UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques
i Informàtica

GRAU DE MATEMÀTIQUES

Treball final de grau

Spatial point processes: from the mathematical basis to its applications

Autor: Arnau García

Director:	**Dr. Carles Rovira Escofet**
Codirector:	**Dr. Jorge Mateu Mahiques**
Realitzat a:	**Departament de Matemàtiques i Informàtica**

Barcelona, **12 de juny de 2023**

Abstract

This work is a study about the spatial point processes. We study the mathematical basis of this object, we expose statistic tools which are used in the analysis of spatial point patterns and, finally, we apply all the exposed theory in a real case study with real data.

In the first and second chapter we present the mathematical theory behind the spatial point processes. In the starting chapter we find the most general and abstract definitions, and the very definition of a spatial point process. In this chapter we have used Stoyan et al. 2013 [17]. In the second chapter, using as a reference Diggle 2013 [10], we explain the mathematical theory of the point processes in tha plane. We define and study the properties of several types of processes, and different quantities which are hugely important in the study of this kind of objects.

In the third chapter, based mainly in Baddeley et al. (2015) [2], we present, giving examples, the statistic tools used in the analysis of point processes in the plane. The tools exposed are related with the theory exposed previously and are used in the last chapter of the project.

Finally, in the last chapter, we put into practice all the knowledge we have acquired in a real case study. Using the database employed in Jorge Mateu, P. Diggle and I. Tamayo-Uria (2014) [18], shared by Jorge Mateu, we perform a study about the rat and cockroach sightings in Madrid city. This constitutes an application in a real public health case of the concepts seen during the work.

Resum

En aquesta memòria s'estudia un tipus de procés estocàstic en concret, es tracta dels processos estocàstics puntals espacials. S'estudien les bases matemàtiques d'aquest objecte, s'exposen eines estadístiques que es fan servir per tractar problemes amb patrons puntuals espacials, i finalment, es posa tota la teoria exposada en pràctica amb una aplicació en un problema real, amb dades reals.

En el primer i segon capítol del treball es dona la teoria matemàtica darrere dels processos puntuals espacials. En el capítol que obre el projecte, hi trobem les definicions més generals i abstractes, i la mateixa definició de procés puntual. Hem seguit en aquest capítol Stoyan et al. 2013 [17]. En el segon capítol, usant com a referència Diggle 2013 [10], hi ha explicada la teoria matemàtica sobre processos puntuals espacials en el pla. Es defineixen i s'estudien les propietats de diferents tipus de processos, i de diferents quantitats que resulten clau en l'estudi d'aquest tipus d'objectes.

En el tercer capítol, basat principalment en Baddeley et al. (2015) [2], es presenten, tot donant exemples, eines estadístiques que es fan servir per a l'estudi de patrons puntuals a regions del pla. Les eines exposades lliguen amb la teoria tractada en el segon capítol i es fan servir en el darrer capítol d'aquest treball.

Finalment, en el darrer capítol del treball, posem en pràctica tots els coneixements adquirits i exposats durant les seccions anteriors en un cas d'estudi real. Utilitzant la base de dades emprada a Jorge Mateu, P. Diggle i I. Tamayo-Uria (2014) [18], dades que m'han estat cedides pel mateix Jorge Mateu, duem a terme un estudi sobre els avistaments de rates i paneroles a la ciutat de Madrid. Això, constitueix una aplicació en un cas real de salut pública dels conceptes vists durant el treball.

2020 Mathematics Subject Classification. 60G55, 60G10, 62P10, 62M30.

Acknowledgments

Firstly, I would like to thank my director Carles Rovira for his availability, support, and the freedom he has given me to develop this work. I would like to make a special mention to Jorge Mateu, who has helped me in a totally disinterested way throughout the work and has lent me the database of one of his papers. Without Jorge, this work as it is presented here today, would not have been possible.

Of course, I thank my family for their support throughout my studies. Especially my grandfather, to whom I hope to dedicate many more works. I also thank especially Maria, for being by my side and supporting me day by day. Many hours of work have been with her by my side.

Finally, I would like to thank my faculty friends Òscar Burés and Néstor Zafrilla and all my friends Sergi, Adrià, Pablo, Pau, Mario, Cesc, Nil and Jordi. They are the family I have chosen, and they are part of everything I do.

Arnau Garcia

Contents

Chapter 1

Point processes general theory

This chapter will be strongly based in the point process theory given in the book '*Stochastic Geometry and Its Applications*' (Stoyan et al., 2013 [17]). In this book one can find an exhaustive study around point processes with a pure mathematical approach. Then, I use this book in order to be familiar with the mathematics behind this kind of processes. I add this chapter on the project with the aim of give some mathematics fundamentals before work in the practical cases.

The objective in this chapter is give a formal mathematical theory about the so-called point processes. Here we are working in an Euclidean space $\mathbb{R}^d$, in this space we will have a pattern of points. In this context we denote by φ a sequence of points in $\mathbb{R}^d$. Sometimes we write $\varphi = \{x_n\}$, were x_n is a point of $\mathbb{R}^d$, in order to emphasize the sequential nature of φ.

The first step will be define point process. Prior to define this, we should focus on some key concepts. In the mathematical literature a point process in $\mathbb{R}^d$ is defined as a random variable taking values in some measurable space. Then, first of all, we should define this measurable space.

Definition 1.1. We define/denote $\mathbb{N}$ to the family of all the sequences φ of points in $\mathbb{R}^d$ satisfying:

(i) The sequence φ is locally finite. That is, each bounded subset of $\mathbb{R}^d$ must contain only a finite number of points of φ.

(ii) The sequence φ is simple. That is, there are not two equal points in φ.

Definition 1.2. We define/denote as $\mathcal{N}$ to the smallest σ-algebra on $\mathbb{N}$ to make all mappings $\varphi \to \varphi(B)$ measurable, for B running through the bounded Borel sets. Here, $\varphi(B)$ is the number of points in the set B.

Now we have a well defined measurable space, that is the pair $(\mathbb{N}, \mathcal{N})$. This is the measurable space that we need to be able to define, formally, a point process.

Definition 1.3. A point process Φ is a measurable mapping of a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ into the measurable space $(\mathbb{N}, \mathcal{N})$.

More intuitively a point process is a random choice of one of the sequences φ in $\mathbb{N}$. It generates a distribution on $(\mathbb{N}, \mathcal{N})$, the *distribution* P of the point process.

Notation 1.4. Is important to introduce the following notation:

$x \in \Phi$ asserts that the point x belongs to the random sequence Φ.

$\Phi(B) = n$ asserts that the set B contains n points of Φ.

Observation 1.5. The word 'process' in the term 'point process' does not imply a dynamic evolution over time. Because of the classical definition of a stochastic process, and also the applications of stochastic processes, it is natural to think that time should be also related with this type of stochastic processes. But this is not the case. The pure spatial point process are not related with time. Notwithstanding, there are the spatio-temporal point processes, which include the time variable as a new dimension. These processes are really related with time, and with a dynamic evolution. These kind of point processes are out of the scope of this work. Anyway, interesting references about this topic are: Jorge Mateu et al. 2016 [13], and the last chapters of Dggile 2013, [10].

Henceforth, the point process Φ is from the probability space $(\Omega, \mathcal{A}, \mathbf{P})$ to the measurable space $(\mathbb{N}, \mathcal{N})$.

Our next step is to define the distribution of a point process. We commented that the definition of points processes generates a distribution.

Definition 1.6. The *distribution* P of a point process Φ is the distribution determined by the probabilities

$$P(Y) = \mathbf{P}(\Phi \in Y) = \mathbf{P}(\{\omega \in \Omega : \Phi(\omega) \in Y\}).$$

Where $Y \in \mathcal{N}$.

Observation 1.7. Let Φ be a point process and $Y \in \mathcal{N}$. Then the term $\Phi \in Y$ means that Φ has some property, for example Φ has no point in the set B. Then $\mathbf{P}(\Phi \in Y)$ denotes the probability that Φ has this property, in the previous example it is the probability that Φ has no point in B.

Another important type of distributions are the so-called *finite-dimensional distributions*. These distributions are used in regular basis in all type of stochastic processes, see in the *'Stochastic processes course'*, of the *Universitat de Barcelona*, notes: Carles Rovira 2021 [16]. Now we see the definition:

Definition 1.8. Let Φ be a point process, we define the *finite-dimensional distributions* as the probabilities of the form

$$\mathbf{P}(\Phi(B_1) = n_1, \ldots, \Phi(B_k) = n_k). \tag{1.1}$$

Where $B_1, \ldots, B_k$ are Borel bounded sets and $n_1, \ldots, n_k \geq 0$.

Here (1.1) denotes the probability that Φ has n_1 points in the set B_1, ..., and n_k points in B_k.

Observation 1.9. The distribution of Φ on $(\mathbb{N}, \mathcal{N})$ is uniquely determined by the system of all this values for all $k = 1, 2, \ldots$. In fact, the distribution is determined by the subsystem for which the constituent B_i are pairwise disjoint.

Two important notions related with point processes are stationarity and isotropy. A point process is said to be *stationary* if its characteristics are invariant under translation. On the other hand, a point process is said to be *isotropic* if its characteristics are invariant under rotation. Now, we see this definitions formally. First, we should introduce some notation.

Notation 1.10. Let Φ be a point process. Recall that we can write $\Phi = \{x_n\}$, because a point process is a sequence of points on $\mathbb{R}^d$. Then, we denote the translation of Φ like $\Phi_x = \{x_n + x\}$, for some $x \in \mathbb{R}^d$. Let $\mathbf{r}$ be a rotation around the origin. We denote the rotation of Φ like $\mathbf{r}\Phi$.

Definition 1.11. A point process $\Phi = \{x_n\}$ is *stationary* if it has the same distribution as the process $\Phi_x = \{x_n + x\}$ for all x in $\mathbb{R}^d$. So

$$\mathbf{P}(\Phi \in Y) = \mathbf{P}(\Phi_x \in Y).$$

For all $Y \in \mathcal{N}$ and for all $x \in \mathbb{R}^d$.

Definition 1.12. A point process $\Phi = \{x_n\}$ is *isotropic* if Φ and $\mathbf{r}\Phi$ have the same distribution for every rotation around the origin $\mathbf{r}$. So

$$\mathbf{P}(\Phi \in Y) = \mathbf{P}(\mathbf{r}\Phi \in Y).$$

For all $Y \in \mathcal{N}$ and for all $\mathbf{r}$.

Observation 1.13. Notice that if a given point process Φ is stationary and isotropic, then the properties of the process only depends on the distance between points. Because the properties of the process do not change under translation and rotation.

On the point processes context one have the analogous to the mean of a real-valued random variable, this is the *intensity measure* of a point process. Like the mean of a real-valued random variable, the intensity measure is a very important quantity.

Definition 1.14. The *intensity measure* Λ of a point process Φ is defined as

$$\Lambda(B) = E(\Phi(B)) = \int \varphi(B) P(d\varphi).$$

For Borel sets B. So $\Lambda(B)$ is the mean number of points in B.

With the previous definition the chapter is closed. Now, our objective is find powerful tools for study point processes and see different type of point processes. With this objective in mind in the following sections we will restrict the study to the plane, where the analysis is easier and where exists a great variety of potent tools.

Chapter 2

Point processes in the plane

2.1 Introduction

In this part of the project we will study in detail the point processes in the plane (i.e. in $\mathbb{R}^2$). This chapter is based on '*Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*' by Diggle 2013 [10]. Everyone I have talked about my goals in this project recommended me this book. Then, Diggle's book will be the most important reference in this part of the project.

From now on we will be working in the Euclidean space $\mathbb{R}^2$, with the Euclidean topology. It is also important to mention that until now we use the term '*point process*' to refer to spatial-point processes. From now on the reader will see the term '*point pattern*' usually. It is important to understand the difference beetwen point pattern and point process. A point pattern is a set of observed points, and a point process is the mathematical object defined in Chapter 1. During all the work the planar regions with which we will work will be Borel bounded sets.

2.1.1 Software

In this work, we will use the R software. In this project, we will be working with the *Spatstat* library, written by Adrian Baddeley and Rolf Turner. This R library has got a big amount of useful tools for the point patterns analysis. I am using the book '*Spatial Point Patterns. Methodology and Applications with R*' (Baddeley et al. 2015 [2]) in order to get acquainted with the resolution of point pattern problems using R. Of course, I am using the manuals of the *Spatstat* library to learn how to use the library correctly and be able to do properly the statistical analysis of point patterns (see this manuals in [3], [5] and [4]).

Furthermore, I will be using several R libraries which are necessary for the practical part of the work. These packages and each manuals are: '*gstat*' (Pebesma

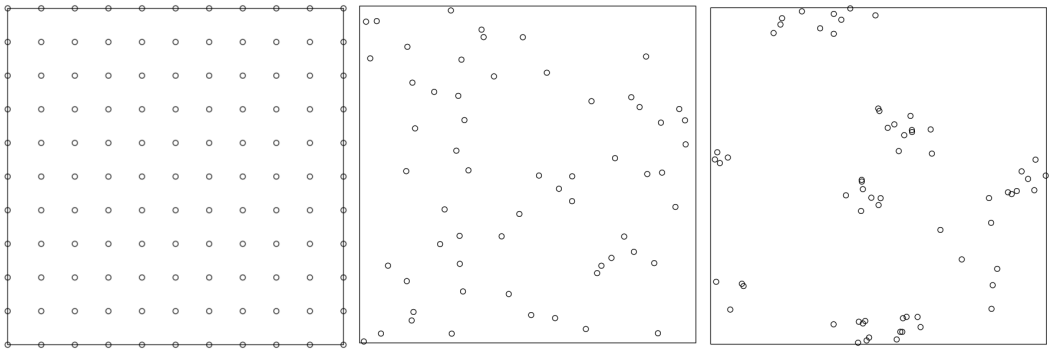

Figure 2.1: Regular (left), independent (middle) and clustered (right) point patterns of 121, 100 and 100 points.

2015, [15]), '*rgdal*' (Bivand 2015, [7]), '*splancs*' (Bivand 2017, [8]), '*spdep*' (Bivand 2015, [6]).

2.1.2 Complete spatial randomness

In this section we will introduce the most important hypothesis in the point processes analysis, it is the hypothesis of *complete spatial randomness* (CSR). The hypothesis of complete spatial randomness for a spatial point pattern asserts that

(i) The number of points of the pattern in any planar region A with area $m_2(A)$ follows a Poisson distribution with mean $\lambda m_2(A)$.

(ii) Given n points of the pattern x_i in a region A, the x_i are an independent random sample from the uniform distribution on A.

Where $m_2(\cdot)$ denotes the Lebesgue measure. The constant λ is the so-called *intensity*. The hypothesis (i) implies that the intensity of points not vary over the plane, (ii) implies that there are no interactions amongst the points. In the Figure 2.1 the reader can see a regular (left), under CSR hypothesis (middle) and aggregated (right) pattern. In this example one can easily understand the behavior of a point pattern under the CSR hypothesis.

It is always a starting point, in the practical cases, to verify if our pattern satisfies (i) and (ii), or not. Because if one point pattern is under CSR hypothesis then it is a completely random pattern. That is, our pattern has no study interest because it is absolutely random, and then we can't draw conclusions about the underlying process.

In addition, the complete spatial randomness hypothesis have some other applications: they are used for the simulation of other more complex models, also,

they are used like a dividing hypothesis to distinguish between different types of patterns.

2.2 Second-order properties

The *summary descriptions* are quantities that we can assign to each point pattern. This quantities allows us to understand how is the point pattern in question, for example, if the point pattern has some clustering, if the points are (or not) uniformly distributed throughout the planar region, the intensity of points per unit area, etc.

In the following sections we will study these summary descriptions. One useful, and usually used, approach for the statistical analysis of point patterns is to compare the empirical summary descriptions between the theoretical summary descriptions.

We can now define the *first-order* and *second-order* properties of a point process. First-order properties are described by an *intensity function*. The intensity function is an interesting quantity because shows us the average number of points per unit of area.

Definition 2.1. Let Φ be a point process in $\mathbb{R}^2$, we define the *intensity function* of Φ as

$$\lambda(x) = \lim_{m_2(dx)\to 0} \frac{E(\Phi(dx))}{m_2(dx)}.$$

Whenever the limit above exists and where $x \in \mathbb{R}^2$.

Notation 2.2. In the previous definition dx denotes an infinitesimal planar region that contains the point $x \in \mathbb{R}^2$. From now on we will be using this notation. The reader can think in dx as $\mathbf{B}(x, \epsilon)$, the ball (i.e. disc) centered in x with radius ϵ, which is arbitrary small.

A variant of the intensity function, which is the first second-order property that we see, is the so-called *second-order intensity function*. It is defined in a similar way than the intensity function:

Definition 2.3. Let Φ be a point process in $\mathbb{R}^2$, we define the *second-order intensity function* of Φ like (when the limit exists)

$$\lambda_2(x, y) = \lim_{m_2(dx), m_2(dy)\to 0} \frac{E(\Phi(dx)\Phi(dy))}{m_2(dx)m_2(dy)}.$$

We will usually be dealing with processes that are *stationary* (see Definition 1.11) and *isotropic* (see Definition 1.12). For the study in finite planar regions it is not an overly restrictive condition.

Observation 2.4. For a stationary process the intensity function assumes a constant value, it is $\lambda(x) = \lambda$. In addition, for stationary processes we have $\lambda_2(x, y) \equiv \lambda_2(x - y)$. And for a stationary and isotropic process $\lambda_2(x - y)$ reduces further to $\lambda_2(t)$ where $t = \|x - y\|$ ($\|\cdot\|$ denotes the Euclidean distance in $\mathbb{R}^2$). It is because (see Observation 1.13) under isotropy and stationarity the properties of the point process only depends on the distances between points. The reader can observe that we are dealing with several abuse of notation. Strictly, it have no sense the expression $\lambda_2(x - y)$ or $\lambda_2(t)$. Because by definition we have $\lambda_2 : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ and $x - y \in \mathbb{R}^2$, $t \in \mathbb{R}$. Usually we will use $\lambda_2(t)$, because we will be dealing with stationary and isotropic processes.

The *conditional intensity* is another important quantity. Heuristically the conditional intensity corresponds to the intensity in $x \in \mathbb{R}^2$ conditional on the information that there is a point of the process at $y \in \mathbb{R}^2$, thus $y \in \mathbb{R}^2 \cap \Phi$.

Definition 2.5. Let Φ be a point process. Assume that $y \in \mathbb{R}^2 \cap \Phi$, then we define the *conditional intensity* at $x \in \mathbb{R}^2$ as

$$\lambda_c(x|y) = \frac{\lambda_2(x, y)}{\lambda(y)}.$$

An other characterisation of the second-order properties of stationary and isotropic processes is given for the *K-function*. This is an important quantity, very useful in several statistical approaches for analyze case studies.

Definition 2.6. Let Φ be a stationary and isotropic point process in $\mathbb{R}^2$ with intensity λ, we define the function $K(t)$ as

$$K(t) = \frac{E(\Phi(\mathbf{B}(*, t)))}{\lambda}.$$

Where $\Phi(\mathbf{B}(*, t))$ is the number of points within the ball of radius t, centered in an arbitrary point of the pattern.

In the mathematical literature we can find (as in Stoyan et al. 2013, [17]) the previous definition with $\mathbf{B}(o, t)$ instead of $\mathbf{B}(*, t)$, where o denotes the origin. This is because we are assuming stationarity and isotropy, and then we can apply a translation and a rotation and take o as a point of the process.

From the previous definition (Definition 2.6) we can deduce that $\lambda K(t)$ is the mean number of further points in a ball of radius t and centered at the typical point.

Now, we want to establish a link between the K-functions and the second order intensity. From now on we assume stationarity and isotropy. In order to establish

a link we have to assume that our process is simple (see Definition 1.1). In the mathematical literature the reader can find this notion called *orderly* (for example in Diggle 2013, [10]). The reader can see that this assumption is not very restrictive in practice, because, normally, the coincidence of two (or more) points does not make physic sense. Under these conditions, the expected number of further points within distance t of an arbitrary point can be computed by integrating the conditional intensity over the disc with centre the origin and radius t. Thus,

$$K(t) = \frac{1}{\lambda} \int_0^{2\pi} \int_0^t \lambda_c(s|o) s ds d\theta. \tag{2.1}$$

Where $o \in \mathbb{R}^2$ is the origin.

Now, remember that we are assuming statonarity and isotropy, then if we use what we saw in Observation 2.4 and we use the conditional intensity definition (see Definition 2.5) we can obtain:

$$\lambda_c(t|o) = \frac{\lambda_2(t,o)}{\lambda(o)} = \frac{\lambda_2(t)}{\lambda}. \tag{2.2}$$

And then, if we replace (2.2) on (2.1), and we apply the Fubini's theorem we finally have:

$$K(t) = \frac{1}{\lambda} \int_0^{2\pi} \int_0^t \frac{\lambda_2(s)}{\lambda} s ds d\theta = \frac{2\pi}{\lambda^2} \int_0^t \lambda_2(s) s ds. \tag{2.3}$$

Thus, we obtained a link between $K(t)$ and $\lambda_2(t)$ that may be useful in order to study stationary, isotropic and orderly processes.

An important and useful property of the K-function is that it is invariant under random thinning. By random thinning, we mean that the point of a process is retained or not according to a series of mutually independent Bernoulli trials. This result follows from the Definition 2.6, where we have that the K-function is

$$K(t) = \frac{E(\Phi(\mathbf{B}(*,t)))}{\lambda}.$$

The effect of thinning is to multiply p (the parameter of the Bernoulli trials) above and below. It is:

$$K(t) = \frac{pE(\Phi(\mathbf{B}(*,t)))}{p\lambda}.$$

Thus, clearly, the K-function remains equal.

There is an interesting, and useful, transformation of the K-function. It is the L-function:

Definition 2.7. Let Φ be a point process. Let K be the K-function of Φ. We define the *L-function* as

$$L(t) = \sqrt{\frac{K(t)}{\pi}}. \tag{2.4}$$

All that we saw in this section is about univariate point processes. Now, we will extend the previous notions to multivariate processes, which are an interesting object of study. This is because in practice, in many cases, we observe different types of point patterns in the same region. For example, in spatial epidemiology (see in Section 2.8) we have a set of cases (people who suffer a disease) and a set of controls (people in risk for suffer the disease, but not currently suffering). These are two different point patterns. Then, this is an example of a bivariate point process.

Definition 2.8. Let Φ be a multivariate point process, such that is stationary and isotropic. Let $A \subset \mathbb{R}^2$ be a planar bounded region. We define the *intensities* as the constants

$$\lambda_j = \frac{E(\Phi_j(A))}{m_2(A)}.$$

And the *second-order intensities* are functions with scalar argument,

$$\lambda_{ij}(t) = \lim_{m_2(dx), m_2(dy) \to 0} \frac{E(\Phi_i(dx)\Phi_j(dy))}{m_2(dx)m_2(dy)}.$$

Where $t = \|x - y\|$ and $\Phi_j(A)$ is the number of type j points in the planar region A.

Of course, it holds from the previous definition that $\lambda_{ij}(t) = \lambda_{ji}(t)$. Also, we can generalize the K-function for multivariate processes.

Definition 2.9. Let Φ be a stationary and isotropic multivariate point process. Then we define the *multivariate K-functions* like

$$K_{ij}(t) = \frac{E(\Phi_j(\mathbf{B}(*_i, t)))}{\lambda_j}.$$

Where $E(\Phi_j(\mathbf{B}(*_i, t)))$ denotes the expected number of type j points within distance t of an arbitrary type i point.

Now, using similar arguments than above, we can extend (2.3) to a multivariate process obtaining

$$K_{ij}(t) = \frac{2\pi}{\lambda_i \lambda_j} \int_0^t \lambda_{ij}(s) s ds. \tag{2.5}$$

From (2.5), $K_{ij}(t) = K_{ji}(t)$ holds.

2.3 The homogeneous Poisson process

The homogeneous planar Poisson process (from now on we will call it Poisson process) is the mainstay on which the point processes theory is built. It represents the simplest possible stochastic mechanism for the generation of spatial point patterns, and in practice it is used as a standard of complete spatial randomness.

Definition 2.10. Let Φ be a stationary and isotropic point process. We say that Φ is a *Poisson process* if Φ satisfy:

(i) For some $\lambda > 0$, and any finite planar region A, $\Phi(A)$ follows a Poisson distribution with mean $\lambda m_2(A)$.

(ii) Given $\Phi(A) = n$, the n points in A form an independent random sample from the uniform distribution in A.

(iii) For any two disjoint finite planar regions A and B, the random variables $\Phi(A)$ and $\Phi(B)$ are independent.

Notice that the hypothesis for the Poisson processes are exactly the hypothesis for complete spatial randomness. We add in Definition 2.10 the hypothesis (iii), which ensure the self-consistency of (i) and (ii). Clearly, the parameter λ of the Poisson process is its intensity.

Proposition 2.11. Let Φ be a Poisson process with rate λ, then the intensity of Φ is exactly λ.

Proof. Using the intensity definition, and the statement (i) of the Poisson process definition we have

$$\lambda(x) = \lim_{m_2(dx)\to 0} \frac{E(\Phi(dx))}{m_2(dx)} = \lim_{m_2(dx)\to 0} \frac{\lambda m_2(dx)}{m_2(dx)} = \lambda.$$

□

We can do several simulations of a Poisson process in the unit square with, for example, rates $\lambda = 10$, $\lambda = 25$ and $\lambda = 50$ (see in Appendix A.1) in order to know how is the behavior of this class of processes. Notice that the resulting patterns have a similar behavior than the pattern in the middle of Figure 2.1. Now, we see interesting properties of the Poisson processes.

Proposition 2.12. Let Φ be a Poisson process with parameter λ. Then, the second-order intensity function of Φ is

$$\lambda_2(t) = \lambda^2. \tag{2.6}$$

for $t > 0$.

Proof. We apply the second-order intensity definition (see Definition 2.3), let $x, y \in \mathbb{R}^2$ be arbitrary points of the plane, then:

$$\lambda_2(x,y) = \lim_{m_2(dx),m_2(dy)\to 0} \frac{E(\Phi(dx)\Phi(dy))}{m_2(dx)m_2(dy)} = \lim_{m_2(dx),m_2(dy)\to 0} \frac{E(\Phi(dx))E(\Phi(dy))}{m_2(dx)m_2(dy)}$$
$$= \lim_{m_2(dx)\to 0} \frac{E(\Phi(dx))}{m_2(dx)} \lim_{m_2(dy)\to 0} \frac{E(\Phi(dy))}{m_2(dy)} = \lambda(x)\lambda(y) = \lambda^2.$$

Clearly dx and dy are two disjoint planar regions, so for the hypothesis (iii) of the Poisson processes we have that $\Phi(dx)$ and $\Phi(dy)$ are independent random variables. Thus, we can separate the expectation in product of expectations. Once we did this, we obtain the intensity function (see Definition 2.1) evaluated in x and y. Finally, using the assumption that the Poisson process is stationary we know that the intensity function is a constant, as we saw above this constant is the parameter λ of the Poisson process. Finally, using again that the Poisson process is stationary and isotropic, and taking into account that we have proven the equation (2.6) for an arbitraries $x, y \in \mathbb{R}^2$, we have that the equation holds for an arbitrary $t \in \mathbb{R}$. □

Proposition 2.13. Let Φ be a Poisson process with parameter λ. Then, the K-function of Φ is

$$K(t) = \pi t^2. \tag{2.7}$$

For $t > 0$.

Proof. In this proof we will use the equation (2.3) and the previous result (Proposition 2.12), using these and the Barrow's rule we obtain the statements easily.

$$K(t) = \frac{2\pi}{\lambda^2} \int_0^t \lambda_2(s) s ds = \frac{2\pi}{\lambda^2} \int_0^t \lambda^2 s ds = \frac{2\pi\lambda^2}{\lambda^2} \int_0^t s ds = \pi t^2.$$

□

Notice that in the last result we take an interesting result: in a Poisson process the K-function does not depend on λ, it only depends on t, which is a distance between points.

Corollary 2.14. Let Φ be a Poisson process with parameter λ. Then, the L-function of Φ is

$$L(t) = t.$$

Proof. Using the definition of the L-function (Definition 2.7) and the previous result we have

$$L(t) = \sqrt{\frac{K(t)}{\pi}} = \sqrt{\frac{\pi t^2}{\pi}} = t.$$

Then, the result has been proven. □

2.4 Bivariate point processes

Now we focus in bivariate processes. This kind of processes will be important in this work. In order to asses the spatial association between two types of points in a bivariate process, we can consider two hypothesis:

(i) *Independence*: The two types of points are generated by a pair of independent univariate processes.

(ii) *Random labelling*: The two types of points are generated by labelling the points of a univariate process in a series of mutually independent Bernoulli trials.

Both hypothesis above are essential if one want to analyze the behavior of a bivariate point process. In practical cases we usually use these hypothesis as null-hypothesis, which is a powerful statistical approach that allows us to study how is our point pattern nature.

Hereinafter, we will see that under this hypothesis we obtain different K-functions. First we will study the K-function under the hypothesis (i).

Proposition 2.15. Let Φ be a bivariate, stationary and isotropic point process such that satisfies the hypothesis (i). We denote the two types of processes within Φ type 1 and type 2, respectively. Then, the multivariate K-function of Φ is

$$K_{12}(t) = \pi t^2. \tag{2.8}$$

In particular, also $K_{21}(t) = \pi t^2$, because as we saw $K_{12}(t) = K_{21}(t)$.

Proof. Under the hypothesis (i) the two component processes are independent, then a points of type 1 has the same status, with respect to points of type 2, as an arbitrary point. In addition, we have the independence of the random variables $\Phi_1(A)$ and $\Phi_2(A)$ where $A \subset \mathbb{R}^2$ is a region of the plane. Then, using the extension of the second order intensities for multivariate processes (see Definition 2.8) we have, if $t = \|x - y\|$,

$$\begin{aligned}\lambda_{12}(t) &= \lim_{m_2(dx),m_2(dy)\to 0} \frac{E(\Phi_1(dx)\Phi_2(dy))}{m_2(dx)m_2(dy)} = \lim_{m_2(dx)\to 0} \frac{E(\Phi_1(dx))}{m_2(dx)} \lim_{m_2(dy)\to 0} \frac{E(\Phi_2(dy))}{m_2(dy)} \\ &= \lambda_1(x)\lambda_2(y) = \lambda_1\lambda_2.\end{aligned}$$

And now, using the expression (2.5) we obtain what we want

$$K_{12}(t) = \frac{2\pi}{\lambda_1\lambda_2} \int_0^t \lambda_{12}(s) s ds = \frac{2\pi\lambda_1\lambda_2}{\lambda_1\lambda_2} \int_0^t s ds = \pi t^2.$$

□

Notice that we have an interesting result in the previous Proposition. Under the hypothesis (i) the multivariate K-function of a bivariate, stationary and isotropic process is exactly the K-function of a univariate Poisson process.

Now we study the K-function under the hypothesis (ii). We will see that we obtain different results.

Proposition 2.16. Let Φ be a bivariate, stationary and isotropic point process such that satisfies the hypothesis (ii). Denoting the two types of processes within Φ type 1 and type 2, respectively. Then,

$$K_{11}(t) = K_{22}(t) = K_{12}(t) = K(t). \tag{2.9}$$

Where $K(t)$ is the K-function for the unlabelled univariate process.

Proof. We have $K(t)$, which is the K-function of the unlabelled, univariate process consisting of all points, irrespective of type. Under the hypothesis (ii) the univariate processes of type 1 and 2 points are each random thinnings of the bigger unlabelled process. Then, using that the K-functions are invariant under random thinning we have

$$K_{11}(t) = K_{22}(t) = K(t).$$

Using the same argument we have also

$$K_{12}(t) = K(t).$$

□

Observation 2.17. Note that independent and random labelling have the same K-functions if and only if the types 1 and 2 processes are both Poisson processes.

2.5 Inhomogeneous Poisson processes

In this section we will see the first example of a non-stationary point process. This process is obtained if we replace the constant intensity λ of the Poisson process by a spatially varying intensity function $\lambda(x)$. It is, the so-called, *inhomogeneous Poisson process*.

Definition 2.18. Let Φ be a point process. We say that Φ is an inhomogeneous Poisson process with intensity function $\lambda(x)$ if

(i) For any finite planar region A, $\Phi(A)$ has a Poisson distribution with mean

$$\int_A \lambda(s)ds.$$

(ii) Given a finite planar region A, such that $\Phi(A) = n$, the n points in A form an independent random sample from the distribution on A with probability distribution function proportional to $\lambda(x)$.

We have defined this process because will be necessary in the following section.

2.6 Cox processes

The *Cox processes* are an important type of point processes. They appear, in regular basis, on the mathematical literature related with point processes, and they are also known as *doubly stochastic processes*. The reader will understand the rationale of this last name when we expose the definition of Cox processes.

Definition 2.19. Let Φ be a point process. Then, Φ is said a Cox process driven by $\{\Lambda(x) : x \in \mathbb{R}^2\}$ if it satisfies:

(i) $\{\Lambda(x) : x \in \mathbb{R}^2\}$ is a non-negative-valued stochastic process.

(ii) Conditional on $\{\Lambda(x) = \lambda(x) : x \in \mathbb{R}^2\}$, the point form an inhomogeneous Poisson process with intensity function $\lambda(x)$.

Thus, indeed, a Cox process is a doubly stochastic process, due the fact that the intensity function is also stochastic. A Cox process Φ driven by $\{\Lambda(x) : x \in \mathbb{R}^2\}$ is stationary and isotropic if and only if $\Lambda(x)$ is stationary and isotropic. Assuming that the Cox process Φ is stationary and isotropic we can obtain the first-order and second-order properties by taking expectations with respect to $\Lambda(x)$. Then we have the intensity of Φ:

$$\lambda = E(\Lambda(x)). \tag{2.10}$$

The second-order intensity is given by

$$\lambda_2(t) = E(\Lambda(x)\Lambda(y)). \tag{2.11}$$

Where $t = \|x - y\|$. Clearly, using the theoretical covariance definition we can rewrite the previous equation as

$$\lambda_2(t) = \lambda^2 + C(t). \tag{2.12}$$

Where $C(t) = Cov(\Lambda(x), \Lambda(y))$.

Finally, with the equations exposed above we can deduce the K-function for a Cox process under the stationarity and isotropy assumptions.

Proposition 2.20. Let Φ be a Cox process driven by $\{\Lambda(x) : x \in \mathbb{R}^2\}$, assume that the point process is stationary and isotropic. The K-function of Φ is given by:

$$K(t) = \pi t^2 + \frac{2\pi}{\lambda^2} \int_0^t C(s) s ds.$$

Proof. We can, easily, demonstrate this result using (2.12) and (2.3).

$$\begin{aligned} K(t) &= \frac{2\pi}{\lambda^2} \int_0^t \lambda_2(s) s ds = \frac{2\pi}{\lambda^2} \int_0^t (\lambda^2 + C(s)) s ds = \frac{2\pi}{\lambda^2} \int_0^t \lambda^2 s ds + \frac{2\pi}{\lambda^2} \int_0^t C(s) s ds \\ &= \pi t^2 + \frac{2\pi}{\lambda^2} \int_0^t C(s) s ds. \end{aligned}$$

□

Cox processes are often used for modeling biological processes which are aggregated. When the source of this aggregation is the environmental heterogeneity and this heterogeneity is stochastic in nature, the best path for modelling the biological processes is using Cox processes or point processed derived from the Cox one.

2.7 Log-Gaussian Cox processes

An interesting, and commonly used, approach for modelling using Cox processes is take the logarithm of the driving stochastic process, and take some extra assumptions. This path leads to the so-called *Log-Gaussian Cox processes*. Firstly we expose the formal definition, and then we discuss the rationale and the advantages of using this approach. Following Moller et al 1998 [14], we have built the definition:

Definition 2.21. Let Φ be a Cox process driven by $\{\Lambda(x) : x \in \mathbb{R}^2\}$. We say that Φ is a *Log-Gaussian Cox process* if the driving process can be written as

$$\Lambda(x) = \exp\left(Z(x)\right). \tag{2.13}$$

Where $\{Z(x) : x \in \mathbb{R}^2\}$ is a real-valued Gaussian process.

Observation 2.22. A real-valued stochastic process $\{Z(x) : x \in \mathbb{R}^2\}$ is a Gaussian process if the finite-dimensional distributions (see Definition 1.8) of the process are Gaussian laws. It is, equivalently, the joint distribution of any finite vector $(Z(x_1), \ldots, Z(x_n))$, where $x_1, \ldots, x_n \in \mathbb{R}^2$, is Gaussian.

Assuming stationarity and isotropy, the distribution of Z is determined by the mean and the variance:

$$\mu = E(Z(x)), \quad \sigma^2 = Var(Z(x)). \tag{2.14}$$

And also for the covariance function, which for some $x, y \in \mathbb{R}^2$ is given by

$$C(t) = Cov(Z(x), Z(y)) = E\left(Z(x)Z(y)\right) - \mu^2. \tag{2.15}$$

Since we are assuming stationarity and isotropy, the covariance function only depends on the distance between points $t = \|x - y\|$, and this is why we write the dependence of the covariance function only on t.

Now, our goal is to define a model with a specific covariance function. There are several models in the mathematical literature (see these models in Moller et al 1998 [14]), but probably, one of the simplest models and one of the most commonly used due to its good results is the so-called *exponential model*. In this model we assume that the Gaussian process $\{Z(x) : x \in \mathbb{R}^2\}$ is centered (i.e. $\mu = 0$) and we take the covariance function (see Baddeley 2015 [2]):

$$C(t) = \sigma^2 \exp\left(-\frac{t}{\delta}\right). \tag{2.16}$$

Where δ is a scale parameter depending on the point pattern. Thus, as the reader can observe, the exponential model only depends on the two parameters σ^2 and δ. And then, if we can estimate these parameters we can control the model.

Now, observe that we can deduce the first-order and second-order properties in this specific case. Firstly, we will see the intensity, but previously we see a probability lemma.

Lemma 2.23. Let X be a real-valued random variable with normal law with mean μ and variance σ^2 (i.e. $N(\mu, \sigma^2)$) then the expectation of $\exp(X)$ is

$$E(exp(X)) = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

Proof. To demonstrate this lemma we will use the theoretical definition of expectation of a real-valued random variable, the probability density function of a normal

and we will perform several calculus

$$\begin{aligned}
E(\exp(X)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp(x) \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} + x\right) dx = \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{-x^2 + 2x\mu - \mu^2 + 2x\sigma^2}{2\sigma^2}\right) dx = \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{-x^2 + 2x(\mu+\sigma^2) - (\mu+\sigma^2)^2 + (\sigma^2)^2 + 2\mu\sigma^2}{2\sigma^2}\right) dx = \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{(x-(\mu+\sigma^2))^2}{2\sigma^2}\right) \exp\left(\frac{(\sigma^2)^2 + 2\mu\sigma^2}{2\sigma^2}\right) dx = \\
&= \left(\exp\left(\frac{(\sigma^2)^2 + 2\mu\sigma^2}{2\sigma^2}\right)\right) \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} \exp\left(-\frac{(x-(\mu+\sigma^2))^2}{2\sigma^2}\right) dx = \\
&= \exp\left(\mu + \frac{\sigma^2}{2}\right).
\end{aligned}$$

In the last equality we have used that the term in the right is the integral of the probability density function of a real-valued random variable with law $N(\mu + \sigma^2, \sigma^2)$ throughout $\mathbb{R}$, then it is exactly 1. Also, in the last equality, we have simplified the left term. □

Now, with the support of the previous Lemma, we can demonstrate the following result.

Proposition 2.24. Let Φ be a stationary and isotropic Log-Gaussian Cox process driven by $\{\Lambda(x) : x \in \mathbb{R}^2\}$. Using all the notation introduced above, we have that the intensity of Φ is

$$\lambda = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

Proof. As we saw before the intensity of a Cox process is given by (2.10). Thus, applied to the Log-Gaussian Cox process we have

$$\lambda = E(\Lambda(x)) = E(\exp(Z(x))).$$

Since Φ is a Log-Gaussian Cox process we know that $\{Z(x) : x \in \mathbb{R}^2\}$ is a Gaussian process with mean μ and variance σ^2, using the previous notation. Then, $Z(x)$ has normal law with mean μ and variance σ^2 for all $x \in \mathbb{R}^2$. Thus, applying the Lemma 2.23 we obtain

$$\lambda = E(\Lambda(x)) = E(\exp(Z(x))) = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

And the result, indeed, holds. □

Observation 2.25. Taking the statements of the previous preposition, and adding the hypothesis that Φ is under the exponential model, it is easy to see that the intensity is

$$\lambda = \exp\left(\frac{\sigma^2}{2}\right). \tag{2.17}$$

We only have to take into account that under the exponential model the Gaussian process $\{Z(x) : x \in \mathbb{R}^2\}$ is centered.

Finally, we want to determine the shape of the K-function of a Log-Gaussian Cox process under the exponential model. Using the Proposition 2.20, we only have to substitute the expression for the intensity under the exponential model (2.17) and the covariance function taken in this model (2.16). Doing this we obtain:

$$K(t) = \pi t^2 + 2\pi \exp\left(-\sigma^2\right) \int_0^t s\sigma^2 \exp\left(\frac{-s}{\delta}\right) ds.$$

We can develop the expression above by integrate. Doing this we finally obtain

$$K(t) = \pi t^2 + 2\pi\sigma^2 \exp\left(-\sigma^2\right) \delta \left(\delta - \exp\left(\frac{-t}{\delta}\right)(\delta + t)\right). \tag{2.18}$$

2.7.1 Parameter estimation: method of minimum contrast

In this section we will be using Diggle et al 2013 [12], Moller 1998 [14] and Diggle 2013 [12]. Here we will explain how to estimate the parameters (σ^2, δ) of the exponential model. We will expose the so-called *minimum contrast method* (see Moller 1998 [14]). Using the approach of that we can describe the point pattern with the K-function, we have as a objective find the theoretical K-function under the model. Thus, we just have to estimate the parameters and put them into (2.18). Let $\hat{K}(t)$ be the empirical K-function from the data, and let $K(t; \sigma^2, \delta)$ the K-function under the exponential model (2.18). The path of this estimation is simple, we want to make $K(t; \sigma^2, \delta)$ as close to $\hat{K}(t)$ as we can. For achieve it, we define a measure of discrepancy between $K(t; \sigma^2, \delta)$ and $\hat{K}(t)$, the area between the two curves or the integrated squared difference between curves. It is

$$\int_0^{t_0} \left((\hat{K}(s))^q - (K(s; \sigma^2, \delta))^q\right)^2 ds. \tag{2.19}$$

And then, we estimate the parameters (σ^2, δ) to be the values $(\hat{\sigma^2}, \hat{\delta})$ which minimizes (2.19). By doing this, as the reader can observe, we obtain $K(t; \hat{\sigma^2}, \hat{\delta})$ which is near as possible to $\hat{K}(t)$. In the literature we can find that $q = 1/2$ is a good choice for regular point patterns, and $q = 1/4$ for aggregated patterns. Also t_0 is to be chosen, there are no clear reasons for a specific election of t_0 in the literature.

2.8 Point processes in spatial epidemiology: spatial clustering

In this section we have followed Diggle 2013 [10], Waller and Gotway 2004 [19] and Diggle and Chetwynd 1991 [11], Bivand et al. 2008 [9] and Baddeley et al. 2015 [2].

By *spatial clustering* we mean a tendency for cases to occur more closely together than would be possible with random sampling of population at risk. For analyze the spatial clustering we will use a *case-control* scheme, which is an approach used on a regular basis in spatial epidemiology. Case-control data involve locations for each cases reported (cases) and a collection of noncases (controls). Sometimes, in practice, controls are chosen following an specific criteria with the aim of match the characteristics of the set of cases. However, in this work, we assume that controls are an independent random sample.

In order to asses spatial clustering one can take the null hypothesis:

$$H_0 : \text{There are no clusters of cases.} \tag{2.20}$$

Let Φ_1 be the process of cases, and Φ_2 be the process of controls. Thus, if we superpose these processes we obtain a bivariate point process $\Phi = \Phi_1 + \Phi_2$. Under the null hypothesis of no clustering, cases form a spatially random sample from the underlying population. Thus, controls necessarily form a spatially random sample from the same population. Hence, no spatial clustering is equivalent to random labelling of the bivariate process of cases and controls. Then, (2.20) is equivalent to

$$H_0 : \Phi = \Phi_1 + \Phi_2 \quad \text{is a bivariate point process under random labelling.} \tag{2.21}$$

In a case-control study the null hypothesis of completely random labelling implies constant disease risk. This means, the probability of contracting the disease does not depend on spatial location. Observe that, under (2.21)

$$D(t) = K_{11}(t) - K_{22}(t), \tag{2.22}$$

where K_{11} and K_{22} denotes the K-functions of cases and controls respectively, is identically zero (it is followed from Proposition 2.16). Positive values of $D(t)$ represents spatial aggregation of type 1 points (cases). An interesting approach is establish a statistic to test the null hypothesis. The natural path is base this statistic in the corresponding empirical function

$$\hat{D}(t) = \hat{K}_{11}(t) - \hat{K}_{22}(t). \tag{2.23}$$

In following sections (see Section 3.4) we will develop a statistic test for asses spatial clustering.

Chapter 3

Statistics in spatial point patterns

In this chapter we will study several useful statistic tools, related with the mathematical concepts exposed in the previous chapters. These statistics will be used in the case studies at the last chapter of the work. In the following pages, the reader will find mechanisms for estimate different quantities and some statistical tests.

For this chapter I have used Baddeley et al. (2015) [2]. Also Diggle (2014) [10] have been consulted.

3.1 Monte Carlo tests

Monte Carlo methods use random simulation to replace complicated calculations in algebra and calculus. Furthermore, Monte Carlo methods, in general, involves a huge number of simulations in order to achieve accuracy. On the other hand, *Monte Carlo test* uses a smaller number of simulations from a given null hypothesis, and appeals to a symmetry principle instead of the law of large numbers. We study the Monte Carlo tests from Diggle (2013) [10] and Baddeley el al. (2015) [2]. Nevertheless, we will focus in Baddeley et al. (2015) [2], in order to expose here this topic.

Let Φ be the observed point pattern. And let H_0 be the null hypothesis. The simplest Monte Carlo test follows the scheme:

1. Generate m simulated random point patterns $\Phi^{(1)}, \ldots, \Phi^{(m)}$ from the null hypothesis. These are random point pattern generated by computer, similar to the observed pattern Φ, but under the assumption that H_0 is true. The random point patterns generated should be independent of each other and of the observed data Φ (independent in the sense of probability).

2. Reduce the point patterns to a single numerical value using a test statistic T. Thus, Φ is reduced to $t_{obs} = T(\Phi)$. And $\Phi^{(1)}, \ldots, \Phi^{(m)}$ to $t_1 = T(\Phi^{(1)}), \ldots, t_m = T(\Phi^{(m)})$.

3. Assuming that larger values of T are more favourable to the alternative hypothesis, the test rule is to reject H_0 at significance level $1/(m+1)$ if the observed value t_{obs} is larger than all of the simulated values $t_1, \ldots, t_m$.

The basis for Monte Carlo tests is symmetry. Assuming that H_0 is true, we have that the original data and the m simulated patterns must be statistically equivalent. Thus, in addition, t_{obs} and $t_1, \ldots, t_m$ must be statistically equivalent. Using a symmetry argument, there is a 1 in $(m+1)$ chance that the statistic t_{obs} is the largest of these $m+1$ values. If this happens, the result is statistically significant at level $\alpha = 1/(m+1)$.

There are several variants from the basic Monte Carlo test. One interesting alternative is a *two-sided* test. Instead of a *one-sided* test, which rejects the null hypothesis if t_{obs} is large, we could have a two-sided test which rejects the null hypothesis if t_{obs} is either largest or smallest of the $m+1$ values. This will have significance level $\alpha = 2/(m+1)$. Another important variant, that we will use during the work, is perform the Monte Carlo test with a test statistic which is a function, not a number. This type of Monte Carlo tests works exactly like the explained above.

3.2 K-function estimation

The goal of this section is show how the K-function (see in Definition 2.6) is estimated in a given point pattern. Let $\|x_i - x_j\|$ be the distance between two different points in the point pattern, let A be the region where the point pattern is and n the number of points of the pattern. We can define the number of t-neighbours for the point x_i as

$$n_i(t) = \sum_{j \neq i} \mathbb{1}\{\|x_i - x_j\| \leq t\}. \tag{3.1}$$

Where we used the indicator notation. It is, $\mathbb{1}\{\text{statement}\}$ is 1 if the statement is true and 0 else. Notice that the number of t-neighbours counts the number of points with distance less than or equal to t with respect the point of the pattern x_i. Equivalently, $n_i(t)$ is the number of data points which fall inside a circle of radius t centred at x_i, not counting x_i itself.

The empirical cumulative distribution function of the pairwise distances is

$$\hat{H}(t) = \frac{1}{n(n-1)} \sum_{i=1}^{n} n_i(t). \tag{3.2}$$

The denominator $n(n-1)$ is the total number of pairs of distinct points. So, $\hat{H}(t)$ is the fraction of pairs for which the distance is less than or equal to t. Thus, using the measure of the region where the patterns is, we have the function $m_2(A)\hat{H}(t)$ which is the standardised average number of t-neighbours of a typical data point. With the aim to be fully able to compare different datasets observed in different regions, we also need to take account of edge effects (see in Diggle 2013 [10]). Thus, we obtain the so called empirical K-function

$$\hat{K}(t) = \frac{m_2(A)}{n(n-1)} \sum_{i=1}^{n} n_i(t) e_{ij}(t). \tag{3.3}$$

Where e_{ij} is an edge correction weight. The explanation of this kind of edge correction is out of the scope of the work. For more information about this see Baddeley et al. 2015 [2].

The estimation of the K-function given by (3.3) is used in the *spatstat* library in order to estimate the K-function in a given spatial point pattern.

3.2.1 Example

In this example we will estimate the K-functions of the point patterns exposed in the Figure 2.1. We use the *spatstat* library for achieve it. We plot (see in Figure 3.1) the estimations and also the theoretical K-function for a random independent pattern, which is $K(t) = \pi t^2$ as we proved in Proposition 2.13. The reader can observe that the K-function is a good benchmark that help us to understand the behavior of the point pattern. And, if we compare the empirical K-function with the theoretical one, it provides a thought about how random is the pattern.

Looking in Figure 3.1 we can, easily, see the discrepancies between the empirical K-function and the theoretical in the regular and clustered cases. This indicates that these are not random patterns. On the other hand, in the independent pattern simulation we obtain an empirical K-function which is very similar to the theoretical one. This suggest that this is a random pattern. Anyway, in the next section we will expose a statistic method to determine if a point pattern is under the CSR hypothesis or not.

3.3 Complete spatial randomness test

As we studied in the beginning of this memory (see Section 2.1.2) the hypothesis of Complete Spatial Randomness (CSR) is an important initial benchmark in a point process analysis. One can find several CSR tests in the mathematical literature. Here we describe a method based in the K-function. Specifically in the

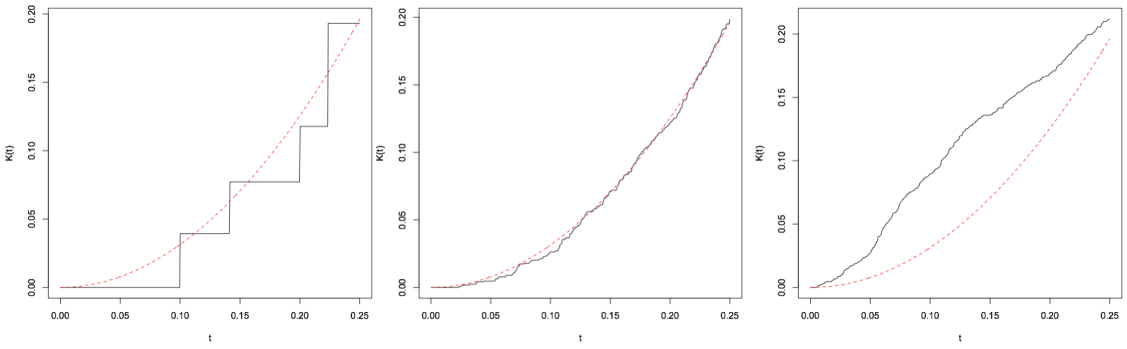

Figure 3.1: K-function estimation of the point patterns in Figure 2.1. K-function of the regular pattern (left), the independent pattern (middle) and the clustered pattern (right). In dashed red lines, the theoretical K-function.

empirical and theoretical K-function. This method is the so-called *global envelopes* (Baddeley et al. 2015 [2]).

Let Φ be a point pattern. We want to test if Φ verify the CSR hypothesis. As we mentioned in Section 2.3, in practice, the Poisson processes are used as a standard CSR. Thus, under the hypothesis of CSR, we are dealing with a Poisson process. We take the null hypothesis

$$H_0 : \Phi \quad \text{is a Poisson process.} \tag{3.4}$$

We know that the K-function for any Poisson process is $K_{pois}(t) = \pi t^2$ (see the Proposition 2.13). From now on, K_{pois} will denote the theoretical K-function. Evidently, if Φ satisfies the CSR hypothesis, the observed K-function (denoted as $\hat{K}$) must be near to K_{pois}.

The global envelopes method is an example of a Monte Carlo test (see in Section 3.1) and shows us if the discrepancy between $\hat{K}$ and K_{pois} is statistically significant. Global envelopes delimit a zone of constant width. The width is determined as follows: assume that we do m simulations, for each simulated dataset, we compute the maximum vertical deviation d_i between graphs of $\hat{K}$ and K_{pois}, over some range of distances. The maximum $d_{max} = max(d_1, \ldots, d_m)$ is taken. And then, the global envelopes are

$$\begin{aligned} Env_-(t) &= K_{pois}(t) - d_{max} \\ Env_+(t) &= K_{pois}(t) + d_{max}. \end{aligned} \tag{3.5}$$

If the graph of the empirical K-function lies outside these limits, at any value of t, we reject the null hypothesis H_0 with an exact significance level of $\alpha = 1/(m+1)$. Thus, with $m = 19$ we have a test with significance level 0.05.

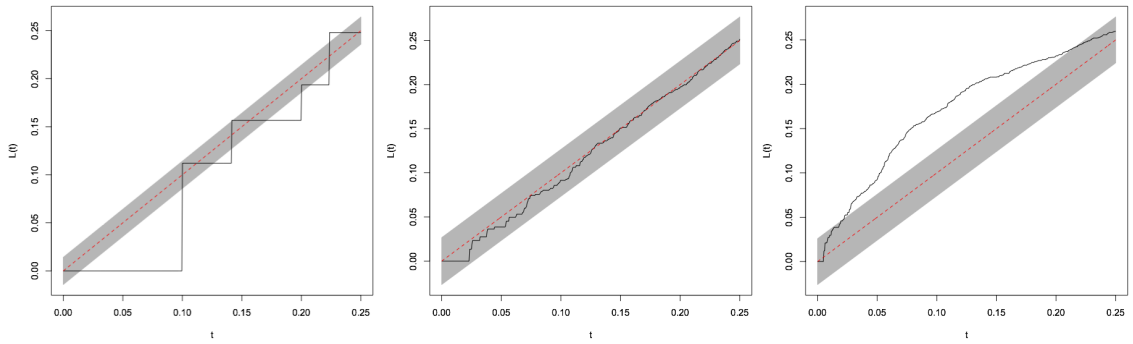

Figure 3.2: Global envelopes with $m = 19$ simulations, for the L-function, relating to the point patterns in Figure 2.1. The regular (left), the independent (middle) and the clustered (right) point pattern. In dashed red lines, the theoretical L-function.

Another interesting alternative, is use the L-function instead of the K-function. Because, as we saw in Corollary 2.14, $L_{pois}(t) = t$. Clearly, it is easier compare the observed function $\hat{L}$ with the theoretical one, because of the simplicity of the last of the two. Thus, this alternative gives us a more powerful test.

3.3.1 Example

Returning to the example given in Section 2.1.2 with a regular, independent and clustered pattern (see Figure 2.1). We can perform an example using global envelopes in order to determine if these patterns are under the CSR hypothesis, or not.

We will compute global envelopes for the corresponding L-functions of point patterns in Figure 2.1. We use *spatstat* and we obtain (see in Figure 3.2) that the simulation of an independent pattern is under CSR hypothesis with a significance level 0.05. Also, with the same significance level, we reject the null hypothesis for the clustered and the regular pattern. Thus, both point patterns, are not under the CSR hypothesis.

3.4 Test of random labelling

In this section we present a test for accept, or reject, the null hypothesis:

$$H_0 : \Phi = \Phi_1 + \Phi_2 \quad \text{is a bivariate point process under random labelling.} \tag{3.6}$$

Remember the definition saw in Section 2.4, a bivariate point process $\Phi = \Phi_1 + \Phi_2$ is random labelled if the two type of points are generated by labelling the points of

a univariate process in a series of mutually independent Bernoulli trials. Our aim in this section is to expose a statistical test of random labelling. We have consulted for this section: Baddeley 2010 [1], Baddeley et al. 2015 [2], Bivand et al 2008 [9].

A *permutation test* is a good path for asses random labelling. A permutation test is a Monte Carlo test based on randomly relabelled versions of the original data. In this test, the observed data spatial locations are kept constant but the marks (type labels) associated with those locations are randomly permuted with an equal probability for each variation. These randomly relabelled datasets are statistically equal to the original data if (3.6) is valid, in which case the rationale of the Monte Carlo test holds.

Using the permutation test approach, we will develop a envelope-based test, similar (but not identical) than the explained in the previous section. In this case we will choose the statistic

$$D(t) = K_{11}(t) - K_{22}(t). \tag{3.7}$$

Where K_{11}, K_{22} are the respective K-functions for the type 1 and 2 points from the bivariate process $\Phi = \Phi_1 + \Phi_2$. Using the Preposition 2.16, it is easy to deduce that under (3.6), necessarily, $D(t) = 0$. Thus, $D(t)$ different to 0, clearly suggests that random labelling is not satisfied.

In this test we will simulate m point patterns, using random relabelling, and we will compute the $D(t)$ functions. Then, we will plot the minimum and the maximum values of the $D(t)$ functions of the simulated patterns (the so-called upper and lower envelopes), and we will shade the region between these upper and lower envelopes. By doing this we have made a two-sided Monte Carlo test, with a significance level $\alpha = 2/(m+1)$, that rejects (3.6) if the observed $D(t)$ function lies outside the shaded region.

3.4.1 Example

We have done two simulations in R in order to expose an example of a test of random labelling. Firstly we simulate a random labelled point pattern. We simulate a Poisson process of rate $\lambda_1 = 75$ in the unit square, and we put the labels "Type 1" and "Type 2" randomly in the points generated. Then, we perform the test with $m_1 = 39$ simulations. In the Figure 3.3 the reader can see the result obtained. In the right graphic, we can observe that the observed $D(t)$ function remains inside the shaded zone. Thus, indeed, we can accept the null hypothesis (3.6) with a significance level $\alpha_1 = 2/(m_1+1) = 2/(39+1) = 0.05$. Also, the reader can observe that the mean of the $D(t)$ functions of the $m_1 = 39$ simulations (in dashed red lines in the right graphic) is near to 0, as we expected following the statements of the Preposition 2.16.

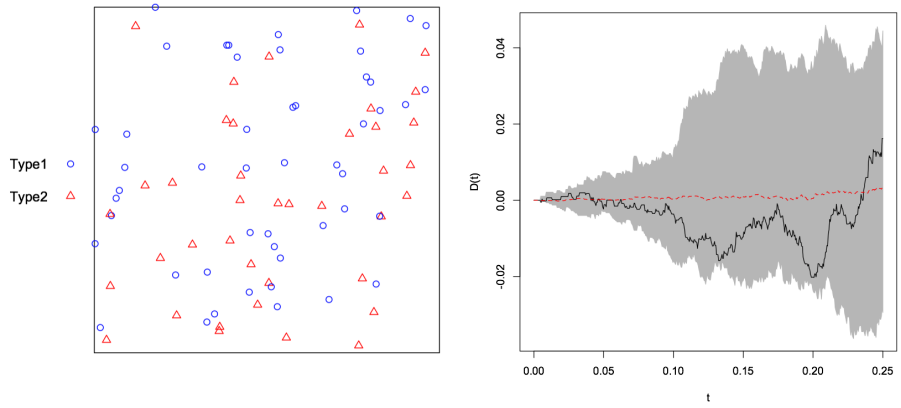

Figure 3.3: In the left: simulation of a random labelled point process with labels "Type 1" and "Type 2". In the right: random labelling test with $m_1 = 39$ simulations. In solid black line the observed $D(t)$ function. The dashed red line is the mean of the $D(t)$ functions of the simulations.

On the other hand, we simulate a bivariate point process which is clearly not random labelled. We superpose a Poisson process of rate $\lambda_2 = 70$, and a clustered point process (left panel of Figure 3.4). As we can observe in the right graphic of the Figure 3.4, the $D(t)$ function observed is outside the shaded region. Then, we can reject the null hypothesis (3.6) with a significance level of $\alpha_2 = 2/(m_2 + 1) = 2/(39 + 1) = 0.05$.

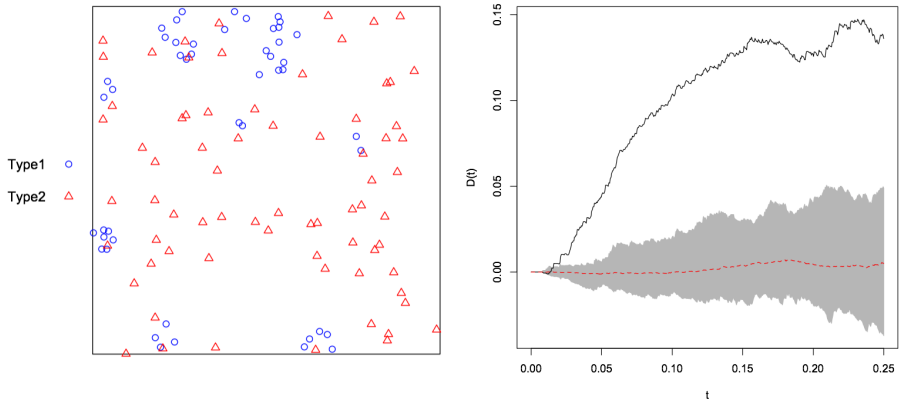

Figure 3.4: In the left: simulation of a bivariate point process with labels "Type 1" and "Type 2". In the right: random labelling test with $m_2 = 39$ simulations. In solid black line the observed $D(t)$ function. The dashed red line is the mean of the $D(t)$ functions of the simulations.

Chapter 4

Case studies

This is the last chapter of the work. In this chapter our main goal is apply all concepts learned and see several case studies that use all the mathematical concepts exposed until now.

Jorge Mateu Mahiques, who is one of the best researchers in this area, decided to help me in my work and also passed me and allowed me to use one of his databases. Thus, this part of the work would not be possible without the support of Jorge. The dataset shared by Jorge is used in the paper Ibon Tamayo-Uria, Jorge Mateu and Peter J. Diggle 2014 [18]. This paper will be very important in this chapter, because will be a great benchmark.

4.1 Introduction

An important problem of public health are the rodent population in urban environments. The brown rat (*Rattus norvegicus*) shares habitat with humans in all the cities in the world. The bigger metropolis trough the sphere provide shelter to a huge number of these rodents. Several sources states that in New York there are at least 8 million of rats, in Paris there are approximate 3 million and also other cities like Barcelona has a public health problem with rats populations. Similar public health issues are caused for the cockroaches (*Blattodea*). It has been proved in several studies that cockroaches can cause different diseases to the human population.

In this part of the work we will analyze the spatial distribution of rat and cockroach sightings in Madrid (Spain). We have access to the location of rat and cockroach sightings in Madrid between the years 2010 and 2013, these sightings are reported week to week. The plan in this study cases will be study two point patterns, one of rats and one of cockroaches, in different time windows. The procedure in each point pattern will be the same one. This chapter finishes with a

conclusion of the study cases done.

4.2 Dataset description

In this section we describe, briefly, the dataset used. For more details about the dataset see Jorge Mateu et al. (2014) [18]. Our study takes place in Madrid city (coordinates $4030'N3°40'O$, average altitude $657m$, surface $604km^2$). Madrid has more than 3 million population (source *Instituto Nacional de Estadística*, 2018). In this data we have, in total, 4 years of observations, 14495 cases which 6702 are rats sightings and 7793 are cockroaches sightings. The reader can see in Table 4.1 a summary of the different sightings per year and classified by rats and cockroaches.

Year	**Rats**	**Cockroaches**	**Total**
2010	1361	2145	3506
2011	2051	2067	4118
2012	1612	2123	3745
2013	1659	1458	3117

Table 4.1: Sightings observations per year.

All these observations were reported for citizens by telephone, fax, email or in person at the front desk. Direct sightings or signs of their presence (e.g. droppings, burrows, gnaw marks, etc) can be reported by the citizens.

4.3 Procedures

In order to follow the same procedure and to carry out rigorous studies, it is important to determine a standard procedure. Thus, in this section we will outline the procedure to be followed in the case studies. The case studies will have two main parts. The first one is descriptive. In this part we will develop the steps:

- Plot the observed pattern of sightings.
- Estimate the K-function.
- Develop a CSR test.
- Perform a spatial clustering assessment.

Figure 4.1: Rat sightings during the first half of 2010 (left) and the sub-region of Madrid city where all the rat sightings observed lies (right).

And the second one is about modeling. In this part we will follow the steps:

- Fit a model to the observed pattern.
- Perform simulations with the fitted model.

4.4 Case study 1: rat sightings in the first half of 2010

In this first Case study we will analyze rat sightings from the first half of the year 2010. We have selected a total amount of 600 cases from the dataset. Firstly we plot the observed cases (see left panel of Figure 4.1). The reader can see in the left panel of Figure 4.1 that there is a lot of surface without observations. Then, we will select the sub-region of Madrid where are the observations (see in right panel of Figure 4.1). From now on, we will perform the point pattern analysis with this sub-region. Before start doing the study described on the Section 4.3 we include a summary of the Euclidean distances of the observed point pattern (see Table 4.2). This summary may help the reader to understand how the sightings are distributed through the surface of Madrid city. It is important to mention that all the calculus and estimations are made taking $t_{max} = 0.2d_{max} = 0.2 \times 20627 = 4125.4$ (i.e. the 20% of the maximum distance between points), because it is recommended in the literature in order to obtain stable results. That is why the x-axis in the following graphics arrives approximately to $t_{max} = 4125.4$.

Our first two steps will be estimate the K-function (using Section 3.2) and develop a CSR test (using Section 3.3). We use *spatstat* library in R and we obtain

	Euclidean distance (in meters)
Minimum	8
First quartile	4351
Median	6911
Mean	7122
Third quartile	9686
Maximum	20627

Table 4.2: Summary of Euclidean distances (in meters) between points.

the graphics exposed in Figure 4.2. As we can see in left graphic, the empirical K-function is far from the theoretical one. Notice that the values in the axis are in the order of thousands meters, this is because the region in what the pattern lives is Madrid city ($604km^2$ of surface, as we mentioned before). In addition, in the right graphic, we can see that the empirical L-function is clearly outside the shaded area. Thus, we can reject the CSR hypothesis for this pattern with a significance level of $1/(19+1) = 0.05$. Notice that, for perform the CSR test, we have used the L-function instead of the K-function. As we commented in Section 3.3 the L-function gives a powerful test. And, as we saw in Definition 2.7, from K-function to L-function there is only a simply transformation. Moreover, the 19 simulations were performed with the same number of points, 600.

Now, we develop a spatial clustering assessment. We prepare a case-control scheme. It is, we simulate a independent random sample of 600 points within the sub-region of Madrid city and we superpose this simulation of points with the observed pattern (see in the left panel of Figure 4.3). Then, we have a bivariate point pattern with points of class Cases and points of class Controls. Using the statistical tools exposed in the previous sections we are able to perform a formal test. Here we use the theory provided in Section 2.8 and the random labelling test explained in Section 3.4. Thus, we take the null hypothesis (2.20) which, as we studied before, is equivalent to (3.6). And we develop a random labelling test with 39 simulations using the function

$$D(t) = K_{11}(t) - K_{22}(t).$$

As the reader can see in Figure 4.3 (in the right part), the observed $D(t)$ function is far to the 0 for almost all t values. In addition, the $D(t)$ function is outside the

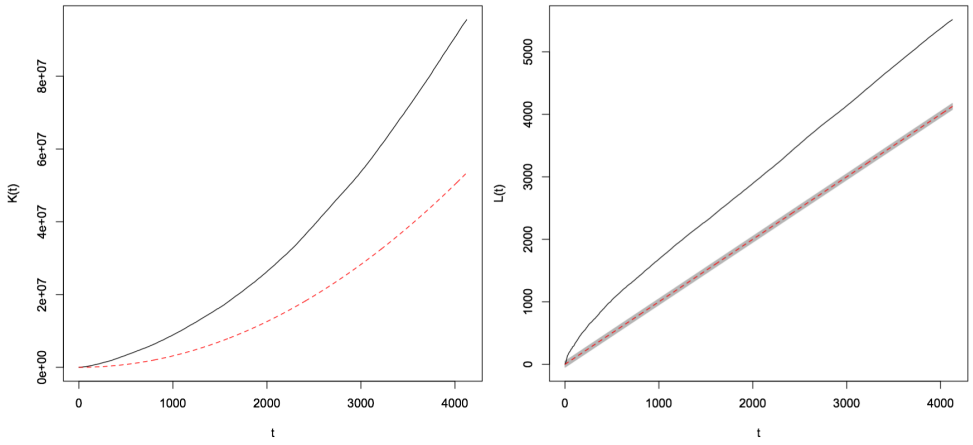

Figure 4.2: In the left: K-function estimation (solid black line) compared with the theoretical function K_{pois} (dashed red line). In the right: global envelopes with 19 simulations performed with fixed number of point, 600 like the observed point pattern. The empirical L-function is the solid black line, and the theoretical L-function is the dashed red line.

shaded zone in almost all t values, then, we can reject the null hypothesis of no spatial clustering (2.20) with a significance level of $2/(39+1) = 0.5$. Moreover, we can see that the observed $D(t)$ seems to be an increasing function. And that, the greater the distance, the farther away the function is from the shaded area. Thus, the degree of clustering increases with the distance between points. The values of $D(t)$ are positive for all t, clearly it means that $K_{11}(t) > K_{22}(t)$ for all t values. This fact shows us that the Type 1 (i.e. type Cases) points are the aggregated ones.

Now, is the moment to fit a model to the observed pattern, perform simulations and see the results of this model. We will be following what we exposed in Section 2.7, and we will use the exponential model. In addition, we will estimate the parameters of the model using the minimum contrast method exposed in Section 2.7.1. We take (2.19) using the parameters $t_0 = t_{max}$ and $q = 1/4$, it is

$$\int_0^{t_{max}} \left((\hat{K}(s))^{1/4} - (K(s;\sigma^2,\delta))^{1/4} \right)^2 ds. \tag{4.1}$$

Using a function implemented in *spatstat* we obtain, using as a initial parameters $(\sigma^2, \delta) = (1, 1)$, the parameter estimation

$$\hat{\sigma^2} = 1.471686, \quad \hat{\delta} = 2150.178436.$$

We have assessed the robustness of the parameter estimation changing the parameter t_0 for others nearby. We take $t_0 = 0.2d_{max} + \epsilon d_{max}$ where $\epsilon \in [-0.02, 0.02]$.

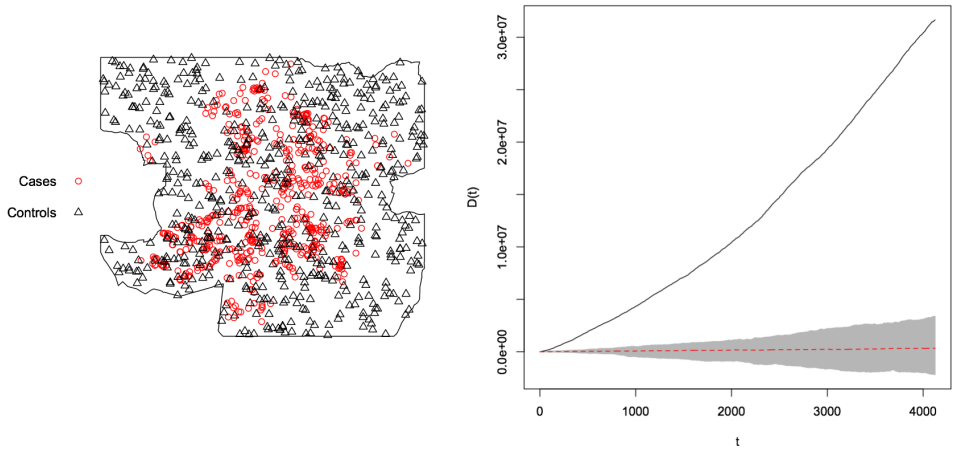

Figure 4.3: In the left: Case-control map. In the right: test of random labelling with 39 simulations. There is the observed $D(t)$ function (solid black line) and the theoretical $D(t)$ (dashed red line).

We will select several values of ϵ in the previous interval, all the values with a distance of 0.0025 with the nearest one. The reader can see the results obtained in Appendix A.2. Indeed, according to the results obtained, the parameter estimation is robust.

Now we have the parameters of the exponential model, and then we can control the model. We can see the results obtained with the fitted model. Firstly, we can show the fitted K-function (see in left panel of Figure 4.4) it is, the K-function under the exponential model (2.18) with parameters $(\hat{\sigma^2}, \hat{\delta})$. As the reader can observe we have obtained a K-function similar to the empirical one. Moreover, we can develop several simulations of a Log-Gaussian Cox process under the exponential model with the estimated parameters, in the right panel of Figure 4.4 we can see four simulations (see more simulations in Appendix A.3).

4.5 The Case study 1 in the unit square

In this section we will perform the analysis of the Case study 1, but with the point pattern translated to the unit square. In all the work we have dealt with stationary and isotropic point processes. This is, we can apply translations and rotations without change the process properties. Now, we expose the results for the rat sightings in the first half of the year 2010, but translating the point pattern observed to the unit square. Work in the unit square implies lower computational cost. As we will see, we obtain equivalent results than the obtained in Section 4.4.

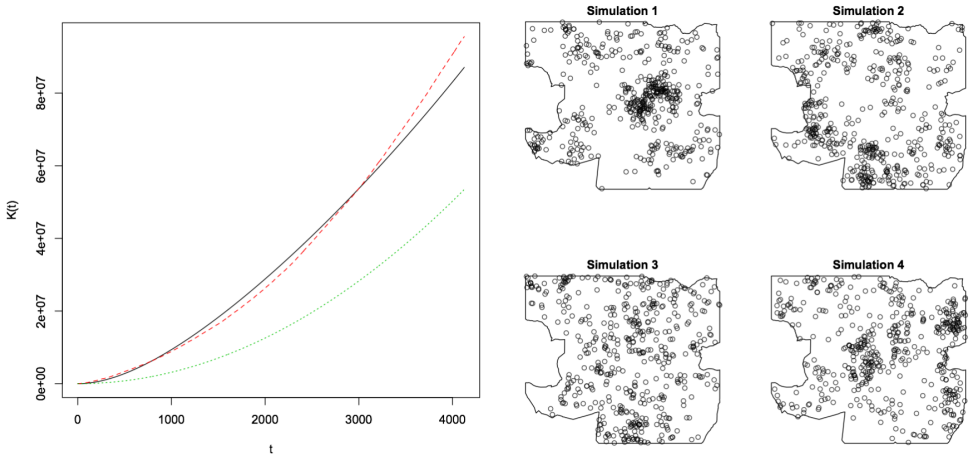

Figure 4.4: In the left: empirical K-function (red dashed line), fitted K-function (solid black line) and theoretical K-function (green dashed line). In the right: four simulations of a Log-Gaussian Cox process under the exponential model with parameters $\hat{\sigma^2} = 1.471686, \hat{\delta} = 2150.178436$.

Thus, it is a little empirical demonstration of the stationarity and isotropy.

Let $\{(x_1, y_1), \ldots, (x_{600}, y_{600})\}$ be the 600 points of the observed rat sightings. We want to translate this points to the unit square. Let $x_{max}, y_{max}, x_{min}, y_{min}$ be the maximums and the minimums of the first and second coordinate respectively (i.e. $x_{max} = max\,(x_1, \ldots, x_{600})$ and the analogous for the rest of the quantities). Then, we define the sequence of points:

$$\left\{ \left(\frac{x_1 - x_{min}}{x_{max} - x_{min}}, \frac{y_1 - y_{min}}{y_{max} - y_{min}} \right), \ldots, \left(\frac{x_{600} - x_{min}}{x_{max} - x_{min}}, \frac{y_{600} - y_{min}}{y_{max} - y_{min}} \right) \right\}. \tag{4.2}$$

This sequence of points is the corresponding to the observed pattern, but in the square unit instead in the sub-window of Madrid city. See the obtained pattern in Figure 4.5. As the reader can see, with the support of the axes added, we have the same point pattern but translated to the unit square. Now, we will perform the same analysis than the developed in Section 4.4. In the unit square the maximum distance between points is $d_{max} = 1.0660120$. The maximum distance between points in the unit square pattern appears as greater than one, this is because there are points in the boundary of the unit square and R computes the distances as greater than one, but that does not affect to our study. We will be working with $t_{max} = 0.2 d_{max} = 0.2132024$.

Therefore, we show the results obtained. In Figure 4.6 the reader can observe the empirical K-function (left panel) and the CSR test performed with the L-function (right panel). In this case we also can reject the hypothesis of CSR with

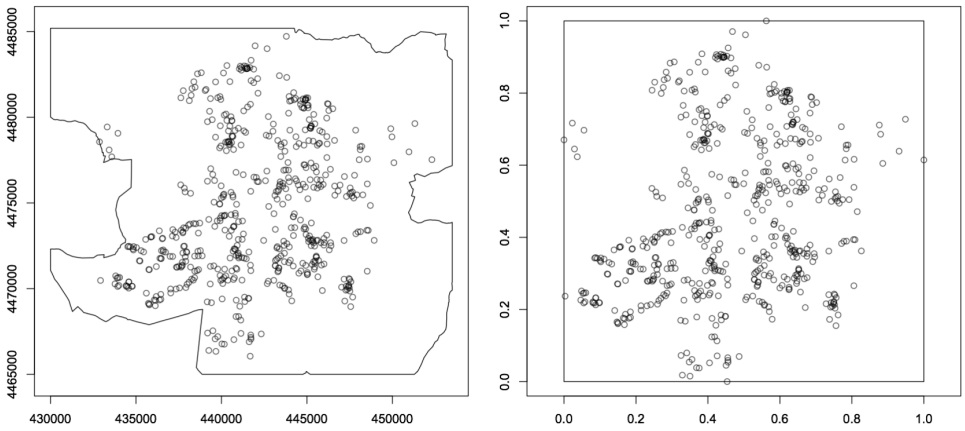

Figure 4.5: In the left: point pattern generated from rat sightings locations in the first half of 2010, data plotted in the sub-region of Madrid city where the sightings are located. In the right: observed point pattern of rat sightings translated to the unit square.

a significance level of $1/(1+19) = 0.5$. Moreover, we can see that the graphics obtained are fairly similar than the exposed in Figure 4.2. We also develop a spatial clustering assessment, following the same procedure used in Section 4.4. In Figure 4.7 we can see the Cases VS Control map (in the left) and the test of random labelling using the $D(t)$ function (in the right). Again, we obtain a similar $D(t)$ function to the one in Section 4.4. Thus, we reject the no clustering of cases hypothesis (2.20) with a significance level $2/(1+39) = 0.5$.

Finally we fit the Log-Gaussian Cox process with exponential covariance model to the pattern in the unit square. Now we take $t_0 = t_{max} = 0.2d_{max} = 0.2132024$, and $q = 1/4$:

$$\int_0^{t_{max}} \left((\hat{K}(s))^{1/4} - (K(s;\sigma^2,\delta))^{1/4} \right)^2 ds.$$

And via minimum contrast method we have obtained the parameters estimation:

$$\hat{\sigma^2} = 1.46731248, \quad \hat{\delta} = 0.09049584.$$

As we was expecting, we obtain a variance close to the obtained in Section 4.4. On the other hand, of course, the scale parameter δ has changed with respect to the Section 4.4, because the distances between the observations are different than the distances in the point pattern in the sub-region of Madrid. The robustness of the parameter estimation is followed from the previous case. In Figure 4.8 the reader can see the fitted K-function obtained, and four simulations of a Log-Gaussian Cox process under the exponential model with the estimated parameters (see more simulations in Appendix A.3).

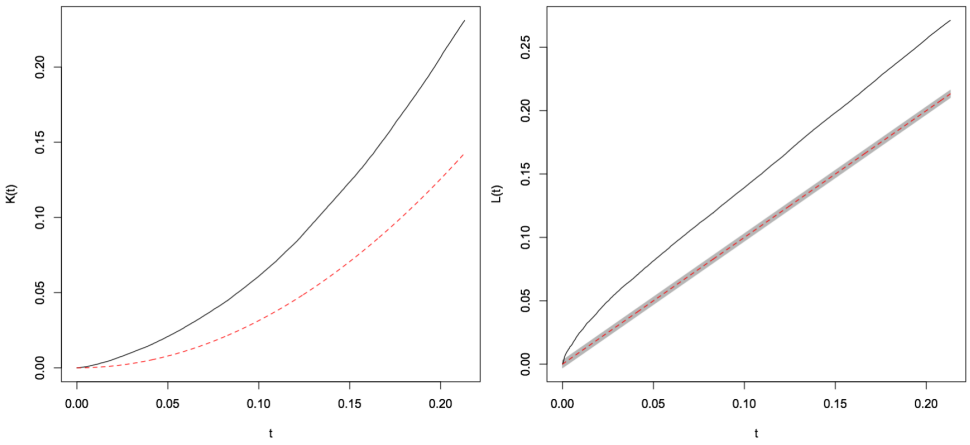

Figure 4.6: In the left: K-function estimation (solid black line) compared with the theoretical function K_{pois} (dashed red line). In the right: global envelopes with 19 simulations performed with fixed number of point, 600 like the observed point pattern. The empirical L-function is the solid black line, and the theoretical L-function is the dashed red line.

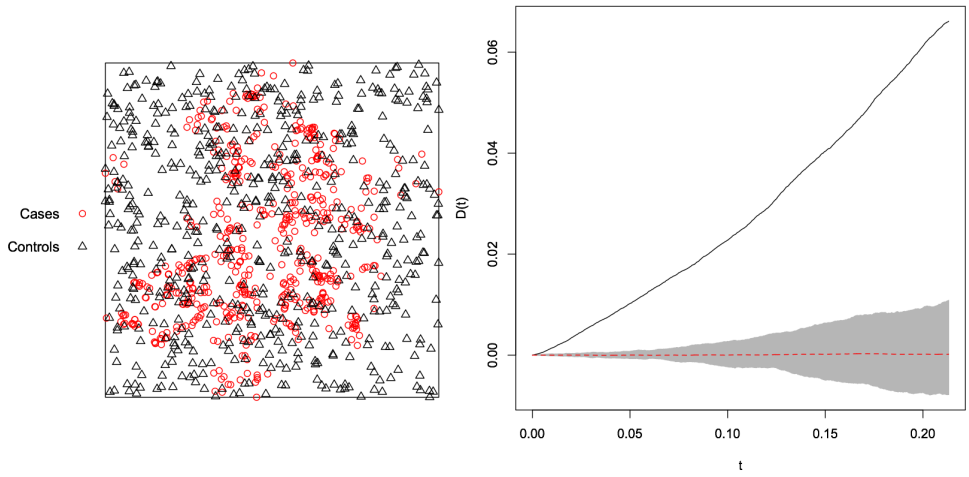

Figure 4.7: In the left: Case-control map. In the right: test of random labelling with 39 simulations. There is the observed $D(t)$ function (solid black line) and the theoretical $D(t)$ (dashed red line).

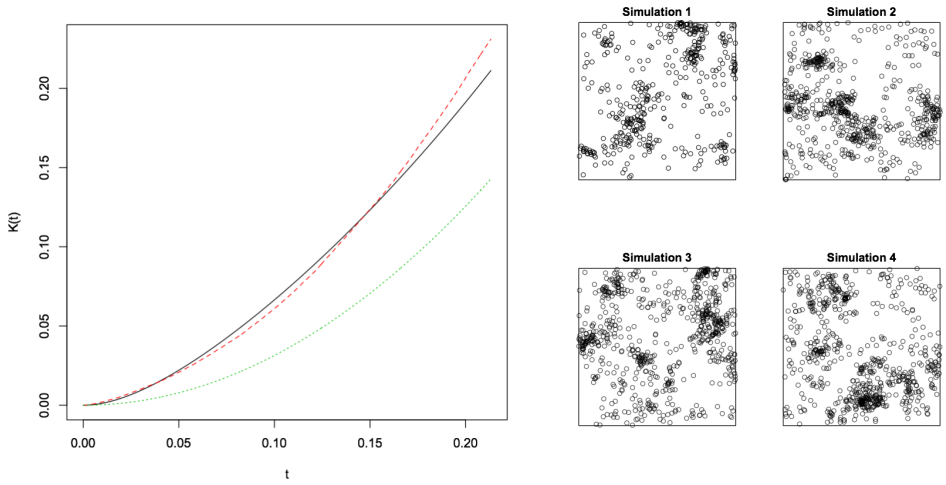

Figure 4.8: In the left: empirical K-function (red dashed line), fitted K-function (solid black line) and theoretical K-function (green dashed line). In the right: four simulations of a Log-Gaussian Cox process under the exponential model with parameters $\hat{\sigma^2} = 1.46731248, \hat{\delta} = 0.09049584$.

4.6 Case study 2: cockroaches sightings in the second half of 2013

Now, we will perform a second practical study. In this case, we will analyze the sightings of cockroaches during the second half of 2013 in the Madrid city. This time, we will be more ambitious taking observations from the dataset, and we will select a total amount of 1022 points. With this quantity of points we deal with a too high computational cost. Perform the same calculus and the tests done in the previous case requires a lot of computation time, and sometimes R is not able to perform the calculations. Thus, we will analyze the point pattern translated to the unit square, where the distances between points are lower and the computational cost is acceptable.

Let $\{(x_1, y_1), \ldots, (x_{1022}, y_{1022})\}$ be the 1022 points of the observed cockroaches sightings. We want to translate this points to the unit square. Let $x_{max}, y_{max}, x_{min}, y_{min}$ be the maximums and the minimums of the first and second coordinate respectively (i.e. $x_{max} = max\,(x_1, \ldots, x_{1022})$ and the analogous for the rest of the quantities). Then, we define the sequence of points:

$$\left\{ \left(\frac{x_1 - x_{min}}{x_{max} - x_{min}}, \frac{y_1 - y_{min}}{y_{max} - y_{min}} \right), \ldots, \left(\frac{x_{1022} - x_{min}}{x_{max} - x_{min}}, \frac{y_{1022} - y_{min}}{y_{max} - y_{min}} \right) \right\}. \tag{4.3}$$

Again, we have obtained the sequence of the observed points translated to the unit

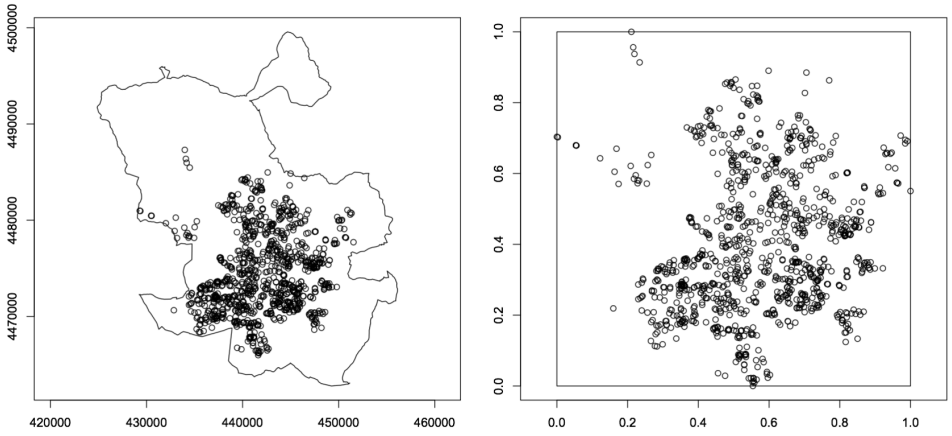

Figure 4.9: In the left: point pattern formed for the rat sightings during the second half of the 2013, plotted in Madrid city region. In the right: point pattern formed for the rat sightings during the second half of the 2013, plotted in the unit square.

square (see in Figure 4.9).

From now on, we will be working with the point process in the unit square. With this pattern we will be able to do computations faster, and we will obtain equivalent results than the analysis done with the entire region (see in Section 4.5 the Case study 1 performed in the unit square, and check that we obtain equivalent results to those of Section 4.4). First, we include a summary about the distances between points in the patterns (see in Table 4.3). We will combine distances of both patterns in Figure 4.9 in the summary, so that the reader can compare distances in each pattern.

As we did in Section 4.4, we will perform the computation with the 20% of the maximum distance between points, it is $t_{max} = 0.2d_{max} = 0.2 \times 1.0770672 = 0.2154134$. We first develop the descriptive part of the analysis. The estimation of the K-function results are in the left graphic in Figure 4.10. Again, like in the previous Case Study, we obtain a K-function above the theoretical one, which is a sign of aggregation. We perform a CSR test, using the L-function, with 19 simulations and a fixed number of 1022 points. In the right panel of Figure 4.10 we have the results of this test. Clearly, the L-function lies outside the shaded area, thus we can reject the null hypothesis of complete spatial randomness with a significance level of $1/(19+1) = 0.5$.

Now, we want to asses the spatial clustering. We assemble a Cases VS Controls scheme by simulating a random point pattern through the unit square and superimposing with the observed point pattern (see in the left part of Figure 4.11). And, using the random labelling test we see in the right part of Figure 4.11 that

	Madrid city	**Unit square**
Minimum	2	0.0000937
First quartile	4376	0.2007808
Median	6898	0.3164438
Mean	7200	0.3308691
Third quartile	9706	0.4462139
Maximum	23331	1.0770672

Table 4.3: Summary of Euclidean distances (in meters) between points in both patterns (Madrid city and Unit square).

the function $D(t) = K_{11}(t) - K_{22}(t)$ lies outside the shaded area. Thus, we can reject the null hypothesis of no spatial clustering of cases with a significance level of $2/(39+1) = 0.5$, and conclude that there is a spatial clustering of cases. Again, the reader can observe that the theoretical $D(t)$-function is near zero, which is what we expected. Wrapping this theoretical $D(t)$-function (dashed red line) we have the shaded area which is the area between the maximum and minimum $D(t)$-function of the 39 simulations under random labelling.

Finally, our goal is to fit a model to the observed pattern. Again we will fit a Log-Gaussian Cox process under the exponential model as we did in the previous study case. We will use the minimum contrast method with the aim of estimate the parameters of the model. We take $t_0 = t_{max}$ and $q = 1/4$. Then, we are working with (4.1), where σ^2, δ are the parameters to estimate. Using R, and taking $(\sigma^2, \delta) = (1, 1)$ as a initial parameters, we obtain the parameters estimation

$$\hat{\sigma^2} = 1.3307893, \quad \hat{\delta} = 0.1765721.$$

We have assessed again the robustness of the parameters estimation, we have followed the same scheme explained before, the reader can see the results in Appendix A.2. With this parameters the fitted K-function is quite similar to the observed one as we can see in the left of Figure 4.12. Finally, we do a simulation of four Log-Gaussian Cox processes under the exponential model with the estimated parameters (see in the right of Figure 4.12). We have added more simulations in Appendix A.3.

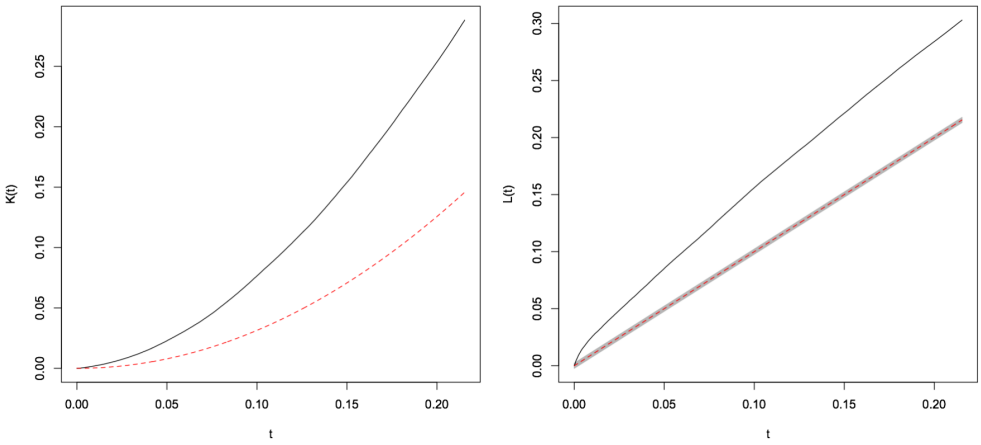

Figure 4.10: In the left: K-function estimation (solid black line) compared with the theoretical function K_{pois} (dashed red line). In the right: global envelopes with 19 simulations performed with fixed number of point, 1022 like the observed point pattern. The empirical L-function is the solid black line, and the theoretical L-function is the dashed red line.

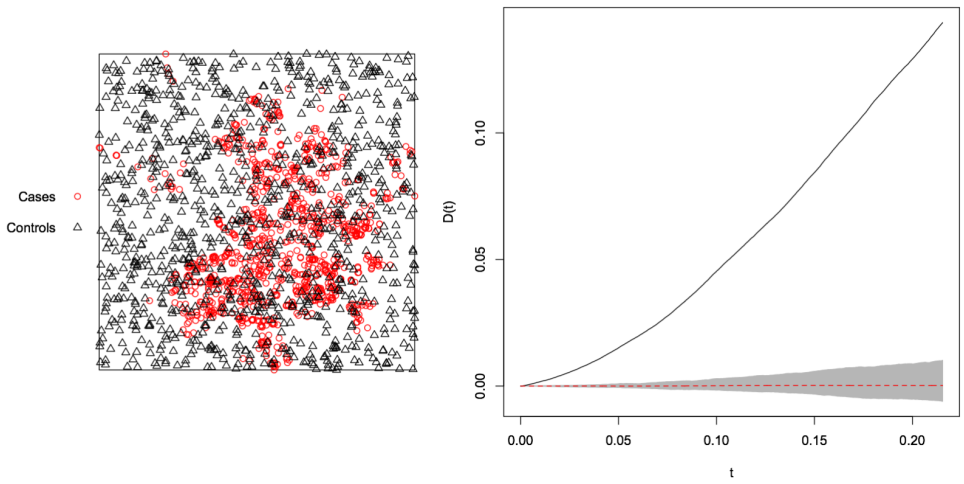

Figure 4.11: In the left: Case-control map. In the right: test of random labelling with 39 simulations. There is the observed $D(t)$ function (solid black line) and the theoretical $D(t)$ (dashed red line).

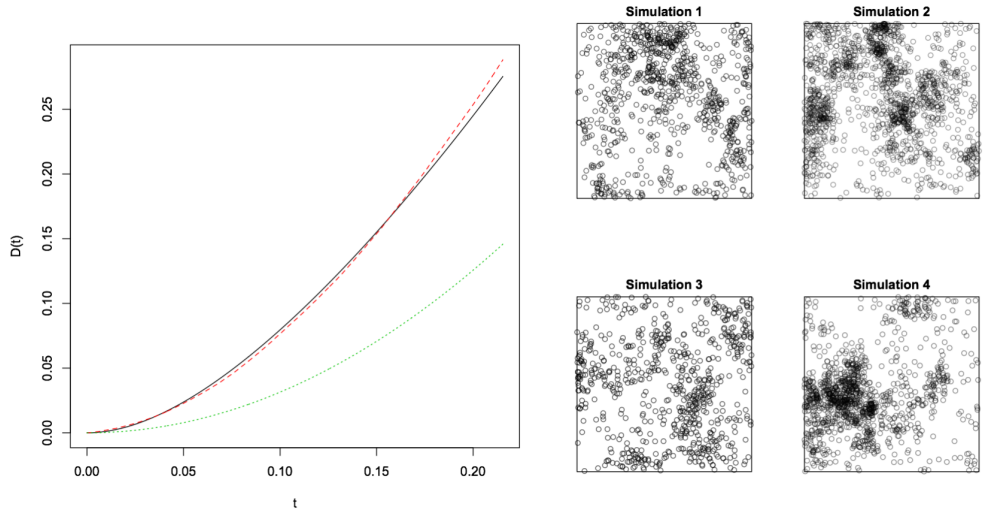

Figure 4.12: In the left: empirical K-function (red dashed line), fitted K-function (solid black line) and theoretical K-function (green dashed line). In the right: four simulations of a Log-Gaussian Cox process under the exponential model with parameters $\hat{\sigma^2} = 1.3307893, \hat{\delta} = 0.1765721$.

4.7 Conclusions of the case studies

After developing the spatial point analysis in the previous sections, we can draw conclusions about how rat and cockroach population is distributed through the surface of Madrid city. In both case studies, the observed sightings were in a similar sub-region of Madrid (see in Figure 4.1 and Figure 4.9). If we compare the obtained patterns with a Madrid city map, we can see that the observations are located in the districts of Madrid city with more population. Thus, we can deduce that in the core of the city, where there is the most human activity, is where we find most sightings.

With respect to the study performed, we can conclude that the distribution of rats and cockroaches in Madrid city is not random. It is a clustered distribution. It is important to mention that in these kind of studies there are covariates included (see Jorge Mateu et al 2014 [18]). This covariates are information about the environment in where we are performing the study, this information allows us to draw conclusions and explain the results obtained. We have not include covariates in this work because it was out of the scope of the project. Notwithstanding, the analysis developed enables us to conclude that the sightings are not randomly distributed. That the observations are spatially clustered. In addition, the fitted model and the simulations performed (see in previous sections and Appendix A.3) shows us the possible evolution of the distribution of sightings in Madrid city.

Chapter 5

Conclusions and future research

In this final chapter of the work, I would like to reflect on the objectives I had before I started and whether they have been met. I will also discuss the difficulties I have encountered, and possible future work related to this one.

The most important objectives were writing in a mathematical way the theory of spatial point processes found in different references. Spatial point processes and spatial statistics are not very popular topics. The references that I have consulted are not totally rigorous (in a mathematical way), because there are directed to scientific from different backgrounds. Thus, my aim in the work was understanding the theory set forth in these references and expose it mathematically, adding definitions, propositions and proofs. On the other hand, my other main goal of the work was develop an application of the theory exposed, and apply all my new knowledge in a real case. I believe that these objectives have been successfully met in this work. Finally, a wish I had was developing an application related with the health world. This is also done in this work.

It is important to mention the difficulties that I have found during the development of this work. The first difficulty was the "unpopularity" of the spatial point processes. The kind of stochastic processes studied in this project are not very widespread, and there are no professors in my faculty who deal with this topic or who have published on it. However I was lucky with this. After several attempts to contact mathematicians who have published on the subject, I contacted Jorge Mateu, who has helped and supported me throughout the work, and who is a great expert on the subject. The other huge difficulty I have found, has been achieving my goal of developing an application. For running an application like the performed in this work is necessary to have a dataset, and have it in the right shape. Having a database at your disposal is a very difficult thing to do because of the privacy involved with the data. Moreover, having it in the right format to be able to analyze it with R software is even more complicated. Here, I was really

lucky again, because Jorge Mateu gave me a database from one of his papers, a very complete database and in the right format. In addition, I had to learn how to use R, a software of which I did not have a very extensive knowledge. And I had to learn to use the R libraries necessary to develop the work presented here.

Finally, I would like to reflect on future work related to what we have seen in this project. I believe that spatial statistics and spatial point processes are a branch of mathematics with great potential, especially considering the evolution of information systems and the rise of data capture and analysis. I would like to continue deepening in this subject and try to exploit all the possibilities it has. It would be very interesting to be able to make a more complete case study than the one seen here, adding covariates and using other methods. I have also left pending to study the case of spatio-temporal point processes, which I think is a topic of great interest and that can allow very powerful studies.

I have invested a lot of time in this project, but every hour of work has been truly enjoyable. This work has been an opportunity to test myself in the world of research, trying to do a work on my own on a topic of interest to me. And the experience has been amazing. I hope that everyone who reads this work can enjoy it as much as I have enjoyed doing it.

Bibliography

[1] Adrian Baddeley et al. Analysing spatial point patterns in r. Technical report, Technical report, CSIRO, 2010. Version 4. Available at www. csiro. au ..., 2008.

[2] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial point patterns: methodology and applications with R*. CRC press, 2015.

[3] Adrian Baddeley and Rolf Turner. Introduction to spatstat. *Help manual for the R package spatstat*, pages 1–42, 2003.

[4] Adrian Baddeley and Rolf Turner. Package spatstat. *The Comprehensive R Archive Network ()*, page 146, 2014.

[5] Adrian Baddeley, Rolf Turner, and Ege Rubak. Getting started with spatstat. *For spatstat version*, pages 1–36, 2014.

[6] Roger Bivand, Micah Altman, Luc Anselin, Renato Assuncao, Olaf Berke, Andrew Bernat, and Guillaume Blanchet. Package spdep. *The comprehensive R archive network*, 604:605, 2015.

[7] Roger Bivand, Tim Keitt, Barry Rowlingson, Edzer Pebesma, Michael Sumner, Robert Hijmans, Even Rouault, and Maintainer Roger Bivand. Package rgdal. *Bindings for the Geospatial Data Abstraction Library. Available online: https://cran. r-project. org/web/packages/rgdal/index. html (accessed on 15 October 2017)*, page 172, 2015.

[8] Roger Bivand, Barry Rowlingson, Peter Diggle, Giovanni Petris, Stephen Eglen, and Maintainer Roger Bivand. Package splancs. *R package version*, pages 2–01, 2017.

[9] Roger S Bivand, Edzer J Pebesma, Virgilio Gomez-Rubio, and Edzer Jan Pebesma. *Applied spatial data analysis with R*, volume 747248717. Springer, 2008.

[10] Peter J Diggle. *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press, 2013.

[11] Peter J Diggle and Amanda G Chetwynd. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, pages 1155–1163, 1991.

[12] Peter J Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor. Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical modelling*, 2013.

[13] Jonatan A González, Francisco J Rodríguez-Cortés, Ottmar Cronie, and Jorge Mateu. Spatio-temporal point process statistics: a review. *Spatial Statistics*, 18:505–544, 2016.

[14] Jesper Moller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

[15] Edzer Pebesma, Benedikt Graeler, and Maintainer Edzer Pebesma. Package gstat. *Comprehensive R Archive Network (CRAN)*, pages 1–0, 2015.

[16] Carles Rovira. Processos estocàstics: un curs bàsic. *Dipòsit Universitat de Barcelona*, 1:1–0, 2021.

[17] Dietrich Stoyan, Wilfrid S Kendall, Sung Nok Chiu, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[18] Ibon Tamayo-Uria, Jorge Mateu, and Peter J Diggle. Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spatial Statistics*, 9:192–206, 2014.

[19] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. John Wiley & Sons, 2004.

Appendix A

Appendix

A.1 Poisson process simulations

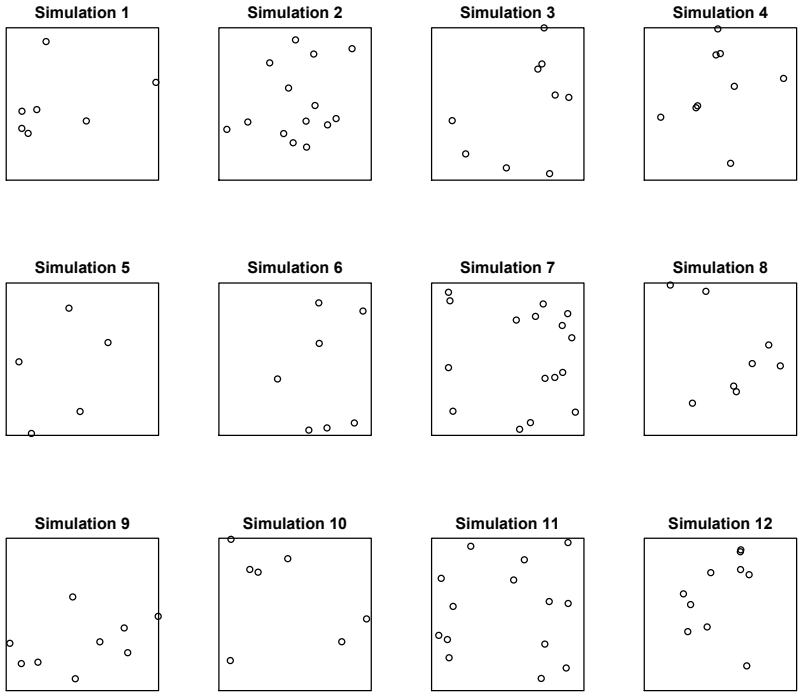

Figure A.1: Twelve simulations of a Poisson process with rate $\lambda = 10$ in the unit square.

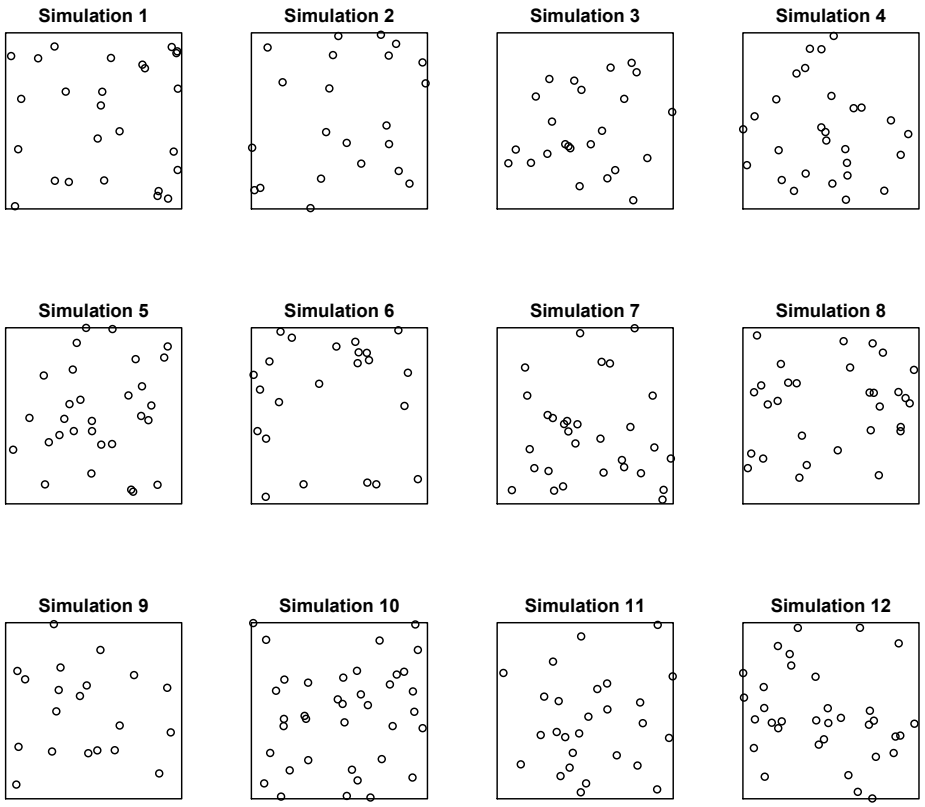

Figure A.2: Twelve simulations of a Poisson process with rate $\lambda = 25$ in the unit square.

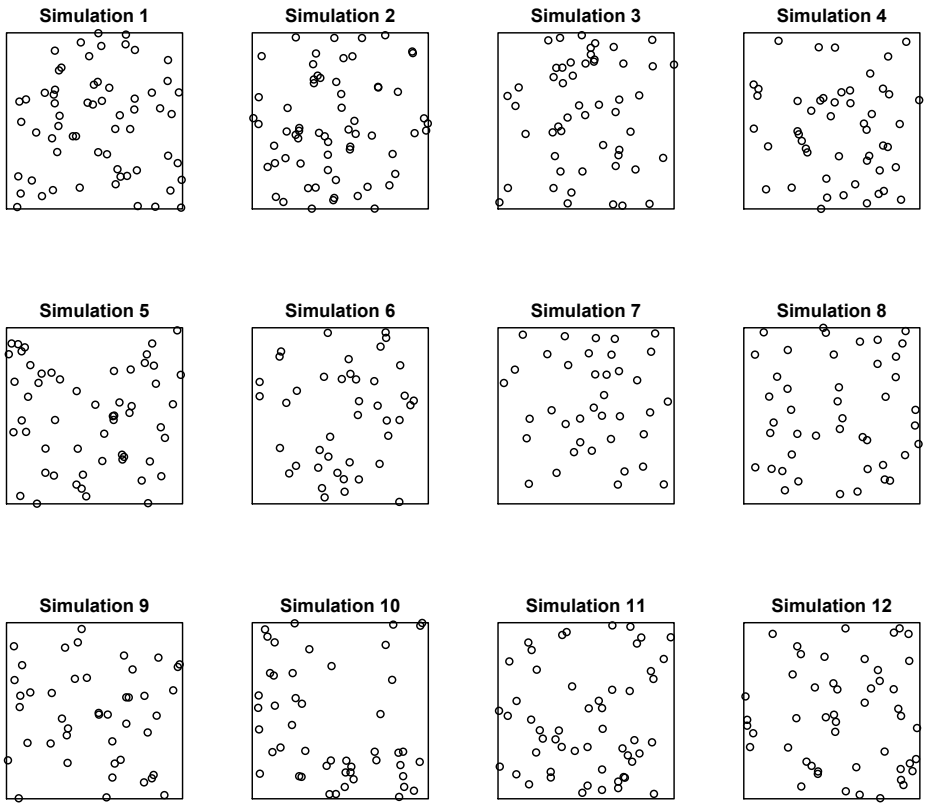

Figure A.3: Twelve simulations of a Poisson process with rate $\lambda = 50$ in the unit square.

A.2 Robustness of parameters estimation

In this appendix we develop again the parameter estimation of the study cases via the minimum contrast method. Now, we change some parameters for others nearby and we expose the obtained results. We want to see that for similar initial parameters we obtain similar estimations.

A.2.1 Case study 1: rat sightings in the first half of 2010

ϵ	t_0	$\hat{\sigma^2}$	$\hat{\delta}$
0.02	$0.22d_{max} = 4537.94$	1.400636	2424.622856
0.0175	$0.2175d_{max} = 4486.373$	1.409867	2386.004497
0.015	$0.215d_{max} = 4434.805$	1.4182	2352.7137
0.0125	$0.2125d_{max} = 4383.238$	1.426321	2319.969030
0.01	$0.21d_{max} = 4331.67$	1.434968	2286.528965
0.0075	$0.2075d_{max} = 4280.102$	1.444729	2248.549761
0.005	$0.205d_{max} = 4228.535$	1.453518	2215.806206
0.0025	$0.2025d_{max} = 4176.968$	1.462888	2182.082609
-0.0025	$0.1975d_{max} = 4073.832$	1.481286	2117.171700
-0.005	$0.195d_{max} = 4022.265$	1.493424	2076.410536
-0.0075	$0.1925d_{max} = 3970.698$	1.502898	2044.012228
-0.01	$0.19d_{max} = 3919.13$	1.513065	2011.067088
-0.0125	$0.1875d_{max} = 3867.562$	1.523925	1977.003723
-0.015	$0.185d_{max} = 3815.995$	1.537047	1936.250265
-0.0175	$0.1825d_{max} = 3764.427$	1.548069	1903.767867
-0.02	$0.18d_{max} = 3712.86$	1.559861	1870.964053

Table A.1: Results of parameter estimation via minimum contrast method changing t_0 for $t_0 = 0.2d_{max} + \epsilon d_{max}$ where $\epsilon \in [-0.02, 0.02]$. With ϵ such that all the values are with a distance of 0.0025 with the nearest one.

	$\hat{\sigma^2}$	$\hat{\delta}$
Minimum	1.401	1871
First quartile	1.433	2003
Median	1.472	2150
Mean	1.476	2147
Third quartile	1.516	2295
Maximum	1.560	2425

Table A.2: Summary of the results exposed in Table A.1.

A.2.2 Case study 2: cockroaches sightings in the second half of 2013

	$\hat{\sigma^2}$	$\hat{\delta}$
Minimum	1.301	0.1616
First quartile	1.315	0.1688
Median	1.331	0.1764
Mean	1.332	0.1762
Third quartile	1.349	0.1839
Maximum	1.367	0.1900

Table A.3: Summary of the results exposed in Table A.4.

ϵ	t_0	$\hat{\sigma^2}$	$\hat{\delta}$
0.02	$0.22d_{max} = 0.2369548$	1.301448	0.189966
0.0175	$0.2175d_{max} = 0.2342621$	1.3051449	0.1882284
0.015	$0.215d_{max} = 0.2315694$	1.3082686	0.1867694
0.0125	$0.2125d_{max} = 0.2288768$	1.3121088	0.1849865
0.01	$0.21d_{max} = 0.2261841$	1.3153254	0.1834947
0.0075	$0.2075d_{max} = 0.2234914$	1.3193977	0.1816354
0.005	$0.205d_{max} = 0.2207988$	1.3229068	0.1800804
0.0025	$0.2025d_{max} = 0.2181061$	1.3270974	0.1781949
-0.0025	$0.1975d_{max} = 0.2127208$	1.3352989	0.1746057
-0.005	$0.195d_{max} = 0.2100281$	1.3391478	0.1729594
-0.0075	$0.1925d_{max} = 0.2073354$	1.3438271	0.1709753
-0.01	$0.19d_{max} = 0.2046428$	1.3477056	0.1693336
-0.0125	$0.1875d_{max} = 0.2019501$	1.3526342	0.1672842
-0.015	$0.185d_{max} = 0.1992574$	1.3568711	0.1655671
-0.0175	$0.1825d_{max} = 0.1965648$	1.3620911	0.1634383
-0.02	$0.18d_{max} = 0.1938721$	1.3667297	0.1616046

Table A.4: Results of parameter estimation via minimum contrast method changing t_0 for $t_0 = 0.2d_{max} + \epsilon d_{max}$ where $\epsilon \in [-0.02, 0.02]$. With ϵ such that all the values are with a distance of 0.0025 with the nearest one.

A.3 More simulations of the fitted models

In this appendix we include more simulations of the fitted models in the case studies. Thus, our goal in this appendix is simulate a Log-Gaussian Cox process under the exponential model with the corresponding parameters estimated for each case study.

The Figures A.4, A.5, A.6 are more simulations of the fitted model saw in Section 4.4. Figures A.7, A.8, A.9 of the model in Section 4.5. Figures A.10, A.11, A.12 of the model in Section 4.6.

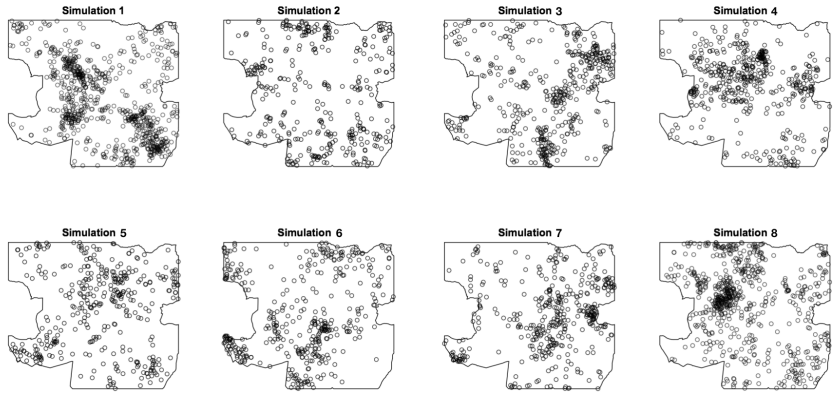

Figure A.4: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.471686, \hat{\delta} = 2150.178436$ in a sub-region of Madrid city.

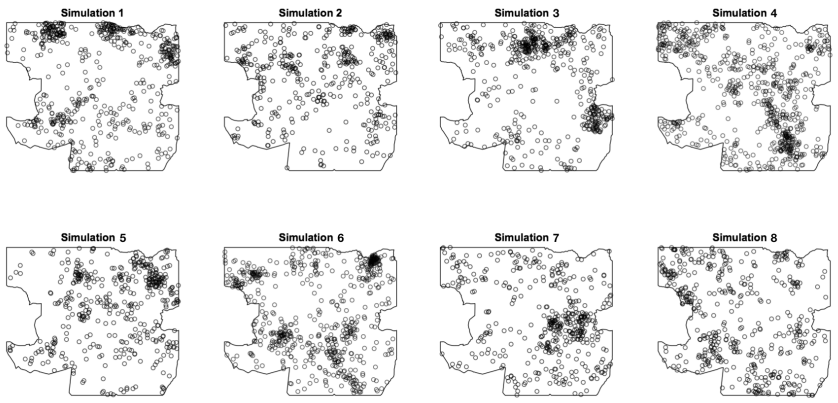

Figure A.5: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.471686, \hat{\delta} = 2150.178436$ in a sub-region of Madrid city.

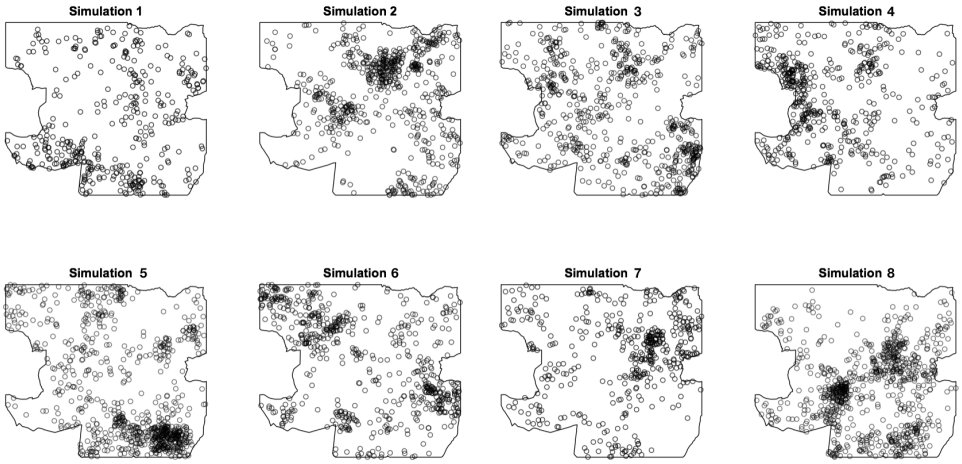

Figure A.6: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.471686, \hat{\delta} = 2150.178436$ in a sub-region of Madrid city.

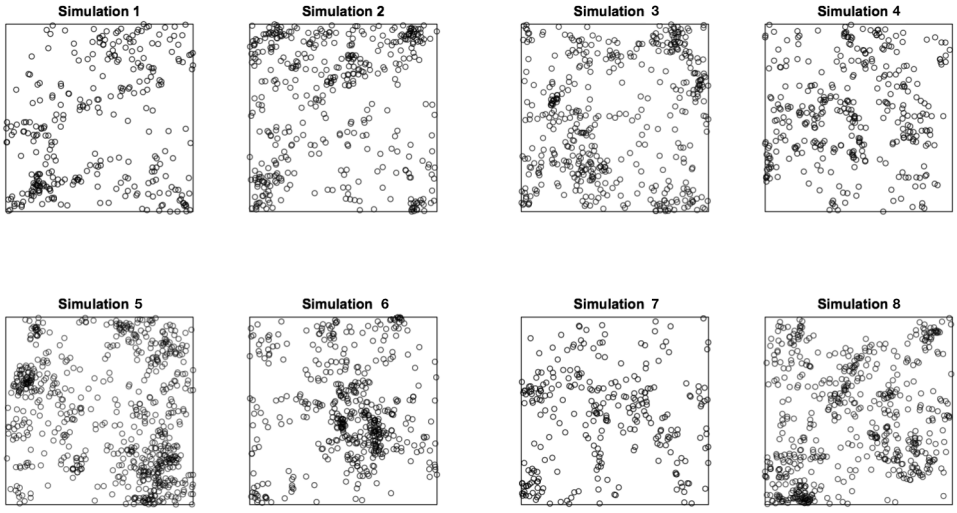

Figure A.7: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.46731248, \hat{\delta} = 0.09049584$ in the unit square.

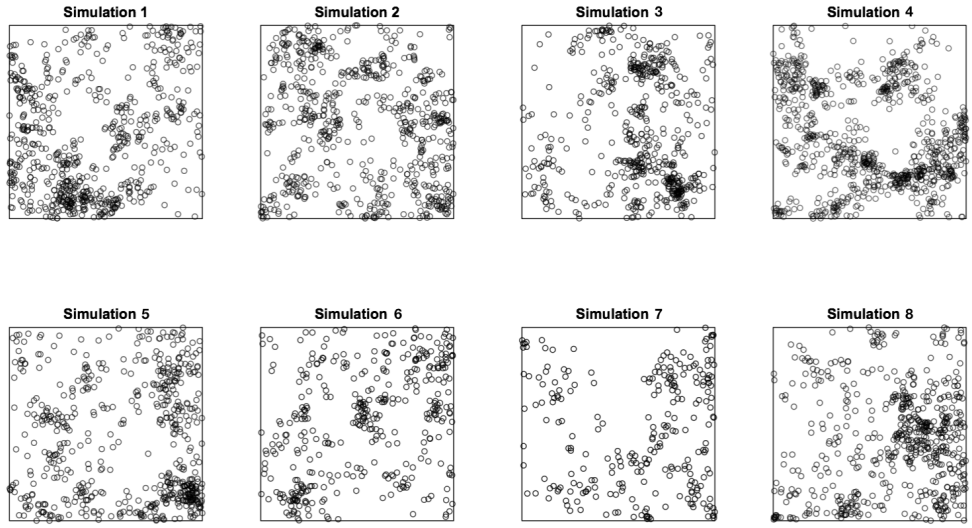

Figure A.8: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.46731248, \hat{\delta} = 0.09049584$ in the unit square.

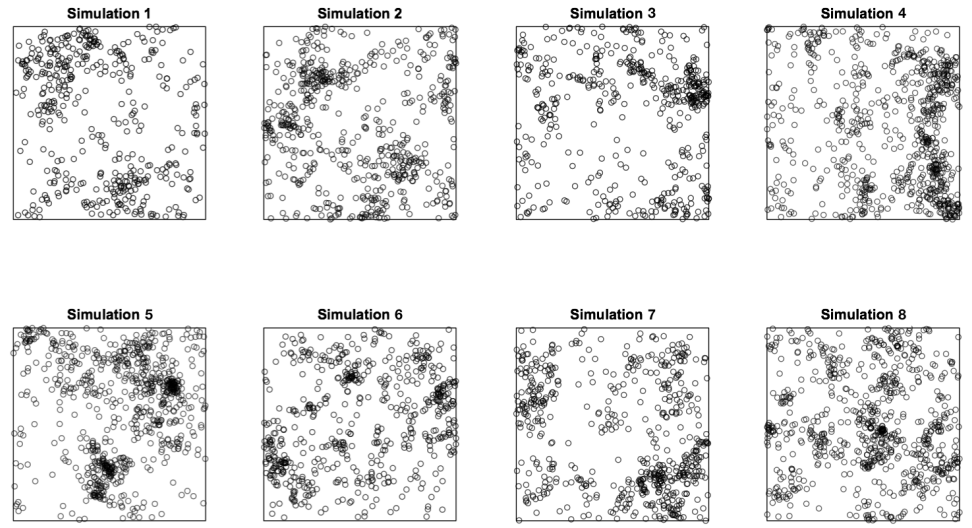

Figure A.9: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.46731248, \hat{\delta} = 0.09049584$ in the unit square.

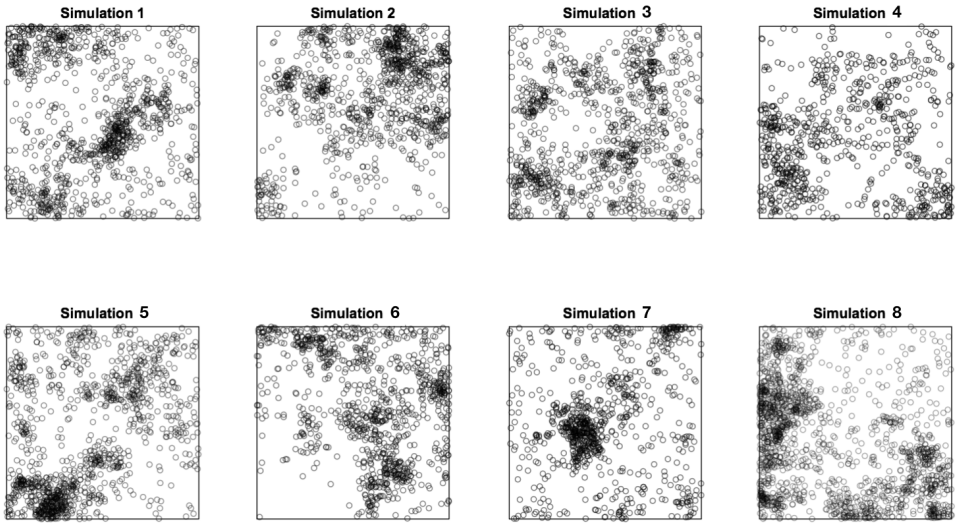

Figure A.10: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.3306416, \hat{\delta} = 0.1766353$ in the unit square.

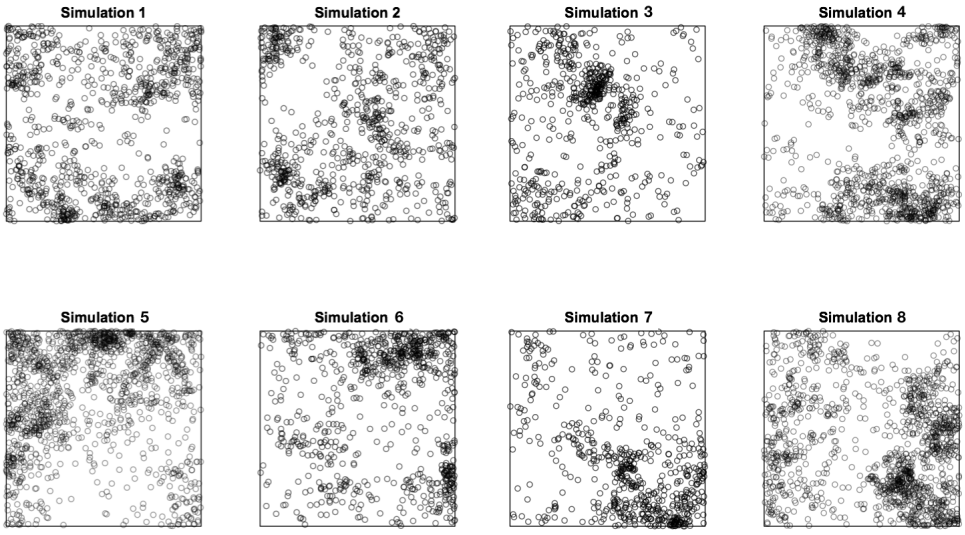

Figure A.11: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.3306416, \hat{\delta} = 0.1766353$ in the unit square.

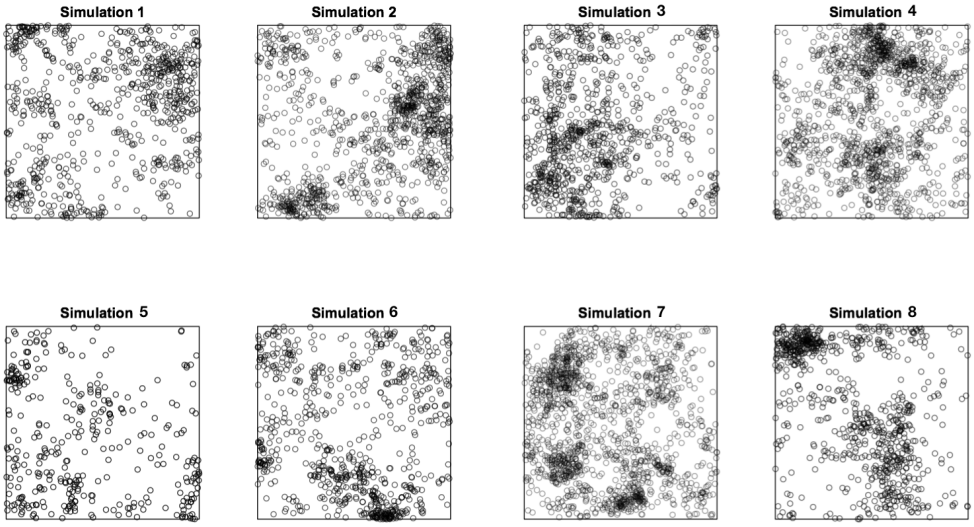

Figure A.12: Eight simulations of a Log-Gaussian Cox process under the exponential model and with parameters $\hat{\sigma^2} = 1.3306416, \hat{\delta} = 0.1766353$ in the unit square.

Appendix B

R code

In this appendix we will include all the R code used in the work. In order to let the reader track the relation between the memory and this code, we will explain on where was the coded used during the work.

During all programming we will use the *spatstat* package, thus it is mandatory to execute the command:

```
> library(spatstat)
```

To plot the point patterns in Figure 2.1, estimate the *K*-function of these patterns (see in Figure 3.1) and perform a CSR test (see in Figure 3.2) we have used:

```
> y <- seq(from=0, to=1, by=0.1)
> i<-0
> X0 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> plot(X0)
> i<-0.1
> X1 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.2
> X2 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.3
> X3 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.4
> X4 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.5
> X5 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.6
> X6 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.7
> X7 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
```

```
> i<-0.8
> X8 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-0.9
> X9 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> i<-1
> X10 <- ppp(c(i,i,i,i,i,i,i,i,i,i), y)
> S<-superimpose(X1,X2,X3,X4,X5,X6,X7,X8,X9,X10,X0)
> plot(S,main="")
> k1<-Kest(S, correction="isotropic")
> plot(k1, main="", xlab="t", ylab="K(t)", legend=FALSE)
> LE1<-envelope(S,Lest,nsim=19,nrank=1,global=TRUE,fix.n=TRUE)
> plot(LE1, main="", xlab="t", ylab="L(t)", legend=FALSE)
> indep<-rpoispp(lambda=50, win=square(1))
> plot(indep, main="")
> k2<-Kest(indep, correction="isotropic")
> plot(k2, main="", xlab="t", ylab="K(t)", legend=FALSE)
> LE2<-envelope(indep,Lest,nsim=19,nrank=1,global=TRUE,fix.n=TRUE)
> plot(LE2, main="", xlab="t", ylab="L(t)", legend=FALSE)
> clust <- rMatClust(12, 0.07, 5)
> plot(clust,main="")
> k3<-Kest(clust, correction="isotropic")
> plot(k3, main="", xlab="t", ylab="K(t)", legend=FALSE)
> LE3<-envelope(clust,Lest,nsim=19,nrank=1,global=TRUE,fix.n=TRUE)
> plot(LE3, main="", xlab="t", ylab="L(t)", legend=FALSE)
```

In the Section 3.4 we exposed a test of random labelling. In this section we presented an example of this test (see Figure 3.3 and Figure 3.4). For this we had to use the code:

```
> alpha<-rpoispp(75, win=owin(c(0:1),c(0:1)))
> X<-rlabel(alpha,labels=factor(c("Type1","Type2")), permute=FALSE)
> plot(X, cols=c(4,2), main="")
> beta1<-rMatClust(10, 0.05, 4, win=square(1))
> beta2<-rpoispp(70, win=owin(c(0:1),c(0:1)))
> S<-superimpose(Type1=beta1, Type2=beta2)
> plot(S, cols=c(4,2), main="")
> Kdif<- function(X, ..., i, j, k, l){
    Kicross <- Kcross(X, ..., i=i, j=j)
    Kjcross <- Kcross(X, ..., k=k, l=l)
    dif <- eval.fv(Kicross - Kjcross)
    return(dif)
```

```
}
> E1 <- envelope(S, Kdif, nsim=39, i="Type1", j="Type1", k="Type2",
    l="Type2", simulate = expression(rlabel(S)))
> E2 <- envelope(X, Kdif, nsim=39, i="Type1", j="Type1", k="Type2",
    l="Type2", simulate = expression(rlabel(X)))
> plot(E1, main="", legend=FALSE, xlab="t", ylab="D(t)")
> plot(E2, main="", legend=FALSE, xlab="t", ylab="D(t)")
```

In the case studies we have used the dataset shared by Jorge Mateu, thus, in order to develop this part of the work it is necessary to load the *.RData* file. Then, we will have charged in our workspace in R the data. It is also important mention that for this part of the work another R libraries have to be used:

```
> library(gstat)
> library(sp)
> library(rgdal)
> library(maptools)
> library(spdep)
> library(splancs)
> library(stpp)
```

Once the data is prepared, to perform the analysis exposed in the Case 1 (see Section 4.4) we have used the following code:

```
> dibujo<-as(rg,"SpatialPolygons")
> madrid<-as(dibujo,"owin")
> A<-rats[c(1:600),]
> Ax<-A[,2]
> Ay<-A[,3]
> W <- owin(c(430000, 453500), c(4465000, 4485200))
> phi<-ppp(Ax,Ay,madrid[W])
> phi<-unique(phi)
> plot(phi, main="", cols=1)
> summary(as.vector(dist(cbind(phi$x,phi$y))))
> dt<-0.2*20627
> Kphi<-Kest(phi, correction="isotropic", rmax=dt)
> plot(Kphi, main="", xlab="t", ylab="K(t)", legend=FALSE)
> LEphi<-envelope(phi,Lest,nsim=19,nrank=1, global=TRUE,
fix.n=TRUE, funargs=c(rmax=dt))
> plot(LEphi, main="", xlab="t", ylab="L(t)", legend=FALSE)
> Cphi<-rpoint(600,win=madrid[W])
> Sphi<-superimpose(Cases=phi,Controls=Cphi)
```

```
> plot(Sphi, main="",cols=c(2,1))
> Kdif<- function(X, ..., i, j, k, l){
          Kicross <- Kcross(X, ..., i=i, j=j)
          Kjcross <- Kest(X, ..., k=k, l=l)
          dif <- eval.fv(Kicross - Kjcross)
          return(dif)
}
> EK <- envelope(Sphi, Kdif, nsim=39, i="Cases", j="Cases",
k="Controls", l="Controls", simulate = expression(rlabel(Sphi)),
funargs=c(rmax=dt))
> plot(EK, main="", xlab="t", ylab="D(t)", legend=FALSE)
> modelphi<-lgcp.estK(phi, startpar=c(var=1, scale=1),
covmodel=list(model="exponential"), rmax=dt)
> plot(modelphi, main="", xlab="t", ylab="K(t)", legend=FALSE)
> varphi<-1.471686
> scalephi<-2150.178436
> muphi<-log(npoints(phi)/area(madrid[W])) - varphi/2
> Xphi<-rLGCP(model="exp", mu=muphi, var=varphi, scale=scalephi,
win=madrid[W])
> plot(Xphi, main="")
```

Similarly, we use the following code for develop the Case 2 (see Section 4.6), now taking into account that we translate the point pattern observed to the unit square:

```
> dibujo<-as(rg,"SpatialPolygons")
> madrid<-as(dibujo,"owin")
> B<-cocks[c(6650:7793),]
> Bx<-B[,2]
> By<-B[,3]
> phi<-ppp(Bx, By, madrid)
> phi<-unique(phi)
> plot(phi, main="", cols=1)
> D<-density(phi)
> summary(as.vector(dist(cbind(phi$x,phi$y))))
> minx<-min(Bx)
> maxx<-max(Bx)
> miny<-min(By)
> maxy<-max(By)
> u<-Bx
> v<-By
```

```
> u2<-(u-minx)/(maxx-minx)
> v2<-(v-miny)/(maxy-miny)
> unit<-ppp(u2,v2,win=square(1))
> unit<-unique(unit)
> plot(unit, main="", cols=1)
> summary(as.vector(dist(cbind(unit$x,unit$y))))
> dist<-0.2*1.0770672
> kunit<-Kest(unit, correction="isotropic", rmax=dist)
> plot(kunit, main="", legend=FALSE, xlab="t", ylab="K(t)")
> LEunit<-envelope(unit,Lest,nsim=19,nrank=1, global=TRUE,
fix.n=TRUE, funargs=c(rmax=dist))
> plot(LEunit, main="", legend=FALSE, xlab="t", ylab="L(t)")
> C2<-rpoint(1022,win=square(1))
> S2<-superimpose(Cases=unit,Controls=C2)
> plot(S2, main="", cols=c(2,1))
> Kdif<- function(X, ..., i, j, k, l){
          Kicross <- Kcross(X, ..., i=i, j=j)
          Kjcross <- Kcross(X, ..., k=k, l=l)
          dif <- eval.fv(Kicross - Kjcross)
          return(dif)
}
> EKunit <- envelope(S2, Kdif, nsim=39, i="Cases", j="Cases",
k="Controls", l="Controls", simulate = expression(rlabel(S2)),
funargs=c(rmax=dist))
> plot(EKunit, main="", legend=FALSE, xlab="t", ylab="D(t)")
> modelunit<-lgcp.estK(unit, startpar=c(var=1, scale=1),
covmodel=list(model="exponential"), rmax=dist)
> plot(modelunit, main="", legend=FALSE, xlab="t", ylab="K(t)")
> varunit<-1.3307893
> scaleunit<-0.1765721
> muunit<-log(npoints(unit))-varunit/2
> Xunit<-rLGCP(model="exp", mu=muunit, var=varunit,
scale=scaleunit, win=square(1))
> plot(Xunit, main="", cols=1)
```

In Section 4.5 we develop the Case study 1 in the unit square. Thus, we have used a similar R code but translating the point pattern to the unit square:

```
> dibujo<-as(rg,"SpatialPolygons")
> madrid<-as(dibujo,"owin")
> A<-rats[c(1:600),]
```

```
> Ax<-A[,2]
> Ay<-A[,3]
> W <- owin(c(430000, 453500), c(4465000, 4485200))
> phi<-ppp(Ax, Ay, madrid[W])
> phi<-unique(phi)
> D<-density(phi)
> summary(as.vector(dist(cbind(phi$x,phi$y))))
> minx<-min(Ax)
> maxx<-max(Ax)
> miny<-min(Ay)
> maxy<-max(Ay)
> u<-Ax
> v<-Ay
> u2<-(u-minx)/(maxx-minx)
> v2<-(v-miny)/(maxy-miny)
> unit<-ppp(u2,v2,win=square(1))
> unit<-unique(unit)
> plot(unit, main="", cols=1)
> summary(as.vector(dist(cbind(unit$x,unit$y))))
> dist<-0.2*1.0660120
> kunit<-Kest(unit, correction="isotropic", rmax=dist)
> plot(kunit, main="", legend=FALSE, xlab="t", ylab="K(t)")
> LEunit<-envelope(unit,Lest,nsim=19,nrank=1, global=TRUE,
fix.n=TRUE, funargs=c(rmax=dist))
> plot(LEunit, main="", legend=FALSE, xlab="t", ylab="L(t)")
> C2<-rpoint(600,win=square(1))
> S2<-superimpose(Cases=unit,Controls=C2)
> plot(S2, main="", cols=c(2,1))
> Kdif<- function(X, ..., i, j, k, l){
          Kicross <- Kcross(X, ..., i=i, j=j)
          Kjcross <- Kcross(X, ..., k=k, l=l)
          dif <- eval.fv(Kicross - Kjcross)
          return(dif)
}
> EKunit <- envelope(S2, Kdif, nsim=39, i="Cases", j="Cases",
k="Controls", l="Controls", simulate = expression(rlabel(S2)),
funargs=c(rmax=dist))
> plot(EKunit, main="", legend=FALSE, xlab="t", ylab="D(t)")
> modelunit<-lgcp.estK(unit, startpar=c(var=1, scale=1),
```

```
covmodel=list(model="exponential"), rmax=dist)
> plot(modelunit, main="", legend=FALSE, xlab="t", ylab="K(t)")
> varunit<-1.46731248
> scaleunit<-0.09049584
> muunit<-log(npoints(unit))-varunit/2
> Xunit<-rLGCP(model="exp", mu=muunit, var=varunit,
scale=scaleunit, win=square(1))
> plot(Xunit, main="", cols=1)
```

In order to do the eight simulations of a Poisson process with rate $\lambda = 10$, $\lambda = 25$ and $\lambda = 50$ (see Appendix A.1), we have used:

```
> PP1 <- rpoispp(10, win=square(1), nsim=12)
> plot(PP1, main="")
> PP2 <- rpoispp(25, win=square(1), nsim=12)
> plot(PP2, main="")
> PP3 <- rpoispp(50, win=square(1), nsim=12)
> plot(PP3, main="")
```

Finally, in Appendix A.3 we only use several times the command:

```
> X<-rLGCP(model="exp", mu=muunit, var=varunit,
scale=scaleunit, win=window, nsim=r)
> plot(X, main="")
```

Where window, depending on the point pattern analyzed, is the sub-region of Madrid city or the unit square. And r is the number of simulations pf the Log-Gaussian Cox process under the exponential model. Also the parameters mu, var, scale, changes depending on the pattern in question.