$\hat{h}_{\text{ISE}}$	$\hat{h}_{\text{Silv}}$	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{Boot}}$
		$m = 50$		
e_1	0.0541	0.1371	0.1008	0.0773
e_2	0.0550	0.1536	0.1275	0.1132
e_3	—	-0.6443	-0.4836	-0.0430
e_4	—	0.1676	0.2270	0.4558
		$m = 100$		
e_1	0.0371	0.0801	0.0602	0.0470
e_2	0.0512	0.0951	0.0817	0.0773
e_3	—	-0.6341	-0.4759	0.0299
e_4	—	0.1487	0.1996	0.4277
		$m = 200$		
e_1	0.0228	0.0465	0.0349	0.0265
e_2	0.0240	0.0429	0.0329	0.0279
e_3	—	-0.6202	-0.4618	0.1018
e_4	—	0.1332	0.1804	0.4062
		$m = 500$		
e_1	0.0115	0.0231	0.0173	0.0131
e_2	0.0085	0.0159	0.0120	0.0093
e_3	—	-0.6166	-0.4588	0.1857
e_4	—	0.1164	0.1555	0.4044

Table 3.1: ISE_{rel} with criteria e_1 to e_4 for Model 1.

	Model 2			
	$\hat{h}_{\text{ISE}}$	$\hat{h}_{\text{Silv}}$	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{Boot}}$
		$m = 50$		
e_1	0.2222	0.4006	0.3697	0.3693
e_2	0.0974	0.2287	0.2118	0.2150
e_3	—	-0.7612	-0.6806	-0.4441
e_4	—	0.2653	0.3249	0.5545
		$m = 100$		
e_1	0.2147	0.3181	0.3053	0.3213
e_2	0.0848	0.1430	0.1367	0.1429
e_3	—	-0.7608	-0.6903	-0.4424
e_4	—	0.2934	0.3626	0.6461
		$m = 200$		
e_1	0.2055	0.2717	0.2658	0.2867
e_2	0.0408	0.0769	0.0717	0.0763
e_3	—	-0.7735	-0.7028	-0.4517
e_4	—	0.3046	0.3922	0.7202
		$m = 500$		
e_1	0.2002	0.2404	0.2386	0.2564
e_2	0.0217	0.0402	0.0392	0.0418
e_3	—	-0.7883	-0.7390	-0.4979
e_4	—	0.3215	0.3893	0.7515

Table 3.2: ISE_{rel} with criteria e_1 to e_4 for Model 2.

	Model 3			
	$\hat{h}_{\text{ISE}}$	$\hat{h}_{\text{Silv}}$	$\hat{h}_{\text{RT}}$	$\hat{h}_{\text{Boot}}$
		$m = 50$		
e_1	0.0937	0.1327	0.1161	0.1128
e_2	0.0893	0.1240	0.1122	0.1055
e_3	—	-0.3221	-0.1214	0.3603
e_4	—	0.3235	0.3900	0.6527
		$m = 100$		
e_1	0.0636	0.0774	0.0718	0.0740
e_2	0.0436	0.0586	0.0529	0.0482
e_3	—	-0.2130	-0.0120	0.4459
e_4	—	0.2978	0.3610	0.5789
		$m = 200$		
e_1	0.0416	0.0460	0.0447	0.0487
e_2	0.0220	0.0251	0.0234	0.0227
e_3	—	-0.0958	0.1220	0.5459
e_4	—	0.2750	0.3453	0.5244
		$m = 500$		
e_1	0.0231	0.0244	0.0251	0.0283
e_2	0.0085	0.0093	0.0088	0.0090
e_3	—	0.0306	0.2637	0.6034
e_4	—	0.2536	0.3166	0.4432

Table 3.3: ISE_{rel} with criteria e_1 to e_4 for Model 3.

An overview of the values in Tables 3.1, 3.2 and 3.3 indicates that in general, the bootstrap bandwidth seems to perform better than the others in most of the cases, and when this does not occur, our procedure is still competitive. Any of the other methods are not far away from it, even though the rule-of-thumb specifically designed for spatial point processes shows a slightly better behaviour than the Silverman's rule-of-thumb, specially in small sample sizes.

In terms of variability, measured by criterion e_2, $\hat{h}_{\text{RT}}$ and $\hat{h}_{\text{Boot}}$ are similar, even though the bootstrap shows in general smaller values. Silverman's bandwidth shows higher values in term of variability for the error, meanwhile, the values of the bandwidths, measured by e_4, are more equal and absolutely comparable.

The bias in bandwidth selection is measured through criterion e_3. In this case, in Model 1 and Model 2, bootstrap bandwidth outperforms far more better than the

others; while in Model 3 the bootstrap selector shows higher variability. Note also that the rule-of-thumb and Silverman's procedures show the bias in the same direction, to be more specific all of them tend to choose smaller bandwidths than the optimal one, while the bootstrap selector does the opposite except for Model 2.

To complete our analysis, we have also carried out another parallel simulation study to include a comparison between our proposals and the one in Guan (2008); remark that Baddeley's estimator behaviour is already studied when we use Silverman's selection method in our approach. To make the comparison simpler we have only included in the summary, Table 3.4, the theoretical benchmark for Guan's estimator, $\hat{h}_{\mathrm{ISE}_{\mathrm{Guan}}}$, and for ours, $\hat{h}_{\mathrm{ISE}}$. For Guan's proposal we have implemented his own bandwidth selection criterion, a least-squares cross-validation, $\hat{h}_{\mathrm{CV}}$ and not to flood with much data we only included our bootstrap bandwidth selector, $\hat{h}_{\mathrm{Boot}}$. In Table 3.4, the results obtained from 500 Monte Carlo samples of two sample sizes are shown.

	$m = 50$				$m = 100$			
	$\hat{h}_{\mathrm{ISE}_{\mathrm{Guan}}}$	$\hat{h}_{\mathrm{CV}}$	$\hat{h}_{\mathrm{ISE}}$	$\hat{h}_{\mathrm{Boot}}$	$\hat{h}_{\mathrm{ISE}_{\mathrm{Guan}}}$	$\hat{h}_{\mathrm{CV}}$	$\hat{h}_{\mathrm{ISE}}$	$\hat{h}_{\mathrm{Boot}}$
				Model 1				
e_1	0.0703	0.1855	0.0542	0.0786	0.0427	0.1885	0.0372	0.0471
e_2	0.0693	0.1367	0.0550	0.1142	0.0395	0.1034	0.0512	0.0814
e_3	—	1.6244	—	-0.0697	—	2.6333	—	0.0180
e_4	—	1.6454	—	0.4661	—	1.8395	—	0.3926
				Model 2				
e_1	0.3217	0.4479	0.2225	0.3702	0.2726	0.4027	0.2148	0.3248
e_2	0.1786	0.2474	0.0978	0.2159	0.1061	0.1756	0.0852	0.1464
e_3	—	1.2894	—	-0.4534	—	1.6405	—	-0.4310
e_4	—	1.5094	—	0.5575	—	1.4820	—	0.6487
				Model 3				
e_1	0.0367	0.0737	0.0938	0.1130	0.0231	0.0442	0.0637	0.0742
e_2	0.0425	0.0822	0.0893	0.1056	0.0252	0.0457	0.0436	0.0484
e_3	—	0.0201	—	0.3606	—	0.2143	—	0.4402
e_4	—	0.6255	—	0.6708	—	0.7085	—	0.5923

Table 3.4: $\mathrm{ISE}_{\mathrm{rel}}$ with criteria e_1 to e_4 for Guan's estimator.

To sum up these values we have to take into account some facts. Note that in the practical definition of Guan's estimator a numerical integration is included which is computationally hard, and moreover, it is known that using a cross-validation criterion implies also a higher computational cost, specially for large sample sizes; this justifies the inclusion of only two sample sizes in this study.

Apart from this, but intimately related to the cross-validation criterion, we have observed that it is not guaranteed for some samples to have a global minimum for the CV function, which leads to an incorrect election of the bandwidth with a local minimum or a boundary of the minimisation interval; this has occurred between 45 and 213 times out of 500 depending on the model and the sample size (it is more likely for smaller sizes).

The numbers in Table 3.4 show that in general, in Model 1 and Model 2, our estimator is theoretically and practically better than Guan's; while Guan's outperforms better in Model 3. In terms of variability, measured by criteria e_2 and e_4, the bootstrap selector provides in general much smaller values, even for medium sample sizes in Model 3. Bias seems also be a problem for Guan's proposal which is not common for cross-validation.

3.7 Canadian wildfires

Forest fires are one of the most important natural disturbances since the last Ice Age and they represent a huge social and economic problem. Canada has quite a long tradition on recording information about their wildfires; and also studies from many different perspectives have been carried out: Walter et al. (2014), Rogers et al. (2013), Di Iorio et al. (2013), Flannigan and Harrington (1988). It is quite well known that fire activity in Canada mostly relies on meteorological elements such as long periods without rain, high temperatures and also lightnings.

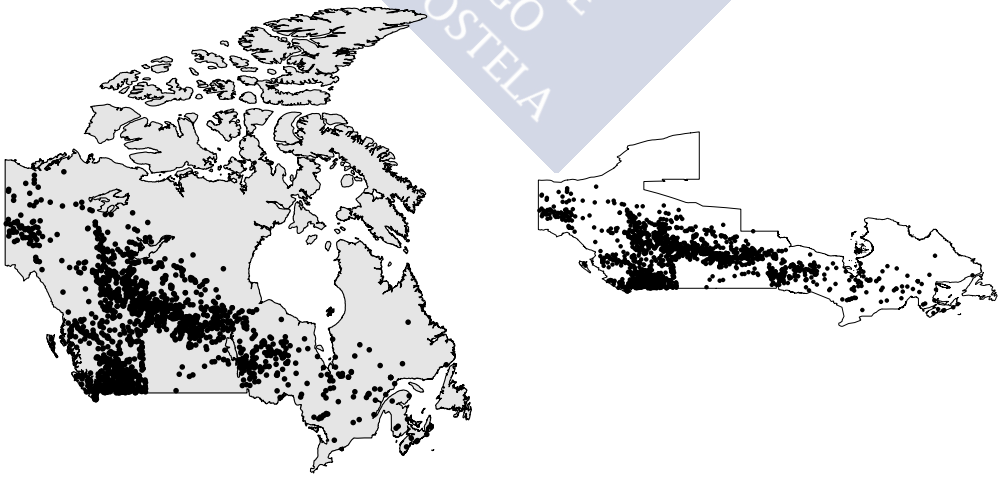


Figure 3.4: Locations of wildfires in Canada during June 2015, over the whole country (left) and only on the observation region (right).

It is important to note that for inferential purposes we have removed two regions (Northwest Territories and Nunavut) from the whole observation window (Canada), because there are no fires registered on those regions and we cannot do any inference with such a lack of information.

We are interested in studying the spatial influence on some of meteorological variables on the distribution of wildfires. The wildfire dataset and also a complete meteorological information from the last decades is available at the Canadian Wildland Fire Information System website (`http://cwfis.cfs.nrcan.gc.ca/home`). The fire season in Canada lasts from late April until August, with a peak of activity in June and July, hence we are interesting in analysing the influence of meteorological covariates on wildfires during June 2015 (see Figure 3.4), and we focus our attention on precipitation and temperature (see Figure 3.5) as the informative covariates.

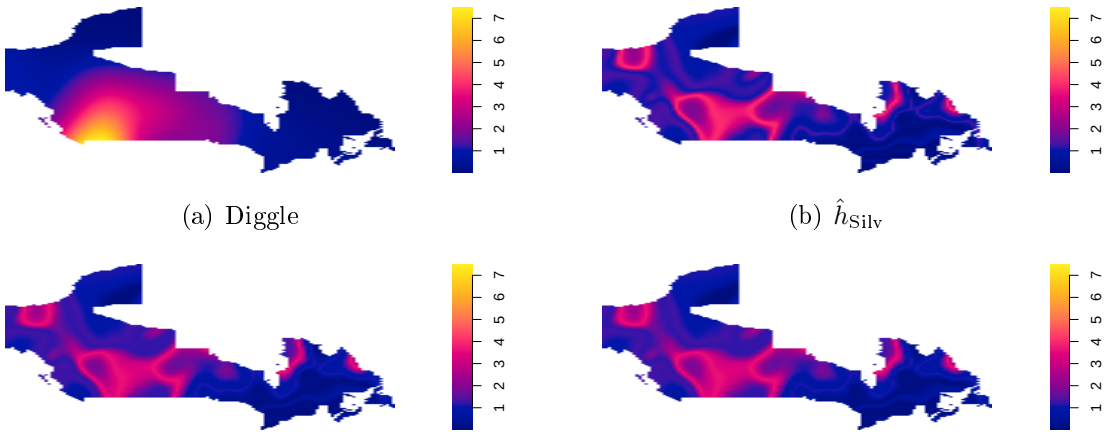
(a) Mean noon-24 hour precipitation registered in June 2015 after a gaussian smoothing with $\sigma = 2$ (in millimetres).

(b) Third quartile of the temperature registered in June 2015 after a gaussian smoothing with $\sigma = 2$ (in Celsius degrees).

Figure 3.5: Covariates to be used in the intensity estimation.

First we start the analysis with the precipitation. We want to compare the results coming from the different bandwidth selectors but also, in order to detect the influence of the covariate, we have included the nonparametric kernel intensity estimation proposed in P. Diggle (1985) that uses only the point pattern coordinates to compute the intensity estimation.

In Figure 3.6 we can see that the results of the estimation obtained with and without the covariate are quite different, while the estimations obtained with the three bandwidth selectors are almost the same. Regarding the point pattern in Figure 3.4 we can confirm that the estimation with the precipitation information is able to reproduce the areas with a higher concentration of events; meanwhile Diggle's proposal is only capable of capture the highest intensity area in south west Canada, and it expands circularly from it, seeming not to follow the point pattern.

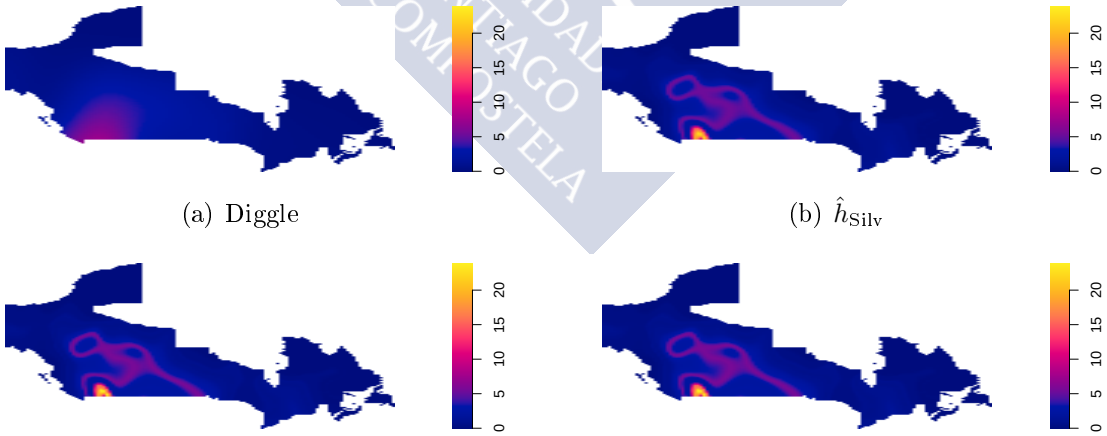


(a) Diggle (b) $\hat{h}_{\text{Silv}}$

(c) $\hat{h}_{\text{RT}}$ (d) $\hat{h}_{\text{Boot}}$

Figure 3.6: Estimation without covariate information (a), and estimations with the different bandwidth selectors using the precipitation as covariate (b), (c) and (d).

We reproduce again the same methods using the temperature as covariate. Instead of considering the maximum value during June in every point of the region, we have computed the third quartile in order not to deal with extreme values. Remark that the scale values of the intensities are different from the ones above so, we have reproduced again Diggle's estimator, which is obviously the same (the covariate is not used) but with a different range of values, in order to allow a fair comparison.



(a) Diggle (b) $\hat{h}_{\text{Silv}}$

(c) $\hat{h}_{\text{RT}}$ (d) $\hat{h}_{\text{Boot}}$

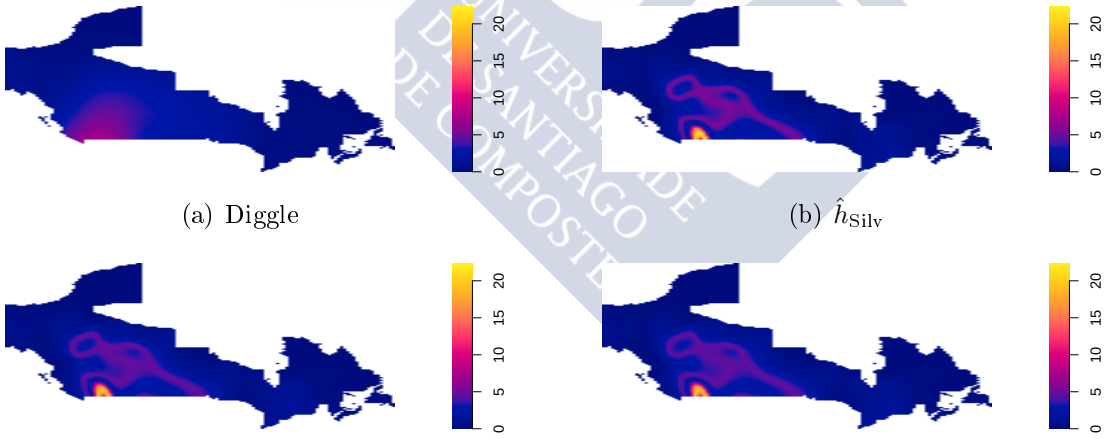
Figure 3.7: Diggle's estimation without covariate information (a), and estimations with the different bandwidth selectors using the temperature as covariate (b), (c) and (d).

Remark that our theoretical model (3.1), allows only the use of one single

unidimensional covariate in the model, that is why we have so far performed the analysis of the data with each of the covariates at a time. However, we can look now for a way to join a bunch of them.

In Section 3.9 we detail some possible procedures to be able to use several covariates in this context. One possibility is to perform a principal component analysis (PCA) using the first principal component in the estimation. In this particular example of wildfire in Canada, the result is $\text{PCA1} = 0.991 * \texttt{Temp} + 0.131 * \texttt{Precip}$, which explains the 93% of the total variance and the correlations with temperature and precipitation were, respectively, 0.9993 and 0.4323. A representation of this new resulting covariate is plotted in Figure 3.8.

Figure 3.8: First principal component resulting of the PCA analysis performed with precipitation and temperature.



(a) Diggle (b) $\hat{h}_{\text{Silv}}$

(c) $\hat{h}_{\text{RT}}$ (d) $\hat{h}_{\text{Boot}}$

Figure 3.9: Estimation without covariate information (a), and estimations with the different bandwidth selectors and using the first PCA component as the covariate (b), (c) and (d).

Figure 3.9 draw similar conclusions as the previous analysis, this is, the covariate, in this case the first principal component, is really informative in outperforming the estimation. Moreover, and as we have been able to include both covariates, we can

confirm that actually both are useful even though the temperature seems to be more influential (remark the expression of the first PCA defined above). The three bandwidth selectors compute similar values so the results of the estimations are almost the same.

Hence, regarding Figure 3.4, Figure 3.7 and Figure 3.9 we can assure that the information given by the covariates is relevant for the estimation, because as we have pointed out Diggle's proposal can barely identify the area with more fires, while when using this extra information, the estimation seems to be more suitable according to the point process pattern. Among the three bandwidth selectors we cannot identify in this specific example one of them outperforming better than the others, actually the resulting estimations look like very similar because the bandwidth values are quite close.

3.8 Conclusions

We have considered kernel intensity estimation in the context of spatial point processes with covariates. We have set up a theoretical framework that has allowed us to develop in detail the asymptotic expansions of the MSE, MISE and AMISE of our intensity estimator. Furthermore we have proposed a consistent smooth bootstrap procedure, and two new data-driven bandwidth selection methods. We have also studied their behaviour, and compare them with the previous methods used in this context, through an extensive simulation study; the overall conclusion being a better performance of the bootstrap based bandwidth. The application to a real dataset also shows that our proposals are competitive with the existing ones, and even better than kernel intensity estimators based only on information provided by the locations of the events. Also in terms of computational costs our proposals are faster or competitive with the existing ones.

3.9 Extensions

3.9.1 Spatio-temporal point processes

Spatio-temporal point processes are the most common extension of the spatial ones. In the context of point processes with covariates almost nothing has been done in the spatio-temporal field, even though the extension seems still natural.

Let us define a spatio-temporal intensity depending on a covariate:

$$\lambda(u,t) = \rho(Z(u,t)),$$

where u and t are, respectively, the spatial and temporal coordinates, $Z : W \times T \subset \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R}$ is the covariate and ρ is a unknown real function. As it has been done before, we assume that Z is known so, we only need to estimate ρ in order to obtain an estimator of λ.

As the spatio-temporal influence is gathered all together through the covariate, all the theory previously developed in this chapter can be immediately applied to this new situation, just taking into account the change of dimension in the domain of the covariate.

Another possible framework in this spatio-temporal context can be that the covariate has only a spatial or a temporal dependence, i.e, that Z is either a function of space or either a function of time but not both together. Hence, the intensity function is

$$\lambda(z,t) = \rho(Z(x),t)) \text{ or } \lambda(x,z) = \rho(x, Z(t)).$$

Remark that now ρ is still a real function but with multivariate domain. In this situation, following the literature about kernel intensity estimation in general spatio-temporal point processes, we propose the following estimators:

$$\hat{\rho}(z,t) = \sum_{i=1}^{N} \frac{1}{g^\star(Z_i)} K_h\left(z - Z_i\right) K_s(t - t_i) \text{ and}$$

$$\hat{\rho}(x,z) = \sum_{i=1}^{N} \frac{1}{g^\star(Z(t_i))} |H|^{-1/2} L\left(H^{-1/2}(x - X_i)\right) K_s(Z(t) - Z(t_i)),$$

where $(X_1, t_1), \ldots, (X_N, t_N)$ is the spatio-temporal pattern, s is a univariate bandwidth parameter, L is a radially simmetric bivariate density function, H is a two dimensional bandwidth matrix, and Z_i and $g^\star$ are previously defined in the manuscript. The theoretical developments done in this work can be extended to these situations with the appropriate regularity conditions.

3.9.2 Increase covariate dimension

Although the framework we have set up in this chapter is defined to use one single one-dimensional covariate, it is of interest to think about how this can be extended to the multivariate case.

Let suppose that we have $\mathbf{Z} = (Z_1, \ldots, Z_p)$ a p-dimensional covariate providing possible significant information about the process. If we want to include it in the model, a first natural approach is defining a linear combination $a_1 Z_1 + \ldots + a_p Z_p$, where a procedure to estimate the coefficients needs to be determined, for example performing a PCA, as it has been shown in Section 3.7, choosing the first component to use its information in the estimation procedure and following our methodology.

Another possible solution to include several covariates is linked to the idea we have taken advantage from of the relationship between the density and the intensity function. Hence, a multivariate version of (3.5) can be defined, where the use of several covariates would be allowed; for sure the technical details must also be adapted to this new estimation procedure. Here we will briefly introduce the results needed to perform this extension, even though we are not developing all the asymptotics in detail.

We need to introduce some notation to adapt the definition of the spatial cumulative distribution function in (3.3) to the multivariate context. Hence, let define

$$G(z) = \frac{1}{|W|} \int_W 1_{\{\mathbf{Z}(u) \leq z\}} du,$$

where $\mathbf{Z}(u) \leq z$ refers to $(Z_1(u) \leq z_1) \cap \ldots \cap (Z_p(u) \leq z_p)$ with $z \in \mathbb{R}^p$. We still use g to denote its first derivative, and $G^\star$ and $g^\star$ their unnormalised versions.

Now we want to extend Theorem 3.1 and Theorem 3.2 to the multivariate case, for which we recall the results in Daley and Vere-Jones (1988) and Reiss (2012).

Theorem 3.8. *Let X be a spatial point process in $W \subset \mathbb{R}^2$ with intensity function of the form $\lambda(u) = \rho(\mathbf{Z}(u))$ for some real-value function ρ and $\mathbf{Z} : W \subset \mathbb{R}^2 \to \mathbb{R}^p$ a continuous function, then $\mathbf{Z}(X)$ is a p-dimensional point process in $\mathbb{R}^p$ with intensity function $\rho g^\star$. Moreover, if the original point process is Poisson, then the transformed one preserve this property and it is also Poisson.*

Theorem 3.9. *Let $W \subset \mathbb{R}^2$ be a bounded subset, $\mathbf{Z} : W \subset \mathbb{R}^2 \to \mathbb{R}^p$ a measurable, Lipschitz and differentiable function with non-zero Jacobian in every point of W, $J\mathbf{Z}(u) \neq 0$. Then, for any integrable function $l : W \to \mathbb{R}$, in our particular case $l(u) = \lambda(u)(J\mathbf{Z}(u))^{-1} : W \to \mathbb{R}^+$, it holds:*

$$\int_W \lambda(u)du = \int_W \lambda(u)(J\mathbf{Z}(u))^{-1} J\mathbf{Z}(u)du = \int_{\mathbb{R}^p} \int_{\mathbf{Z}^{-1}(y)} \lambda(u)(J\mathbf{Z}(u))^{-1} dH^p(u)dy,$$

where $\mathbf{Z}^{-1}(y) = \{u \in W / \mathbf{Z}(u) = y\}$ and dH^p is the p-dimensional Haussdorf measure.

Applying this result to the unnormalised version of the spatial cumulative distribution function defined above, we have:

$$G^\star(z) = \int_{-\infty}^{z_1} \cdots \int_{-\infty}^{z_p} \int_{\mathbf{Z}^{-1}(y)} (J\mathbf{Z}(u))^{-1} dH(u)dy_1, \ldots dy_p,$$

and deriving with respect to z, we get $g^\star(z) = \int_{\mathbf{Z}^{-1}(z)} \rho(\mathbf{Z}(u))(J\mathbf{Z}(u))^{-1} dH^p(u)$.

We can now rewrite the relationship between the original spatial point process intensity and the transformed one through an integral, in a similar way we have done previously in this chapter for the univariate case, implying that the expected number

of events in the corresponding region is the same in both processes:

$$\begin{aligned} m &= \int_W \lambda(u)du = \int_{\mathbb{R}^p}\int_{\mathbf{Z}^{-1}(y)} \lambda(u)(J\mathbf{Z}(u))^{-1}dH(u)dy \\ &= \int_{\mathbb{R}^p}\int_{\mathbf{Z}^{-1}(y)} \rho(\mathbf{Z}(u))(J\mathbf{Z}(u))^{-1}dH(u)dy = \int_{\mathbb{R}^p}\rho(y)\int_{\mathbf{Z}^{-1}(y)}(J\mathbf{Z}(u))^{-1}dH(u)dy \\ &= \int_{\mathbb{R}^p}\rho(y)g^{\star}(y)dy. \end{aligned}$$

Once we have established the appropriate framework, we propose the following estimator for the associated "artificial" multivariate density, $f(\cdot) = \frac{\rho(z)g^{\star}(z)}{m}$:

$$\hat{f}_h(z) = g^{\star}(z)\frac{1}{N}\sum_{i=1}^{N}\frac{1}{g^{\star}(Z_i)}\mathbf{K}_H\left(z - Z_i\right)1_{\{N\neq 0\}},$$

and the extension of Baddeley's proposal to estimate ρ can be:

$$\hat{\rho}(z) = \sum_{i=1}^{N}\frac{1}{g^{\star}(Z_i)}\mathbf{K}_H(z - Z_i),$$

where now $\mathbf{K}$ should be a multivariate radially symmetric kernel function and H a p-dimensional bandwidth matrix.

The analogous theoretical developments presented in Section 3.3 and Section 3.4 can be extended easily to the multivariate situation following the steps we have already detailed for the univariate case, and also applying some specifically designed statistical tools from multivariate analysis used in Scott (1992), Cucala (2006) and Fuentes-Santos et al. (2015).

Chapter 4

Testing covariate significance in spatial point processes first-order intensity

Contents

Modelling the first-order intensity function in one of the main aims in point process theory, and it has been approached so far from different perspectives. One appealing model describes the intensity as a function of a spatial covariate. In the recent literature, estimation theory and several applications have been developed assuming this hypothesis, but without formally checking the goodness-of-fit of the model. In this chapter we address this problem and we test whether the model is appropriate. We propose a test statistic based on a L^2-distance; we prove the asymptotic normality of the statistic and suggest a bootstrap procedure to calibrate the test. We present two applications with real data and a simulation study to better understand the performance of our proposals.

4.1 Introduction

The understanding of the spatial distribution of point patterns is crucial in many different fields as we have already illustrated in this manuscript and is also shown in many references: ecology (Illian et al. (2009) and Law et al. (2009)); epidemiology (Lawson (2013)); seismology (Ogata and Zhuang (2006) and Schoenberg (2011)); forestry (Stoyan and Penttinen (2000)); geology (Foxall and Baddeley (2002)); and also plays an important role in statistical theory, see for example Daley and Vere-Jones (1988),P. J. Diggle (2013) and Cressie (1993).

Initially, from the point of view of statistical inference, the main interest of spatial point process theory was to test the complete spatial randomness (CSR) hypothesis, i.e., to determine if a pattern comes from a homogeneous Poisson point process, in which case the underlying intensity is constant over the observation region; distance-based and quadrat counts are the classical methods on this topic, see P. J. Diggle (2013) and Cressie (1993) for more details. Moreover, some applications have been developed under this assumption, as it can be seen in Baddeley and Van Lieshout (1995) and Dasgupta and Raftery (1998).

The first order characteristics, among which the intensity function is the most important one, have been also studied from different perspectives, such as Bayesian, see Altieri et al. (2015) and Mrkvička and Soubeyrand (2017), parametric and nonparametric techniques. Recall the intensity is formally defined as:

$$\lambda(x) = \lim_{|dx|\to 0} \frac{E[N(dx)]}{|dx|},$$

where N is the counting measure, i.e., N is the random variable that counts the number of points lying on a given set. Assumed parametric models and using for instance a likelihood score, see Van Lieshout (2000), Moller and Waagepetersen (2003) and P. J. Diggle (2013) or pseudolikelihood procedures, see Waagepetersen (2007) may generate unreliable estimates when it deviates from the true one. Hence the

nonparametric approach is a well-known alternative. P. Diggle (1985) proposed the first kernel intensity estimator which has been widely used in exploratory analysis and that we present in Chapter 3. The main disadvantage of this estimator is its lack of consistency, for which Cucala (2006) proposed a modification based on the relationship between the intensity and the density functions.

Recently, this lack of consistency and some real applications requirements induced a new scenario based on the inclusion of covariates in the model. Guan (2008) proposed a kernel intensity estimator, assuming that the intensity function depends on some observed spatially varying covariates through an unknown continuous function. The consistency of his estimator was proved under an increasing domain asymptotic framework, and he also deals with the problem of high dimensional covariates. Recall also that later, Baddeley et al. (2012) considered spatial covariates postulating that

$$\lambda(x) = \rho(Z(x)),\ x \in W \subset \mathbb{R}^2, \tag{4.1}$$

where $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ is a spatial continuous covariate that is exactly known in every point of the region of interest W. Their proposal includes some intensity estimators based on the local likelihood as well as some kernel estimators, but without proper theoretical developments. In Chapter 3 we have developed in detail the theoretical properties of a kernel intensity estimator in this context of point processes with covariates; proposing a specifically designed bootstrap procedure and data-driven bandwidth selectors.

The inclusion of spatially varying covariates has been a big step forward in point process theory; however, a little attention has been paid to test the significance of these covariates. To the extent of our knowledge, only something similar has been done in a slightly different context in Díaz-Avalos et al. (2014), where the authors assume that the intensity depends on a linear combination of several covariates and they test, using the conditional intensity function, if any of the coefficients is null. In this work we try to fulfil the existing gap on checking the goodness-of-fit of model (4.1) by defining a new testing procedure.

This chapter is organised as follows. Section 4.2 is devoted to present in detail the two motivating examples we use all along the chapter. In Section 4.3 the statistical test is introduced, detailing its asymptotic null distribution and a bootstrap procedure which is used to improve its calibration. The methodology proposed is applied to the real datasets in Section 4.4. The performance of the test is analysed in Section 4.5 through a simulation study based on real data analyses in order to make it more realistic. Finally we draw some remarks in Section 4.6.

4.2 Motivating examples

In this section we motivate our proposal using two real dataset. The first one is the Murchison dataset, that has been widely used by Baddeley in many of his papers, books and software; and a new dataset made up of wildfires in Canada.

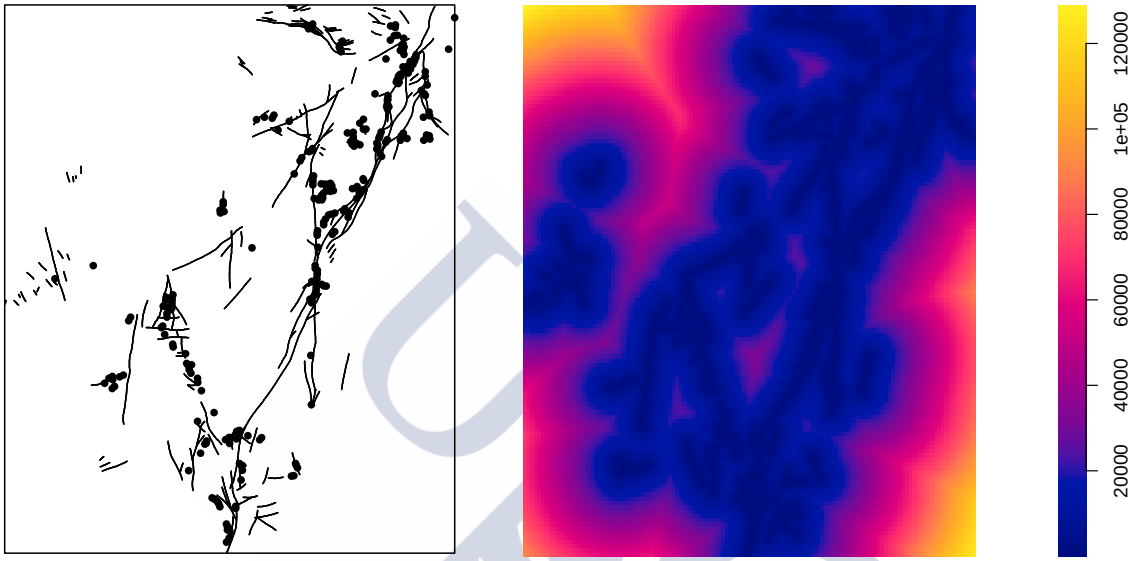


Figure 4.1: On the left: Murchison geological survey data, gold deposits (points) and geological faults (lines); on the right: Covariate information, i.e, distance to the nearest geological fault (in meters)

The Murchison geological survey data shown in Figure 4.1 record the spatial locations of gold deposits (a total number of 255) and the surrounding geological faults. These data came from a 330×394 km region in the Murchison area of Western Australia and have been obtained from Watkins and Hickman (1990). At this scale (1:500000) the gold deposits spatial extension is negligible and they can be considered as points without losing generality. Note that the real gold deposits and faults are three-dimensional while here we use a two-dimensional projection. Moreover, some geological faults may have been missed because they are not recorded by direct observation but in magnetic field surveys or geologically inferred from discontinuities in the rock sequences.

Once we have the locations of the gold deposits and the faults, the construction of the covariate is simple, we have to compute the distance from every point in the observation region to the nearest fault, see Figure 4.1 (right). We use this covariate to model the intensity of the process under assumption (4.1).

In Baddeley et al. (2015) their aim on studying this dataset is to "specify zones of hight prospectivity to be explored for gold", so they have already assumed that the

influence of the fault information is relevant and it may actually explain the localisation of gold deposits under model (4.1). Our goal is to check the adequacy of this model and test the hypothesis that the distance to geological faults is enough to explain the spatial distribution of gold deposits.

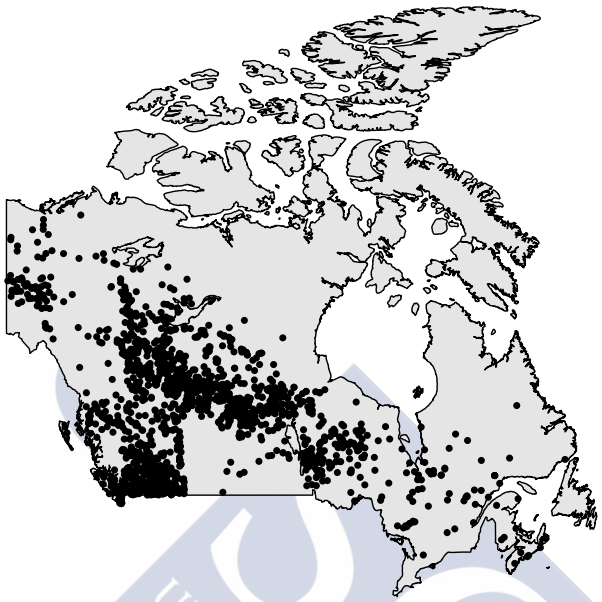


Figure 4.2: Wildfires in Canada during June 2015.

We know that forest fires are one of the most important natural disturbances since the last Ice Age and they represent a huge social and economic problem. Canada has quite a long tradition on recording information about their wildfires; and also studies from many different perspectives have been carried out: Walter et al. (2014), Rogers et al. (2013), Di Iorio et al. (2013), Flannigan and Harrington (1988). We have already pointed out that fire activity in Canada mostly relies on meteorological elements such as long periods without rain, high temperatures and also lightnings.

We are interested in studying the spatial influence of some of this meteorological covariates on the distribution of wildfires. Recall that the fire season in Canada lasts from late April until August, with a peak of activity in June and July, hence we are interesting in analysing the influence of meteorological covariates on wildfires during June 2015 (a total number of 1841), see Figure 4.2, and we focus our attention on temperature (see Figure 4.3). It is important to note again that for inferential purposes we have removed two regions (Northwest Territories and Nunavut) from the whole observation window (Canada) because there are no fires registered on those regions and we cannot do any inference with such a lack of information.

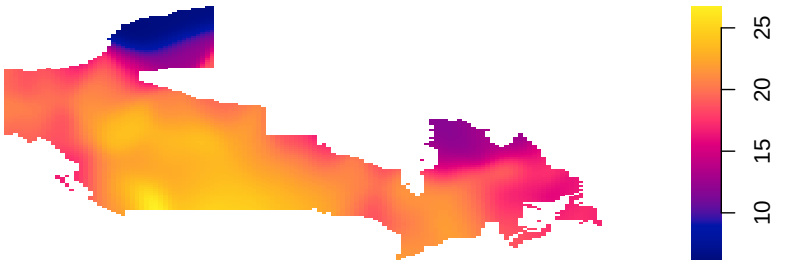


Figure 4.3: Third quartile of the temperature registered in June 2015 in Canada, after a gaussian smoothing with $\sigma = 2$ (in Celsius degrees).

Our aim is try to determine if the temperature is the main covariate influencing the generating process of the wildfires in Canada, in the form detailed in (4.1), and then if it can explain itself the spatial distribution of wildfires.

To check if a selected covariate provides with enough information to determine de intensity model, we present in the next section a testing procedure based on nonparametric techniques.

4.3 The proposed method

The objective of the test we are proposing here is to check whether the model we have previously assumed in Chapter 3 for the intensity function is or not appropriate. In practice testing this assumption should the first step to follow before applying any of the already defined methodology that form Chapter 3.

4.3.1 The test

Let X be a point process defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area. Let $X_1, \ldots, X_N$ be a realisation of the process where N is the random variable counting the number of events. Let again $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ be the spatial continuous covariate that is exactly known in every point of the region of interest W.

We want to test a null hypothesis $H_0 : \lambda(x) = \rho(Z(x)), x \in W$ versus a general alternative in which the intensity function is not explained completely through the covariate. The idea is to define a test statistic based on a L^2-distance between the classical kernel intensity estimator defined by P. Diggle (1985) and the intensity estimator proposed by Baddeley et al. (2012) and Borrajo et al. (2017b). Due to

the lack of consistency of Diggle's proposal we have decided to do an equivalent comparison using the concept of "density of events location" of Cucala (2006) instead of using the intensities, i.e., the null hypothesis can be equivalently rewritten as $H_0 : \lambda_0(x) = \rho(Z(x))/m$, with $\lambda_0(x) = \lambda(x)/m$ and $m = \int_W \lambda(x)dx$.

Hence, the test statistic is defined as:

$$S = \int_W \left(\hat{\lambda}_{0,H}(x) - \hat{\rho}_{0,b}(Z(x))\right)^2 dx, \tag{4.2}$$

where $\hat{\lambda}_{0,H}(x) = \frac{1}{Np_H(x)} \sum_{i=1}^{N} K_H(x - X_i) 1_{\{N\neq 0\}}$ is the bivariate estimation for the density of events location proposed by Fuentes-Santos et al. (2015), H is a bandwidth matrix, $p_H(x)$ is the edge correction term, $\hat{\rho}_{0,b}(Z(x)) = \frac{\hat{\rho}_b(x)}{N} 1_{\{N\neq 0\}}$ with $\hat{\rho}_b(x) = \sum_{i=1}^{N} \frac{1}{g^*(Z(X_i))} L_b(Z(x) - Z(X_i))$, b is a bandwidth parameter, K and L are kernel functions, $K_H(u) = |H|^{-1/2} K(H^{-1/2}u)$, $L_b(u) = \frac{1}{b} L\left(\frac{u}{b}\right)$ and g^* is the unnormalised version of the derivative of the spatial cumulative distribution function $G(z) = \int_W 1_{\{Z(u)\leq z\}du}$.

4.3.2 Asymptotic properties and calibration

Hereafter we assume that $W = \mathbb{R}^2$ to avoid the edge effects in the theoretical developments, and we need to introduce some regularity conditions:

(C.1) $\int_{\mathbb{R}} L(z)dz = 1$; $\int_{\mathbb{R}} zL(z)dz = 0$ and $\mu_2(L) := \int_{\mathbb{R}} z^2 L(z)dz < \infty$,

(C.2) $\lim_{m\to\infty} b = 0$ and $\lim_{m\to\infty} \frac{A(m)}{b} = 0$, where $A(m) = E\left[\frac{1}{N} 1_{\{N\neq 0\}}\right]$,

(C.3) The bandwidth matrix H is symmetric and positive-definite, and such that all entries of H tends to zero, and $m^{-1}|H|^{-1/2} \to 0$ as m increases,

(C.4) K is a continuous, symmetric, square integrable bivariate density function such that $\int_{\mathbb{R}^2} uu^T K(u)du = \mu_2(K) I_2$ with $\mu_2(K) < \infty$ and I_2 denoting the two-dimensional identity matrix and

(C.5) $Z(x)$ is a continuity point of ρ for all $x \in W$.

Hall (1984) proposed a central limit theorem for the integrated square error of multivariate kernel density estimators, which have a similar structure to our test statistic. The following theorem provides an analogous result to the one in Hall (1984) in the context of spatial point processes with covariates for the statistic S.

Theorem 4.1. *Under conditions (C.1) to (C.5) and assuming the null hypothesis* H_0 :

$\lambda(x)/m = \rho(Z(x))/m$, $\forall x \in W$, *it holds*

$$\frac{S-\mu_S}{\sigma_S} \longrightarrow N(0,1),$$

where

$$\mu_S = A(m)|H|^{-1/2}R(K)+\frac{1}{2}\mu_2(K)\int \lambda_0(x)tr(HD^2\lambda_0(x))dx+\frac{1}{4}\mu_2^2(K)\int tr^2(HD^2\lambda_0(x))dx,$$

$$\sigma_S^2 = A(m)|H|^{-1/2}\int\int \lambda_0^2(x)\lambda_0(y)(K\circ K)(H^{-1/2}(x-y))dxdy+2A(m)|H|^{-1/2}R(\lambda_0)R(K),$$

with $\circ$ *denoting the convolution between two functions,* $tr(\cdot)$ *the trace of a matrix,* D^2 *the Hessian matrix and* $R(K)=\int K^2(x)dx$.

– Proof –

Along this proof we obtain the mean and variance of the statistic S as well as its asymptotic normality. We start rewwriting the statistic (4.2) in a more convenient way:

$$\begin{aligned} S &= \int_W \left(\hat{\lambda}_{0,H}(x)-\hat{\rho}_{0,b}(Z(x))\right)^2 dx = \int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx \\ &+ \int_W \left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2 dx + 2\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right) dx. \end{aligned} \tag{4.3}$$

Mean and variance of S

From expression (4.3) into account, we immediately obtain that

$$\begin{aligned} E[S] &= E\left[\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx\right] + E\left[\int_W \left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2 dx\right] \\ &+E\left[\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right) dx\right] \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}[S] &= \mathrm{Var}\left[\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx\right] + \mathrm{Var}\left[\int_W \left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2 dx\right] \\ &+\mathrm{Var}\left[\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right) dx\right] \\ &+2\mathrm{Cov}\left(\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx, \int_W \left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2 dx\right) \\ &+2\mathrm{Cov}\left(\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx, \int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right) dx\right) \\ &+2\mathrm{Cov}\left(\int_W \left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2 dx, \int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right) dx\right). \end{aligned}$$

So, we first of all compute the mean and the variance of each of the addends in (4.3), and finally we will deal with the covariances between the different terms. For all cases we proceed as follows: first obtain the explicit expressions for the squares, then swap the mean operator and the integrals, and then compute several means of product terms of the estimators involved. In this last step we use properties of conditional mean as well as some Taylor expansions.

First addend

$$\begin{aligned} E\left[\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx\right] &= \int E\left[\hat{\lambda}_{0,H}(x)\right]dx - 2\int \lambda_0(x)\hat{\lambda}_{0,H}(x)dx + R(\lambda_0) \\ &= A(m)|H|^{-1/2}R(K) + \tfrac{1}{4}\mu_2^2(K)\int tr^2(HD^2\lambda_0(x))dx \\ &+ o\left(A(m)|H|^{-1/2}\right) + o\left(tr(H)\right) \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}\left[\int_W \left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2 dx\right] &= \int\int \Big(E\left[\hat{\lambda}_{0,H}^2(x)\hat{\lambda}_{0,H}^2(y)\right] + 2\lambda_0^2(y)E\left[\hat{\lambda}_{0,H}^2(x)\right] \\ &-4\lambda_0(y)E\left[\hat{\lambda}_{0,H}^2(x)\hat{\lambda}_{0,H}(y)\right] \\ &-4\lambda_0^2(x)\lambda_0(y)E\left[\hat{\lambda}_{0,H}(y)\right] \\ &+4\lambda_0(x)\lambda_0(y)E\left[\hat{\lambda}_{0,H}(x)\hat{\lambda}_{0,H}(y)\right] \\ &+\lambda_0^2(x)\lambda_0^2(y)\Big)\,dxdy \\ &= R(K)o(A(m)|H|^{-1/2}) \\ &-2R(K)R(\lambda_0)o(A(m)|H|^{-1/2}) \\ &-6R(\lambda_0)o(tr(H)), \end{aligned}$$

where we have used the following results:

$$E\left[K_H(x-X_1)\right] = \int K_H(x-u)\lambda_0(u)du = \lambda_0(x) + \tfrac{1}{2}\mu_2(K)tr(HD^2\lambda_0(x)) + o(tr(H)),$$

$$E\left[K_H^2(x-X_1)\right] = \int K_H^2(x-u)\lambda_0(u)du = |H|^{-1/2}\lambda_0(x)R(K) + o(|H|^{-1/2}),$$

$$
\begin{aligned}
E\left[K_H(x-X_1)K_H(y-X-1)\right] &= \int K_H(x-u)K_H(y-u)\lambda_0(u)du \\
&= |H|^{-1/2}\lambda_0(x)(K\circ K)(H^{-1/2}(x-y)) + o(|H|^{-1/2}),
\end{aligned}
$$

$$
\begin{aligned}
E\left[K_H^2(x-X_1)K_H(y-X_1)\right] &= \int K_H^2(x-u)K_H(y-u)\lambda_0(u)du \\
&= |H|^{-1}\lambda_0(x)(K^2\circ K)(H^{-1/2}(x-y)) + o(|H|^{-1}),
\end{aligned}
$$

$$
\begin{aligned}
E\left[K_H^2(x-X_1)K_H^2(y-X_1)\right] &= \int K_H^2(x-u)K_H^2(y-u)\lambda_0(u)du \\
&= |H|^{-3/2}\lambda_0(x)(K^2\circ K^2)(H^{-1/2}(x-y)) + o(|H|^{-3/2}),
\end{aligned}
$$

$$
E\left[\hat{\lambda}_{0,H}(x)\right] = (1-e^{-m})E\left[K_H(x-X_1)\right] = \lambda_0(x) + \tfrac{1}{2}\mu_2(K)tr(HD^2\lambda_0(x)) + o(tr(H)),
$$

$$
\begin{aligned}
E\left[\hat{\lambda}_{0,H}^2(x)\right] &= A(m)E\left[K_H^2(x-X_1)\right] + (1-e^{-m}-A(m))E^2\left[K_H(x-X_1)\right] \\
&= A(m)|H|^{-1/2}\lambda_0(x)R(K) + \lambda_0^2(x) + \mu_2(K)\lambda_0(x)tr(HD^2\lambda_0(x)) \\
&+\tfrac{1}{4}\mu_2^2(K)tr^2(HD^2\lambda_0(x)) + A(m)\lambda_0^2(x) + A(m)\mu_2(K)\lambda_0(x)tr(HD^2\lambda_0(x)) \\
&+\tfrac{1}{4}A(m)\mu_2^2(K)tr^2(HD^2\lambda_0(x)) + 2\lambda_0(x)o(tr(H)) + R(K)o(A(m)|H|^{-1/2}) \\
&+\mu_2(K)tr(HD^2\lambda_0(x))o(tr(H)) + o(tr^2(H)) + o(A(m)),
\end{aligned}
$$

$$
\begin{aligned}
E\left[\hat{\lambda}_{0,H}(x)\hat{\lambda}_{0,H}(y)\right] &= A(m)E\left[K_H(x-X_1)K_H(y-X_1)\right] \\
&+ (1-e^{-m}-A(m))E\left[K_H(x-X_1)\right]E\left[K_H(y-X_1)\right] \\
&= A(m)|H|^{-1/2}\lambda_0(x)(K\circ K)(H^{-1/2}(x-y)) \\
&+ (K\circ K)(H^{-1/2}(x-y))o(A(m)|H|^{-1/2}) + \lambda_0(x)\lambda_0(y) \\
&+ \frac{1}{2}\lambda_0(x)\mu_2(K)tr(HD^2\lambda_0(y))\lambda_0(x)o(tr(H)) \\
&+ \frac{1}{2}\lambda_0(y)\mu_2(K)tr(HD^2\lambda_0(x)) + \frac{1}{4}\mu_2^2(K)tr(HD^2\lambda_0(x))tr(HD^2\lambda_0(y)) \\
&+ \frac{1}{2}\mu_2(K)tr(HD^2\lambda_0(x))o(tr(H)) + \lambda_0(y)o(tr(H)) \\
&+ \frac{1}{2}\mu_2(K)tr(HD^2\lambda_0(y))o(tr(H)) + A(m)\lambda_0(x)\lambda_0(y) \\
&+ \frac{1}{2}A(m)\lambda_0(x)\mu_2(K)tr(HD^2\lambda_0(y)) + A(m)\lambda_0(x)o(tr(H))
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{2}\lambda_0(y)\mu_2(K)tr(HD^2\lambda_0(x)) \\
&+ \frac{1}{4}A(m)\mu_2^2(K)tr(HD^2\lambda_0(x))tr(HD^2\lambda_0(y)) + o(A(m)tr(H)), \\
E\left[\hat{\lambda}^2_{0,H}(x)\hat{\lambda}_{0,H}(y)\right] &= B(m)E\left[K_H^2(x-X_1)K_H(y-X_1)\right] \\
&+ A(m)E\left[K_H^2(x-X_1)\right]E\left[K_H(y-X_1)\right] \\
&+ 2A(m)E\left[K_H(x-X_1)K_H(y-X_1)\right]E\left[K_H(x-X_1)\right] \\
&+ (1-e^{-m})E^2\left[K_H(x-X_1)\right]E\left[K_H(y-X_1)\right] \\
&= A(m)|H|^{-1/2}\lambda_0(x)\lambda_0(y)R(K) \\
&+ \frac{1}{2}A(m)|H|^{-1/2}\lambda_0(x)\mu_2(K)R(K)tr(HD^2\lambda_0(x)) \\
&+ \lambda_0(x)R(K)o(A(m)|H|^{-1/2}tr(H)) + \lambda_0(y)R(K)o(A(m)|H|^{-1/2}) \\
&+ \frac{1}{2}R(K)\mu_2(K)tr(HD^2\lambda_0(y))o(A(m)|H|^{-1/2}) \\
&+ R(K)o(A(m)|H|^{-1/2}tr(H)) \\
&+ 2A(m)|H|^{-1/2}\lambda_0^2(x)(K\circ K)(H^{-1/2}(x-y)) \\
&+ A(m)|H|^{-1/2}\mu_2(K)(K\circ K)(H^{-1/2}(x-y))\lambda_0(x)tr(HD^2\lambda_0(x)) \\
&+ 2(K\circ K)(H^{-1/2}(x-y))o(A(m)|H|^{-1/2}tr(H)) \\
&+ 2\lambda_0(x)(K\circ K)(H^{-1/2}(x-y))o(A(m)|H|^{-1/2}) \\
&+ \mu_2(K)(K\circ K)(H^{-1/2}(x-y))tr(HD^2\lambda_0(x))o(A(m)|H|^{-1/2}) \\
&+ o(A(m)|H|^{-1/2}tr(H)) \\
&+ \lambda_0^2(x)\lambda_0(y) + \frac{1}{2}\mu_2(K)\lambda_0^2(x)tr(HD^2\lambda_0(y)) \\
&+ \lambda_0^2(x)o(tr(H)) + \mu_2(K)\lambda_0(x)\lambda_0(y)tr(HD^2\lambda_0(x)) \\
&+ \frac{1}{2}\mu_2^2(K)\lambda_0(x)tr(HD^2\lambda_0(x))tr(HD^2\lambda_0(y)) \\
&+ \mu_2(K)\lambda_0(x)tr(HD^2\lambda_0(x))o(tr(H)) + 2\lambda_0(x)\lambda_0(y)p(tr(H)) \\
&+ \mu_2(K)\lambda_0(x)tr(HD^2\lambda_0(y))o(tr(H)) + 2\lambda_0(x)o(tr^2(H))
\end{aligned}
$$

$$+ \lambda_0(y)o(tr^2(H)) + \frac{1}{2}\mu_2(K)tr(HD^2\lambda_0(y))o(tr^2(H)) + o(tr^3(H)) \text{ and}$$

$$\begin{aligned}
E\left[\hat{\lambda}^2_{0,H}(x)\hat{\lambda}^2_{0,H}(y)\right] &= C(m)E\left[K^2_H(x-X_1)K^2_H(y-X_1)\right] \\
&+ 2B(m)E\left[K^2_H(x-X_1)K_H(y-X_1)\right]E\left[K_H(y-X_1)\right] \\
&+ 2B(m)E\left[K_H(x-X_1)K^2_H(y-X_1)\right]E\left[K_H(x-X_1)\right] \\
&+ B(m)E\left[K^2_H(x-X_1)\right]E\left[K^2_H(y-X_1)\right] \\
&+ 2B(m)E^2\left[K_H(x-X_1)K_H(y-X_1)\right] \\
&+ A(m)E\left[K^2_H(x-X_1)\right]E^2\left[K_H(y-X_1)\right] \\
&+ 4A(m)E\left[K_H(x-X_1)K_H(y-X_1)\right]E\left[K_H(x-X_1)\right]E\left[K_H(y-X_1)\right] \\
&+ A(m)E\left[K^2_H(y-X_1)\right]E^2\left[K_H(x-X_1)\right] \\
&+ (1-e^{-m})E^2\left[K_H(x-X_1)\right]E^2\left[K_H(y-X_1)\right] \\
&= A(m)|H|^{-1/2}\lambda_0(x)\lambda^2_0(y)R(K) \\
&+ A(m)|H|^{-1/2}R(K)\mu_2(K)\lambda_0(x)\lambda_0(y)tr(HD^2\lambda_0(y)) \\
&+ R(K)\lambda^2_0(y)o(A(m)|H|^{-1/2}) \\
&+ R(K)\mu_2(K)\lambda_0(y)tr(HD^2\lambda_0(y))o(A(m)|H|^{-1/2}) \\
&+ 4A(m)|H|^{-1/2}\lambda^2_0(x)\lambda_0(y)(K\circ K)(H^{-1/2}(x-y)) \\
&+ 2A(m)|H|^{-1/2}\lambda^2_0(x)\mu_2(K)(K\circ K)(H^{-1/2}(x-y))tr(HD^2\lambda_0(y)) \\
&+ 4A(m)\lambda^2_0(x)(K\circ K)(H^{-1/2}(x-y))o(|H|^{-1/2}tr(H)) \\
&+ 2A(m)|H|^{-1/2}\lambda_0(x)\lambda_0(y)\mu_2(K)(K\circ K)(H^{-1/2}(x-y))tr(HD^2\lambda_0(x)) \\
&+ 4\lambda_0(x)\lambda_0(y)(K\circ K)(H^{-1/2}(x-y))o(A(m)|H|^{-1/2}) \\
&+ o(A(m)|H|^{-1/2}tr(H)) + A(m)|H|^{-1/2}\lambda^2_0(x)\lambda_0(y)R(K) \\
&+ A(m)|H|^{-1/2}R(K)\mu_2(K)\lambda_0(x)\lambda_0(y)tr(HD^2\lambda_0(x)) \\
&+ R(K)\lambda^2_0(x)o(A(m)|H|^{-1/2}) \\
&+ R(K)\mu_2(K)\lambda_0(x)tr(HD^2\lambda_0(x))o(A(m)|H|^{-1/2})
\end{aligned}$$

$$
\begin{aligned}
&+ \lambda_0^2(x)\lambda_0^2(y) + \lambda_0^2(x)\lambda_0(y)\mu_2(K)tr(HD^2\lambda_0(y))2\lambda_0^2(x)\lambda_0(y)o(tr(H)) \\
&+ \frac{1}{4}\lambda_0^2(x)\mu_2^2(K)tr^2(HD^2\lambda_0(y)) + \lambda_0^2(x)\mu_2(K)tr(HD^2\lambda_0(y))o(tr(H)) \\
&+ \lambda_0^2(y)\lambda_0(x)\mu_2(K)tr(HD^2\lambda_0(x)) \\
&+ \mu_2^2(K)\lambda_0(x)\lambda_0(y)tr(HD^2\lambda_0(x))tr(HD^2\lambda_0(y)) \\
&+ 2\mu_2(K)\lambda_0(x)\lambda_0(y)tr(HD^2\lambda_0(x))o(tr(H)) \\
&+ 2\lambda_0(x)\lambda_0^2(y)o(tr(H)) + 2\mu_2(K)\lambda_0(x)\lambda_0(y)tr(HD^2\lambda_0(y))o(tr(H)) \\
&+ \frac{1}{4}\mu_2^2(K)\lambda_0^2(y)tr^2(HD^2\lambda_0(x)) + \lambda_0^2(y)\mu_2(K)tr(HD^2\lambda_0(x))o(tr(H)) \\
&+ o(A(m)) + o(tr^2(H)),
\end{aligned}
$$

with $B(m) = E\left[\frac{1}{N^2}1_{\{N\neq 0\}}\right]$ and $C(m) = E\left[\frac{1}{N^3}1_{\{N\neq 0\}}\right]$.

Second addend

We have also computed the mean and variance of the second addend using the same tools as in the first one, and also the relationship established in Theorem 3.1 and Theorem 3.2 in Chapter 3. We finally obtain that both, mean and variance, are negligible in comparison with the terms obtained for the first addend, in particular we have found that

$$E\left[\int_W \left(\lambda_0(x) - \hat{\rho}_{0,b}(Z(x))\right)^2 dx\right] = o(A(m))$$

and

$$\mathrm{Var}\left[\int_W \left(\lambda_0(x) - \hat{\rho}_{0,b}(Z(x))\right)^2 dx\right] = 0(A(m)),$$

which are both smaller than $o(A(m)|H|^{-1/2} + tr(H))$ corresponding to the first addend.

Third addend

$$
\begin{aligned}
&E\left[\int_W \left(\hat{\lambda}_{0,H}(x) - \lambda_0(x)\right)\left(\lambda_0(x) - \hat{\rho}_{0,b}(Z(x))\right) dx\right] = \tfrac{1}{2}\mu_2(K)\int \lambda_0(x)tr(HD^2\lambda_0(x))dx \\
&+o(tr(H))
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{Var}\left[\int_W \left(\hat{\lambda}_{0,H}(x) - \lambda_0(x)\right)\left(\lambda_0(x) - \hat{\rho}_{0,b}(Z(x))\right) dx\right] \\
&= A(m)|H|^{-1/2}\int\int \lambda_0^2(x)\lambda_0(y)(K\circ K)(H^{-1/2}(x-y))dxdy \\
&+ \int\int \lambda_0(x)\lambda_0(y)(K\circ K)(H^{-1/2}(x-y))o(A(m)|H|^{-1/2})dxdy + o(A(m)),
\end{aligned}
$$

where we have used that $E\left[\hat{\lambda}_{0,H}(x)\hat{\lambda}_{0,H}(y)\hat{\rho}_{0,b}(x)\hat{\rho}_{0,b}(y)\right]$, $E\left[\hat{\lambda}_{0,H}(x)\hat{\lambda}_{0,H}(y)\hat{\rho}_{0,b}(x)\right]$, $E\left[\hat{\lambda}_{0,H}(x)\hat{\rho}_{0,b}(x)\hat{\rho}_{0,b}(y)\right]$, $E\left[\hat{\lambda}_{0,H}(x)\hat{\rho}_{0,b}(x)\right]$, $E\left[\hat{\rho}_{0,b}(x)\hat{\rho}_{0,b}(y)\right]$ and $E\left[\hat{\rho}_{0,b}(x)\right]$ are smaller than the main term in the first addend's variance. And that the only terms that still contribute here to the global variance are: $E\left[\hat{\lambda}_{0,H}(x)\right]$ and $E\left[\hat{\lambda}_{0,H}(x)\hat{\lambda}_{0,H}(y)\right]$, which have already been calculated in a previous step of this proof.

Covariances

$$\begin{aligned}&\mathrm{Cov}\left[\int_W\left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2dx,\int_W\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2dx\right]\\&\quad=A(m)|H|^{-1/2}R(\lambda_0)R(K)+R(\lambda_0)R(K)o(A(m)|H|^{-1/2})+o(A(m)),\end{aligned}$$

$$\mathrm{Cov}\left[\int_W\left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)^2dx,\int_W\left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)dx\right]$$

and

$$\mathrm{Cov}\left[\int_W\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)^2dx,\int_W\left(\hat{\lambda}_{0,H}(x)-\lambda_0(x)\right)\left(\lambda_0(x)-\hat{\rho}_{0,b}(Z(x))\right)dx\right]$$

are smaller than the main variance term of the first addend.

Asymptotic normality

Our test statistic can be expanded and written as:

$$\begin{aligned}S\ &=\int_W\left(\hat{\lambda}_{0,H}(x)-\hat{\rho}_{0,b}(Z(x))\right)^2dx=\tfrac{1}{N^2}\textstyle\sum_{i=1}^N\int K_H(x-X_i)dx\\&+\tfrac{1}{N^2}\textstyle\sum_{i=1}^N\sum_{j\neq i}\int K_H(x-X_i)K_H(x-X_j)dx\\&+\tfrac{1}{N^2}\textstyle\sum_{i=1}^N\int\frac{1}{g^\star(Z(X_i))}L_b(Z(x)-Z(X_i))dx\\&+\tfrac{1}{N^2}\textstyle\sum_{i=1}^N\sum_{j\neq i}\int\frac{1}{g^\star(Z(X_i))}\frac{1}{g^\star(Z(X_j))}L_b(x-X_i)L_b(x-X_j)dx\\&-\tfrac{2}{N^2}\textstyle\sum_{i=1}^N\int\frac{1}{g^\star(Z(X_i))}K_H(x-X_i)L_b(Z(x)-Z(X_i))dx\\&-\tfrac{2}{N^2}\textstyle\sum_{i=1}^N\sum_{j\neq i}\int\frac{1}{g^\star(Z(X_j))}K_H(x-X_i)L_b(Z(x)-Z(X_j))dx\end{aligned}$$

where each of the addends is a U-statistic on a Poisson point process, remarking that the sums does not allow duplicated points in the same expression. Moreover, every of the addends is absolutely convergent in the sense defined by Reitzner and Schulte (2013), hence following its Theorem 4.7 we can assure the normality of each term. And therefore the normality of our test statistic, with the same mean and variance detailed

in the main body of Theorem 4.1.

□

In practice, the asymptotic distribution given in Theorem 4.1 can be approximated estimating m by the sample size, n, and $A(m)$ by $1/n$, as it has been extensively justified. However, this asymptotic distribution may not be the best way to calibrate our test. This requires some extra estimations and, as the convergence rate may be slow, it is not suitable for small patterns. Our proposal to deal with this inaccuracy is to use a bootstrap procedure for the calibration of the test.

We have chosen a smooth bootstrap procedure inspired in Cao (1993) and Cowling et al. (1996) to resample under the null hypothesis. Hence, let us assume the null hypothesis, i.e., $\hat{\lambda}(x) = \hat{\rho}_b(Z(x))$ with b a pilot bandwidth. And now, conditional on the pattern $X_1, \ldots, X_N$, let $N^* \sim Pois\left(\int_W \hat{\rho}_b(Z(x))dx\right)$. Generate n^* as a realisation of this random variable N^* and then draw $X_1^*, \ldots, X_{n^*}$ by sampling randomly n^* times from the distribution with density proportional to $\hat{\lambda}(x) = \hat{\rho}_b(Z(x))$. Then compute the test statistic under the bootstrap distribution using the expression in (4.2) applied to the bootstrap sample. Repeat this procedure B times and compute the quantile. Finally compare the value of the statistic using the initial data with the quantile to decide whether we reject or not the null hypothesis.

Following Cowling et al. (1996) we allow the bandwidths in the smooth bootstrap and in the test statistic to be different.

4.4 Data analyses

In this section, the testing procedure to check the intensity dependence on the covariate is illustrated in practice. As mentioned in Section 4.2, our real data examples come from two sources, a geological survey in the Murchison area of Western Australia and wildfire tracking by the national Wildland Fire Information System in Canada. The test is applied to decide if, in the first case, the distance to geological faults has an influence on the spatial distribution of gold deposits, and in the second one, if temperature can be the main cause of wildfires in Canada.

Our test statistic relies on two bandwidth parameters that need to be selected, the first one, b, in (4.2) and the second, t, in the bootstrap calibration. In Chapter 3 we present different data-driven bandwidth selection procedures defined for our estimator that could be, in principle, used for the test.

As an application, we have chosen the bootstrap bandwidth selector to compute the value of the statistic while, in order to thoughtfully validate our results, we have used different bandwidth values in a suitable range in the calibration of the test. This

appropriate range has been chosen after looking into the scale of the data and the covariate values.

Murchison gold deposits
Recall that here we have the gold deposit locations in an area of 330 × 394 kilometers and the distance to the nearest fault as covariate represented in Figure 4.1 (right). After having meticulously analysed the data and taking into account the bandwidth values and results obtained in the simulations presented later in Section 4.5, we have concluded that an interval around 0.6 is an appropriate range for the bandwidth parameter; if we consider smaller values we obtain an undersmoothed estimations, with lots of unreal features; and with larger values we are on the opposite end with clearly oversmoothed intensities. In Table 4.1 we show the p-values obtained for different bandwidth values in that range, using $B = 500$ replications for the bootstrap calibration and in Figure 4.4 we represent this variation of the p-value depending on the one-dimensional bandwidth.

	$b = 0.2$	$b = 0.3$	$b = 0.4$	$b = 0.5$	$b = 0.6$	$b = 0.7$	$b = 0.8$	$b = 0.9$	$b = 1$
p-value	0.940	0.938	0.894	0.804	0.702	0.602	0.456	0.358	0.236

Table 4.1: P-values of the test statistic for different bandwidths in the bootstrap calibration.

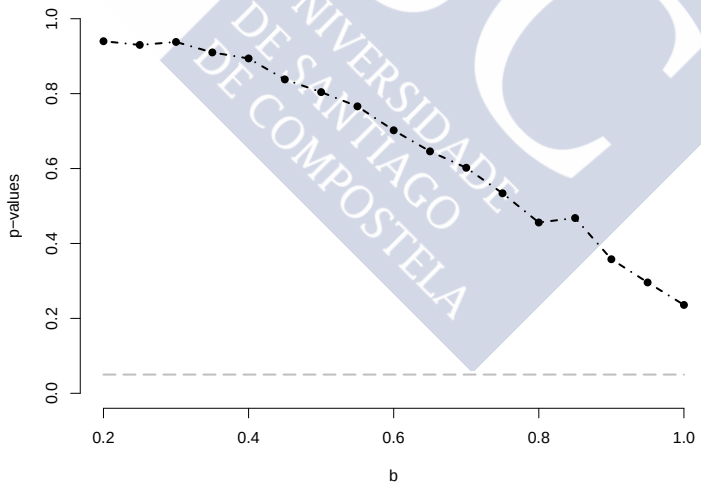


Figure 4.4: Representation of the p-values of the test statistic depending on the one-dimensional bandwidth parameter (black) and the level value $\alpha = 0.05$ (gray).

Regarding Table 4.1 we can accept the null hypothesis for this dataset not having any proofs against it with similar p-values for all the bandwidths. Hence, there is statistical evidence supporting that the geological faults have an essential role in the location of the gold deposits, and then, this information is enough to determine its spatial intensity in the form shown in (4.1).

Wildfires in Canada
Now, we want to test if the temperature plays a fundamental role in the intensity of the wildfires in Canada during June 2015. Again we use a data-driven bandwidth procedure to compute the value of the statistic and different bandwidth values in a suitable range for the bootstrap calibration, in this case around 0.4. The results were the same for all the possibilities, rejecting the null hypothesis with high evidence against it, indeed the p-value was always around zero, even for some trials of the bandwidth outside that appropriate range. Hence, we can conclude that, even if the temperature is likely to have an effect on the distribution of the wildfires in Canada, it is not enough to explain them alone. This suggests that maybe another meteorological covariates or some indexes (gathering several variables) should be used to analyse this process.

4.5 Simulated illustrative examples

This section is devoted to analyse the performance of our proposal through Monte Carlo simulations. The models we use are based on the real datasets previously presented in Section 4.2 and analysed in Section 4.4. To evaluate the power of the test we define multiplicative models (based on the initial ones) that depend on a parameter regulating the discrepancy from the null hypothesis. The intensity is defined as

$$\lambda(x) = \lambda_{\text{ini}}(x)r(x),$$

where λ_{ini} denotes the intensity of the initial model under the null and r is the function to perturb it depending on a parameter that we will explain in detail later in this section.

The first model is based on the Canadian dataset, so, our theoretical model is the intensity obtained after applying the kernel intensity estimator proposed in Chapter 3 to the wildfire dataset during June 2015 with the temperature (see Figure 4.3) as covariate. This intensity function is represented in Figure 4.5.

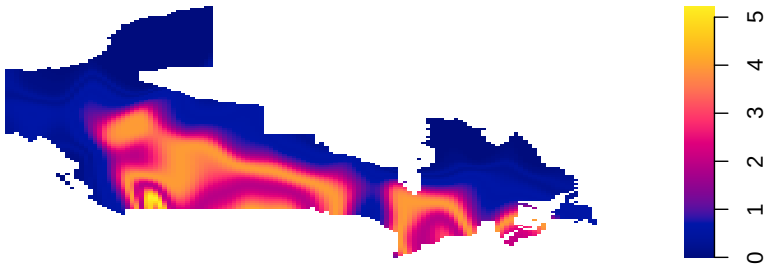


Figure 4.5: Theoretical intensity function for the first model analysed in the simulation study, that has been obtained applying a kernel intensity estimator to the Canadian wildfire dataset.

The second model has been constructed in a similar way but using the Murchison dataset. The intensity is represented in Figure 4.6 and the covariate used is the distance to the nearest geological fault, see Figure 4.1 (right).

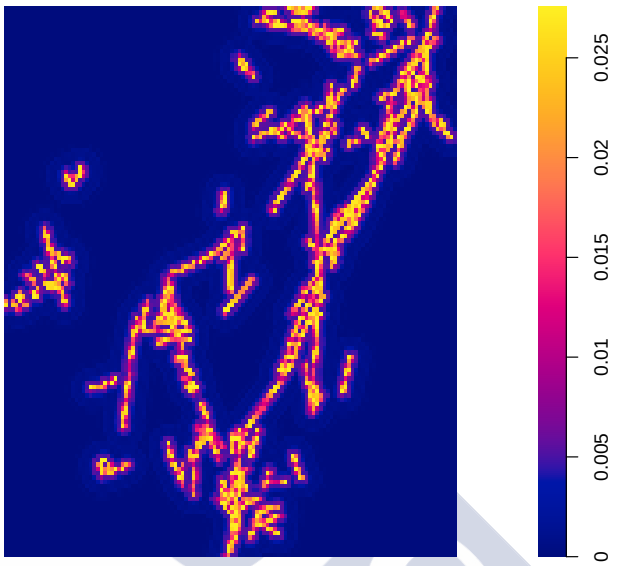


Figure 4.6: Theoretical intensity function for the second model analysed in the simulation study, that has been obtained applying a kernel intensity estimator to the Murchison dataset.

These two models lie under the null hypothesis, indeed their intensity function depends on a covariate through a univariate function, in both cases the one given by the expression of the nonparametric kernel intensity estimator developed in Chapter 3.

The function depending on a one-dimensional parameter that determines its discrepancy from the null hypothesis is a diagonal band that crosses the observation region nullifying the extension out of it (with smooth change). This band is wider or thinner depending on the parameter; when its wide enough it covers the whole observation region, so we approach the null hypothesis, and as it becomes thinner the model gets away of it.

In the two models we have considered two different discrepancy functions because the observation regions of each model are not the same. However, we preserve the same idea so, in both cases the diagonal band is based on a univariate Normal density, ϕ, depending on a parameter, d_C in the Canada dataset and d_M in the Murchison dataset. For the Canada dataset, $r_C(u,v) = \phi(u, 15 - v - v_{0C}, d_C)$, where $v_{0C} = 60.40$ is the middle point of the y-axis in the observation region, and d_C is the parameter we have been talking about which takes the values $6, 12, 20$ and 30. In Figure 4.7 we represent, for each value of the parameter d_C, both the r_C function (first row) and the final intensity function after multiplying the initial one by r_C (second row).

For the Murchison dataset, the function is defined as $r_M(u,v) = \phi(u - v_{0M}, u - v_{0M}, d_M)$, where $u_{0M} = 517.69$ and $v_{0M} = 6900.61$ are the middle point of the x and y-axis respectively, and d_M is the parameter that, in this case, may take values $10, 20, 40$ and 60. We have decided to rotate the diagonal band due to the distribution of the data and the covariate information, but it could have been done in the other way. Take also into account that due to this rotation, even a thin band collect a lot of information from the covariate which may cause smaller rejection proportions. In Figure 4.8 we represent the four r_M functions (first row) as well as the associated intensities (second row).

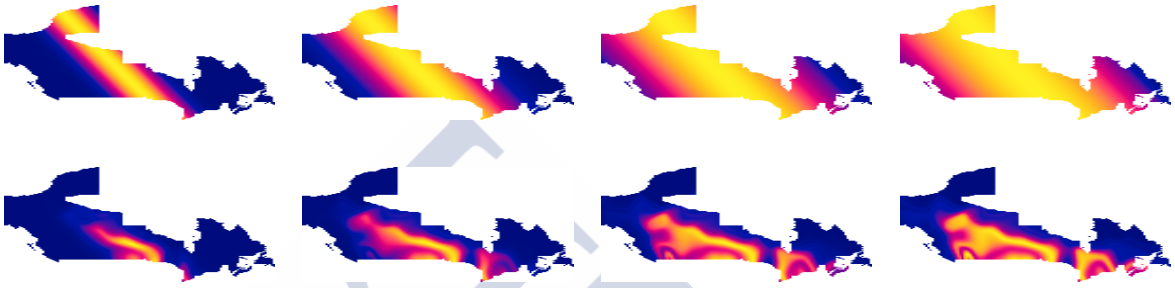


Figure 4.7: Representation of the r_C functions (first row) and the resulting intensity (second row) for the four values of the parameter $d_C = 6, 12, 20, 30$.

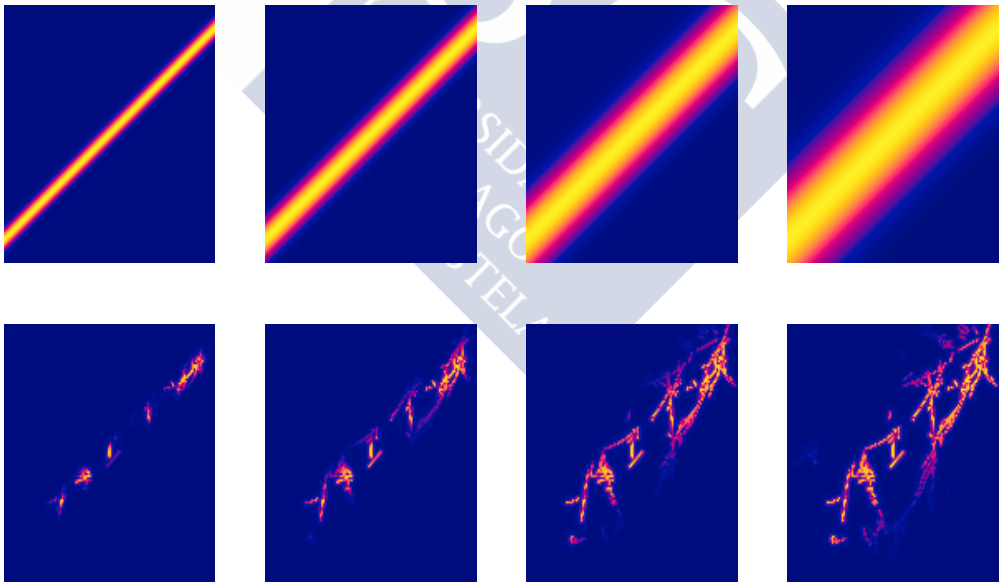


Figure 4.8: Representation of the r_M functions (first row) and the resulting intensity (second row) for the four values of the parameter $d_M = 10, 20, 40, 60$.

Remark that we have not included the scales in Figure 4.7 and 4.8 because the values of the intensity depend on the expected sample size so they will change in each of the situations considered in the study. In the tables below we denote by $d_C = \infty$

and $d_M = \infty$ the initial model without any band restriction on them, so the situation fulfilling the null hypothesis.

In Table 4.2 and Table 4.3 we show the rejection proportions for different situations that go, from the null hypothesis ($d_\bullet = \infty$) to the further situation away from it ($d_C = 6$ and $d_M = 10$). We can see that the power of the test seem to be better for the first model, where even for $d_C = 20$, which is a situation near to the null, the power values are high for medium and large sample sizes. In the second model, the values do not reach those levels. This may be due to the reason we have already pointed out that, if we keep in mind the spatial distribution of the gold deposits, even for thin bands we are gathering a lot of information from the covariate, and hence we are not really as far from the null hypothesis as we may think.

	$d_C = 6$	$d_C = 12$	$d_C = 20$	$d_C = 30$	$d_C = \infty$
$m = 50$	1	0.6852	0.1454	0.0722	0.0480
$m = 100$	1	0.9308	0.2232	0.0826	0.0502
$m = 200$	1	0.9986	0.4076	0.1060	0.0516
$m = 500$	1	1	0.8266	0.1744	0.0520

Table 4.2: Rejection proportions for the Canadian wildfire model, with different values of the parameter controlling the discrepancy from the null hypothesis, d_C, and four expected sample sizes, m.

	$d_M = 10$	$d_M = 20$	$d_M = 40$	$d_M = 60$	$d_M = \infty$
$m = 50$	0.9584	0.3858	0.1048	0.0636	0.0490
$m = 100$	0.9926	0.4774	0.1049	0.0680	0.0496
$m = 200$	0.9998	0.6404	0.1136	0.0610	0.0504
$m = 500$	1	0.9376	0.1727	0.0633	0.0492

Table 4.3: Rejection proportions for the Murchison model, with different values of the parameter controlling the discrepancy from the null hypothesis, d_M and four expected sample sizes, m.

4.6 Conclusions

In the context of spacial point process we have used nonparametric techniques to define a test statistic that allows to determine whether a given covariate is or not

decisive in the model. We have used the theoretical framework proposed in Chapter 3 to detail the asymptotic normality of the test, as well as a bootstrap method to improve its calibration. We have provided a couple of motivating examples to show the practical behaviour of our proposal, and we have also accomplished a simulation study based on these two real situations to better analyse its performance, that in general turn out to be satisfactory. In this sense we have also shown really competitive values in terms of level and power of our test.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

Nonparametric comparison of first-order intensities in point processes with covariates

Contents

Testing problems have been one of the main interests in statistical inference. In the point process context, this type of methodology is almost reduced to the complete spatial randomness testing, see Section 1.1.2, stationarity, isotropy, and in the spatio-temporal case separability testing. In Chapter 4, we have addressed the problem of goodness-of-fit of the model previously presented in Chapter 3, where the intensity is described as a function of a spatial covariate. Now, assuming this model, we want to solve the well-known two sample problem, that has already been addressed in other contexts such as multivariate density or the general spatial point process context without covariate information. Hence, we propose a L^2-distance based test statistic, we prove its asymptotic normality and suggest a bootstrap procedure to improve the calibration of the test. We also perform a simulation study with models based on real data to show the good behaviour of our proposals.

5.1 Introduction

Testing problems in point processes have mainly focused on some specific features. Initially, the main purpose of statistical testing in this field was complete spatial randomness (CSR), because rejecting CSR is a minimal prerequisite to any attempt of modelling. Different techniques were developeded to this aim such as neighbour distances and quadrat counts, see P. J. Diggle (2013) for more details. Also stationarity (the distribution of the point process is invariant under translations) and isotropy (the distribution of the point process is invariant under rotations) testing have been of interest in the last decades, see Cabana (1987), Rosenberg (2004) and Guan et al. (2006). Moreover, the separability testing problem has been addressed in the analysis of spatio-temporal point processes, see for example Schoenberg (2004). Furthermore, in the last decades, the problem of testing second-order characteristics has also been of interest.

Recall the model we have previously introduced in Chapter 3, where the intensity depends on a spatial covariate as follows:

$$\lambda(u) = \rho(Z(u)),\ u \in W \subset \mathbb{R}^2, \tag{5.1}$$

with $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ the spatial continuous covariate exactly known in every point of the region of interest W. In Chapter 3 a new kernel estimator was defined with a complete and detailed theoretical framework, bandwidth selection procedures as well as a bootstrap resampling method. The testing of this assumption (5.1), was addressed in Chapter 4, where we propose an L^2-statistic comparing two kernel intensity estimators: the general one without covariate information defined by P. Diggle (1985) and the one developed in Chapter 3.

One can assume the hypothesis of the covariate dependence as it is indicated in (5.1), then a natural subsequent step is the two sample problem. Without assuming the

covariate dependence model (5.1) this problem has been addressed by Fuentes-Santos et al. (2017), as an extension of the multivariate density proposed in Duong et al. (2012). Also, different recent studies have introduced area-based tests to measure discrepancies between two patterns, see Andresen (2009) and Alba-Fernández et al. (2016).

In this chapter we formulate the two sample comparison problem under model (5.1). This has some advantages with respect to the existing proposals and specially in comparison with our more natural competitor in Fuentes-Santos et al. (2017): first it is simpler due to the dimension reduction that influences not only the computational cost but also the complexity of the bandwidth selection problem, and second it has superior power when the assumed model is correct, because we are taking important extra information into account. Moreover, this model may explain some phenomena that without covariate information are harder to analyse from a statistical point of view.

5.2 The proposed method

Our aim in this work has been briefly introduced by the end of the previous section: test whether two given patterns are originated by the same process, assuming that the theoretical intensity depends on a known covariate in the way shown in (5.1). To check this hypothesis, we define an L^2-distance based test statistic, following similar ideas to those considered in Chapter 4.

5.2.1 The test

Let X_i with $i = 1, 2$ be two point processes defined in a region $W \subset \mathbb{R}^2$, where W is assumed to have finite positive area. Let $X_{11}, \ldots, X_{1N_1}$ and $X_{21}, \ldots, X_{2N_2}$ be two realisations of the processes where N_i are the random variables counting the number of events. Let again $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ be the spatial continuous covariate that is exactly known in every point of the region of interest W. Recall that in practice this covariate will commonly be known in an enough amount of points spread over the region, so the values for the rest of the points can be interpolated and it can be assumed that these values are indeed the real ones. We denote by $Z_{11}, \ldots, Z_{1N_1}$ and $Z_{21}, \ldots, Z_{2N_2}$ the realisations transformed through the covariate, i.e., $Z_{ij} = Z(X_{ij})$ and Z_i will be used to denote the processes $Z(X_i)$.

Under model (5.1) let denote by $\lambda_i(x) = \rho_i(Z(x))$ the intensity functions corresponding to the processes X_i with $i = 1, 2$. We want to test the null hypothesis $H_0 : \lambda_1(x) = \lambda_2(x)$, $x \in W$ versus the two-sided alternative. Following again what it has been done in Cucala (2006), Borrajo et al. (2017b) and Fuentes-Santos et al.

(2017), we use the density of events location to define an equivalent null hypothesis. Hence, let $f_i(z) = \frac{g^\star(z)\rho_i(z)}{m_i}$, where $g^\star$ is the unnormalised version of the density derived from the spatial cumulative distribution function $G(z) = \frac{1}{|W|}\int_W 1_{\{Z(u)\leq z\}}du$, and $m_i = \int_W \lambda_i(x)dx$ is the expected number of events in each of the processes. Then, our final null hypothesis is $H_0 : f_1(z) = f_2(z), z \in \mathbb{R}$; remark that this does not really need to be in $\mathbb{R}$ but in a subset of it covering the range of values of the covariate Z.

To define an appropriate test, we need first of all to choose a discrepancy measure, i.e., a distance between the two theoretical densities used to define our statistic. We have chosen an L^2-distance, that can be written as follows:

$$D = \int_{\mathbb{R}} (f_1(z) - f_2(z))^2 dz = \int f_1^2(z)dz + \int f_2^2(z)dz - \int f_1(z)f_2(z)dz - \int f_2(z)f_1(z)dz$$
$$= E_{Z_1}[f_1(Z_1)] + E_{Z_2}[f_2(Z_2)] - E_{Z_2}[f_1(Z_2)] - E_{Z_1}[f_2(Z_1)] \equiv \psi_{11} + \psi_{22} - \psi_{12} - \psi_{21}. \tag{5.2}$$

Now, we can define our test statistic, S, as

$$S = \widehat{\psi}_{11} + \widehat{\psi}_{22} - \widehat{\psi}_{12} - \widehat{\psi}_{21}, \tag{5.3}$$

where taking into account the estimator defined in Borrajo et al. (2017b), we can write

$$\widehat{\psi}_{11} = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i} - Z_{1j})1_{\{N_1\neq 0\}},$$
$$\widehat{\psi}_{22} = \frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\frac{g^\star(Z_{2i})}{g^\star(Z_{2j})}K_{h_2}(Z_{2i} - Z_{2j})1_{\{N_1\neq 0\}},$$
$$\widehat{\psi}_{12} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(Z_{2j})}{g^\star(Z_{1i})}K_{h_1}(Z_{2j} - Z_{1i})1_{\{N_1\neq 0,N_2\neq 0\}},$$
$$\widehat{\psi}_{21} = \frac{1}{N_1N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(Z_{1i})}{g^\star(Z_{2j})}K_{h_2}(Z_{1i} - Z_{2j})1_{\{N_1\neq 0,N_2\neq 0\}},$$

with h_i scalar bandwidths, K a univariate kernel function, and $1_{\{.\}}$ denoting the indicator function.

5.2.2 Asymptotic properties and calibration

Hereafter we will assume that $W = \mathbb{R}^2$ to avoid the edge effects in the theoretical developments. We need to introduce some regularity conditions:

(D.1) $\int_{\mathbb{R}} K(z)dz = 1$; $\int_{\mathbb{R}} zK(z)dz = 0$ and $\mu_2(K) := \int_{\mathbb{R}} z^2L(z)dz < \infty$,

(D.2) $\lim_{m\to\infty} h_i = 0$ and $\lim_{m\to\infty} \frac{A(m_i)}{h_i} = 0$, where $A(m_i) := E\left[\frac{1}{N_i} 1_{\{N_i \neq 0\}}\right]$,

(D.3) G and the densities of events location are three times differentiable and

(D.4) $Z(x)$ is a continuity point of ρ_i for all $x \in W$.

The following theorem establishes the asymptotic null distribution of S in our framework of spatial point processes with covariate information.

Theorem 5.1. *Under conditions (D.1) to (D.4) and assuming the null hypothesis* $H_0 : f_1(z) = f_2(z)\ \forall z \in \mathbb{R}$, *it holds*

$$\frac{S - \mu_S}{\sigma_S} \longrightarrow N(0,1),$$

where

$$\mu_S = (A(m_1)h_1 + A(m_2)h_2)\, K(0) + o(A(m_1)) + o(A(m_2)) \text{ and}$$

$$\begin{aligned}\sigma_S^2 = {} & 2B(m_1)\frac{1}{h_1}R(K)\psi + 2B(m_2)\frac{1}{h_2}R(K)\psi + A(m_1)A(m_2)\psi R(K)\left(\frac{1}{h_1} + \frac{1}{h_2}\right) \\ & + A(m_1)A(m_2)\psi\left(\frac{1}{h_1}\int K(u)K_{h_2/h_1}(u)du + \frac{1}{h_2}\int K(u)K_{h_1/h_2}(u)du\right) \\ & + O(B(m_1) + O(B(m_2))),\end{aligned}$$

with $\psi \equiv \psi_{11} \equiv \psi_{22} \equiv \psi_{12} \equiv \psi_{21}$ *under the null,* $K_l(\cdot) = \frac{1}{l}K\left(\frac{\cdot}{l}\right)$, $R(K) := \int K^2(u)du$ *and* $B(m_i) = E\left[\frac{1}{N_i^2} 1_{\{N_i \neq 0\}}\right]$.

– Proof –

Along this proof we will obtain the mean and variance of the statistic S as well as assure its asymptotic normality. It is easy to see, using common properties of the mean and variance operators, that

$$E[S] = E\left[\widehat{\psi}_{11}\right] + E\left[\widehat{\psi}_{22}\right] - E\left[\widehat{\psi}_{12}\right] - E\left[\widehat{\psi}_{21}\right]$$

and

$$\begin{aligned}\mathrm{Var}[S] = {} & \mathrm{Var}\left[\widehat{\psi}_{11}\right] + \mathrm{Var}\left[\widehat{\psi}_{22}\right] + \mathrm{Var}\left[\widehat{\psi}_{12}\right] + \mathrm{Var}\left[\widehat{\psi}_{21}\right] + 2\mathrm{Cov}\left[\widehat{\psi}_{11}, \widehat{\psi}_{12}\right] \\ & + 2\mathrm{Cov}\left[\widehat{\psi}_{11}, \widehat{\psi}_{21}\right] + 2\mathrm{Cov}\left[\widehat{\psi}_{22}, \widehat{\psi}_{12}\right] + 2\mathrm{Cov}\left[\widehat{\psi}_{22}, \widehat{\psi}_{21}\right] - 2\mathrm{Cov}\left[\widehat{\psi}_{12}, \widehat{\psi}_{21}\right].\end{aligned}$$

Remark that we haven't considered the covariance between $\widehat{\psi}_{11}$ and $\widehat{\psi}_{22}$ because it is zero due to the independence between the processes X_1 and X_2.

Mean of S

$$E\left[\widehat{\psi}_{11}\right] = E\left[E\left[\widehat{\psi}_{11}|N_1\right]\right] = E\left[E\left[\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i}-Z_{1j})1_{\{N_1\neq 0\}}|N_1\right]\right]$$
$$= \sum_{l=1}^{\infty}\frac{\mathbb{P}(N_1=l)}{l^2}E\left[\sum_{i=1}^{l}K_{h_1}(0)+\sum_{i=1}^{l}\sum_{j\neq i}\frac{g^\star(Z_{1i})}{g^\star(Z_1 j)}K_{h_1}(Z_{1i}-Z_{1j})\right]$$
$$= A(m_1)K_{h_1}(0)+\left(1-e^{-m_1}-A(m_1)\right)\left(\psi_{11}+\frac{1}{2}h_1^2\mu_2(K)\int\frac{f_1}{g^\star}(y)(g^\star f_1)''dy\right.$$
$$\left.+\int\frac{f_1}{g^\star}(y)dy\, o(h_1^2)\right), \tag{5.4}$$

where we have used a second order Taylor expansion to obtain

$$E\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right] = \int\int\frac{g^\star(x)}{g^\star(y)}K_{h_1}(x-y)f_1(x)f_1(y)dxdy$$
$$= \int\frac{f_1}{g^\star}(y)\int K(u)(g^\star f_1)(y-h_1u)dudy = \psi_{11}+\frac{1}{2}h_1^2\mu_2(K)\int\frac{f_1}{g^\star}(y)(g^\star f_1)''dy$$
$$+\int\frac{f_1}{g^\star}(y)dy\, o(h_1^2).$$

Similarly, we obtain

$$E\left[\widehat{\psi}_{22}\right] = A(m_2)K_{h_2}(0)+\left(1-e^{-m_2}-A(m_2)\right)\left(\psi_{22}+\frac{1}{2}h_2^2\mu_2(K)\int\frac{f_2}{g^\star}(y)(g^\star f_2)''dy\right.$$
$$\left.+\int\frac{f_2}{g^\star}(y)dy\, o(h_2^2)\right). \tag{5.5}$$

Now, we compute the expectations of the other two terms:

$$E\left[\widehat{\psi}_{12}\right] = E\left[E\left[E\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] = \sum_{l=1}^{\infty}\sum_{k=1}^{\infty}E\left[\widehat{\psi}_{12}|N_1,N_2\right]\mathbb{P}(N_1=l)\mathbb{P}(N_2=k)$$
$$= (1-e^{-m_1})(1-e^{-m_2})\left(\psi_{12}+\frac{1}{2}\frac{h_1^2}{m_1}\mu_2(K)\int(g^\star f_2)(y)\rho_1''(y)dy\right.$$
$$\left.+\int(g^\star f_2)(y)dy\, o\left(\frac{h_1^2}{m_1}\right)\right), \tag{5.6}$$

where we have taken into account that

$$E\left[\frac{g^\star(Z_{21})}{g^\star(Z_{12})}K_{h_1}(Z_{21}-Z_{12})\right] = \int\int\frac{g^\star(y)}{g^\star(x)}K_{h_1}(y-x)f_1(x)f_2(y)dxdy$$
$$= \int\frac{f_2}{g^\star}(y)\int K(u)\left(\frac{f_1}{g^\star}\right)(y-h_1u)dudy = \psi_{12}+\frac{1}{2}\frac{h_1^2}{m_1}\mu_2(K)\int(g^\star f_2)(y)\rho_1''(y)dy$$
$$+\int(g^\star f_2)(y)dy\, o\left(\frac{h_1^2}{m_1}\right).$$

And again, similarly, we get

$$E\left[\widehat{\psi}_{21}\right] = (1-e^{-m_1})(1-e^{-m_2})\left(\psi_{21} + \frac{1}{2}\frac{h_2^2}{m_2}\mu_2(K)\int (g^\star f_1)(y)\rho_2''(y)dy + \int (g^\star f_1)(y)dy\, o\left(\frac{h_2^2}{m_2}\right)\right). \tag{5.7}$$

Hence, gathering equations (5.4) to (5.7), and taking into account that we are under the null hypothesis, i.e., $f_1 = f_2 := f$ and $\psi_{11} = \psi_{22} = \psi_{12} = \psi_{21} := \psi$, we have that

$$E[S] = (A(m_1)h_1 + A(m_2)h_2)\,K(0) + o(A(m_1)) + o(A(m_2)).$$

Variance of S

We need to compute the variances of every addend in the statistic as well as the covariances detailed at the beginning of this proof. Let us start with the variances:

$$\begin{aligned}\mathrm{Var}\left[\widehat{\psi}_{11}\right] &= E\left[\mathrm{Var}\left[\widehat{\psi}_{11}|N_1\right]\right] + \mathrm{Var}\left[E\left[\widehat{\psi}_{11}|N_1\right]\right] = 4A(m_1)\left(\int f_1^3(x)dx - \psi_{11}^2\right) \\ &+ 2B(m_1)\frac{1}{h_1}R(K)\psi_{11} + O(A(m_1)h_1^2) + O(B(m_1)).\end{aligned} \tag{5.8}$$

To obtain this result we use the following calculation in the first addend of the variance decomposition

$$\begin{aligned}&E\left[\mathrm{Var}\left[\widehat{\psi}_{11}|N_1\right]\right] \\ &= \sum_{l=1}^{\infty}\mathrm{Cov}\left[\frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i}-Z_{1j}), \frac{1}{N_1^2}\sum_{s=1}^{N_1}\sum_{t=1}^{N_1}\frac{g^\star(Z_{1s})}{g^\star(Z_{1t})}K_{h_1}(Z_{1s}-Z_{1t})\right] \\ &= B(m_1)\mathrm{Var}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right] \\ &+ B(m_1)\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{12})}{g^\star(Z_{11})}K_{h_1}(Z_{12}-Z_{11})\right] \\ &+ A(m_1)\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{11})}{g^\star(Z_{13})}K_{h_1}(Z_{11}-Z_{13})\right] \\ &+ A(m_1)\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{13})}{g^\star(Z_{11})}K_{h_1}(Z_{13}-Z_{11})\right] \\ &+ A(m_1)\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{12})}{g^\star(Z_{13})}K_{h_1}(Z_{12}-Z_{13})\right] \\ &+ A(m_1)\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{13})}{g^\star(Z_{12})}K_{h_1}(Z_{13}-Z_{12})\right] \\ &= 4A(m_1)\left(\int f_1^3(x)dx - \psi_{11}^2\right) + O(A(m_1)h_1^2) + O(B(m_1)h_1),\end{aligned}$$

where after some second order Taylor expansions and reducing the negligible terms, we have

$$\text{Var}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right] = \frac{1}{h_1}R(K)\psi_{11} - \psi_{11}^2 - \mu_1(K^2)\int \frac{f_1}{g^{\star^2}}(g^{\star^2}f_1)'(x)dx + O(h_1^2),$$

$$\text{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{12})}{g^\star(Z_{11})}K_{h_1}(Z_{12}-Z_{11})\right] = \frac{R(K)}{h_1}\psi_{11} - \mu_1(K^2)\int \frac{f_1}{g^{\star^2}}(g^{\star^2}f_1)' - \psi_{11}^2 + O(h_1),$$

$$\text{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{11})}{g^\star(Z_{13})}K_{h_1}(Z_{11}-Z_{13})\right] = \int f_1^3(x)dx - \psi_{11}^2 + O(h_1),$$

$$\text{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{13})}{g^\star(Z_{11})}K_{h_1}(Z_{13}-Z_{11})\right] = \int f_1^3(x)dx - \psi_{11}^2 + O(h_1),$$

$$\text{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{12})}{g^\star(Z_{13})}K_{h_1}(Z_{12}-Z_{13})\right] = \int f_1^3(x)dx - \psi_{11}^2 + O(h_1) \text{ and}$$

$$\text{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}), \frac{g^\star(Z_{13})}{g^\star(Z_{12})}K_{h_1}(Z_{13}-Z_{12})\right] = \int f_1^3(x)dx - \psi_{11}^2 + O(h_1).$$

The second addend of the variance decomposition is computed as follows:

$$\begin{aligned}
&\text{Var}\left[E\left[\widehat{\psi}_{11}|N_1\right]\right] = \text{Var}\left[\frac{1}{N_1}K_{h_1}(0)1_{\{N_1\neq 0\}} + \frac{N_1-1}{N_1}1_{\{N_1\neq 0\}}E\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12})\right]\right] \\
&= K_{h_1}^2(0)\text{Var}\left[\frac{1}{N_1}1_{\{N_1\neq 0\}}\right] + E^2\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})K_{h_1}(Z_{11}-Z_{12})}\right]\text{Var}\left[\frac{N_1-1}{N_1}1_{\{N_1\neq 0\}}\right] \\
&+ 2K_{h_1}(0)E\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})K_{h_1}(Z_{11}-Z_{12})}\right]\text{Cov}\left[\frac{1}{N_1}1_{\{N_1\neq 0\}}, \frac{N_1-1}{N_1}1_{\{N_1\neq 0\}}\right] \\
&= K_{h_1}^2(0)\left(B(m_1) - A^2(m_1)\right) + \left(\psi_{11}^2 + O(h_1)\right)\left(e^{-m_1} + B(m_1) + A^2(m_1) - 2e^{-m_1}A(m_1)\right).
\end{aligned}$$

And gathering both of them we obtain the expression in (5.8). Using the same developments we get

$$\begin{aligned}
\text{Var}\left[\widehat{\psi}_{22}\right] &= 4A(m_2)\left(\int f_2^3(x)dx - \psi_{22}^2\right) + 2B(m_2)\frac{1}{h_2}R(K)\psi_{22} + O(A(m_2)h_2^2) \\
&+ O(B(m_2)). \qquad (5.9)
\end{aligned}$$

Now, we will compute the variance of $\widehat{\psi}_{12}$,

$$\begin{aligned}\mathrm{Var}\left[\widehat{\psi}_{12}\right] &= E_{N_2}\left[E_{N_1}\left[\mathrm{Var}\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] + E_{N_2}\left[\mathrm{Var}_{N_1}\left[E\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] \\ &\quad + \mathrm{Var}_{N_2}\left[E_{N_1}\left[E\left[\widehat{\psi}_{12}|N_1,N_2\right]\right]\right] \\ &= A(m_1)A(m_2)\left(\frac{1}{h_1}\psi_{12}R(K) - \mu_2(K^2)\int\frac{(g^{\star^2}f_2)'(x)f_1(x)}{g^{\star^2}}dx\right. \\ &\quad \left. - \psi_{12}^2 + O(h_1)\right) + A(m_1)(1-e^{-m_2})\left(\int f_1(x)f_2^2(x)dx - \psi_{12}^2 + O(h_1^2)\right) \\ &\quad + A(m_2)(1-e^{-m_1})\left(\int f_1^2(x)f_2(x)dx - \psi_{12}^2 + O(h_1^2)\right). \end{aligned} \tag{5.10}$$

To reach this final expression we expand separately each of the three addends, but only the first one is not null. We have used the development of the conditional variance in terms of covariances, as well as the independence between N_1 and N_2.

Similarly,

$$\begin{aligned}\mathrm{Var}\left[\widehat{\psi}_{21}\right] &= A(m_1)A(m_2)\left(\frac{1}{h_2}\psi_{21}R(K) - \mu_2(K^2)\int\frac{(g^{\star^2}f_1)'(x)f_2(x)}{g^{\star^2}}dx\right. \\ &\quad \left. - \psi_{21}^2 + O(h_2)\right) + A(m_2)(1-e^{-m_1})\left(\int f_1^2(x)f_2(x)dx - \psi_{21}^2 + O(h_2^2)\right) \\ &\quad + A(m_1)(1-e^{-m_2})\left(\int f_1(x)f_2^2(x)dx - \psi_{21}^2 + O(h_2^2)\right). \end{aligned} \tag{5.11}$$

We have to obtain finally the expressions of the covariances:

$$\begin{aligned}\mathrm{Cov}\left[\widehat{\psi}_{11},\widehat{\psi}_{12}\right] &= E_{N_1}\left[E_{N_2}\left[\mathrm{Cov}\left[\widehat{\psi}_{11},\widehat{\psi}_{12}|N_1,\ N_2\right]\right]\right] \\ &\quad + E_{N_1}\left[\mathrm{Cov}_{N_2}\left[E\left[\widehat{\psi}_{11}|N_1,\ N_2\right],E\left[\widehat{\psi}_{12}|N_1,\ N_2\right]\right]\right] \\ &\quad + \mathrm{Cov}_{N_1}\left[E_{N_2}\left[E\left[\widehat{\psi}_{11}|N_1,\ N_2\right]\right],E\left[E\left[\widehat{\psi}_{12}|N_1,\ N_2\right]\right]\right] \\ &= 2A(m_1)\left(\int f_1^2(x)f_2(x)dx - \psi_{11}\psi_{12} + O(h_1^2)\right), \end{aligned} \tag{5.12}$$

where we use that the two last addends are negligible and the first one is obtained as:

$$\begin{aligned}
&\mathrm{Cov}\left[\widehat{\psi}_{11},\widehat{\psi}_{12}|N_1,\ N_2\right]\\
&=\frac{1}{N_1^3N_2}\sum_{i,j,k=1}^{N_1}\sum_{l=1}^{N_2}\mathrm{Cov}\left[\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i}-Z_{1j}),\frac{g^\star(Z_{1k})}{g^\star(Z_{2l})}K_{h_1}(Z_{1k}-Z_{2l})\right]\\
&=\frac{N_1-1}{N_1^2}\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{11})}{g^\star(Z_{21})}K_{h_1}(Z_{11}-Z_{21})\right]\\
&+\frac{N_1-1}{N_1^2}\mathrm{Cov}\left[\frac{g^\star(Z_{11})}{g^\star(Z_{12})}K_{h_1}(Z_{11}-Z_{12}),\frac{g^\star(Z_{12})}{g^\star(Z_{21})}K_{h_1}(Z_{12}-Z_{21})\right],
\end{aligned}$$

where after applying expectations, we easily get (5.12).

Following the same steps we obtain

$$\mathrm{Cov}\left[\widehat{\psi}_{11},\widehat{\psi}_{21}\right]=2A(m_1)\left(\int f_1^2(x)f_2(x)dx-\psi_{11}\psi_{21}+O(h_1^2)+O(h_2^2)\right), \tag{5.13}$$

$$\mathrm{Cov}\left[\widehat{\psi}_{22},\widehat{\psi}_{12}\right]=2A(m_2)\left(\int f_1(x)f_2^2(x)dx-\psi_{22}\psi_{12}+O(h_1^2)+O(h_2^2)\right)\ \text{and} \tag{5.14}$$

$$\mathrm{Cov}\left[\widehat{\psi}_{22},\widehat{\psi}_{21}\right]=2A(m_2)\left(\int f_1(x)f_2^2(x)dx-\psi_{22}\psi_{21}+O(h_2^2)\right). \tag{5.15}$$

The only term left is the covariance involving $\widehat{\psi}_{12}$ and $\widehat{\psi}_{21}$. The computation of this one is closely to the ones above but with more non-zero terms in the expression of the conditional covariance. Applying the same mathematical tools we get:

$$\begin{aligned}
\mathrm{Cov}\left[\widehat{\psi}_{12},\widehat{\psi}_{21}\right]&=A(m_1)A(m_2)\left(\frac{\psi_{12}}{2h_1}\int K(u)K_{h_2/h_1}(u)du\right)\\
&+A(m_1)\left(\int f_1(x)f_2^2(x)dx-\psi_{12}\psi_{21}\right)+A(m_2)\left(\int f_1^2(x)f_2(x)dx-\psi_{12}\psi_{21}\right)\\
&+O(A(m_1)h_1^2)+O(A(m_1)h_2^2)+O(A(m_2)h_1^2)+O(A(m_2)h_2^2).
\end{aligned} \tag{5.16}$$

Finally, gathering (5.8) to (5.16), taking into account that following Borrajo et al. (2017c) the order of the optimal bandwidths h_j is $A(m_j)^{1/5}$ and that we are under the null hypothesis, i.e., $f_1=f_2:=f$ and $\psi_{11}=\psi_{12}=\psi_{21}=\psi_{22}:=\psi$, we get the final

result of

$$\begin{aligned}\mathrm{Var}\left[S\right] = &\left(B(m_1)\frac{1}{h_1} + B(m_2)\frac{1}{h_2}\right) 2R(K)\psi + A(m_1)A(m_2)\psi R(K)\left(\frac{1}{h_1}+\frac{1}{h_2}\right)\\ &+ A(m_1)A(m_2)\psi\left(\frac{1}{h_1}\int K(u)K_{h_2/h_1}(u)du + \frac{1}{h_2}\int K(u)K_{h_1/h_2}(u)du\right)\\ &+ O(B(m_1)) + O(B(m_2)) + O(A(m_1)A(m_2)).\end{aligned}$$

Asymptotic normality

Our test statistic can be expanded and written as sums of non-duplicated points, where each of the addends is a U-statistic on a Poisson point process. Moreover, every of the addends is absolutely convergent in the sense defined by Reitzner and Schulte (2013), hence following their Theorem 4.7 we can assure the normality of each term. Then the normality of our test statistic is assured with the mean and variance detailed in the main body of Theorem 5.1.

□

To use the asymptotic distribution given in Theorem 5.1 we need to estimate some quantities to obtain μ_S and σ_S^2: m_i which would be replaced by the sample size n_i and $A(m_i)$ by $1/n_i$ as it has been extensively justified.

However, this asymptotic distribution may not be the best way to calibrate our test, because it requires some extra estimations and, as the convergence rate may be slow, it is not suitable for small patterns. Our proposal to deal with this inaccuracy is to use a bootstrap procedure to improve the calibration of the test.

The bootstrap we propose is based on the one defined in Chapter 3, which was based on Cao (1993) and Cowling et al. (1996). To resample under the null hypothesis we can assume that $X_{11},\ldots,X_{1N_1}$ and $X_{21},\ldots,X_{2N_2}$ are a unique sample coming from the same process, and then, conditional on this joint sample, $X_1,\ldots,X_{N_1+N_2}$, let $N^* \sim Pois\left(\int_W \hat{\rho}_t(Z(x))dx\right)$, where $\hat{\rho}_t = \sum_{i=1}^{N_1+N_2}\frac{1}{g^\star(Z(X_i))}K_t\left(Z(x)-Z(X_i)\right)$ and it has been defined in Chapter 3. We generate two realisations, n_1^* and n_2^*, of this random variable and then we draw $X_{11}^*,\ldots,X_{1n_1^*}^*$ and $X_{21}^*,\ldots,X_{2n_2^*}^*$ by sampling randomly in two steps from the distribution with density proportional to $\lambda(x) = \hat{\rho}_t(Z(x))$, i.e, $\frac{\hat{\rho}_t(Z(x))}{\int_W \hat{\rho}_t(Z(x))dx}$.

5.3 Simulations

This section is devoted to analyse the performance of our proposal through Monte Carlo simulations. We simulate the same models as in Chapter 4, based on real datasets in order to obtain more realistic results. Recall the first one is the well known Murchison

dataset, consisting of 255 locations of gold deposits and the surrounding geological faults in a region of $330 \times 394\text{km}^2$ located in the Murchison area of Western Australia. At this scale (1:500000) the gold deposits spatial extension is negligible and they can be considered as points without losing generality. Note that the real gold deposits and faults are three-dimensional while here we use a two-dimensional projection. Moreover, some geological faults may have been missed because they are not recorded by direct observation but in magnetic field surveys or geologically inferred from discontinuities in the rock sequences. A representation of the data can be seen in Figure 4.1.

All along this chapter we are assuming model (5.1), initially presented in Chapter 3, so we need to define an appropriate covariate for our intensity. In this case, and as we have done in Chapter 4, it is not complicated, once we have the locations of the faults, we have to compute the distance from every point in the observation region to the nearest fault, see Figure 4.1 (right).

The second model is related to one of the most important natural disturbances, wildfires; in this occasion we are using again the wildfire records in Canada during June 2015 (see Figure 4.2), that are available at the Canadian Wildland Fire Information System website (`http://cwfis.cfs.nrcan.gc.ca/home`). Recall that the fire season in Canada lasts from late April until August, with a peak of activity in June and July, and based on the existing literature, the main cause of this fires relies on meteorological issues. Hence, we have obtain meteorological data of Canada, in particular, we obtained diary temperature data for the whole month, and we have constructed a covariate for the model that is the third quartile of the maximum temperature (in order to avoid extreme values that may interfere in the analysis), see Figure 4.3.

Once we have presented the data and the respective covariates, we need to build the intensity function to define our models. We are going to simply use again the nonparametric kernel intensity estimator presented in Chapter 3, with the bootstrap bandwidth defined in the same chapter, see also Borrajo et al. (2017b). The intensity function for both models can be seen in Figure 4.6 and Figure 4.5, respectively.

This simulation study is divided in two parts, the first one devoted to analyse the level of the test, and the second to study its power. In the first part, we have run $M = 10000$ Monte Carlo samples to better estimate the low quantile $\alpha = 0.05$, while in the second we have found that with $M = 5000$ is enough. We have considered four different expected sample sizes, $m = 50, 100, 200$ and 500.

Note that there is no bandwidth selector specifically designed for this test, so we have two possibilities to compute it: on the one hand we can use one of the selectors proposed in Chapter 3 for the intensity estimation, on the other hand we can study a suitable range of possible values for the bandwidth parameter. We selected the procedure of choosing the bandwidth based on the level approximation, hence for the Murchison dataset, the automatic data-driven bandwidth selectors from Chapter

3 seems to produce accurate values, while for the Canadian data we have used a grid of bandwidths, t, in the range $(0.005, 0.3)$ to be able to adjust the level. The results obtained for the empirical level of the test in each model can be seen in Table 5.1 and Table 5.2.

$m = 50$	$m = 100$	$m = 200$	$m = 500$
0.0518	0.0521	0.0528	0.0512

Table 5.1: Empirical percentage of rejections under the null hypothesis of the proposed test for the Murchison dataset for the different expected samples sizes, m, and using the bootstrap data-driven bandwidth selector proposed in Chapter 3.

	$t = 0.005$	$t = 0.05$	$t = 0.1$	$t = 0.3$
$m = 50$	0.0513	0.0522	0.0526	0.0552
$m = 100$	0.0497	0.0509	0.051	0.0531
$m = 200$	0.0483	0.0487	0.0481	0.0534
$m = 500$	0.0473	0.0477	0.0486	0.0557

Table 5.2: Empirical percentage of rejections under the null hypothesis of the proposed test for the Canadian dataset, for the different expected samples sizes, m, and a suitable range of possible bandwidths, t.

Regarding the values gathered in Table 5.1 and 5.2 we can see that in the first case with the automatic bandwidth selector and in the second for the second for different bandwidths in the suitable range, we perfectly adjust the level value around 0.05.

Now, let us focus on the power. Remember that we are under assumption (5.1), so we have to keep this while constructing the alternative hypothesis. Hence, we have built a new covariate, function of the initial one and varying on a parameter, $d_{\{\cdot\}}$. This parameter determines how far it will be from the null hypothesis; as the parameter increases its value we are further away from it. Then, the intensity is built applying the kernel estimator with the new covariate. In the Murchison example, the new covariate is $d_M e^{-Z(x)} + Z(x)$, and in the Canada data $d_C \frac{1}{Z(x)} + Z(x)$, where $Z(x)$ denotes in each case the corresponding initial covariate, $d_M = 0.5, 2, 5, 10$ and $d_C = 50, 100, 200, 300$. We can see the representations of these new covariates as well as the corresponding intensities in Figure 5.1 and Figure 5.2 for each of the situations considered.

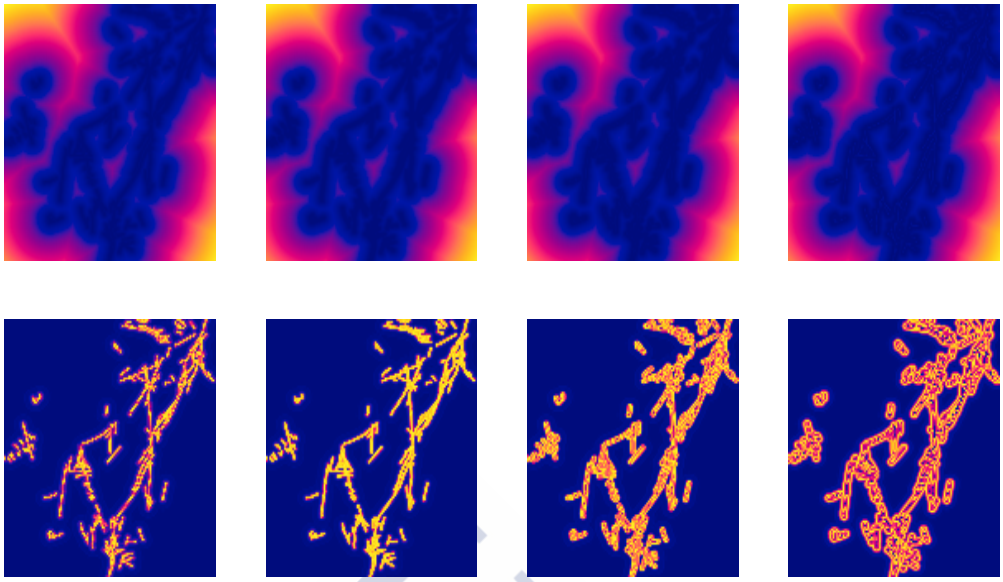


Figure 5.1: Representation of the new covariates (first row) and the corresponding intensities (second row) for the Murchison model with parameter values $d_M = 0.5, 2, 5, 10$.

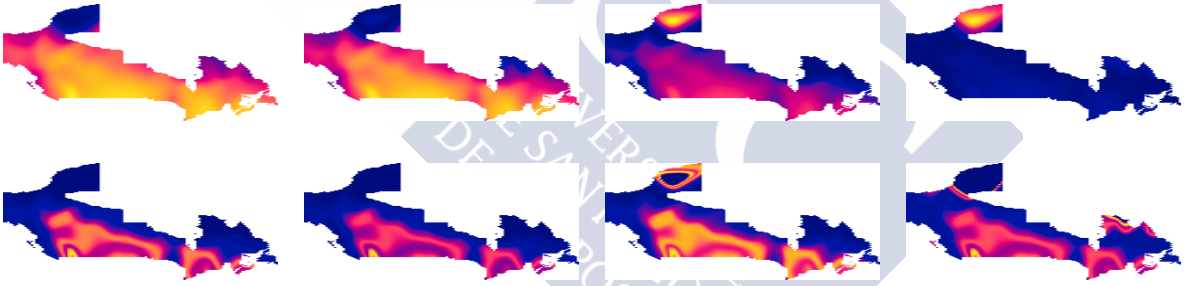


Figure 5.2: Representation of the new covariates (first row) and the corresponding intensities (second row) for the Canadian model with parameter values $d_C = 50, 100, 200, 300$.

To compute the rejection proportions, we generate, in each iteration, two samples, one from the initial model, the one fulfilling the null hypothesis, and another from the model under the alternative. We compute the test and decide whether we accept or reject the null hypothesis. We repeat this step $M = 5000$ times to obtain the rejection proportions.

The results of the power can be seen in Table 5.3 and Table 5.4. Note that in the first model we keep using the data-driven selector for the bandwidth, and in the second situation, we report only one table because, even though we have computed the test for the different possible bandwidths in the suitable range presented above, the values obtained are the same.

	$d_M = 0.5$	$d_M = 2$	$d_M = 5$	$d_M = 10$
$m = 50$	0.0944	0.4823	0.8592	0.9871
$m = 100$	0.1267	0.9234	0.9998	1
$m = 200$	0.2001	1	1	1
$m = 500$	0.4651	1	1	1

Table 5.3: Rejection proportions for the Murchison model, with different values of the parameter controlling the discrepancy from the null hypothesis, d_M, and four expected sample sizes, m.

	$d_C = 50$	$d_C = 100$	$d_C = 200$	$d_C = 300$
$m = 50$	0.0604	0.0661	0.0918	0.3695
$m = 100$	0.0576	0.0714	0.2462	0.8848
$m = 200$	0.0612	0.0794	0.8940	1
$m = 500$	0.0691	0.1128	1	1

Table 5.4: Rejection proportions for the Canadian wildfire model, with different values of the parameter controlling the discrepancy from the null hypothesis, d_C, and four expected sample sizes, m.

In Table 5.3 and Table 5.4 we show the rejection proportions for situations differing gradually from the null. We can see that the power of the test seems to be better for the first model, where even for $d_M = 2$, which is a situation near to the null, the power values are high for medium and large sample sizes. In the second model, the values do not reach those levels. However, we can see that when we are further enough from the null hypothesis the power increases notably, reaching in expected large sample sizes the 100 %.

5.4 Extensions

A possible extension for the two sample problem we address in this chapter, which involves exclusively independent samples, may be the situation where we allow dependence between the two given patterns. In that case, even though the statistic can be defined in an analogous way, its properties must be derived taking into account the existing second order structure.

Another subsequent extension after having addressed the two sample problem for dependent and independent patterns, is the k-sample problem, which consist of determine whether, given k point patterns under model (5.1), they come or not from the same process. To address this problem, the proposal in Vilar-Fernández et al. (2007) for regression functions can be followed as a guide, with the notable differences between both contexts.

5.5 Conclusions

We have addressed the classical two sample problem in the context of point processes with covariates. We have assumed a model where the intensity depends on a known covariate and we propose a test statistic to determine whether two given samples come from the same process. We have used the theoretical framework detailed in Chapter 3 and Chapter 4 to prove the asymptotic normality of the test, as well as a bootstrap procedure to improve its calibration. We have accomplished a simulation study based on real models that confirms the good performance of our proposal which reaches competitive values in terms of level and power.

Resumen en castellano
Summary in Spanish

La estadística espacial es uno de los campos más amplios dentro de la estadística, que se centra en analizar procesos aleatorios con alguna componente espacial, esto es, estudia las localizaciones asociadas a un proceso usando propiedades topológicas, geométricas o geográficas.

Históricamente, la primera mención relacionada con lo que hoy en día conocemos como estadística espacial fue en Halley (1686), trabajo en el que el autor intentaba identificar la causa de ciertos vientos y monzones en torno al trópico dibujando su localización y dirección sobre un mapa. Sin embargo, tenemos que esperar casi dos siglos, hasta 1854, para encontrar una aplicación rudimentaria de estadística espacial propiamente dicha: en ese momento existía en Londres una epidemia de cólera que se propagaba a gran velocidad entre la población, dejando un gran número de fallecidos en un corto período de tiempo, y cuyo origen era totalmente desconocido. Fue el Dr. Jonh Snow, quien hablando con los residentes y representando en un mapa los casos de cólera existentes fue capaz de discernir que el origen de la epidemia era una fuente contaminada en Bond street. Posteriormente, a comienzos del s.XX, Student (1907) se interesó por la distribución de las partículas dentro de un fluido, y su método de análisis consistió en agrupar las partículas y contar cuántas de ellas coexistían por unidad de volumen, que como se comprenderá posteriormente es uno de los precedentes más cercanos del concepto de intensidad. Fue unos años más tarde cuando Fisher (1935) se interesó en aplicar la estadística espacial al ámbito de la agricultura, que luego resultaría ser uno de los contextos más prolíficos en cuanto a aplicaciones.

En estadística espacial se recogen medidas de una variable aleatoria en ciertas localizaciones (que pueden ser aleatorias o estar previamente determinadas), y formalmente esto se representa por

$$\{Q(s) : s \in D \subset \mathbb{R}^d\}. \tag{R.1}$$

Dependiendo de las propiedades de las localizaciones, s, y las medidas, Q, se distinguen tres escenarios diferentes que constituyen los tres subcampos de la estadística

espacial: geoestadística, datos reticulares y procesos puntuales.

- **Geoestadística:** comenzó a principios de los años 80 y estudia procesos que varían de forma continua sobre el espacio, pero que se miden únicamente en localizaciones discretas; por consiguiente el conjunto D en (R.1) está fijo y es continuo. Algunos ejemplos típicos de esta rama son datos meteorológicos o de contaminación de aire. En este contexto, los principales objetivos son determinar un patrón espacial para los datos, modelizar la correlación o la covarianza, hacer predicciones y testear si existe o no una estructura espacial.

- **Datos reticulares:** también conocidos como datos de área, estudian medidas (continuas o discretas) asociadas a regiones, que pueden ser tanto regulares (tipo rejilla) como irregulares (por ejemplo divisiones geopolíticas). Por tanto, en este caso el conjunto D en (R.1) será fijo y discreto. Aquí no existe la posibilidad de medir entre las localizaciones, y el hecho de agrupar los datos reduce en cierta manera la información disponible.

- **Procesos puntuales:** en este caso la principal característica es que se trata de medidas aleatorias que ocurren en localizaciones también aleatorias; por tanto D en (R.1) es un conjunto aleatorio de índices. El interés es determinar el patrón espacial subyacente (si existe), así como otras propiedades o características relevantes como la dependencia de segundo orden, la inclusión de una componente espacial en el proceso o la influencia de información externa a través de marcas y/o covariables.

Hay que tener en cuenta que el comportamiento de los tres tipos de procesos espaciales es muy diferente entre sí, y por consiguiente la técnicas estadísticas y la metodología que se emplean en cada una de ellas son también distintas. Aunque el objetivo principal y general de la estadística espacial sí es común en todos ellos y consiste en modelizar el proceso subyacente con el fin de poder entender de la forma más completa posible el comportamiento de los datos.

La estadística espacial posee un gran abanico de aplicaciones; la geoestadística se emplea por ejemplo en hidrología, medioambiente (análisis de datos de contaminación medidos en localizaciones fijas), climatología (variables como la temperatura, el viento o la lluvia que son medidas en estaciones) y geología (ver Cressie (1993) y Isaaks y Srivastava (1989)). Ejemplos de datos reticulares son fáciles de encontrar en datos agrupados por código postal, censo, fronteras o datos recolectados a distancia y reportados en forma de píxeles; también aparecen este tipo de datos con cierta frecuencia en estadística pública al estudiar factores económicos y sociales por regiones, ver Schabenberger y Gotway (2017). Finalmente, tal y como se puede comprobar en esta tesis, los campos de aplicación de los procesos puntuales son muy variados: ciencias forestales, epidemiología, medioambiente, análsis de crímenes, geología, zoología, astronomía... En P. J. Diggle (2013) se pueden encontrar numerosos ejemplos.

Capítulo 1: Introducción

A lo largo del Capítulo 1 se realiza una introducción a los conceptos fundamentales asociados a los procesos puntuales. Se considera que un proceso puntual X es un modelo probabilístico vinculado con una serie de puntos o localizaciones en un espacio dado. A cada una de las localizaciones asociadas al proceso se la denomina evento y el espacio sobre el que se encuentran es la región de observación, W, que generalmente suele ser un subconjunto del espacio en el que el proceso está definido. Además, existe una función actuando sobre subconjuntos de la región de observación, $N : \mathcal{P}(W) \to \mathbb{N}^+$, que determina el número de eventos que tienen lugar en el correspondiente subconjunto.

Sobre el citado proceso se definen propiedades generales como la estacionaridad (el proceso es invariante por traslaciones), la isotropía (el proceso es invariante por rotaciones) y la aleatoriedad espacial completa. Al proceso de testeo de esta última se dedica una sección completa, pues de aceptarse, estaríamos asumiendo que el proceso es totalmente aleatorio y por tanto no ha lugar modelización alguna del mismo.

En este capítulo también se incluye una explicación detallada sobre los procesos de Poisson, pues se trata de uno de los modelos más relevantes y con mayor importancia en la fundamentación teórica de los procesos puntuales. Por supuesto, también se incluye una descripción de las principales propiedades de primer y segundo orden: intensidad, función de correlación, la K-función..., así como la extensión de los mismos al caso espacio-temporal. En este apartado también se menciona el concepto de separabilidad y algunas referencias en las que se trata el tema.

Además, en la parte dedicada a la intensidad de primer orden, se incluye una sección dedicada a exponer la relación intrínseca existente entre la función de densidad y la función de intensidad, que será ampliamente utilizada a lo largo de la presente tesis doctoral. Esta relación será una herramienta fundamental en los desarrollos teóricos que se van a presentar, permitiendo definir, entre otros, estimadores consistentes de la función de intensidad.

Hay también una sección dedicada a las distintas formas en las que podemos encontrar información extra más allá de las coordenadas de las localizaciones, para un proceso puntual. Se trata fundamentalmente de las marcas (información asociada a los propios eventos, como podría ser la superficie quemada en un incendio forestal) y las covariables (información existente en toda la región de observación del proceso independientemente de que haya o no un evento, por ejemplo la temperatura ambiente). En esta sección se incluyen ejemplos ilustrativos que permiten al lector comprender y visualizar mejor estos conceptos, así como las diferencias fundamentales entre ellos.

Finalmente resaltar que este primer capítulo pretende proporcionar al lector con las herramientas básicas necesarias para poder seguir los desarrollos que posteriormente se

realizan. Posee por tanto un carárter más bien divulgativo, aunque sin perder el grado mínimo de formalismo que consideramos necesario.

Capítulo 2: Estimación de la densidad con datos sesgados

Inspirados en la relación presentada en el Capítulo 1 entre la función de intensidad y la función de densidad, así como la estructura de algunos de los estimadores existentes para la primera de ellas, surge la oportunidad de realizar nuevas aportaciones en el ámbito de la estimación de la densidad con datos sesgados.

El campo de los datos sesgados (o datos ponderados) se origina porque, aunque se supone que una muestra debe poseer las mismas características básicas que la población a la que representa, no siempre ocurre así, y puede que el sesgo se incluya en el esquema de muestreo. Esto ocurre cuando la probabilidad de escoger un dato depende del valor del mismo o de otras covariables de interés. Existen numerosos ejemplos de este tipo de datos en diversos ámbitos como biomedicina, epidemiología, industria textil, economía o control de calidad.

Sea X una variable aleatoria no observada y denotemos por f su función de densidad. Asumamos que únicamente disponemos de información observable de una variable intrínsecamente relacionada con la original que denotaremos por Y, y cuya función de densidad sesgada o ponderada es:

$$f_{Y,\omega}(y) = \frac{\omega(y)f(y)}{\mu_\omega}, \quad y > 0,$$

siendo ω una función conocida y $\mu_\omega = \int \omega(x)f(x)dx < \infty$. El caso particular en el que $\omega(y) = y$ se conoce como datos longitudinalmente sesgados, y en ellos nos centraremos en este capítulo, aunque como se puede ver al final del mismo, la extensión al caso general es casi inmediata. Consideremos por tanto

$$f_Y(y) = \frac{yf(y)}{\mu}, \qquad y > 0,$$

con $\mu = \int yf(y)dy$.

Dada $Y_1, \ldots, Y_n$ una muestra aleatoria simple (m.a.s.) de f_Y, Jones (1991) propuso el siguiente estimador tipo núcleo de la densidad:

$$\hat{f}_h(y) = \frac{1}{n}\hat{\mu}\sum_{i=1}^{n}\frac{1}{Y_i}K_h(y - Y_i), \tag{R.2}$$

con $\hat{\mu} = \left(\frac{1}{n}\sum_{i=n}^{n}\frac{1}{Y_i}\right)^{-1}$, y $K_h(\cdot) = \frac{1}{h}K(\frac{\cdot}{h})$, donde K es una función núcleo simétrica. Bajo ciertas hipótesis, en el documento de tesis se desarrollan las propiedades teóricas de este estimador, obteniendo las expresiones para la media y varianza puntuales, el error cuadrático medio (ECM), su versión integrada (ECMI) y la versión asintótica de éste último (ECMIA), que se detallan a continuación:

$$E\left[\hat{f}_h(y)\right] = \left(K_h \circ f\right)(y) + O\left(\tfrac{1}{n}\right),$$

$$\mathrm{Var}\left[\hat{f}_h(y)\right] = n^{-1}\left[\left(K_h^2 \circ \gamma\right)(y) - \left(K_h \circ f\right)^2(y)\right] + O\left(\tfrac{1}{n}\right),$$

$$\mathrm{ECM}\,(h,y) = \frac{1}{4}h^4\left(f''(y)\right)^2\mu_2^2(K) + \frac{\gamma(y)}{nh}R(K) + o\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{ECMI}\,(h) = \frac{1}{4}h^4\mu_2^2(K)R(f'') + \frac{R(K)\mu c}{nh} + o\left(h^4 + \frac{1}{nh}\right) \text{ y}$$

$$\mathrm{ECMIA}\,(h) = \frac{1}{4}h^4\mu_2^2(K)R(f'') + \frac{R(K)\mu c}{nh}, \text{ con } c = \int\frac{1}{y}f(y)dy,$$

siendo $\circ$ la convolución entre dos funciones, $\gamma(y) = \mu f(y)/y$, $\mu_2(K) = \int u^2K(u)du$ y $R(K) = \int K^2(u)du$.

Como consecuencia de esta última ecuación se obtiene la expresión de la ventana óptima que minimiza el ECMIA:

$$h_{\mathrm{ECMIA}} = \left(\frac{R(K)\mu c}{n\mu_2^2(K)R(f'')}\right)^{1/5}. \tag{R.3}$$

Además en este capítulo se presentan dos procedimientos de remuestreo bootstrap completamente innovadores en el contexto de los datos sesgados. Ambos son consistentes y constituyen la base en la definición de los correspondientes selectores de ventana para el estimador (R.2).

El primero de ellos es un procedimiento bootstrap basado en el estimador de Jones, (R.2), en el que dada una m.a.s. de f_Y, las muestras bootstrap son generadas a partir de la densidad $\hat{f}_{Y,g}(y) = y\hat{f}_g(y)/\hat{\mu}$, en donde g es una ventana piloto. Así, denotando por Y^* la variable aleatoria generada por ese método bootstrap, se define un estimador de la densidad de Y^* como

$$\hat{f}_h^*(y) = \frac{1}{n}\hat{\mu}^*\sum_{i=1}^{n}\frac{1}{Y_i^*}L_h\left(y - Y_i^*\right), \tag{R.4}$$

donde $\hat{\mu}^* = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i^*}\right)^{-1}$ y $L_h(\cdot) = \frac{1}{h}L(\frac{\cdot}{h})$, con L una función núcleo simétrica verificando las mismas propiedades que K.

De manera análoga a lo hecho para el estimador de Jones, detallamos el cálculo de la media y varianza puntuales, el error cuadrático medio, el error cuadrático medio integrado y su versión asintótica para el estimador (R.4):

$$E^*\left[\hat{f}_h^*(y)\right] = \left(L_h \circ \hat{f}_g\right)(y) + O_P\left(\tfrac{1}{n}\right) \text{ and}$$
$$\mathrm{Var}^*\left[\hat{f}_h^*(y)\right] = n^{-1}\left[\left(L_h^2 \circ \hat{\gamma}_g\right)(y) - \left(L_h \circ \hat{f}_g\right)^2(y)\right] + O_P\left(\tfrac{1}{n}\right).$$

$$\mathrm{ECM}^*(h,y) = \frac{1}{4}h^4\left(\hat{f}_g''(y)\right)^2\mu_2^2(L) + \frac{\hat{\gamma}_g(y)}{nh}R(L) + o_P\left(h^4 + \frac{1}{nh}\right),$$

$$\mathrm{ECMI}^*(h) = \tfrac{1}{4}h^4\mu_2^2(L)R(\hat{f}_g'') + \tfrac{R(L)\hat{\mu}\hat{c}}{nh} + o_P\left(h^4 + \tfrac{1}{nh}\right),$$
$$\mathrm{ECMIA}^*(h) = \tfrac{1}{4}h^4\mu_2^2(L)R(\hat{f}_g'') + \tfrac{R(L)\hat{\mu}\hat{c}}{nh}, \quad \text{con } \hat{c} = \hat{\mu}\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{1}{Y_i^2}.$$

Y también análogamente a la situación inicial se obtiene una expresión para la ventana óptima:

$$h_{\mathrm{ECMIA}^*} = \left(\frac{R(L)\hat{\mu}\hat{c}}{n\mu_2^2(L)R(\hat{f}_g'')}\right)^{1/5},$$

que como se puede observar resulta una versión plug-in de (R.3).

El segundo procedimiento bootstrap que se presenta, emplea en lugar del estimador de Jones el estimador tipo núcleo habitual en densidad: dada una m.a.s. de f_Y, las muestras bootstrap se generan aleatoriamente y con reemplazamiento de $\tilde{f}_{K,g}$, que denota al estimador tipo núcleo habitual en densidad con ventana piloto g y función núcleo K. El estimador bootstrap de la densidad se define análogamente a (R.4), con la salvedad de que los datos de la muestra bootstrap se generan de manera diferente.

En este caso, hemos obtenido también la media y varianza puntuales del estimador bootstrap, así como su error cuadrático medio. De las versiones integradas y asintótica no es posible obtener expresiones explícitas debido a la complejidad de las mismas, aunque evidentemente su cómputo en casos concretos sí es posible sin más que integrar el ECM y suprimir los términos de orden inferior, respectivamente.

$$E^*\left[\hat{f}_h^*(y)\right] = \tfrac{1}{\int \frac{1}{z}\tilde{f}_{K,g}(z)dz}\int \tfrac{1}{z}L_h(y-z)\tilde{f}_{K,g}(z)dz + O_P\left(\tfrac{1}{n}\right),$$

$$\mathrm{Var}^*\left[\hat{f}_h^*(y)\right] = \tfrac{1}{n\left(\int \frac{1}{z}\tilde{f}_{K,g}(z)dz\right)^2}\left[\int \tfrac{1}{z^2}L_h^2(y-z)\tilde{f}_{K,g}(z)dz - \int \tfrac{1}{z}L_h(y-z)\tilde{f}_{K,g}(z)dz^2\right]$$
$$+O_P\left(\tfrac{1}{n}\right) \text{ y}$$

$$\begin{aligned}\mathrm{ECM}^*(h,y) &= \left(\frac{\int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz}{\int \frac{1}{z}\tilde{f}_{K,g}(z)dz} - \hat{f}_h(y)\right)^2 \\ &+\frac{1}{n\left(\int \frac{1}{z}\tilde{f}_{K,g}(z)dz\right)^2}\left[\int \frac{1}{z^2} L_h^2(y-z)\tilde{f}_{K,g}(z)dz - \int \frac{1}{z} L_h(y-z)\tilde{f}_{K,g}(z)dz^2\right] \\ &+O_P\left(\frac{1}{n}\right).\end{aligned}$$

Otra de las aportaciones relevantes de este capítulo son los selectores de ventana definidos específicamente para el estimador de Jones. Hasta el momento existía únicamente un selector de validación cruzada propuesto por Guillamón et al. (1998). En el presente capítulo se incluyen tres nuevas propuestas: una regla del pulgar al estilo de Silverman (1986), cuya base es la asunción de la normalidad de la distribución subyacente, y dos ventanas bootstrap basadas en los respectivos procedimientos anteriormente detallados. En estas dos últimas propuestas se necesita una ventana piloto, en el caso basado en el estimador de Jones, empleamos las ideas de Cao (1993) para establecer unas condiciones que nos garantizan la expresión óptima con el correspondiente orden para la ventana piloto; en el caso de usar el estimador tipo núcleo habitual, y visto que no se puede llegar a expresiones explícitas generales, empleamos la propuesta de Dutta (2016) para la elección del piloto.

Además se realiza un completo estudio de simulación, con modelos que abarcan una amplia variedad de características, lo que permite analizar el comportamiento de las distintas propuestas. Las conclusiones generales del estudio permiten afirmar que tanto nuestra regla del pulgar como las ventanas bootstrap funcionan bien en la mayoría de los modelos, mejorando la única propuesta existente que corresponde al selector de validación cruzada. Si bien es cierto que cuando la complejidad de los modelos se incrementa notablemente, dicha propuesta, como era de esperar, proporciona mejores resultados, aunque nuestros selectores siguen siendo competitivos.

Capítulo 3: Estimación tipo núcleo de la intensidad en procesos espaciales con covariables

En este capítulo nos dedicamos ya al campo de los procesos puntuales, más concretamente a la estimación de la intensidad desde una perspectiva noparamétrica. Dicho problema estuvo prácticamente limitado durante años al ámbito exploratorio, debido a la falta de consistencia de las propuestas existentes. Este capítulo se centra en la definición de un estimador tipo núcleo de la intensidad usando covariables, se presenta un procedimiento de remuestreo bootstrap y se proponen dos nuevos selectores de ventana específicamente diseñados para el citado estimador. Además se lleva a cabo un completo estudio de simulación con varios modelos con características diferenciadas entre sí, y se detalla una interesante aplicación a datos reales sobre incendios forestales en Canadá durante junio de 2015, empleando datos meteorológicos como covariables.

Se comienza por definir la función de intensidad de primer orden:

$$\lambda(x) = \lim_{|dx|\to 0} \frac{\mathbb{E}[N(dx)]}{|dx|},$$

donde $|dx|$ denota el área de una región infinitesimal, dx, que contiene a $x \in \mathbb{R}^2$ y $N(dx)$ el número de puntos del proceso en esa región.

Existen diversos modelos paramétricos para la función de intensidad, sin embargo es de sobra conocido que, cuando el modelo escogido se desvía de la verdadera intensidad, los estimadores que se obtienen no son fiables, es por ello que esta tesis se centra en técnicas noparamétricas. Inicialmente, en los años 80, Diggle propuso el primer estimador tipo núcleo de la intensidad en la recta real, que fue posteriormente extendido al espacio euclídeo bidimensional:

$$\hat{\lambda}_H^D(x) = \frac{\sum_{i=1}^N K_H(x - X_i)}{p_H(x)}, \quad x \in \mathbb{R}^2,$$

donde H is la matriz de ventanas, K denota una función núcleo bidimensional, $K_H(x) = |H|^{-1/2} K\left(H^{-1/2}x\right)$, $p_H = \int_W |H|^{-1/2} K(H^{-1/2}(x-y))dy$ es la corrección de efecto frontera y $X_1, \ldots, X_N$ es una realización del proceso.

Como ya se avanzaba anteriormente, la falta de consistencia de este estimador provocó que su uso fuera meramente exploratorio, y que décadas después se intentasen buscar posibles soluciones a ese problema. Una de las propuestas en Cucala (2006) fue la introducción del concepto de "densidad de localización de eventos", $\lambda_0(x) = \lambda(x)/m$, donde $m = \int_W \lambda(x)dx$ es el número esperado de puntos en W. A partir de esto se propuso el estimador:

$$\hat{\lambda}_{0,h}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) 1_{\{N\neq 0\}}, \; x \in \mathbb{R},$$

con $1_{\{\}}$ denotando la función indicadora y h un parámetro ventana escalar. El marco teórico de este estimador está desarrollado con detalle en Cucala (2006) y existe una extensión al espacio euclídeo bidimensional en Fuentes-Santos et al. (2015).

La segunda propuesta para atajar la problemática de la falta de consistencia en la estimación noparamétrica de la intensidad, es el uso de información externa en forma de covariables, en cuyo desarrollo nos centramos a lo largo de esta tesis doctoral. Sea $Z : W \subset \mathbb{R}^2 \to \mathbb{R}$ una covariable espacial continua, conocida en cada punto de la región de observación W (o al menos en una cantidad suficiente de puntos que permita hacer una interpolación) y $Z_1, \ldots, Z_N$ una realización del proceso transformado, esto es, $Z_i = Z(X_i)$. En ciertos casos se puede asumir un modelo en el que la intensidad es función de dicha covariable:

$$\lambda(u) = \rho(Z(u)), \; u \in W \subset \mathbb{R}^2, \tag{R.5}$$

donde ρ es una función desconocida. Por consiguiente, dado que Z sí es conocida, bastaría con estimar ρ para obtener un estimador de λ. Bajo este supuesto, existen hasta el momento dos estimadores tipo núcleo diferentes de la intensidad, uno de ellos propuesto en Guan (2008) cuya consistencia es probada bajo el marco asintótico del "infill structure", y el otro en Baddeley et al. (2012) con varias variantes dentro de la estimación tipo núcleo, así como otras propuestas basadas en verosimilitud local. En esta segunda aproximación los desarrollos teóricos son mínimos.

El objetivo en este capítulo es, no solo definir un estimador de la intensidad en procesos puntuales con covariables, sino también desarrollar el marco teórico asociado en el que probar sus buenas propiedades, así como posibles selectores de ventana. Para ello necesitamos establecer formalmente la relación entre el proceso inicial, X, y el transformado a través de la covariable, $Z(X)$. Omitiendo los detalles que pueden consultarse en el documento de tesis, el resultado final se resume en que si el proceso original X es un proceso de Poisson bidimensional no homogéneo con intensidad $\lambda(x) = \rho(Z(x))$, el proceso transformado, $Z(X)$ es también un proceso de Poisson, ahora unidimensional, con intensidad $\rho(z)g^\star(z)$, donde $g^\star$ es la derivada sin normalizar de la función de distribución espacial asociada a Z.

Definimos ahora la densidad "artificial"

$$f(\cdot) = \frac{\rho(\cdot)g^\star(\cdot)}{m}, \tag{R.6}$$

y proponemos el estimador

$$\hat{f}_h(z) = g^\star(z)\frac{1}{N}\sum_{i=1}^{N}\frac{1}{g^\star(Z_i)}K_h\left(z - Z_i\right)1_{\{N\neq 0\}}, \tag{R.7}$$

donde K es una función núcleo univariante y $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$. Éste es un estimador de f, por tanto, una vez obtenido, es sencillo llegar a la intensidad haciendo un plug-in en la expresión (R.5) a través de ρ.

De manera similar a lo realizado en el capítulo anterior, pero teniendo en cuenta la doble aleatoriedad existente por tratarse de un proceso puntual (tamaño muestral y localización son ambas variables aleatorias), se han obtenido expresiones para la media y varianza puntuales del estimador anterior, su error cuadrático medio, la versión integrada de éste y su análogo asintótico; a partir del cual también se llega a una expresión de la ventana óptima tal y como puede verse a continuación:

$$E\left[\hat{f}_h(z)\right] = \frac{g^\star(z)(K_h\circ\rho)(z)}{m}\left(1 - e^{-m}\right),$$

$$\mathrm{Var}\left[\hat{f}_h(z)\right] = A(m)\frac{(g^\star(z))^2}{n}\left(K_h^2\circ\frac{\rho}{g^\star}\right)(z) - (A(m) + e^{-2m} - e^{-m})(g^\star(z))^2(K_h\circ\rho)^2(z),$$

$$\begin{aligned} ECM(h,z) &= e^{-2m}f^2(z) + (1-e^{-m})^2\tfrac{h^4}{4}\left(\tfrac{\rho''(z)g^\star(z)}{m}\right)^2\mu_2^2(K) \\ &\quad -e^{-m}(1-e^{-m})h^2\mu_2(K)\tfrac{(g^\star(z))^2\rho(z)\rho''(z)}{m^2} + \tfrac{A(m)}{h}f(z)R(K) \\ &\quad +o(h^2(1-e^{-m})e^{-m}) + o(h^4(1-e^{-m})^2) + o\left(\tfrac{A(m)}{mh}\right), \\ ECMI(h) &= e^{-2m}R(f) + (1-e^{-m})^2\tfrac{h^4}{4}R\left(\tfrac{\rho''g^\star}{m}\right)\mu_2^2(K) \\ &\quad -e^{-m}(1-e^{-m})h^2\mu_2(K)\textstyle\int_{\mathbb{R}}\tfrac{g^\star(z)\rho''(z)f(z)}{m}dz + \tfrac{A(m)}{h}R(K) \\ &\quad +o(h^2(1-e^{-m})e^{-m}) + o(h^4(1-e^{-m})^2) + o\left(\tfrac{A(m)}{mh}\right), \\ ECMIA(h) &= (1-e^{-m})^2\tfrac{h^4}{4}R\left(\tfrac{\rho''g^\star}{m}\right)\mu_2^2(K) + \tfrac{A(m)}{h}R(K) \qquad \text{y} \end{aligned}$$

$$h_{ECMIA} = \left(\frac{A(m)R(K)}{\mu_2^2(K)(1-e^{-m})^2R\left(\frac{\rho''g^\star}{m}\right)}\right)^{1/5} = \left(\frac{R(K)}{\mu_2^2(K)R(\rho''g^\star)}\frac{A(m)m^2}{(1-e^{-m})^2}\right)^{1/5}. \tag{R.8}$$

Además de los resultados arriba enumerados, en este capítulo se propone un innovador procedimiento de remuestreo bootstrap para el contexto de procesos puntuales con covariables. Esta metodología está inspirada en los trabajos previos de Cao (1993) y Cowling et al. (1996) y consiste en: dada $X_1,\ldots,X_n$ una realización del proceso puntual original, construimos $Z_1,\ldots,Z_n$ la realización asociada correspondiente al proceso transformado. Sea además $\hat{f}_b$ el estimador de la densidad en (R.7) y $\hat{\rho}_b(z) = \sum_{i=1}^N \frac{1}{g^\star(Z_i)}K_b(z-Z_i)$ con b una ventana piloto. Condicionalmente a $Z_1,\ldots,Z_n$, $N^* \sim Poiss\,(\hat{m})$ con $\hat{m} := \int_{\mathbb{R}}\hat{\rho}_b(z)g^\star(z)dz$, generamos n^* una realización de la variable aleatoria N^* y entonces $Z_1^*,\ldots,Z_{n^*}^*$ se obtienen por muestreo aleatorio simple con reemplazamiento de una densidad proporcional a $g^\star\hat{\rho}_b$. Si denotamos por Z^* la variable aleatoria generada por el procedimiento bootstrap, se puede definir el estimador:

$$\hat{f_h}^*(z) = g^\star(z)\frac{1}{N^*}\sum_{i=1}^{N^*}\frac{1}{g^\star(Z_i^*)}K_h\left(z-Z_i^*\right)1_{\{N^*\neq 0\}}, \tag{R.9}$$

obteniendo el estimador asociado para ρ:

$$\hat{\rho}_h^*(z) = \sum_{i=1}^{N^*}\frac{1}{g^\star(Z_i^*)}K_h\left(z-Z_i^*\right),$$

y finalmente por plug-in se tiene un estimador de λ.

A continuación, y asumiendo ciertas hipótesis explicitadas en el capítulo, se han obtenido la media y varianza puntuales del estimador bootstrap, su error cuadrático medio, el error cuadrático medio integrado, la versión asintótica de éste y a partir de

ella la expresión de la ventana óptima:

$$E^*\left[\hat{f}_h^{\,*}(z)\right] = \tfrac{g^\star(z)}{\hat{m}}(K_h \circ \hat{\rho}_b)(z)(1-e^{-\hat{m}}),$$

$$\mathrm{Var}^*\left[\hat{f}_h^*(z)\right] = \tfrac{(g^\star(z))^2}{\hat{m}}\left(K_h^2 \circ \tfrac{\hat{\rho}_b}{g^\star}\right)(z)A(\hat{m}) - \tfrac{(g^\star(z))^2}{\hat{m}^2}(K_h \circ \hat{\rho}_b)^2(z)(A(\hat{m}) + e^{-2\hat{m}} - e^{-\hat{m}}),$$

$$\begin{aligned} ECM^*(h,z) &= e^{-2\hat{m}}(\tilde{f}_b(z))^2 + \frac{h^4}{4}(\hat{\rho}_h''(z))^2\frac{(g^\star(z))^2}{\hat{m}^2}\mu_2^2(L)(1-e^{-\hat{m}})^2 \\ &\quad - e^{-\hat{m}}(1-e^{-\hat{m}})h^2\tilde{f}_b(z)\frac{\hat{\rho}_b''(z)g^\star(z)}{\hat{m}}\mu_2(L) + \frac{A(\hat{m})}{h}R(L) + o_P(h^4(1-e^{-\hat{m}})^2) \\ &\quad + o_P(h^2(1-e^{-\hat{m}})e^{-\hat{m}}) + o_P\left(\frac{A(\hat{m})}{\hat{m}h}\right), \end{aligned} \tag{R.10}$$

$$\begin{aligned} ECMI^*(h) &= e^{-2\hat{m}}R(\tilde{f}_b) + \tfrac{h^4}{4}R\left(\tfrac{\hat{\rho}_b''g^\star}{\hat{m}}\right)\mu_2^2(L)(1-e^{-\hat{m}})^2 \\ &\quad -e^{-\hat{m}}(1-e^{-\hat{m}})h^2\mu_2(L)\int\tfrac{\tilde{f}_b(z)\hat{\rho}_b''(z)g^\star(z)}{\hat{m}}dz + \tfrac{A(\hat{m})}{h}R(L) \\ &\quad +o(h^4(1-e^{-\hat{m}})^2) + o_P(h^2(1-e^{-\hat{m}})e^{-\hat{m}}) + o_P\left(\tfrac{A(\hat{m})}{\hat{m}h}\right), \\ ECMIA^*(h) &= \tfrac{h^4}{4}R\left(\tfrac{\hat{\rho}_b''g^\star}{\hat{m}}\right)\mu_2^2(L)(1-e^{-\hat{m}})^2 + \tfrac{A(\hat{m})}{h}R(L) \qquad \text{y} \\ h_{ECMIA^*} &= \left(\frac{A(\hat{m})R(L)}{\mu_2^2(L)(1-e^{-\hat{m}})^2R(\frac{\hat{\rho}_b''g^\star}{\hat{m}})}\right)^{1/5}. \end{aligned} \tag{R.11}$$

En base a estos resultados, queda fácilmente probada la consistencia del bootstrap en el sentido de que tanto el ECMI^* como ECMIA* son estimadores consistentes del ECMI y ECMIA, respectivamente.

Una vez definido el estimador y probadas formalmente sus buenas propiedades, nos interesa definir selectores de ventana automáticos. Se proponen dos selectores diferentes, uno de ellos basado en las ideas de Silverman y otro en el procedimiento bootstrap. Además, se ha realizado un estudio de simulación para analizar el comportamiento de ésos selectores en muestras finitas, y se han comparado de nuestras propuestas con las de Guan (2008) y Baddeley et al. (2012).

Las conclusiones generales del estudio de simulación permiten afirmar que, en general, la ventana bootstrap se comporta mejor que los restantes competidores en la mayor parte de los modelos, y cuando esto no ocurre sigue siendo competitivo. También se puede ver como la regla del pulgar que proponemos y que está especialmente diseñada para el contexto de procesos puntuales, funciona mejor que aplicar la regla de Silverman de densidad a los datos tal y como sugieren en Baddeley et al. (2012). En la comparativa con la propuesta de Guan (2008) hay que tener en cuenta varias

consideraciones: en primer lugar su propuesta incluye una integración numérica que es computacionalmente muy costosa, en segundo lugar su selector de ventana automático es del tipo validación cruzada que también posee un coste computacional elevado, y por último, hemos detectado que en varias muestras (entre 45 y 213 sobre 500) la función de validación cruzada no posee mínimo global, lo que incurre en una elección incorrecta del parámetro ventana.

También se ha incluido en este capítulo una aplicación a datos reales. Para ello hemos recogido datos de incendios forestales en Canadá durante el mes de junio de 2015, junto con variables meteorológicas: temperatura y precipitación. Inicialmente empleamos cada una de ellas por separado y aplicamos nuestro estimador para obtener las intensidades correspondientes, además proponemos hacer un análisis incluyendo las dos covariables conjuntamente mediante componentes principales.

Para finalizar el capítulo, hacemos una breve disertación sobre las posibles extensiones al caso espacio-temporal de la metodología desarrollada en procesos puntuales espaciales, y también sobre diferentes técnicas que permitiesen incluir información de más de una covariable en el proceso.

Capítulo 4: Contraste de significación de la covariable para la función de intensidad de primer orden en procesos puntuales espaciales

Previamente se ha visto que la modelización de la función de intensidad ha sido objeto de estudio desde diferentes perspectivas. En el capítulo anterior se ha trabajado bajo la suposición de un modelo para la misma en el que la intensidad dependía de una covariable espacial a través de una función desconocida, y dicho modelo fue asumido sin comprobación alguna. En este capítulo nos centramos en analizar si el citado modelo es o no apropiado, para lo que proponemos un test estadístico basado en la distancia L^2. A lo largo del capítulo se demuestra con detalle la normalidad asintótica de esta propuesta, así como un procedimiento bootstrap para mejorar su calibración. Además, se presentan dos aplicaciones a datos reales, una de ellas empleando de nuevo los datos de incendios forestales en Canadá en junio de 2015, y la otra con el conjunto de datos de las minas de Murchison, que ya fue utilizado previamente en la literatura existente.

Recordemos que el modelo a contrastar viene especificado por

$$\lambda(x) = \rho(Z(x)),\ x \in W \subset \mathbb{R}^2, \tag{R.12}$$

donde Z es una covariable espacial conocida en casi todo punto de la región de observación W. Debemos destacar que, hasta el momento, no se ha desarrollado ningún

otro test con esta misma finalidad, aunque sí se han hecho desarrollos en una línea similar en Díaz-Avalos et al. (2014).

Para comenzar, vamos a presentar los conjunto de datos que empleamos en este capítulo. Los datos del estudio geológico de Murchison, ver Figura R.1, recogen las localizaciones espaciales de 255 depósitos de oro, junto con las coordenadas de las fallas geológicas de la zona, un área de 330×394 km^2 en la región de Murchison en Australia Occidental, y que fueron inicialmente recogidos en Watkins y Hickman (1990). La covariable que emplearemos se construye a partir de los datos como la distancia desde cada punto de la región de observación a la falla geológica más cercana.

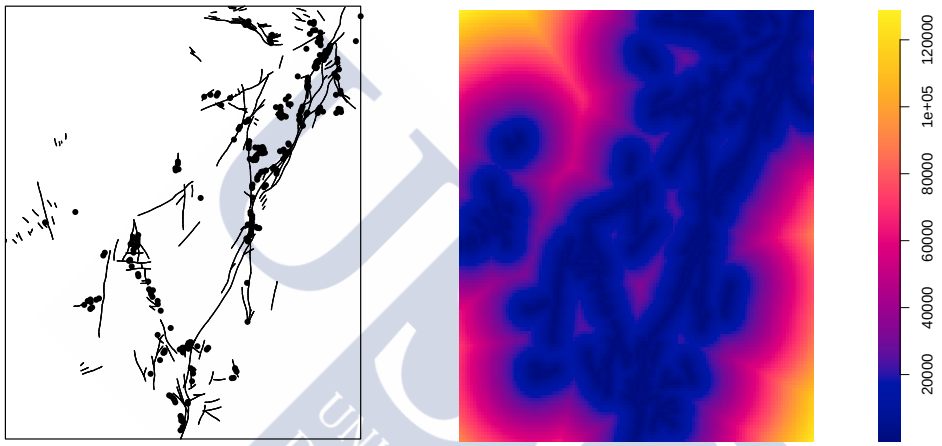


Figura R.1: Datos del estudio geológico de Murchison: depósitos de oro (puntos) y fallas geológicas (líneas) (dcha), junto con la covariable construida como la distancia a la falla más cercana (izq).

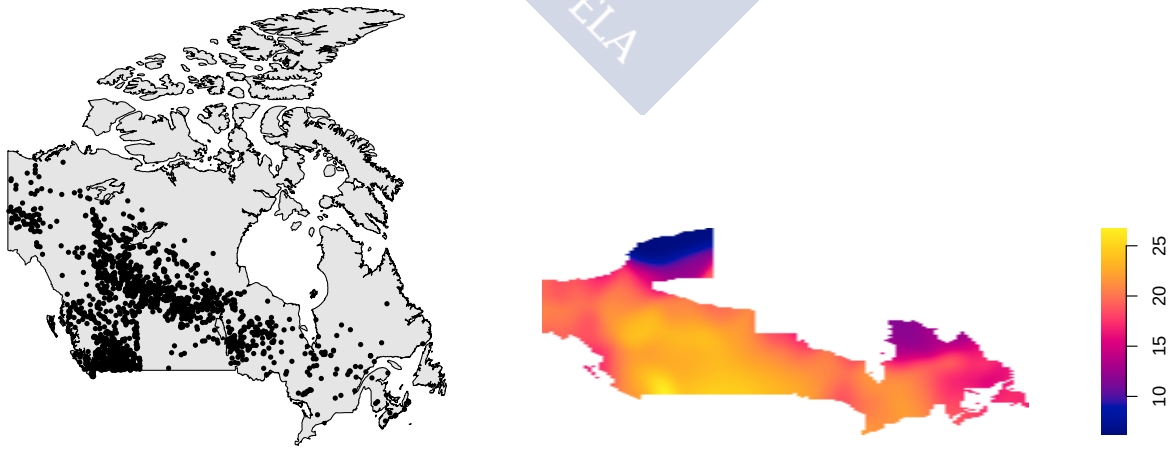


Figura R.2: Datos de incendios forestales en Canadá en 2015 (dcha), junto con el tercer cuartil de la temperatura registrada en Canadá en junio de 2015 tras aplicar una suavización gaussiana con $\sigma = 2$ (izq).

El segundo conjunto de datos se corresponde con las localizaciones de los incendios forestales que tuvieron lugar en Canadá durante junio de 2015, ver Figura R.2. En este caso, y teniendo en cuenta la información proporcionada por diversos estudios relacionados con los incendios en Canadá, que afirma que generalmente son producidos por causas meteorológicas, consideraremos la temperatura como covariable. Para evitar la utilización de datos extremos, nos quedaremos con el tercer cuartil de la temperatura, y dado que esos valores se recogen en estaciones meteorológicas fijas, suavizaremos los datos para obtener una covariable continua en lugar de discreta. Nótese que esto puede hacerse sin pérdida de generalidad debido al gran número de estaciones meteorológicas empleadas para recabar la información.

Nuestro objetivo último es intentar determinar si, en un caso, la distancia a las fallas geológicas es suficiente para explicar la distribución espacial de los depósitos de oro; y en el otro, si la temperatura explica por sí misma la distribución de incendios forestales en Canadá.

Formalmente, nuestra propuesta busca contrastar la hipótesis nula $H_0 : \lambda(x) = \rho(Z(x)), x \in W$, frente a una alternativa general en la que la función de intensidad no pueda ser totalmente explicada a través de la covariable. La idea es comparar un estimador tipo núcleo de la intensidad no dependiente de la covariable, P. Diggle (1985), y otro que sí lo sea, Borrajo et al. (2017b); y para ello emplearemos una distancia del tipo L^2. Sin embargo, y debido a la falta de consistencia del estimador de Diggle sobre la que ya hemos hablado anteriormente, en lugar de emplear directamente estimadores de la intensidad, usaremos la densidad relativa, ya que la hipótesis nula puede ser transformada de manera equivalente en $H_0 : \lambda_0(x) = \rho(Z(x))/m$, con $\lambda_0(x) = \lambda(x)/m$ y $m = \int_W \lambda(x)dx$.

Así, el test estadístico es

$$S = \int_W \left(\hat{\lambda}_{0,H}(x) - \hat{\rho}_{0,b}(Z(x))\right)^2 dx, \tag{R.13}$$

donde $\hat{\lambda}_{0,H}(x) = \frac{1}{N p_H(x)} \sum_{i=1}^{N} K_H(x - X_i) 1_{\{N \neq 0\}}$ es la estimación bivariante de la densidad relativa propuesta en Fuentes-Santos et al. (2015), H es una matriz de ventanas, $p_H(x)$ es la corrección de efecto frontera, $\hat{\rho}_{0,b}(Z(x)) = \frac{\hat{\rho}_b(x)}{N} 1_{\{N \neq 0\}}$ con $\hat{\rho}_b(x) = \sum_{i=1}^{N} \frac{1}{g^\star(Z(X_i))} L_b(Z(x) - Z(X_i))$, b es un parámetro ventana, K y L funciones núcleo, $K_H(u) = |H|^{-1/2} K(H^{-1/2}u)$, $L_b(u) = \frac{1}{b} L\left(\frac{u}{b}\right)$ y g^* es la versión no normalizada de la derivada de la función de distribución espacial de Z, $G(z) = \int_W 1_{\{Z(u) \leq z\}} du$.

En el capítulo se detallan las hipótesis bajo las cuales se prueba la normalidad asintótica de (R.13), aunque debido a la baja velocidad de convergencia, se ha propuesto también un procedimiento bootstrap para mejorar la calibración del test. Se trata de un bootstrap suavizado, inspirado en Cao (1993) y Cowling et al. (1996): asumida la hipótesis nula y condicionalmente a una realización $X_1, \ldots, X_n$, sea $N^* \sim Pois\left(\int_W \hat{\rho}_t(Z(x))dx\right)$; se genera n^* una realización de la variable aleatoria

N^* y se generan $X_1^*, \ldots, X_n^*$ muestreando aleatoriamente n^* veces de la distribución con densidad proporcional a $\hat{\lambda}(x) = \hat{\rho}_t(Z(x))$. A continuación se computa el test estadístico bajo la distribución bootstrap y repitiendo este proceso B veces, con B suficientemente grande, conseguimos computar el cuantil correspondiente que nos servirá como regla de decisión en el test.

Aplicamos el test a los conjuntos de datos anteriormente presentados. Para la elección de la ventana hemos escogido un rango de valores apropiados que nos permite evaluar el correcto comportamiento del test. En los datos de Murchison obtenemos que, sin lugar a duda, la información dada por las fallas geológicas es suficiente para describir la distribución espacial de los depósitos de oro (p-valores entre 0.602 y 0.804 en el rango de ventanas). Sin embargo, en el caso de los incendios de Canadá la temperatura parece no ser información suficiente, y por consiguiente se debería o bien considerar otra variable meteorológica o idealmente intentar definir algún tipo de índice compendiando varias variables de manera que con el conjunto de todas ellas sí se pueda explicar la distribución de localizaciones de los incendios forestales.

En este capítulo se incluye además un estudio de simulación, con modelos basados en esos mismos datos reales. Las conclusiones extraídas son muy positivas, podemos ver como en ambos modelos el nivel se ajusta perfectamente para diferentes tamaños muestrales, y la potencia va aumentando según nos alejamos de la hipótesis nula, alcanzando valores próximos a 1 (en términos de proporción de rechazo) en situaciones no extremadamente distantes de la misma.

Capítulo 5: Comparación noparamétrica de funciones de intensidad de primer orden en procesos puntuales espaciales con covariable

En los dos capítulos anteriores, se ha establecido un modelo y desarrollado un completo marco teórico en torno a él: se ha abordado tanto la estimación de la intensidad como la bondad de ajuste. En este capítulo se plantea una continuación natural que es el contraste de dos muestras, esto es, asumido y testeado apropiadamente el modelo, disponemos de dos muestras (dos realizaciones de procesos puntuales espaciales con covariable) y queremos determinar si proceden exactamente del mismo modelo, es decir, si son realizaciones de una misma función de intensidad.

Para abordar este nuevo problema, hemos utilizado una metodología similar a la presentada en el Capítulo 4, es decir, construiremos un estadístico de contraste basado en la distancia L^2 entre los estimadores tipo núcleo de las intensidades de cada muestra. En este caso emplearemos en ambas muestras el estimador tipo núcleo definido en el

Capítulo 3. Formalmente, la distancia es

$$D = \int_{\mathbb{R}} \left(f_1(z) - f_2(z)\right)^2 dz = \int f_1^2(z)dz + \int f_2^2(z)dz - \int f_1(z)f_2(z)dz - \int f_2(z)f_1(z)dz$$
$$= \mathbb{E}_{Z_1}\left[f_1(Z_1)\right] + \mathbb{E}_{Z_2}\left[f_2(Z_2)\right] - \mathbb{E}_{Z_2}\left[f_1(Z_2)\right] - \mathbb{E}_{Z_1}\left[f_2(Z_1)\right] \equiv \psi_{11} + \psi_{22} - \psi_{12} - \psi_{21}, \tag{R.14}$$

donde f_1 y f_2 son las densidades teóricas asociadas a cada muestra.

El estadístico de contraste, basado en (R.14), se define como

$$S = \widehat{\psi}_{11} + \widehat{\psi}_{22} - \widehat{\psi}_{12} - \widehat{\psi}_{21}, \tag{R.15}$$

siendo

$$\widehat{\psi}_{11} = \frac{1}{N_1^2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\frac{g^\star(Z_{1i})}{g^\star(Z_{1j})}K_{h_1}(Z_{1i} - Z_{1j})1_{\{N_1\neq 0\}},$$

$$\widehat{\psi}_{22} = \frac{1}{N_2^2}\sum_{i=1}^{N_2}\sum_{j=1}^{N_2}\frac{g^\star(Z_{2i})}{g^\star(Z_{2j})}K_{h_2}(Z_{2i} - Z_{2j})1_{\{N_1\neq 0\}},$$

$$\widehat{\psi}_{12} = \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(Z_{2j})}{g^\star(Z_{1i})}K_{h_1}(Z_{2j} - Z_{1i})1_{\{N_1\neq 0, N_2\neq 0\}},$$

$$\widehat{\psi}_{21} = \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\frac{g^\star(Z_{1i})}{g^\star(Z_{2j})}K_{h_2}(Z_{1i} - Z_{2j})1_{\{N_1\neq 0, N_2\neq 0\}},$$

con h_i ventanas escalares, K una función núcleo real y $1_{\{\cdot\}}$ denotando la función indicadora.

Además, en el manuscrito se detallan las hipótesis bajo las cuales se demuestra la normalidad asintótica de (R.15). Sin embargo, esta aproximación tiene de nuevo una tasa de convergencia bastante lenta, lo que hace que los resultados para muestras de tamaño pequeño o medio no sean especialmente buenos, es por ello que proponemos un procedimiento bootstrap que mejora la calibración del estadístico y proporciona resultados más fiables.

Se realiza también un estudio de simulación para poder analizar el comportamiento de nuestra propuesta. Los modelos empleados son los mismos que en el Capítulo 4, que se construyen a partir de los conjuntos de datos reales de Murchison y de Canadá, y que pueden verse en la Figura R.1 y Figura R.2.

Los resultados del citado estudio han sido muy favorables: hemos conseguido ajustar el nivel en ambos modelos para todos los tamaños muestrales esperados, y los valores de potencia son bastante altos, especialmente para el modelo sobre los datos de Murchison. Los valores específicos obtenidos pueden obtenerse en las tablas resumen incluidas en el Capítulo 5.

Bibliography

Ahmad, I. A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statistics & Probability Letters*, *22*(2), pp. 121–129. doi: 10.1016/0167-7152(94)00057-F

Alba-Fernández, M., Ariza-López, F., Jiménez-Gamero, M. D. and Rodríguez-Avi, J. (2016). On the similarity analysis of spatial patterns. *Spatial Statistics*, *18*, pp. 352–362. doi: 10.1016/j.spasta.2016.07.004

Altieri, L., Scott, E. M., Cocchi, D. and Illian, J. (2015). A changepoint analysis of spatio-temporal point processes. *Spatial Statistics*, *14*, 197–207. doi: 10.1016/j.spasta.2015.05.005

Andresen, M. A. (2009). Testing for similarity in area-based spatial patterns: a nonparametric Monte Carlo approach. *Applied Geography*, *29*(3), pp. 333–345. doi: 10.1016/j.apgeog.2008.12.004

Andrews, J. G., Ganti, R. K., Haenggi, M., Jindal, N. and Weber, S. (2010). A primer on spatial modeling and analysis in wireless networks. *IEEE Communications Magazine*, *48*(11), pp. 156-163. doi: 10.1109/MCOM.2010.5621983

Asgharian, M., M'Lan, C. E. and Wolfson, D. B. (2002). Length-biased sampling with right censoring: an unconditional approach. *Journal of the American Statistical Association*, *97*(457), pp. 201–209. doi: 10.1198/016214502753479347

Babu, G. J. and Feigelson, E. D. (1996). *Astrostatistics* (Vol. 3). CRC Press.

Baddeley, A., Chang, Y. M., Song, Y. and Turner, R. (2012). Nonparametric estimation of the dependence of a spatial point process on spatial covariates. *Statistics and Its Interface*, *5*, pp. 221-236. doi: 10.4310/SII.2012.v5.n2.a7

Baddeley, A., Møller, J. and Waagepetersen, R. (2000). Non and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, *54*(3), pp. 329–350. doi: 10.1111/1467-9574.00144

Baddeley, A., Rubak, E. and Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC Press.

Baddeley, A. and Van Lieshout, M. (1995). Area-interaction point processes. *Annals of the Institute of Statistical Mathematics*, *47*(4), pp. 601–619. doi: 10.1007/BF01856536

Barmi, H. E. and Simonoff, J. S. (2000). Transformation-based density estimation for weighted distributions. *Journal of Nonparametric Statistics*, *12*(6), pp. 861–878. doi: 10.1080/10485250008832838

Berman, M. and Diggle, P. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *Journal of the Royal Statistical Society. Series B*, *51*(1), pp. 81–92. Retrieved from `http://www.jstor.org/stable/2345843`

Bhattacharyya, B. B., Franklin, L. A. and Richardson, G. D. (1988). A comparioson of nonparametric unweighted and length-biased density estimation of fibres. *Communications in Statistics: Theory and Methods*, *17*(11), pp. 3629–3644. doi: 10.1080/03610928808829825

Borrajo, M. I., González-Manteiga, W. and Martínez-Miranda, M. D. (2017a). Bandwidth selection for kernel density estimation for length-biased data. *Journal of Nonparametric Statistics*, *29*(3), pp. 636-668. doi: http://dx.doi.org/10.1080/10485252.2017.1339309

Borrajo, M. I., González-Manteiga, W. and Martínez-Miranda, M. D. (2017b). Bootrapping kernel intensity estimation for nonhomogeneous point processes depending on spatial covariates. *Test (2nd review)*.

Borrajo, M. I., González-Manteiga, W. and Martínez-Miranda, M. D. (2017c). *Testing covariate significance in spatial point process intensity (submitted)*.

Bose, A. and Dutta, S. (2013). Density estimation using bootstrap bandwidth selector. *Statistics & Probability Letters*, *83*(1), pp. 245–256. doi: 10.1016/j.spl.2012.08.027

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, *71*(2), pp. 353–360. doi: 10.2307/2336252

Brooks, M. M. and Marron, J. S. (1991). Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions. *Stochastic Processes and their Applications*, *38*(1), pp. 157-165. doi: 10.1016/0304-4149(91)90076-O

Brunel, E., Comte, F. and Guilloux, A. (2009). Nonparametric density estimation in

presence of bias and censoring. *TEST*, *18*(1), pp. 166–194. doi: 10.1007/s11749-007-0075-5

Cabana, E. (1987). Affine processes: a test of isotropy based on level sets. *SIAM Journal on Applied Mathematics*, *47*(4), pp. 886–891. doi: 10.1137/0147059

Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, *18*(1), pp. 179–189. doi: 10.1007/BF02869528

Cao, R. (1990). *Aplicaciones y nuevos resultados del método bootstrap en la estimación no paramétrica de curvas* (Unpublished doctoral dissertation). Universidade de Santiago de Compostela.

Cao, R. (1993). Bootstrapping the mean integrated squared error. *Journal of Multivariate Analysis*, *45*(1), pp. 137–160. doi: 10.1006/jmva.1993.1030

Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis*, *17*(2), pp. 153–176. doi: 10.1016/0167-9473(92)00066-Z

Chakraborty, R. and Rao, C. R. (2000). Selection biases of samples and their resolutions. In P. K. Sen and C. R. Rao (Eds.), *Handbook of statistics* (Vol. 18, pp. 675–712). Elsevier.

Chesneau, C. (2010). Wavelet block thresholding for density estimation in the presence of bias. *Journal of the Korean Statistical Society*, *39*(1), pp. 43–53. doi: 10.1016/j.jkss.2009.03.004

Chiu, S. (1992). An automatic bandwidth selector for kernel density estimation. *Biometrika*, *79*(4), pp. 771–782. doi: 10.1093/biomet/79.4.771

Chiu, S. (2003). Spatial point pattern analysis by using Voronoi diagrams and Delaunay tessellations: a comparative study. *Biometrical journal*, *45*(3), pp. 367–376. doi: 10.1002/bimj.200390018

Collomb, G. (1976). *Estimation non parametrique de la regression par la méthode du noyau* (Unpublished doctoral dissertation). Université Paul Sabatier de Toulouse.

Comte, F. and Rebafka, T. (2016). Nonparametric weighted estimators for biased data. *Journal of Statistical Planning and Inference*, *174*, pp. 104–128. doi: 10.1016/j.jspi.2016.01.008

Cowling, A., Hall, P. and Phillips, M. J. (1996). Bootstrap confidence regions for the intensity of a poisson point process. *Journal of the American Statistical Association*,

91(436), pp. 1516–1524. doi: 10.2307/2291577

Cox, D. (2005). Some sampling problems in technology. In D. Hand and A. Herzberg (Eds.), *Selected statistical papers of Sir David Cox* (Vol. 1, pp. 81–92). Cambridge University Press.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Cucala, L. (2006). *Espacements bidimensionnels et données entachés d'erreurs dans l'analyse des procesus ponctuels spatiaux* (Doctoral dissertation, Université des Sciences de Toulouse I). Retrieved from `https://tel.archives-ouvertes.fr/tel-00135890/document`

Cutillo, L., De Feis, I., Nikolaidou, C. and Sapatinas, T. (2014). Wavelet density estimation for weighted data. *Journal of Statistical Planning and Inference*, *146*, pp. 1–19. doi: 10.1016/j.jspi.2013.09.015

Daley, D. J. and Vere-Jones, D. (1988). *An introduction to the theory of point processes*. Springer Verlag New York Inc.

Dasgupta, A. and Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, *93*(441), pp. 294–302. doi: 10.1080/01621459.1998.10474110

de Uña-Álvarez, J. (2004). Nonparametric estimation under length-biased sampling and type I censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics*, *56*(4), pp. 667–681. doi: 10.1007/BF02506482

Díaz-Avalos, C., Juan, P. and Mateu, J. (2013). Similarity measures of conditional intensity functions to test separability in multidimensional point processes. *Stochastic Environmental Research and Risk Assessment*, *27*(5), pp. 1193–1205. doi: 10.1007/s00477-012-0654-1

Díaz-Avalos, C., Juan, P. and Mateu, J. (2014). Significance tests for covariate-dependent trends in inhomogeneous spatio-temporal point processes. *Stochastic environmental research and risk assessment*, *28*(3), pp. 593–609. doi: 10.1007/s00477-013-0775-1

Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society. Series C*, *34*(2), pp. 138–147. doi: 10.2307/2347366

Diggle, P. and Marron, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association*, *83*, pp. 793–800. doi: 10.2307/2289308

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A*, *153*(3), pp. 349–362. doi: 10.2307/2982977

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC Press. doi: 10.1002/bimj.201400024

Diggle, P. J., Chetwynd, A. G., Häggkvist, R. and Morris, S. E. (1995). Second-order analysis of space-time clustering. *Statistical methods in medical research*, *4*(2), pp. 124–136. doi: 10.1177/096228029500400203

Di Iorio, T., Anello, F., Bommarito, C., Cacciani, M., Denjean, C., De Silvestri, L., ... others (2013). Long range transport of smoke particles from canadian forest fires to the mediterranean basin during june 2013. In *Agu fall meeting abstracts*.

Duong, T. (2013). Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics*, *25*(3), 635–645. doi: 10.1080/10485252.2013.810217

Duong, T., Goud, B. and Schauer, K. (2012). Closed-form density-based framework for automatic detection of cellular morphology changes. *Proceedings of the National Academy of Sciences*, *109*(22), pp. 8382–8387. doi: 10.1073/pnas.1117796109

Dutta, S. (2016). Cross-validation revisited. *Communications in Statistics-Simulation and Computation*, *45*(2), pp. 472–490. doi: 10.1080/03610918.2013.862275

Efromovich, S. (2004). Density estimation for biased data. *Annals of Statistics*, *32*(3), pp. 1137–1161. doi: 10.1214/009053604000000300

Federer, H. (1969). Geometric measure theory. In *Die grundlehren der mathematischen wissenschaften* (Vol. 153). Springer Verlag New York Inc. doi: 10.1007/978-3-642-62010-2

Fisher, R. A. (1935). *The design of experiments*. Olivier and Boyd.

Flannigan, M. and Harrington, J. (1988). A study of the relation of meteorological variables to monthly provincial area burned by wildfire in canada (1953–80). *Journal of Applied Meteorology*, *27*(4), pp. 441–452. doi: 10.1175/1520-0450(1988)027<0441:ASOTRO>2.0.CO;2

Foxall, R. and Baddeley, A. (2002). Nonparametric measures of association between a spatial point process and a random set with geological applications. *Journal of the Royal Statistical Society: Series C*, *51*(2), pp. 165–182.

Fuentes-Santos, I., González-Manteiga, W. and Mateu, J. (2015). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. *Scandinavian Journal of Statistics*, *43*(2), pp. 416-435. doi: 10.1111/sjos.12183

Fuentes-Santos, I., González-Manteiga, W. and Mateu, J. (2017). A nonparametric test for the comparison of first-order structures of spatial point processes. *Spatial Statistics*. doi: 10.1016/j.spasta.2017.02.007

Gabriel, E. (2014). Estimating second-order characteristics of inhomogeneous spatio-temporal point processes. *Methodology and Computing in Applied Probability*, *16*(2), pp. 411–431. doi: 10.1007/s11009-013-9358-3

Gabriel, E. and Diggle, P. J. (2009). Second-order analysis of inhomogeneous spatio-temporal point process data. *Statistica Neerlandica*, *63*(1), pp. 43–51. doi: 10.1111/j.1467-9574.2008.00407.x

Gatrell, A. C., Bailey, T. C., Diggle, P. J. and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, *21*(1), pp. 256–274. doi: 10.2307/622936

González, J. A., Rodríguez-Cortés, F. J., Cronie, O. and Mateu, J. (2016). Spatio-temporal point process statistics: a review. *Spatial Statistics*, *18*, pp. 505–544. doi: 10.1016/j.spasta.2016.10.002

González-Manteiga, W., Martínez-Miranda, M. D. and Pérez-González, A. (2004). The choice of smoothing parameter in nonparametric regression through wild bootstrap. *Computational Statistics and Data Analysis*, *47*(3), pp. 487–515. doi: 10.1016/j.csda.2003.12.007

Guan, Y. (2008). On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. *Journal of the American Statistical Association*, *103*(483), pp. 1238–1247. doi: 10.1198/016214508000000526

Guan, Y., Sherman, M. and Calvin, J. A. (2006). Assessing isotropy for spatial point processes. *Biometrics*, *62*(1), pp. 119–125. doi: 10.1111/j.1541-0420.2005.00436.x

Guillamón, A., Navarro, J. and Ruiz, J. M. (1998). Kernel density estimation using weighted data. *Communications in Statistics-Theory and Methods*, *27*(9), pp. 2123–2135. doi: 10.1080/03610929808832217

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of multivariate analysis*, *14*(1), pp. 1–16. doi: 10.1016/0047-259X(84)90044-7

Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, *6*(2), pp. 109–115. doi: 10.1016/0167-7152(87)90083-6

Halley, E. (1686). An historical account of trade winds and monsoons observable in the seas between and near the tropicks with an attempt to assign the phisical cause of said winds. *Philosophical Transactions*, *183*, pp. 153-168. doi: 10.1098/rstl.1686.0026

Heckman, J. J. (1990). Selection bias and self-selection. In *Econometrics* (pp. 201–224). Springer.

Heidenreich, N.-B., Schindler, A. and Sperlich, S. (2013). Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *Advances in Statistical Analysis*, *97*(4), pp. 403–433. doi: 10.1007/s10182-013-0216-y

Illian, J., Møller, J. and Waagepetersen, R. P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, *16*(3), pp. 389–405. doi: 10.1007/s10651-007-0070-8

Illian, J., Penttinen, A., Stoyan, H. and Dietrich, S. (2008). *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons. doi: 10.1002/9780470725160

Illian, J., Sørbye, S. H. and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). *The Annals of Applied Statistics*, *6*(4), pp. 1499–1530. Retrieved from `http://www.jstor.org/stable/41713484`

Isaaks, E. H. and Srivastava, R. M. (1989). *An introduction to applied geostatistics*. Oxford University Press.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika*, *78*(3), pp. 511–519. doi: 10.1093/biomet/78.3.511

Jones, M. C. and Karunamuni, R. J. (1997). Fourier series estimation for length biased data. *Australian Journal of Statistics*, *39*(1), pp. 57–68. doi: 10.1111/j.1467-842X.1997.tb00523.x

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, *91*(433), pp. 401–407. doi: 10.1080/01621459.1996.10476701

Kerscher, M. (2000). Statistical analysis of large-scale structure in the universe. In *Statistical physics and spatial statistics* (Vol. 554, pp. 36–71). Springer. doi: 10.1007/3-540-45043-2_3

Krkošek, M., Lewis, M. A. and Volpe, J. P. (2005). Transmission dynamics of parasitic sea lice from farm to wild salmon. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1564), pp. 689–696. doi: 10.1098/rspb.2004.3027

Law, R., Illian, J., Burslem, D. F., Gratzer, G., Gunatilleke, C. and Gunatilleke, I. (2009). Ecological information from spatial patterns of plants: insights from point process theory. *Journal of Ecology*, *97*(4), pp. 616–628. doi: 10.1111/j.1365-2745 .2009.01510.x

Lawson, A. B. (2013). *Statistical methods in spatial epidemiology*. John Wiley & Sons. doi: 10.1002/9780470035771.biblio

Lewis, P. A. and Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics (NRL)*, *26*(3), pp. 403–413. doi: 10.1002/nav.3800260304

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P. and Sperlich, S. (2011). Do-validation for kernel density estimation. *Journal of the American Statistical Association*, *106*(494), pp. 651–660. doi: 10.1198/jasa.2011.tm08687

Mammen, E., Martínez-Miranda, M. D., Nielsen, J. P. and Sperlich, S. (2014). Further theoretical and practical insight to the Do-validated bandwidth selector. *Journal of the Korean Statistical Society*, *43*(3), pp. 355–365. doi: 10.1016/j.jkss.2013.11.001

Marron, J. S. (1988). Automatic smoothing parameter selection: a survey. *Empirical Economics*, *13*(3-4), pp. 187–208. doi: 10.1007/BF01972448

Marron, J. S. (1992). Bootstrap bandwidth selection. In R. Lepage and L. Billard (Eds.), *Exploring the limits of bootstrap* (pp. 249–262). Wiley.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics*, *20*(2), pp. 712–736. Retrieved from `http://www.jstor.org/stable/2241980`

Moller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press. doi: 10.1002/sim.1896

Mrkvička, T. and Soubeyrand, S. (2017). On parameter estimation for doubly inhomogeneous cluster point processes. *Spatial Statistics*, *20*, pp. 191–205. doi: 10.1016/j.spasta.2017.03.005

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, *83*(401), pp. 9–27. doi: 10.1080/01621459.1988.10478560

Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, *50*(2), pp. 379–402. doi: 10.1023/A:1003403601725

Ogata, Y. and Zhuang, J. (2006). Space–time ETAS models and an improved extension. *Tectonophysics*, *413*(1), pp. 13–23. doi: 10.1016/j.tecto.2005.10.016

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*(3), pp. 1065–1076. doi: 10.1214/aoms/1177704472

Patil, G. P. and Rao, C. R. (1977). The weighted distributions: a survey of their applications. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 383–405). North Holland.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ramírez, P. and Vidakovic, B. (2010). Wavelet density estimation for stratified size-biased sample. *Journal of Statistical Planning and Inference*, *140*(2), pp. 419–432. doi: 10.1016/j.jspi.2009.07.021

Reiss, R. D. (2012). *A course on point processes*. Springer Science & Business Media. doi: 10.1007/978-1-4613-9308-5

Reitzner, M. and Schulte, M. (2013). Central limit theorems for U-statistics of poisson point processes. *The Annals of Probability*, *41*(6), pp. 3879–3909. doi: 10.1214/12-AOP817

Richardson, G. D., Kazempour, M. K. and Bhattacharyya, B. (1991). Length biased density estimation of fibres. *Journal of Nonparametric Statistics*, *1*(1-2), pp. 127–141. doi: 10.1080/10485259108832515

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B*, *39*(2), pp. 172–212. Retrieved from `http://www.jstor.org/stable/2984796`

Rogers, B., Randerson, J. and Bonan, G. (2013). High-latitude cooling associated with landscape changes from north american boreal forest fires. *Biogeosciences*, *10*(2), pp. 699–718. doi: 10.5194/bg-10-699-2013

Rosenberg, M. (2004). Wavelet analysis for detecting anisotropy in point patterns. *Journal of Vegetation Science*, *15*(2), pp. 277–284. doi: 10.1658/1100-9233(2004)

015[0277:WAFDAI]2.0.CO;2

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), pp. 832–837. doi: 10.1007/978-1-4419-8339-8_13

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, *9*(2), pp. 65–78. Retrieved from `http://www.jstor.org/stable/4615859`

Schabenberger, O. and Gotway, C. A. (2017). *Statistical methods for spatial data analysis*. CRC press.

Schoenberg, F. P. (2004). Testing separability in spatial-temporal marked point processes. *Biometrics*, *60*(2), pp. 471–481. doi: 10.1111/j.0006-341X.2004.00192.x

Schoenberg, F. P. (2005). Consistent parametric estimation of the intensity of a spatial–temporal point process. *Journal of Statistical Planning and Inference*, *128*(1), pp. 79–93. doi: 10.1016/j.jspi.2003.09.027

Schoenberg, F. P. (2011). Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, *98*, pp. 789-795. doi: 10.1198/016214503000000710

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice and visualisation*. Wiley.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B*, *53*(3), pp. 683–690. Retrieved from `http://www.jstor.org/stable/2345597`

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. CRC press. doi: 10.1002/bimj.4710300745

Simon, R. (1980). Length biased sampling in etiologic studies. *American Journal of Epidemiology*, *111*(4), pp. 444–452. doi: 10.1093/oxfordjournals.aje.a112920

Stoyan, D., Kendall, W. and Mecke, J. (1995). *Stochastic geometry and its applications*. John Wiley & Sons, Chichester.

Stoyan, D. and Penttinen, A. (2000). Recent applications of point process methods in forestry statistics. *Statistical Science*, *15*(1), pp. 61–78. doi: 10.1214/ss/1009212674

Student. (1907). On the error of counting with haemacytometer. *Biometrika*, *5*, pp.

351-360. doi: 10.1093/biomet/5.3.351

Turlach, B. A. (1993). *Bandwidth selection in kernel density estimation: A review* (Tech. Rep.). B-1348 Louvain-la-neuve, Belgium: Université Catholique de Louvain.

Van Lieshout, M. (2000). *Markov point processes and their applications.* World Scientific. doi: 10.1142/9781860949760

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, *10*(2), pp. 616–620. doi: 10.1214/aos/1176345802

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, *13*(1), pp. 178–203. doi: 10.1214/aos/1176346585

Vilar-Fernández, J. M., Vilar-Fernández, J. A. and González-Manteiga, W. (2007). Bootstrap tests for nonparametric comparison of regression curves with dependent errors. *TEST*, *16*(1), pp. 123–144.

Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics*, *63*(1), pp. 252–258. doi: 10.1111/j.1541-0420.2006.00667.x

Walter, C., Freitas, S., Kraut, I., Rieger, D., Vogel, H. and Vogel, B. (2014). Influence of 2010 canadian forest fires on cloud formation on the regional scale. In *Agu fall meeting abstracts.*

Watkins, K. P. and Hickman, A. H. (1990). *Geological evolution and mineralization of the Murchison Province, Western Australia* (Vol. 1). Department of Mines, Western Australia.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, *56*(3), pp. 601–614. doi: 10.1093/biomet/56.3.601

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Notation index

$1_{\{\cdot\}}$	indicator function
$\lvert A\rvert$	area or volume of a set A
$AMISE$	asymptotic mean integrated squared error
$AMSE$	asymptotic mean squared error
Cov	covariance
CSR	complete spatial randomness
CSTR	complete spatiotemporal randomness
CV	cross validation
E	expectation
E^*	expectation under the bootstrap distribution
e_i	error measure based on the $\mathrm{ISE}_{\mathrm{rel}}$ criterion
f	density function
$\hat{f}_h$	kernel density estimator
G	spatial cumulative distribution function
$G^\star$	unnormalised version of the spatial cumulative distribution function
g	spatial density function (G')
$g^\star$	unnormalised version of the spatial density function
g_2	pair correlation function
$Gamma(a, b)$	Gamma distribution with parameters a and b
H	bandwidth matrix
h_{AMISE}	optimal bandwidth parameter for the AMISE criterion
h_{AMSE}	optimal bandwidth parameter for the AMSE criterion
ISE	integrated squared error
$\mathrm{ISE}_{\mathrm{rel}}$	relative integrated squared error
K, L	kernel functions
J	Jacobian

$\|\| \cdot \|\|$	L^2-norm
λ	first-order intensity function
λ_0	density function associated to intensity function λ
λ_2	second-order intensity function
λ_c	conditional first-order intensity function
$\mathcal{M}$	space of the marks
m	expected sample size
$MISE$	mean integrated squared error
MSE	mean squared error
μ_r	r-th order moment
μ_S	mean value of the statistic S
n	deterministic sample size
N	counting measure
$N_0(r)$	number of events within distance r of an arbitrary event
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
∇	gradient
$\mathcal{P}(A)$	parts of set A
$Pois(\cdot)$	Poisson distribution
$\mathbb{R}^d$	d-dimensional euclidean space
σ_S	standard deviation of the statistic S
σ_S^2	variance of the statistic S
S	test statistic
Var	variance
Var*	variance under the bootstrap distribution
W	observation region
X	point process
$X_1, \ldots, X_N$	original spatial point pattern
$Z(\cdot)$	spatially varying continuous one-dimensional covariate
$\mathbf{Z}(\cdot)$	spatially varying continuous multidimensional covariate
$Z_1, \ldots, Z_N$	transformed point pattern through covariate Z

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA