# 4. Spatial distribution of points

**JEAN-MICHEL FLOCH**
*INSEE*

**ERIC MARCON**
*AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana.*

**FLORENCE PUECH**
*RITM, Univ. Paris-Sud, Université Paris-Saclay & CREST, 92330 Sceaux, France.*

**Abstract**

Statisticians carry out close examination of spatialized data, such as the distribution of household income, the location of industrial or commercial establishments, the distribution of schools in cities, etc. Answers can be found through analyses of one or more predefined geographical scales such as neighbourhoods, districts or statistical blocks. However, it is tempting to preserve the individual data and to work with the exact position of the entities that are being studied. If that is the case, statisticians have to conduct analyses based on geolocation data without carrying out any geographical aggregation. Observations are taken as points in space and the objective is to characterise these point distributions.

Understanding and mastering statistical methods that process this individual and spatialized information enables us to work on data that are now increasingly accessible and sought after because they provide very precise analyses of distributions studied (Ellison et al. 2010; Barlet et al. 2013). In this framework of analysis, statisticians who have sets of points to analyse are faced with several important methodological questions: how can such data with thousands or even millions of observations be represented and characterised spatially? What statistical tools exist that can be used to study these observations relating to households, employees, firms, stores, equipment or travel, for example? How can the qualitative or quantitative characteristics of the observations being studied be taken into account? How can any attractions or repulsions between points or between different types of points be highlighted? How can we assess the significance of the results obtained, etc?

The purpose of this chapter is to help statisticians to provide statistically robust results from the study of spatialized data that is not based on predefined zoning. To do this, we will review the literature on the subject of statistical methods used to characterise point distributions and we will explain the associated issues. We will use simple examples to explain the advantages and disadvantages of the most frequently adopted approaches. The code provided in R will be used to reproduce the examples covered.

**Acknowledgements:** The authors would like to thank Gabriel LANG and Salima BOUAYAD AGHA for their careful review of the first version of this chapter and for all their constructive comments. Thanks also to Marie-Pierre de BELLEFON and Vincent LOONIS who provided the initiative for this project: this chapter has undeniably benefited from all their editorial efforts and those of Vianney COSTEMALLE.

## Introduction

The study of spatial distributions of points may seem more removed from the concerns of public statisticians than some other methods. So why give them a place in this manual? The answer is simple: geolocation of data provides numerous localised observations on firms, facilities and housing. This swiftly leads us to consider the possibility of gathering together these observations, the spatial configuration of their random, or non-random setting, and their dependence on other processes (the proximity of industrial establishments with strong *input-output* links may be desirable and therefore lead to spatial interactions between establishments from different sectors). The aim of this chapter is to present an introduction to a body of methods that are sometimes complex in their mathematical foundations, but which often serve to illustrate quite simple questions. The development of these methods was based in the issues facing ecologists, foresters and epidemiologists. P.J. Diggle, the author of the first reference work (Diggle 1983), is known for his extensive epidemiology work (Diggle et al. 1991). As a result, educational examples illustrating point processes often come from forestry or epidemiological data. In this chapter we will use examples of this type provided in certain R packages such as *spatstat* (Baddeley et al. 2005) or *dbmss* (Marcon et al. 2015b). We will also use data on the location of facilities in France.

Unlike zoning or geostatistical methods, when studying spatial distribution, a variable is not measured locally, but the very location of the points is at the heart of the subject in question. We will build models and make inferences based on these points.

The maps in Figure 4.1, produced from data in the permanent database of facilities (BPE), show four examples of the location of activities in the city of Rennes (France). [1]

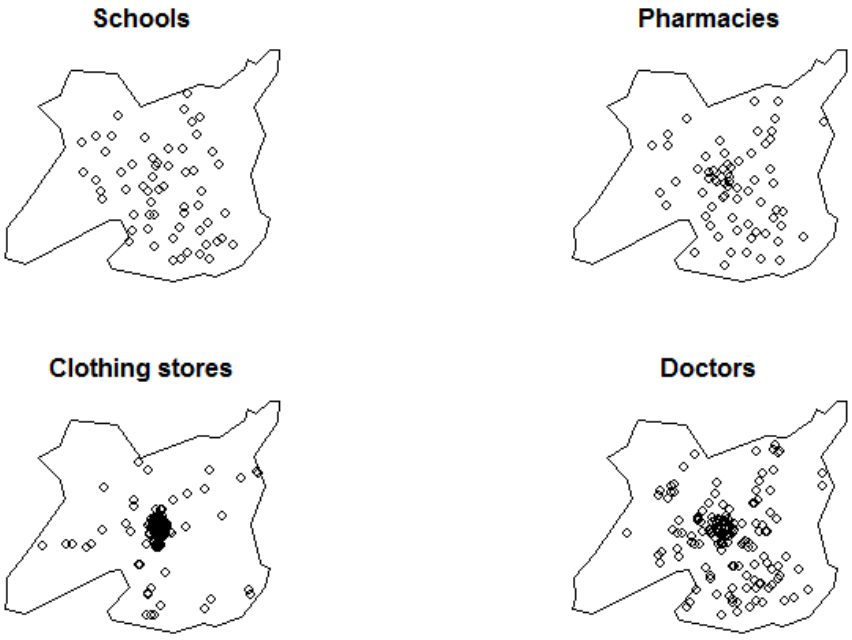


Figure 4.1 – Four examples of the location of activities in the municipality of Rennes in 2015
**Source:** *INSEE-BPE, authors' calculations*

1. If equipment is positioned imprecisely, it is assigned by default to the centroid of the associated IRIS (INSEE zoning in "**I**lots **R**egroupés pour l'**I**nformation **S**tatistique" that can be translated as "aggregated units for statistical information", *see* https://www.insee.fr/en/metadonnees/definition/c1523).

```
library("spatstat")
library("sp")
# BPE file on the INSEE.fr site: https://www.insee.fr
# Data for these examples:
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))
bpe_sch <- bpe[bpe$TYPEQU=="C104", ]
bpe_pha <- bpe[bpe$TYPEQU=="D301", ]
bpe_clo <- bpe[bpe$TYPEQU=="B302", ]
bpe_doc <- bpe[bpe$TYPEQU=="D201", ]
par(mfrow=c(2,2), mar=c(2, 2, 2, 2))
plot(carte, main="Schools") ; points(bpe_sch[, 2:3])
plot(carte, main="Pharmacies") ; points(bpe_pha[, 2:3])
plot(carte, main="Clothing stores") ; points(bpe_clo[, 2:3])
plot(carte, main="Doctors") ; points(bpe_doc[, 2:3])
par(mfrow=c(1,1))
```

These four simple figures provide an initial overview of the major differences in the locations of these facilities. There is a large number of clothing stores, but they are extremely concentrated in the center of Rennes. On the other hand, primary schools seem to be distributed more evenly. Pharmacies are also evenly distributed, but with a greater presence in the city center. The location of doctors is more aggregated than that of pharmacies, but less so than that of clothing stores. These initial conclusions on the distribution of activities could be supplemented by more advanced spatial analyses, for example by applying data for population distribution or accessibility (closer to or further away from the main communication routes). The methods presented in this manual make it possible to go beyond the conclusions of these first maps, which are certainly informative but insufficient to characterise and explain the location of the entities in question.

In this chapter, we have chosen not to deal with methods that discretise space, *i.e.* approaches based on study zoning (such as employment zones in France based on commuting patterns) or administrative zoning (such as the breakdown of the Nomenclature of Territorial Units for Statistics - NUTS - from Eurostat). Specific works (Combes et al. 2008) provide a very good introduction to this subject for any interested readers. This chapter will be limited to methods that take into account the exact geographical position of the entities studied. Our choice is motivated by at least two factors. The first is linked to access to such data on a large scale and the development of appropriate technical methods to analyse them in a meaningful way. Different packages are, for example, accessible in the R software. The second is that by favouring methods that preserve the nature of the individual data analysed (position in space, characteristics), the Modifiable Areal Unit Problem - MAUP, well known to geographers (Openshaw et al. 1979a), will be avoided. MAUP refers to the fact that the discretisation of initially non-aggregated data potentially creates several statistical biases linked to the position of borders, aggregation level etc. (Briant et al. 2010).

## 4.1 Framework of analysis: basic concepts

This section aims to define the fundamental concepts we will use in this chapter to explain statistical methods of spatial analysis of point data.

### 4.1.1 Configurations and processes

To study these empirical **spatial distribution** of points (or set of points), we use the random point process theory. A point process can be used to randomly generate an infinity of outcomes, which share a number of properties.

Usually, we note the point process as $X$ and a realization from this process as $S$. Spatial distributions are modelled using inferential methods that apply to objects that are observed only once. For example, for many data, statisticians only have one set of points observed at a given date. Therefore, there is only one distribution of doctors in the city of Rennes (see figure 4.1), bus stops in London, housing in Friesland in the Netherlands or cinemas in Belgium on a given date. However, the unique observed realization must not alter our analysis: we will, therefore, ensure that the available data is able to provide a good approximation of the point process that generated it. We will come back to this in this chapter.

**Definition 4.1.1 — Spatial distribution.** A distribution of $n$ points, written $C = \{x_1, ..., x_n\}$ is a set of points from $\mathbb{R}^2$ in this chapter: the objects are located on a map. The theory does not limit the dimension of space but applications in three-dimensional spaces are rare, and almost non-existent in $\mathbb{R}^d$, $d > 3$. The number of points in the distribution is noted as $n(C)$. The points are not considered to be duplicated, as this would prevent many methods from being used. **The combined points in the region $B$ is written $C \cap B$ , and $n(C \cap B)$.**

The process $X$ is defined if the number of points $n(X \cap B)$ is known for any region $B$. The number of points is also written $N(B)$ if no confusion is possible. In general, we are limited to locally finite processes, for which $n(X \cap B) < +\infty, \forall B$.

### 4.1.2 Marked processes

One or more characteristics can be associated with each point. These characteristics are known as marked points. In this case, we talk about **marked point processes**. This approach has been widely used in forest studies (see for example Marcon et al. 2012).

The marks used can be qualitative (different tree species) or quantitative (trunk diameter, tree size). If we take the example of clothing shops, qualitative markers could be the type of store (ready-to-wear or made-to-measure) and quantitative marks could be store surface area or number of employees. Marks can be more sophisticated. For example, Florent Bonneu characterised the spatial distribution of incidents in the Toulouse region in 2004 using the associated workload for each fire service intervention (Bonneu 2007). This quantitative mark is obtained by multiplying the duration of the intervention and the number of firefighters mobilised.

To begin, we will limit ourselves to unmarked processes.

### 4.1.3 Observation window

The area to study the location of points is often called the **window** and it is often arbitrary. The authors take an area for study that may be square (Møller et al. 2014), rectangular (Cole et al. 1999), circular (Szwagrzyk et al. 1993), an administrative area (Arbia et al. 2012) or study zone (Lagache et al. 2013).

The indicators used to detect the underlying spatial structures are based on an analysis of the **neighbourhood of points**: for example, for all the points studied, the average number of similar points within a radius of 2 km, 4 km, etc. It may then be necessary to take into account points located on the edge of the area of interest. The risk is to underestimate the neighbourhood of points located on the edge of the area, as some of their neighbours are located outside the area. For example, we can see this in figure 4.2. Let us assume that the area being studied is a square plot within a forest and that the points represent trees. The neighbourhood of points $i$ is described as the circle with a radius $r$, centred on point $i$. If you want to estimate the number of neighbours for point $i$, counting only the points in the circle that are included within the parcel would underestimate the actual number of its neighbours. The reason is simple: a part of the circle is located outside the field of study.

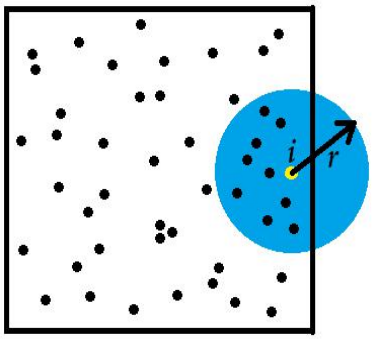


Figure 4.2 – Edge effect example
**Source:** *the authors*

The study by Marcon et al. 2003 illustrates, for example, the importance of not taking this bias into account when estimating the concentration of industrial activities in France. Generally, regardless of the area of application, this potential bias is deemed severe enough for the use of a corrective technique to account for **"edge effects"**. There is a great deal of literature on these edge effects and their correction (overall or individual correction, creation of a buffer zone around the area, use of toroidal correction [2]...) Interested readers may refer to traditional spatial statistics manuals for further information (Illian et al. 2008 ; Baddeley et al. 2015b). From a practical point of view, calculation software (and in particular R) can be used to treat these effects using different correction methods. An example will be provided in chapter 8: "Spatial smoothing".

## 4.2 Point processes: a brief presentation

### 4.2.1 The homogeneous Poisson process

To begin, let's look at the point process that is used to generate completely random spatial point distributions (Complete Spatial Randomness - CSR). To achieve this, we can start with a particularly simple process, $U$, which generates a single point that can be randomly located in an area of interest $W$. If $u_1$ and $u_2$ are the coordinates of the point, it is possible to calculate the probability that the point generated by $U$ is located in a small space $B$, which is selected arbitrarily:

$$P(U \in B) = \int_B f(u_1,u_2)du_1du_2. \tag{4.1}$$

The distribution is uniform over $W$ if $f(u_1,u_2) = \frac{1}{|W|}$ where $|W|$ designates the area of $W$.

Therefore, we have: $P(U \in B) = \int_B f(u_1,u_2)du_1du_2 = \frac{1}{|W|}\int_B du_1du_2 = \frac{|B|}{|W|}$. This process allows another process to be defined - the binomial process. $n$ points are distributed evenly across the region $W$, independently. Traditionally, we would write that:

$$P(n(X \cap B) = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.2}$$

with $p = \frac{|B|}{|W|}$. The `runifpoint` function in the R package *spatstat* generates spatial distributions of points from a uniform binomial process. For example, in figure 4.3, 1,000 points are expected in a 10 x 10 observation window.

2. The toroidal correction can be applied to a rectangular window. The window is folded over onto itself to form a torus: continuity is established between the right and left limits (upper and lower, respectively) of the window, which, therefore, no longer has any edge

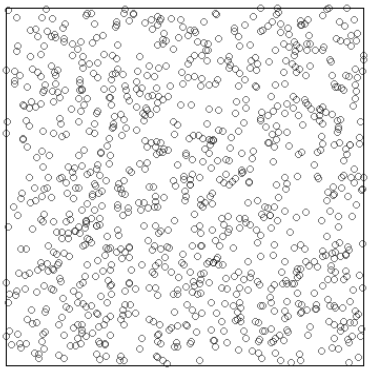

Figure 4.3 – 1 000 points sample using a uniform binomial process
**Source:** *package spatstat, authors' calculations*

```
library("spatstat")
plot(runifpoint(1000, win=owin(c(0, 10),c(0, 10))), main="")
```

Why is such a process, in which each point is placed uniformly at random, not appropriate to define a CSR process? Initially, we require two properties from such a process:

— **homogeneity** which corresponds to the absence of "preference" for a particular location (this is indeed the case for the binomial process).

— **Independence**, to reflect the fact that realizations in one area of the space have no influence on realizations in another region. This is not the case for the binomial process.

If there are $k$ points in the $B$ area of $W$, there are $n-k$ in the rest of the area.
Homogeneity induces that the number of points expected in the $B$ region is either proportional to its surface, or $E\left[n(X \cap B)\right] = \lambda \left|B\right|$. $\lambda$ is a constant that corresponds to the average number of points per unit of surface area. The Poisson law, which will be used to characterise a CSR process, can be introduced heuristically based on the property of independence. This implies that all counts in grids are independent, regardless of the size of the square. When cells, numbered $m$, become extremely small, most of them contain no points and some contain only one. The probability of a region containing more than one point becomes negligible. Based on the hypothesis of independence, $n(X \cap B)$ is the number of successes from a large number of independent drawings, with each drawing having a very low probability of success. This number of successes follows a binomial law of parameters $m$ and $\lambda \left|B\right| / m$, which tends towards the Poisson law for the $\lambda \left|B\right|$ parameter when $m$ becomes large:

$$P(n(X \cap B) = k) = e^{-\lambda |B|} \frac{\lambda^k \left|B\right|^k}{k!}. \tag{4.3}$$

Therefore, we come to this conclusion on the basis of the hypotheses of homogeneity and independence.

**Definition 4.2.1 — CSR process.** The CSR process or homogeneous Poisson process is often defined as follows:

— $P(n(X \cap B) = k) = e^{-\lambda |B|} \frac{\lambda^k |B|^k}{k!}$.
This defines the Poissonian nature of the distribution **(PP1)**;

— $E[n(X \cap B)] = \lambda |B|$.
This defines the homogeneity **(PP2)**;

— $n(X \cap B_1)$,..., $n(X \cap B_m)$ are $m$ independent random variables **(PP3)**;

— once the number of points is set, the distribution is uniform **(PP4)**.

Properties **PP2** and **PP3** are sufficient to define the CSR process (Diggle 1983), and it can be demonstrated that others are consequential. Other properties result from this. Firstly, the superposition of independent Poisson processes with parameters $\lambda_1$ and $\lambda_2$ gives a Poisson process with a parameter of $\lambda_1 + \lambda_2$. If points are eliminated randomly with a constant probability $p$ in a Poisson process (*thinned process*), the resulting process is always a Poisson process with parameter $p\lambda$, where $p$ is the thinning parameter.

The homogeneous Poisson process plays a decisive role in modelling spatial distributions of points [3] Many spatial processes have been defined, and we will give a few examples in this chapter. These can be implemented using package *spatstat*. For example, the `rpoispp` function will be used to simulate homogeneous Poisson processes. Figure 4.4 is a realization of a homogeneous Poisson process in a 1 x 1 observation window: 50 points are expected and the points are distributed completely randomly over the window.

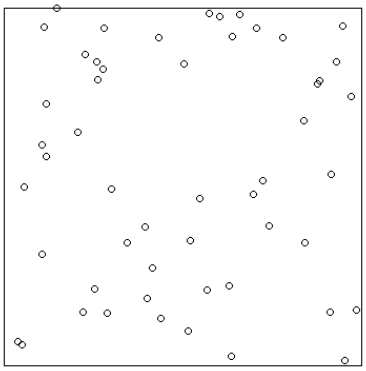

Figure 4.4 – 50 points sample by a homogeneous Poisson process
**Source:** *package spatstat, authors' calculations*

```
library("spatstat")
plot(rpoispp(50), main="")
```

### 4.2.2 Intensity, first-order property

Process laws are very complex (Møller et al. 2004), which in practice leads to the preferred use of indicators that are qualified as first-order or second-order, in the same way as first-order and second-order moments (expectation and variance) are used to identify a random variable of unknown law.

3. A little like the Normal law in classical inferential statistics (although its properties make it closer to the uniform law).

**Definition 4.2.2 — Intensity of a process.** Intensity featured in the presentation of the Poisson process, where it was constant ($\lambda$). There are other processes in which this hypothesis is rejected, and in which the intensity function $\lambda(x)$ is variable. It is defined as $E[n(X \cap B)] = \mu(B) = \int_B \lambda(x)dx$.

By applying the definition of expectation to a small region centred on $x$ and surface $dx$, we can define **the intensity** at this point x as **the number of points expected in this small area when it tends towards 0**, or:

$$\lambda(x) = \lim_{|dx| \to 0} \frac{E[N(dx)]}{|dx|}. \tag{4.4}$$

If it is not constant, it may be **estimated using non-parametric methods** that are used for density estimation. In its simplest version, without correcting edge effects, the intensity estimator is written: $\lambda(u) = \sum_{i=1}^{n} K(u - x_i)$, $K$ designating the kernel, which can be Gaussian, or with finished support (Epanechnikov kernel, Tukey's biweight kernel). They must check that $\int_{\mathbb{R}^2} K(u)du = 1$. As in all non-parametric methods, **the choice of kernel has a limited impact**. In contrast, **the choice of bandwidth is extremely important** (see, for example Illian et al. 2008). A presentation of these estimation methods can be found in chapter 8 of this manual: "Spatial smoothing". The function used in the R software is `density` in package *spatstat*, which provides contours, 3D representations and colour degradations. Several examples will be given in section 4.6.1 of this chapter.

### 4.2.3 The Inhomogeneous Poisson Process

Inhomogeneous Poisson processes are of variable intensity and their points are distributed independently of each other (the **PP3** condition is maintained). The **PP1** condition regarding the Poissonian nature of the distribution, conditional to $n$, is maintained, as the parameter for the law is no longer $\lambda |B|$, but $\mu(B)$ as defined above. The **PP4** condition is modified. Subject to a number of fixed points $n$, the points are independent and identically distributed, with a probability density of $f(x) = \frac{\lambda(x)}{\int_B f(u)du}$.

Figure 4.5 shows two examples of inhomogeneous Poisson processes, characterised by their intensity function (with coordinates $x$ and $y$).

```
library("spatstat")
par(mfrow=c(1, 2))
plot(rpoispp(function(x, y) {500*(x+y)}), main=expression(lambda==500*(x+y)
    ))
plot(rpoispp(function(x,y) {1000*exp(-(x^2+y^2)/.3)}), main=expression(
    lambda==1000*exp(-(x^2+y^2)/.3)))
par(mfrow=c(1,1))
```

### 4.2.4 Second-order properties

To introduce the second-order properties of a point process, we will look at the **variance and covariance of point counts**, defined below:

$$var(n(X \cap B) = E[n(X \cap B)^2] - E[n(X \cap B)]^2 \tag{4.5}$$

$$cov\,[n(X \cap B_1), n(X \cap B_2)] = E[n(X \cap B_1)n(X \cap B_2)] - E[n(X \cap B_1)]E[n(X \cap B_2)] \tag{4.6}$$

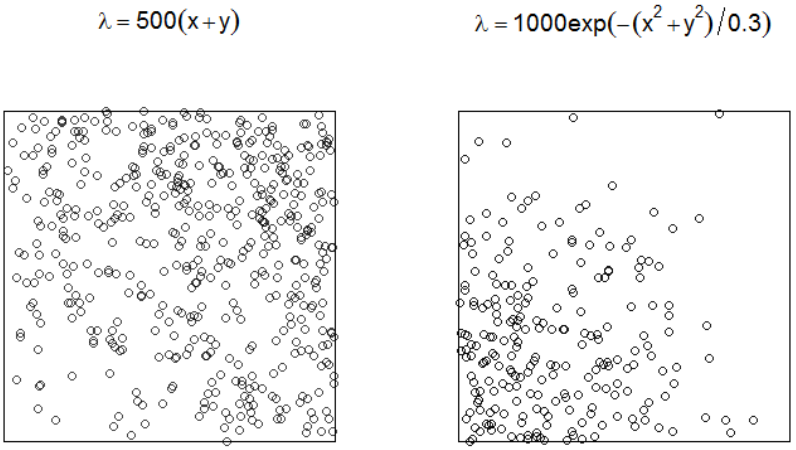


Figure 4.5 – Examples of inhomogeneous processes
**Source:** *package spatstat, authors' calculations*

**Definition 4.2.3 — Second-order moment of a process.** Rather than using these indicators, the second-order moment is defined as follows:

$$\nu_{|2|}(A \times B) = E[n(X \cap A)n(X \cap B)] - E[n(X \cap A \cap B)], \tag{4.7}$$

which, for the Poisson process, gives: $\lambda^2 |A| |B|$. When this measure includes a density, it is called order 2 intensity and noted $\lambda_2$. It is defined as $\nu_{|2|}(C) = \int_C \lambda_2(u,v)dudv$.

This second-order intensity can be interpreted as:

$$\lambda_2(x,y) = \lim_{|dx|\to 0|dy|\to 0,} \frac{E[N(dx)N(dy)]}{|dx|\,|dy|}. \tag{4.8}$$

First- and second-order intensities are used to define a function, called the *point pair correlation function*, as follows:

$$g_2(u,v) = \frac{\lambda_2(u,v)}{\lambda(u)\lambda(v)}. \tag{4.9}$$

In the case of a homogeneous Poisson process,$\lambda_2(u,v) = \lambda^2$, $g_2(u,v) = 1$.

When a process is **stationary (at the second order)**[4], the intensity of the second order is not affected by translation and depends only on the difference between the points: $\lambda_2(x,y) = \lambda_2(x-y)$.

When it is also **isotropic**, the process is not affected by rotation and the second-order intensity depends only on the distance between $x$ and $y$. Note that second-order stationarity and isotropy are essential for many spatial statistical tools.

4. The term stationary, without any further details, is often used for constant order 1 and 2 intensity processes; first-order stationarity is synonymous with homogeneity.

## 4.3 From point processes to observed point distributions

### 4.3.1 Distribution by random, aggregation, regularity

When you look at a distribution of points, two main questions arise: are the observed points distributed randomly or is there an interaction? If there is interdependence, is it aggregate or repellent? Depending on the answers to these questions, **three spatial distributions** are generally found: a so-called completely random distribution, an aggregate and a regular distribution. An example of these three theoretical distributions is shown in figure 4.6. These spatial distributions are obtained from known point processes, simulated using package *spatstat*.

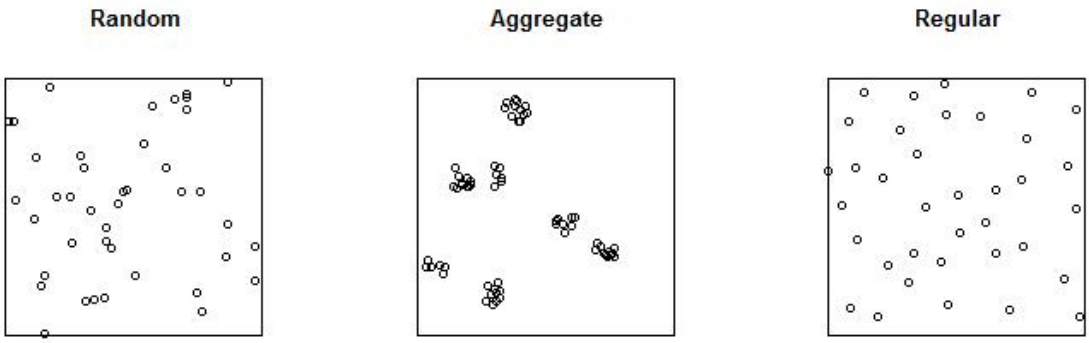


Figure 4.6 – The three standard spatial distributions of points
**Source:** *package spatstat, authors' calculations*

```
library("spatstat")
par(mfrow=c(1, 3))
plot(rpoispp(50), main="Random")
plot(rMatClust(5, 0.05, 10), main="Aggregate")
plot(rMaternII(200,0.1), main="Regular")
par(mfrow=c(1,1))
```

The **completely random** configuration is central to the theory. All spatial distributions, as point process realizations, are random but this corresponds to a "completely random" distribution of points on a surface: points are located everywhere with the same probability and independently of each other. This distribution corresponds to a realization of a homogeneous Poisson process. In this case, there is no interaction between the points but only the use of indicators makes it possible to judge whether the observed distribution differs *significantly* from a completely random distribution. Indeed, it is extremely difficult to identify such a configuration with the naked eye. In this example, we selected the `rpoispp` function in package *spatstat* to simulate the homogeneous Poisson processes.

The second distribution of points is said to be **regular**: consider the spatial distribution of trees in an orchard or along streets in town, the distribution of deckchairs on a beach, etc. In such a configuration, the points are *more regularly spaced* than they would be in a completely random distribution. Points repel each other and create a dispersed points distribution. A dispersion phenomenon can be seen for certain commercial activities, such as gas stations in Lyon (France, see Marcon et al. 2015a). Location constraints can also create dispersions, the geographic distribution of the capitol buildings in the USA is a good example of this (Holmes et al. 2004). In the right-hand chart of figure 4.6, we used a realization of a Matern process to represent a dispersed point distribution. Specifically, two simple examples of repellent processes are provided by the Matern I and II processes (see Baddeley et al. 2015b). In process I, all point pairs located at distances below

a threshold $r$ are deleted. In process II, each point is marked by an arrival time, a random variable in $[0, 1]$. Points located at a lesser distance than $r$ from a previously determined point are deleted. Using package *spatstat*, the `rMaternI` and `rMaternII` functions can be used to simulate these two Matern processes. In the example given in figure 4.6, we used a sample of a Matern type II process obtained using this package. It should be noted that other dispersed distributions can be observed: intuitively, for example, a dispersion phenomenon can be seen in a distribution of points located at the intersections of a honeycomb pattern: in this case the distance between the points is maximum (and it is greater than it would have been if the distribution was random).

Finally, the last possible configuration is known as **aggregated**. In this case, an interaction between the points can be seen. They attract each other, creating aggregates: a geographic concentration can then be detected. Looking at figure 4.1 in the introduction, it seems that the clothing stores in Rennes are mainly located in the city centre. This observation could be shared with other types of shops, such as clothing in specialised stores in the city Lyon (Marcon et al. 2015a). An aggregated configuration corresponds, for example, to the central theoretical case in figure 4.6 which is obtained by drawing a Matern cluster process. The idea of this process to simulate aggregates is quite intuitive. Around each "parent" point, in a circle with radius $r$, "offspring" points are distributed uniformly. In package *spatstat*, the `rMatClust` function can be used to simulate Matern cluster process realizations. We used this function to obtain the aggregated distribution in figure 4.6. In particular, we specified the intensity of the Poisson process for the parent points (equal to 5) and the average number of offspring points (10) drawn around the parent points in a circle of radius $r$ (equal to 0.05).

### 4.3.2 Warnings

These spatial structures (aggregated, random or dispersed) are open to a very intuitive interpretation based on the hypothesis of stationarity of the process: by comparing the distributions of observed points to a random distribution, it seems easy to detect the interactions of repellent or attractions that cause dispersion or spatial concentration phenomena.

However, any conclusions should not be too hasty as it should be kept in mind that the same aggregated or dispersed structures can be obtained with an inhomogeneous Poisson process in which the intensity of the process varies in space but the points are independent of each other (see figure 4.5). A single observation of a spatial distribution does not allow for any distinction between first- and second-order properties of a process in the absence of additional information such as that provided by a model that links a covariable to the intensity. Ellison et al. 1997, showed that natural advantages (involving greater intensity) have an effect on the location of establishments that is indistinguishable from that of positive externalities (causing the aggregation): confusion between these two properties may also concern processes.

One final warning concerns homogeneity. Indeed, initially, the methods developed in spatial statistics consisted of testing for the existence of aggregation or repulsion, assuming the homogeneity of the process: the aim was, therefore, to test a spatial distribution against the null hypothesis of complete spatial randomness (CSR). To analyse such datasets, measurements such as the original $K$ function, proposed by B.D. Ripley (widely used in statistical literature) are adequate. However, if the null hypothesis of a completely random point distribution is considered too strong, other functions must be favoured. This is the case, for example, for earthquake studies (Veen et al. 2006). Figure 4.7 illustrates 5,970 earthquake epicentres in Iran between 1976 and 2016 (of a magnitude greater than 4.5). This data comes from package *etas*.

```
data(iran.quakes, package = "ETAS")
plot(iran.quakes$lat~iran.quakes$long , xlab="Longitude", ylab="Latitude")
```

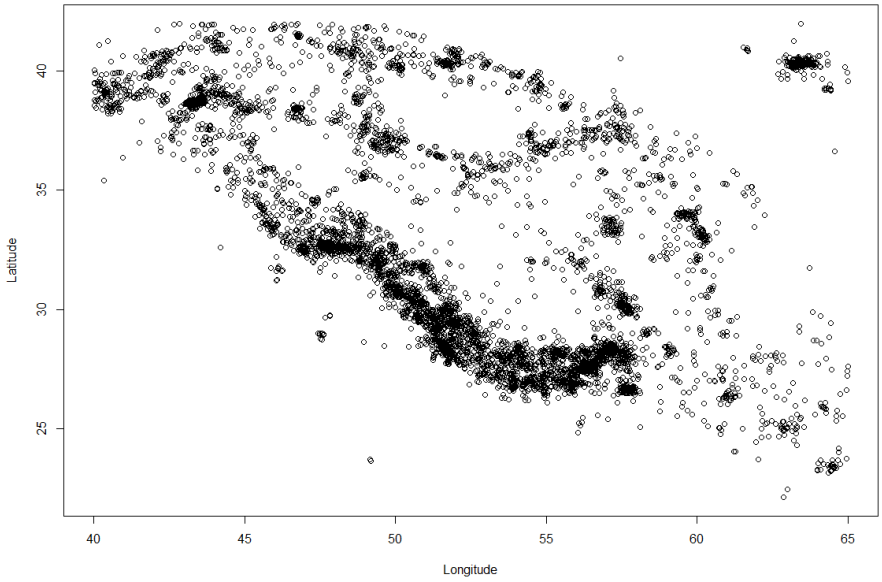


Figure 4.7 – Location of 5,970 earthquake epicentres in Iran from 1976 to 2016
**Source:** *package etas, authors' calculations.*

It is easy to see that any reference to the homogeneity of space is not optimal because there are geological predispositions in this case. B.D. Ripley's $K$ function would be unsuitable for analysis of this type of data and other tools should be used, such as the *inhomogeneous* $K$ function from Baddeley et al. 2000, which we will present in this chapter. Duranton et al. 2005 also highlighted this limitation of homogeneity of space to analyse the distribution of industrial activities and proposed a new function $K_d$.

Thorough knowledge of the available functions is, therefore, essential to characterise point distribution *accurately*. This will be the subject of the next section.

## 4.4 What statistical tools should be used to study spatial distributions?

Unfortunately, the answer to this question is not straightforward. The answer lies in precise analysis of the question that we are attempting to answer, using distance-based measures (particularly with regards to the reference value) and examination the properties of the functions. To fully understand this point and, therefore, the difficulty associated with the choice of the measure, this section will begin with a presentation of the original Ripley's $K$ function and significant developments that have resulted from this work (sections 4.4.1 and 4.4.2). We will then take time to better explain the determining factors in the choice of measure (section 4.4.3). We will then see the advantages and disadvantages of the existing measures. For an overview of the literature or an in-depth and more complete comparison of measurements, please refer to the work of Baddeley et al. 2015b or the typology of distance-based measures proposed by Marcon et al. 2017.

### 4.4.1 Ripley's $K$ function and its variants

The most widely used indicator for illustrating correlation in point processes is the $\widehat{K}$ empirical function, proposed by B.D. Ripley in 1976 (Ripley 1976; Ripley 1977). This function is commonly known as **Ripley's function** and has been the subject of many comments and developments and several variants. Specifically, this function will allow us to estimate the average number of neighbours relative to the intensity.

**Definition 4.4.1 — Ripley's K function.** Its estimator is written as follows:

$$\hat{K}(r) = \frac{|W|}{n(n-1)} \sum_i \sum_{j \neq i} \mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\} c(x_i, x_j; r), \tag{4.10}$$

where $n$ is the total number of points in the observation window, $\mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}$ is an indicator that is worth 1 if points $i$ and $j$ are at least equal to $r$ and 0 otherwise. $c(x_i, x_j; r)$ corresponds to the correction of edge effects and $W$ to the study area.

$K$ is a **cumulative function**, giving the average number of points at a distance less than or equal to $r$ from any point, **standardized by the intensity of the process ($n/|W|$), which is assumed to be homogeneous**.

In practical terms, to study the neighbourhood of points, we will analyse all the distances $r$, by calculating the value of the $K$ function for each of these distances. This is done as follows:

1. for each point and distance $r$, the number of neighbours (other points) located on the circle with radius $r$ is counted;
2. we then calculate the *average* number of neighbours (taking into account any edge effects) for each distance $r$;
3. lastly, these results will be compared to those obtained on the assumption of a homogeneous distribution (completion of a homogeneous Poisson process), which will be the expected reference value.

Finally, we will try to detect if there is a significant difference between the estimates of the observed and expected number of neighbours.

In Figure 4.8 we compare the three typical spatial distributions that we considered previously and the three resulting $K$ function curves. The distance $r$ is represented graphically in abscissa and the value of the $K$ function estimated at this distance is represented in ordinate. With package *spatstat*, the $K$ function is calculated using the `Kest.` function. In Figure 4.8, the estimated $K$ function is shown in black on the three graphs and the reference value in red dotted lines.

Findings:

— **when the process is completely random, the curve deviates relatively little from $\pi r^2$.** This can be seen on the graph at the bottom left of Figure 4.8. The $K$ curve remains close to the reference value $\pi r^2$, for all radii $r$.

— **in the case of a regular process,** we obtain: $\hat{K}(r) < K_{pois}(r)$ because if the points are repulsive, they have fewer neighbours on average in a radius $r$ than they would have based on the assumption of a random distribution of points. Graphically, the $K$ curve reflects this repulsion: we see on the right-hand graph that the $K$ curve is located below the reference value ($\pi r^2$) for all radii.

— **in the case of an aggregated process**, there are on average more points in a radius $r$ around the points than the expected number under a random distribution: consequently, the points attract each other and $\hat{K}(r) > K_{pois}(r)$. Graphically, the $K$ curve is this time located above the reference value for all areas of study, as can be seen on the central graph shown in Figure 4.8.

```
library("spatstat")
par(mfrow=c(2, 3), mar=c(1, 2, 2, 2))
plot(rpoispp(50), main="")
plot(rMatClust(5, 0.05, 10), main="")
```

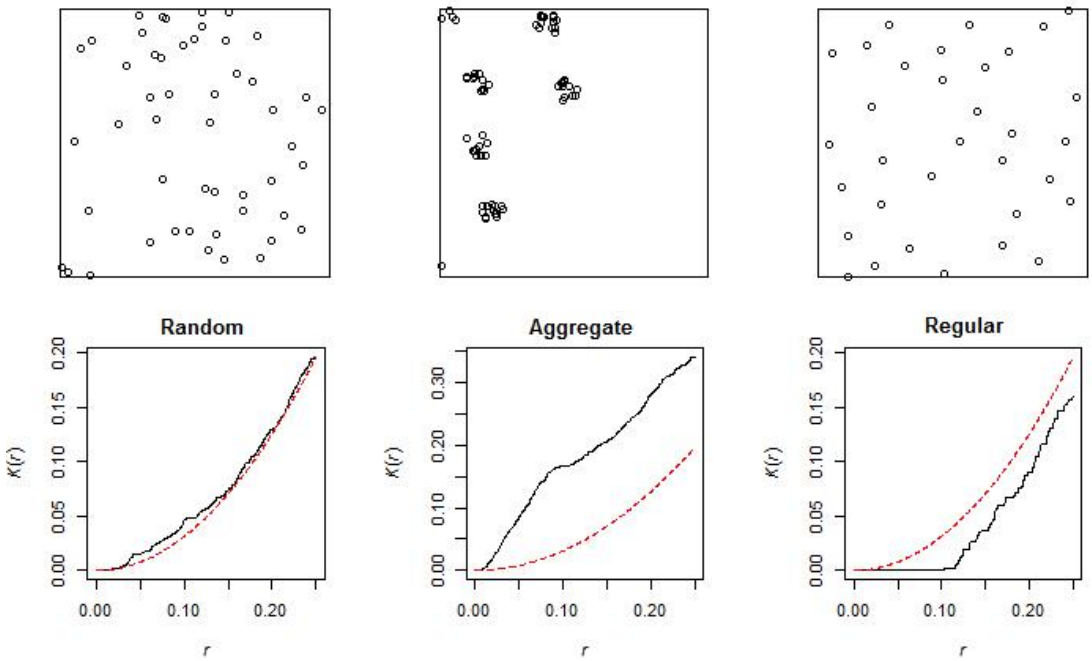


Figure 4.8 – $K$ functions for the three standard configurations of points
**Source:** *package spatstat, authors' calculations*

```
plot(rMaternII(200,0.1), main="")
par(mar=c(4, 4.1, 2, 3))
# Function K calculated by spatstat
plot(Kest(rpoispp(50),correction="isotropic"),legend=FALSE,main="Random")
plot(Kest(rMatClust(5, 0.05, 10),correction="isotropic"),legend=FALSE,main=
    "Aggregate")
plot(Kest(rMaternII(200,0.1),correction="isotropic"),legend=FALSE,main="
    Regular")
par(mfrow=c(1, 1))
```

Let's go over a few important points.

Firstly, the $K$ function is defined using the (strong) stationarity hypothesis. In the case of an inhomogeneous Poisson processes, the difference from the empirical function may be due to the variation in intensity rather than to a phenomenon of attraction, *i.e.* related to the second order property.

Similarly, the interpretation is subject to the same questions as for "conventional" statistics. Correlation does not lead to causality. A lack of correlation does not necessarily lead to independence either.

In addition, the cumulative nature of the function $K$ must be taken into account. A high $K$ value at distance $r_0$ may be due to the combination of phenomena at smaller distances, whereas no interaction exists between points far from $r_0$.

Note that there is a **link between the $K$ function and the point pair correlation function**. This can be approached as follows: draw two concentric circles with radii $r$ and $r+h$, and you count the points in the resulting ring. The expected number is $\lambda K(r+h) - \lambda K(r)$ If the expression

is standardised by the expected value in the ring for a Poisson process, we obtain:

$$g_h(r) = \frac{\lambda K(r+h) - \lambda K(r)}{\lambda \pi (r+h)^2 - \lambda \pi r^2} = \frac{K(r+h) - K(r)}{2\pi r h + \pi h^2}. \tag{4.11}$$

If we make $h$ tend to 0, $g(r) = \frac{K'(r)}{2\pi r}$ or $K(r) = \int_0^1 sg(s)ds$, the link between the $g$ function and the $K$ function is clear.

Finally, the values returned by the $K$ function enable possible interactions to be detected between the points for each of the distances studied, on *the whole* of the analysed territory. However, it may be worthwhile to have local information, as for surface data models for which we calculate local indicators known as LISA alongside spatial autocorrelation indicators (such as Moran) (see Chapter 3: "Spatial autocorrelation indices"). In point models, **there are also local indicators built on the principle of Ripley indicators**. An indicator is calculated for each point $\widehat{K}(r, x_i)$. The only pairs of points taken into account are those that contain the point $x_i$. One of the local values or all the values can then be represented graphically. Different points can be identified graphically or even by using functional data analysis methods.

#### The $L$ function of Besag 1977

The particular interest of the Ripley function and more generally of distance-based methods lies in the fact that they analyse the space studied by running *all distances* and not using just one or a few geographical levels. The spread of points is very carefully studied and no analysis distance is omitted. Consequently, **only these methods can be used to detect *exactly* at what distance(s) attraction or dispersion phenomena are observable, with no scale bias associated with a predefined zone.** If there are, for example, aggregates of aggregates in spatialized data, such functions can detect the distances at which spatial concentrations occur: down to the size of the aggregate and the distance between aggregates. More complex spatial structures may also be detected, such as multiple agglomeration phenomena for certain distances and repulsion for other distances (this will be the case if several aggregates are regularly spaced, for example). An additional benefit is to be able to compare the values produced by the functions between several distances. This can be done with the $K$ function. In the original version of the $K$ function, it is not easy to directly compare the estimated values for several areas because the reference value, $\pi r^2$, requires new calculations (since hyperbolic graphic comparisons are not immediate). As we will see, this has been one of the motivations for development of Ripley's original function.

Two transformations of the Ripley function are frequently used. It is not uncommon to find applications with these variants in statistical literature rather than the original $K$ function (*e.g.* Arbia 1989 concerning the distribution of industrial companies, Goreaud et al. 1999 concerning the distribution of trees or Fehmi et al. 2001 for plants). The first variant is the $L(r)$ function proposed by Besag (Besag 1977), which is defined by: $L(r) = \sqrt{\frac{K(r)}{\pi}}$, and which is valid in a random process $L_{Pois}(r) = r$. With package *spatstat*, the $L$ function can be calculated using the `Lest` function. Another possible version is $L(r) - r$, which is compared to 0 in the case of a completely random distribution. The two advantages to these variants are one the one hand a more stable variance (Goreaud 2000) and, on the other hand, almost immediately interpretable results (Marcon et al. 2003). For example, by using the second variant, if the $L(r) - r$ function reaches 2 for a radius $r$ of 1, this means that on average there are as many neighbours within a radius of 1 around each point in this configuration as there would be in a radius of 3 (=2+1) if the distribution were homogeneous. A better standardisation is $\frac{K(r)}{\pi r^2}$ whose expected value is 1 and whereby the empirical value is the ratio between the number of neighbours observed and neighbours expected (Marcon et al. 2017).

By way of illustration, we have used the example of an aggregated distribution that was given in Figure 4.9, with the four estimated results of the functions $K$, $L$, $L-r$ and $K(r)/\pi r^2$ for this distribution.

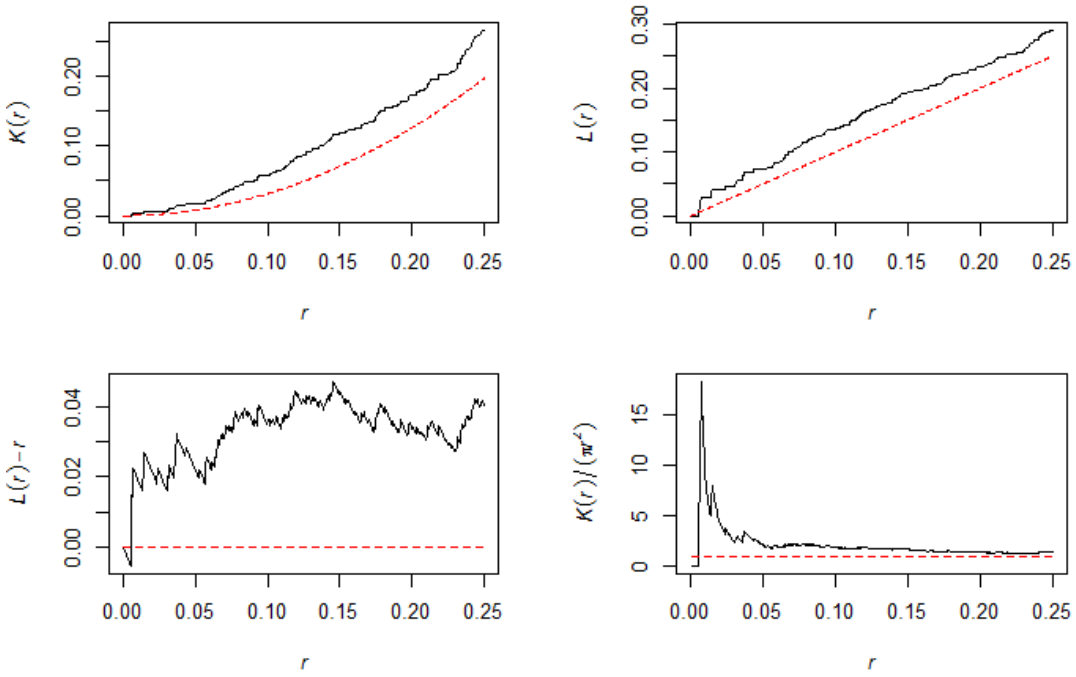


Figure 4.9 – Representation of functions $K$, $L$, $L-r$ and $K(r)/\pi r^2$
**Source:** *package spatstat, authors' calculations*

```
library("spatstat")
AGRE<-  rMatClust(10, 0.08, 4)
K<-  Kest(AGRE,correction="isotropic")
L<-  Lest(AGRE,correction="isotropic")
par(mfrow=c(2, 2))
plot(K,legend =FALSE, main="") # K
plot(L,legend =FALSE, main="") # L
plot(L, .-r~ r, legend =FALSE, main="") # L defined as L(r)-r
plot(K, ./(pi*r^2)~ r, legend =FALSE, main="") # K(r)/(pi r^2)
par(mfrow=c(1, 1))
```

#### The $D$ function of Diggle et al. 1991

$K$ and $L$ functions may be used in the studies if the hypothesis of homogeneity of the analysed space is verified. Another variant of the $K$ function allows the non-homogeneity of space to be taken into account: this is the $D$ function as proposed by Diggle et al. 1991. This indicator is directly derived from the work of epidemiologists, seeking to compare the concentration of "cases" (children with a rare disease in North Britain) and "controls" (healthy children in the same study area). This function is very simply defined as the difference between two Ripley $K$ functions: cases and controls. We obtain:

$$D(r) = K_{cas}(r) - K_{controls}(r) \tag{4.12}$$

The $D$ function enables distributions of two sub-populations to be compared. Intuitively, it is understood that if cases are more localised than controls, a spatial concentration of cases will be

detected by the $D$ function. Conversely, if the distribution of cases is less concentrated than that of controls, the $D$ function will detect that cases will be spatially more dispersed than controls. The benefit of using this function is to be able to detect differences in the distribution being studied compared to a reference distribution. This may be interesting, for example, if we want to know whether a certain type of housing is geographically more concentrated than other types of housing, or whether a type of business is more concentrated within cities than other types of businesses etc. The difference in two $K$ functions gives a comparison value for $D$ of 0 for all areas of study. However, it is impossible to compare the estimated $D$ values due to changes in the reference sub-population. This $D$ function can be implemented in the R software using package *dbmss*: we will then use the function called `Dhat`. Just like the $K$ function, it is also possible to associate a level of significance of the results by randomly labelling points (see below). The `DEnvelope` function will then be favoured. Various applications are available in the literature regarding the spatial concentration of economic activities (such as Sweeney et al. 1998). Interested readers may also find a variant of the $D$ function proposed by Arbia et al. 2008.

#### The $K_{inhom}$ function of Baddeley et al. 2000.

$K_{inhom}$, the version of Ripley's $K$ function in inhomogeneous space was proposed by Baddeley et al. 2000. The estimated value of $K_{inhom}$ therefore involves the estimated values of the intensity (the hypothesis of an identical intensity at any point in the territory being studied must be rejected since the space in question is no longer homogeneous). By noting $\widehat{\lambda}(x_i)$ as the estimation of the process around point $i$ and $\widehat{\lambda}(x_j)$ as the estimation of the process around point $j$, the cumulative function $K_{inhom}$ can be defined as follows:

$$\widehat{K}_{inhom}(r) = \frac{1}{D}\sum_i \sum_{j\neq i} \frac{\mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}}{\widehat{\lambda}(x_i)\widehat{\lambda}(x_j)} e(x_i, x_j; r) \tag{4.13}$$

with $D = \frac{1}{|W|}\sum_i \frac{1}{\widehat{\lambda}(x_i)}$.

We can show that in the case of an inhomogeneous process: $K_{inhom,pois}(r) = \pi r^2$. Estimates of $K_{inhom}$ are therefore interpreted in the same way as in the case of the homogeneous $K$ function. From a practical point of view, the `Kinhom` function in package *spatstat* enables the $K_{inhom}$ function to be calculated.

In theory, the treatment of non-stationary processes could be considered resolved, but the practical difficulty lies in estimating local densities using the kernel method. Beyond the technical difficulties, the theoretical impossibility of separation in a single observation based on first order phenomenon (intensity) and based on aggregation of the phenomenon being studied results in significant biases when the window used to estimate local densities is of the same order of magnitude as the $r$ value in question. There are still few empirical applications for this indicator (Bonneu 2007; Arbia et al. 2012).

### 4.4.2 How can we test the significance of the results?

Several statistical methods can be used to assess the significance of the results obtained using the various, previously presented functions. The most common technique is the use of the Monte Carlo method to simulate a confidence interval, which we will begin by explaining.

#### Monte Carlo methods

Without any knowledge of the theoretical distribution of Ripley's $K$ function under the null hypothesis of a completely random distribution, **the significance of the difference between observed values and theoretical values is tested by the Monte Carlo method**. This method can

be used to determine confidence intervals for all derivative functions of $K$ that have been presented. The function in question will be designated generically by $S$. To do this, we proceed as follows:

1. A number $q$ of datasets is generated, corresponding to the null hypothesis of the test. If the null hypothesis is a completely random process, we generate $q$ Poisson intensity processes, corresponding to the spatial distribution being tested.
2. Curves $U(r) = max\left\{S^{(1)}(r),...,S^{(q)}(r)\right\}$ and $L(r) = min\left\{S^{(1)}(r),...,S^{(q)}(r)\right\}$ are defined, which can be used to define an envelope, represented in grey in the graphs produced with the R software.
3. For a bilateral test, the defined envelope corresponds to a first type of risk $\alpha = \frac{2}{q+1}$, *i.e.* 39 simulations for a test at 5 %.

For each of the functions, we can build this envelope that allows us to compare the statistics built from the data to statistics derived from the simulation of a random process corresponding to the tested null hypothesis (a homogeneous Poisson process of the same intensity for the function $K$). In package *spatstat*, the generic `envelope` command is used to run Monte Carlo simulations and construct curves corresponding to the upper and lower values of the envelope. The envelope should not be interpreted as a confidence interval around the indicator being studied: it indicates the critical values of the test. To give a simple example, let's use dataset `paracou16` relating to the location of trees in the Paracou forest research station in French Guiana. This data is available in package *dbmss*. Let's calculate the confidence interval associated with the $K$ function with 39 simulations. In Figure 4.10, the obtained $K$ curve is shown (full black line), the red dotted curve represents the middle of the confidence interval and the two envelope markers are given as well as the envelope (curves and grey envelope). We can see that, up to a distance of close to 2 metres, we cannot reject the null hypothesis of a CSR process based on the Ripley function.

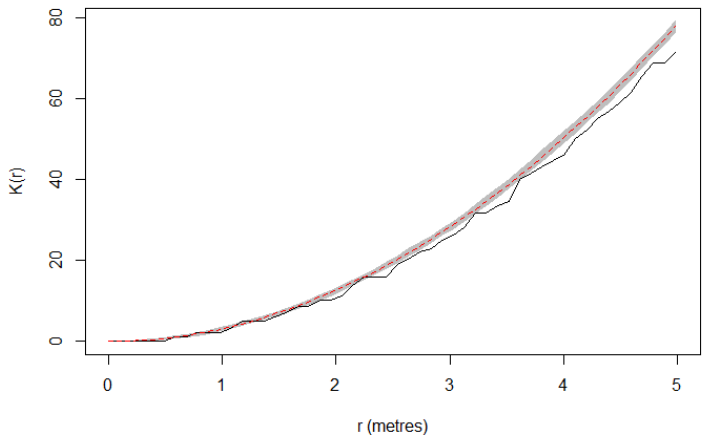


Figure 4.10 – Example of a confidence envelope for the $K$ function
**Source:** *packages spatstat and dbmss, paracou16 data, authors' calculations*

```
library("dbmss")
# Envelope calculated using package dbmss, data: 2,426 points.
env <- KEnvelope(paracou16, NumberOfSimulations=39)
plot(env,legend =FALSE, main="", xlim=c(0,5), xlab = "r (metres)", ylab= "K
    (r)")
```

With increased computing power, a common practice is to simulate the null hypothesis many times (1,000 or 10,000 times rather than 39) and to define the envelope from quantiles $\alpha/2$ and $1-\alpha/2$ for values of $S(r)$.

The test is repeated for each value of $r$: the risk of mistakenly rejecting the null hypothesis is therefore increased beyond $\alpha$. This underestimation of the first type of risk is not very large because the values of the cumulative functions are auto-correlated to a great degree. The test is therefore commonly used without any particular precautions. However, authors such as Duranton et al. 2005 consider this to be serious and try to remedy it. A method to correct the problem is presented in Marcon et al. 2010 and implemented in package *dbmss* under the name of the overall confidence interval of the null hypothesis (as opposed to the local confidence intervals calculated at each $r$ value). It consists of repeatedly removing a part $\alpha$ of simulations of which at least one value contributes to $U(r)$ or $L(r)$.

One important point: when calculating an envelope under R, it is systematically associated with a particular function. In other words, the calculation routines available in the packages take into account the specific nature of the functions: the confidence intervals are therefore simulated by considering the correct null hypothesis. For example, to simulate the envelope for the $K$ function, the null hypothesis is constructed from points that are distributed randomly and independently in the study area. However, for the $D$ function of Diggle et al. 1991, to develop a confidence interval with the same assumptions as for the $K$ function would be incorrect. For $D$, you must take into account variations in intensity in the area studied. What next? Remember that the null hypothesis for this function corresponds to a situation where the sub-population of the cases and the sub-population of the controls have the same spatial distribution. The solution suggested by Diggle et al. 1991 is random labelling which involves, for each simulation, assigning a "case" or "control" label for each location. This random permutation of labels in unchanged locations is a quite intuitive technique that will also be used to develop confidence intervals for other functions that we will study in section 4.5. Under R, packages *spatstat* or *dbmss* have options for calculating functions that allow this hypothesis of random labelling to be simulated.

#### Analytical tests

There are few analytical tests in the literature and they are rarely used in studies, even though they have the advantage of saving calculation time for confidence intervals. For $K$, for example, analytical tests exist in simple areas of study (Heinrich 1991). In the particular case of the CSR character test in a rectangular window, Gabriel Lang and Eric Marcon recently developed a classic statistical test (Lang et al. 2013) available in the `Ktest` function of package *dbmss* (Marcon et al. 2015b). It returns the probability of mistakenly rejecting the null hypothesis of a completely random distribution from a spatial distribution, without using simulations: the distribution of the $K$ function with no correction for side effects follows an asymptomatically normal distribution of known variance. The test can be used from a few dozen points. It should be noted that such tests for lesser known functions are also proposed in the literature (Jensen et al. 2011).

### 4.4.3 Review and focus on important properties for new measurements

Measures derived from the Ripley's $K$ function are useful in many configurations to explain the interactions between the points studied. We have, incidentally, given many references in various areas of application. However, specific developments can still be considered to answer certain questions, such as the location of economic activities. To understand this point, we will consider the strengths and limitations of the measures taken by Ripley's $K$ function as part of this framework of analysis.

**Review: Are the functions derived from Ripley's $K$ suitable for describing the spatial concentration of economic activities?**

The statistical tools presented in the previous sections are valuable, but their use in illustrating data for equipment or companies is not straightforward. To further consider this notion, let's go back to the examples in the introduction (the four facilities) and select Ripley's $K$ function to characterise the spatial structures of each of these facilities. The results are shown in Figure 4.11: the estimated function of Ripley's $K$ is shown as an continuous line, the confidence intervals obtained from 99 simulations by the grey area, the centre of the confidence interval is indicated by the dotted curve and the edge effects were calculated by the Ripley method. This correction of edge effects is based on the idea that, for a given point, the part of the crown outside the area (see Figure 4.2) contains the same density of neighbours as the part located within the study area. This hypothesis is acceptable because, let's remember, we consider there to be a completely random point distribution in the case of Ripley $K$ function. [5]

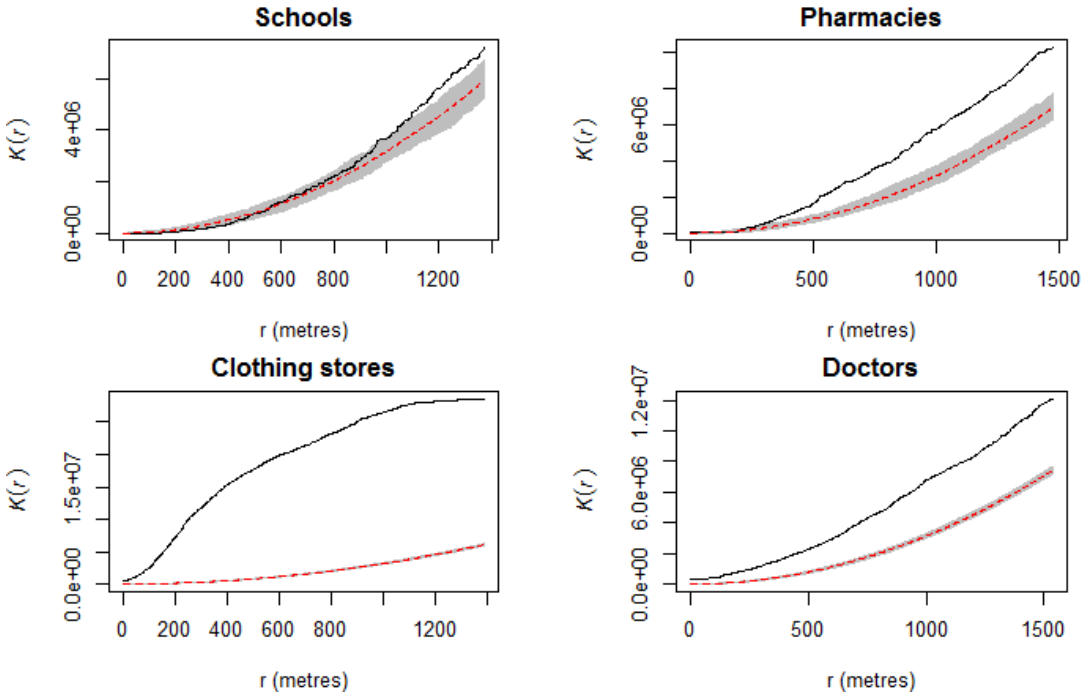


Figure 4.11 – Ripley functions for the four facilities
**Source:** *INSEE-BPE, packages spatstat and dmbss, authors' calculations*

```
library("dbmss")
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))
bpe_sch<- bpe[bpe $TYPEQU=="C104", ]
bpe_pha<- bpe[bpe $TYPEQU=="D301", ]
bpe_clo<- bpe[bpe $TYPEQU=="B302", ]
bpe_doc<- bpe[bpe $TYPEQU=="D201", ]

schools <- as.ppp(bpe_sch[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    sch[,"lambert_x"]),max (bpe_sch[,"lambert_x"])),c (min(bpe_sch[,"
```

5. Technically, let us assume that a neighbour of a given point is located in the crown width (inside the domain). The Ripley correction consists in assigning to this neighbour a weight equal to the inverse of the ratio of the perimeter of the crown over the total perimeter of the crown.

```
    lambert_y"]), max(bpe_sch[,"lambert_y"])))
bpe_schools_wmppp <- as.wmppp(schools)
pharma <- as.ppp(bpe_pha[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    pha[,"lambert_x"]), max(bpe_pha[,"lambert_x"])), c(min(bpe_pha[,"
    lambert_y"]), max(bpe_pha[,"lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
clothing <- as.ppp(bpe_clo[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    clo[,"lambert_x"]), max(bpe_clo[,"lambert_x"])), c(min(bpe_clo[,"
    lambert_y"]), max(bpe_clo[,"lambert_y"]))))
bpe_clothing_wmppp <- as.wmppp(clothing)
doctors <- as.ppp(bpe_doc[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    doc[,"lambert_x"]), max(bpe_doc[,"lambert_x"])), c(min(bpe_doc[,"
    lambert_y"]), max(bpe_doc[,"lambert_y"]))))
bpe_doctors_wmppp <- as.wmppp(doctors)

kenv_schools <- KEnvelope(bpe_schools_wmppp, NumberOfSimulations=99)
kenv_pharma <- KEnvelope(bpe_pharma_wmppp, NumberOfSimulations=99)
kenv_clothing <- KEnvelope(bpe_clothing_wmppp, NumberOfSimulations=99)
kenv_doctors <- KEnvelope(bpe_doctors_wmppp, NumberOfSimulations=99)
par(mfrow=c(2, 2))

plot(kenv_schools, legend=FALSE, main="Schools", xlab = "r (metres)")
plot(kenv_pharma, legend=FALSE, main="Pharmacies", xlab = "r (metres)")
plot(kenv_clothing, legend=FALSE, main="Clothing stores", xlab = "r (metres
    )")
plot(kenv_doctors, legend=FALSE, main="Doctors", xlab = "r (metres)")
par(mfrow=c(1, 1))
```

The results obtained in Figure 4.11 confirm the notions that we had regarding the spatial distribution of each of the facilities in Rennes (see Figure 4.1). For doctors, clothing shops and pharmacies, significant levels of spatial concentration are detected (graphically, the $K$ curves are located above the confidence interval). With regard to schools, the trend towards concentration as well as dispersion is not evident since the $K$ curve for this sector remains within the confidence interval below a radius of one kilometre and then, beyond this radius, the observed distribution of schools in Rennes does not seem to deviate significantly from a random distribution. Finally, note that the spatial concentration is particularly high for clothing stores (the difference between the $K$ curve and the upper band of the confidence interval is greatest in this sector).

**Can we consider these results sufficient to describe the spatial structure of these facilities or should they be pursued further?** The answer is simple: these conclusions are based on statistically correct calculations, but they may seem economically irrelevant. These results come up against several important limits, in particular the hypothesis of homogeneity. First of all, remember that a spatial concentration detected with the Ripley $K$ function satisfies a particular definition here: the distributions observed are more concentrated than they would be under the hypothesis of random distribution. This null hypothesis may seem very strong. Let us consider the case of the location of pharmacies: we know that in France that this has to meet certain regulatory provisions linked to the population. The CSR reference distribution does not, therefore, appear to be the most relevant in this case. A solution would then be to take into account this non-homogeneity of the space, for example by retaining the function $D$ of Diggle et al. 1991 to compare the distribution of pharmacies with that of residents. Provided that the data are available and accessible, this

would allow us to monitor the heterogeneity of the territory. This technique would also make it possible for us to regulate to a certain extent the severe constraints of setting up operations (which would prevent an equal probability of being located at any point in the territory analysed) such as the impossibility of being located in non-buildable areas in Rennes, in urban parks, etc.: the population and shops cannot be located there. It has to be said that although this strategy is attractive, it is not completely satisfactory. For example, in the case of facilities, and even more so in the case of companies, we have observations that are generally weighted very differently (number of employees, etc.). It is therefore difficult to consider that the points analysed all have the same characteristics. However, all the functions presented to date ($K$, $L$, $D$ and $K_{inhom}$) cannot include a weighting of points. This observation may be very problematic, especially considering that studies of industrial concentration in the sense of Ellison et al. 1997, Maurel et al. 1999 brought together economists' and spatial statisticians' concerns in the late 1990's toward zoning-based spatial concentration indicators. Further developments in this regard must therefore be made for the measures resulting from Ripley's $K$.

#### Development of distance-based measures to meet economic criteria

In the 2000s', **lists of economically relevant criteria** were proposed to characterise the spatial concentration of economic activities (Duranton et al. 2005; Combes et al. 2004; Bonneu et al. 2015) as:

— the insensitivity of the measure to a change in the definition of geographical scales;

— the insensitivity to a change in definition of sectoral level (according to the selected sectoral classification);

— the comparability of results between sectors;

— taking into account the productive structure of industries (*i.e.* industrial concentration in the sense of Ellison et al. 1997 which depends on both the number of establishments within the sectors and the workforce);

— a reference must be clearly established.

These questions have been discussed in many studies, in particular to distinguish between **appreciable criteria** such as the comparability of results between sectors, **essential criteria** such as the criterion regarding insensitivity of the measure following a change in the definition of geographical scales (this refers to the previously mentioned MAUP). The benefit of all distance-based measures presented in this chapter is avoiding the pitfall of MAUP. On the other hand, no measure has yet tackled sectoral divisions: the problem raised by the second criterion in the above list therefore remains intact.

#### What research options exist for extending the presented measures?

Several significant developments were proposed in the 2000s. Continued work by spatial statistical specialists and the inclusion of spatial concerns in economic studies have contributed to important innovations in concentration indicators. Not all of the studies will be covered in this context, but we will consider some of the most widely used information. In the first instance, we will introduce a slightly counter-intuitive notion of the **reference value**. When we try to characterise a point distribution, we implicitly compare it with a reference distribution (the statistician's null hypothesis) and it is the difference from this theoretical distribution that makes it possible to assess the geographical concentration, the dispersion or if the difference is not sufficient to conclude if there is any interdependence between the points. Let's re-examine the example of clothing stores and look at three types of indicators (Marcon et al. 2015a; Marcon et al. 2017 ) to characterise their location:

— the **topographical measures** use physical space as a reference value (Brülhart et al. 2005). The number of neighbours of points of interest is relative to the surface area of the neighbourhood in question: this is part of the mathematical framework of point processes. This kind of analysis allows the following question to be answered: is the density of clothing shops high around footwear stores? A positive response, for example, will show a topographical concentration of clothing stores (in the vicinity of these stores, the density of clothing stores is high). The measures presented in $K$, $L$, $D$ and $K_{inhom}$ accommodate this topographical definition of the reference value (depending on the functions, the theoretical density is considered to be constant or not). It is interesting to note that, for this reference value, the hypothesis of a homogenous or inhomogeneous space can be used;

— the **relative measures** use a distribution that is not physical space as a reference value. The number of neighbours is not shown on the surface, but in the number of points in the reference distribution. This is a clear departure from the theory of point processes, except to consider the reference distribution as an estimate of the intensity of the process based on the null hypothesis of independence between the points. In our example, this amounts to testing the existence of an over-representation or under-representation of clothing stores in the vicinity of clothing stores compared to a reference, such as all business activities. Note: the $D$ function is not a relative measure under these hypotheses as it compares density to another density, based on difference. On the other hand, a relative measure would answer the following question: around clothing stores, is the frequency of clothing stores is greater than average, throughout the territory? A positive response leads us to conclude that there is a relative concentration of clothing stores;

— lastly, **absolute measures** do not require any standardisation (by space or by comparison with any other reference). In our example, this amounts to simply counting the number of clothing shops around the clothing stores. The number obtained can then be compared to its value under the chosen null hypothesis, obtained using the Monte Carlo method.

Based on the works presented above, in particular regarding the $K$ function, statistical indicators have been proposed in the statistical literature to characterise these spatial structures under the three reference values, as mentioned above (Marcon et al. 2017). We will develop several indicators in the following sections and we will see that another important difference lies in the notion of neighbourhood. For example, it is possible to study the proximity of the points analysed *up to* a certain distance $r$. In practical terms, this means characterising the proximity of points on discs of radius $r$, which defines cumulative-type functions (such as Ripley's $K$). Another possibility is to assess the proximity of the points not *up to* a distance $r$ but *at* a certain distance $r$. Neighbourhood is assessed in a crown (also called a ring) and density functions are used to characterise it (like the $g$ function that we have already considered). A graphic illustration of these two definitions is given in Figure 4.12. In the figure on the left, the grey area corresponds to the surface of a disc with radius $r$ and, in the figure on the right, to the surface of a crown with a radius $r$.

The choice of neighbourhood is not insignificant. Therefore, density functions are more precise around the study radius but do not provide information on spatial structures at smaller distances, unlike cumulative functions. Only a cumulative function may, for example, detect whether aggregates are randomly located or whether there is a spatial interaction between aggregates (*e.g.* aggregates of aggregates). However, as cumulative functions accumulate spatial information up to a certain distance, local information at the radius $r$ is unclear, unlike density functions. The use of one or other of these neighbourhood concepts has advantages and disadvantages (Wiegand et al. 2004; Condit et al. 2000).

Marcon et al. 2017 proposed an initial classification of distance-based functions according to these two criteria:

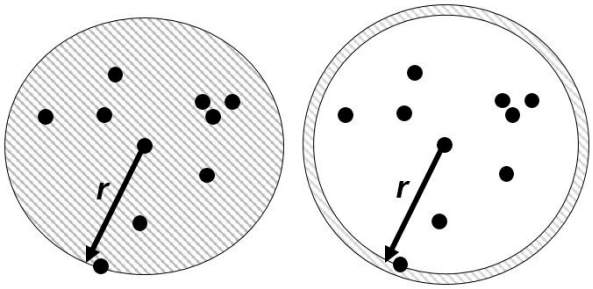


Figure 4.12 – Two possible neighbourhood concepts: on a disk or on a crown
**Source:** *the authors*

— **the type of function**: probability density, like the $g$ function or cumulative , such as Ripley's $K$ function;

— **the reference value** that can be topographical (Ripley functions and their direct variants), related to a reference situation (such as $M$ that we will present in the next section) or absolute (*i.e.* without reference such as the $K_d$ function, also presented in the next section).

It is easy to see why the choice of the correct measure is not immediately clear: first of all, the question being asked must be identified in order to select the most appropriate measure.

## 4.5 Recently proposed distance-based measures

In this section, we will present two measures relating to two references that have not yet been dealt with: the absolute and relative reference.

### 4.5.1 The $K_d$ indicator of Duranton and Overman

Unlike the previously presented functions, this indicator was developed by economists and was drawn up without any direct links with Ripley's work (although it was referred to in the bibliography). The idea of this function is to be able to estimate the probability of finding a neighbour at a distance $r$ from each point.

**Definition 4.5.1 — $K_d$ function of Duranton and Overman.** Through standardisation, Duranton et al. 2005 define $K_d$ as a function of density of probability of finding a neighbour at a distance $r$. This function can therefore be qualified as an absolute measurement of density because it has no reference. The proposed indicator is written:

$$K_d(r) = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} \kappa \left( \left\| x_i - x_j \right\|, r \right) \tag{4.14}$$

with $n$ designating the total number of points of the sample and $\kappa$, the Gaussian kernel as

$$\kappa \left( \left\| x_i - x_j \right\|, r \right) = \frac{1}{h\sqrt{2\pi}} exp \left( - \frac{\left( \left\| x_i - x_j \right\| - r \right)^2}{2h^2} \right).$$

Here we can see the technical difficulty of counting neighbours at a distance $r$ because it requires the use of a smoothing function (hence the use of the Gaussian kernel in the function). This smoothing function allows you to count neighbours whose distance is "around" $r$. The bandwidth can be defined in several ways but the Silverman 1986 method is mentioned in the original article of Duranton and Overman. As with other distance-based functions, a confidence interval of the null hypothesis can be assessed to assess the significance of the results obtained. The marks (weight/type

pairs) are redistributed to all existing locations (positions taken by points): this technique makes it possible to control both the industrial concentration and the general location trends of all types of points (two properties listed in the "correct" concentration index criteria applicable to economic activities). The hypothesis of a random location of type $S$ points is rejected at distances $r$, if the function $K_d$ is located above or below the trust boundary of the null hypothesis. Another version of $K_d$ that takes into account the weighting of points exists - it was proposed in the original article by Duranton et al. 2005. Behrens et al. 2015 used a cumulative function $K_d$. It should be noted that the $K_d$ function has been the subject of many empirical applications in spatial economics (*e.g.* Duranton 2008, Barlet et al. 2008).

The $K_d$ function can be calculated under R using the `Kdhat` function in package *dbmss*. The `KdEnvelope` function that is available in the same package can be used to associate a confidence interval with the results obtained.

### 4.5.2 $M$ function of Marcon and Puech

The $M$ indicator by Marcon et al. 2010 is a cumulative indicator, like Ripley's $K$, as it is calculated by varying a disc of radius $r$ around each point. This is a relative indicator since it compares the proportion of points of interest in a neighbourhood with the proportion of points seen throughout the territory analysed. If we consider that clothing stores are attracted to each other, their proportion around each clothing store will be higher than in the city. In practice, for a radius $r$, we will calculate the ratio between the local proportion of clothing stores around clothing stores and the proportion observed in the city. This calculation is repeated for all clothing stores and the average of these relative proportions is calculated. The reference value for the $M$ function is 1. A higher value reflects a relative spatial concentration, and a lower value shows a tendency towards repulsion (the minimum value being 0). The values of $M$ can also be interpreted in terms of ratio comparisons: for example, if $M(r)$=3, this indicates that on average there is a 3 times higher frequency of points of interest appearing around points of interest within a radius $r$ than the frequency observed over the entire observation window. Finally, like the $K_d$ function, $M$ can include weighting of points.

**Definition 4.5.2 — $M$ function of Marcon and Puech.** Formally, for $S$ type points, Marcon and Puech's $M$ function is defined as:

$$M(r) = \sum_{i \in S} \frac{\sum\limits_{j \neq i, j \in S} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)}{\sum\limits_{j \neq i} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)} \Big/ \frac{n_S - 1}{n - 1}. \tag{4.15}$$

where $n_S$ and $n$ refer respectively to the total number of $S$ type points and the total number of all types of points in the study window. This indicator should be read as the result of two frequency reports. The local average of the frequency of $S$ type points is compared within a radius $r$ around $S$ type points with the frequency of $S$ type points over the entire observation window. Removing a point from the denominator avoids a slight bias, since the centre point cannot always be counted in its neighbourhood.

As for the $K_d$ function, a version exists that takes into account the weighting of points (Marcon et al. 2017). Technically, this means multiplying the indicator by the weight of the neighbouring point in question (for example, by the number of its employees or its turnover if we look at industrial establishments). As with the other indicators, a confidence interval can be generated using Monte Carlo methods. The specific nature of the points is retained (weight/sector pairing). For $M$, as for $K_d$, the control for industrial concentration is not present in the definition of the function but in

the definition of the confidence interval, as the points labels (weight/sector pairs) are redistributed to the existing locations. In their latest work, Lang et al. 2015 offered a non-cumulative version of the $M$ indicator, named $m$, similar to the $g$ function for $K$, see Equation 13.8. As in all the situations we have encountered, the indicators can lead to different analyses: since the reference values are not the same, they answer different questions. The analyses provided are, therefore, complementary (Marcon et al. 2015a; Lang et al. 2015). Finally, note that the $M$ function does not require correction of edge effects and can be calculated in R using the `Mhat` function in package *dbmss*. The `MEnvelope` function in the same package makes it possible to combine a confidence interval to judge the significance of the results obtained.

As an example of application, consider the spatial structures of the four facilities in the introductory example of the city of Rennes. A graphical representation of the results of the $M$ function for schools, pharmacies, doctors and clothing stores is given in Figure 4.13.

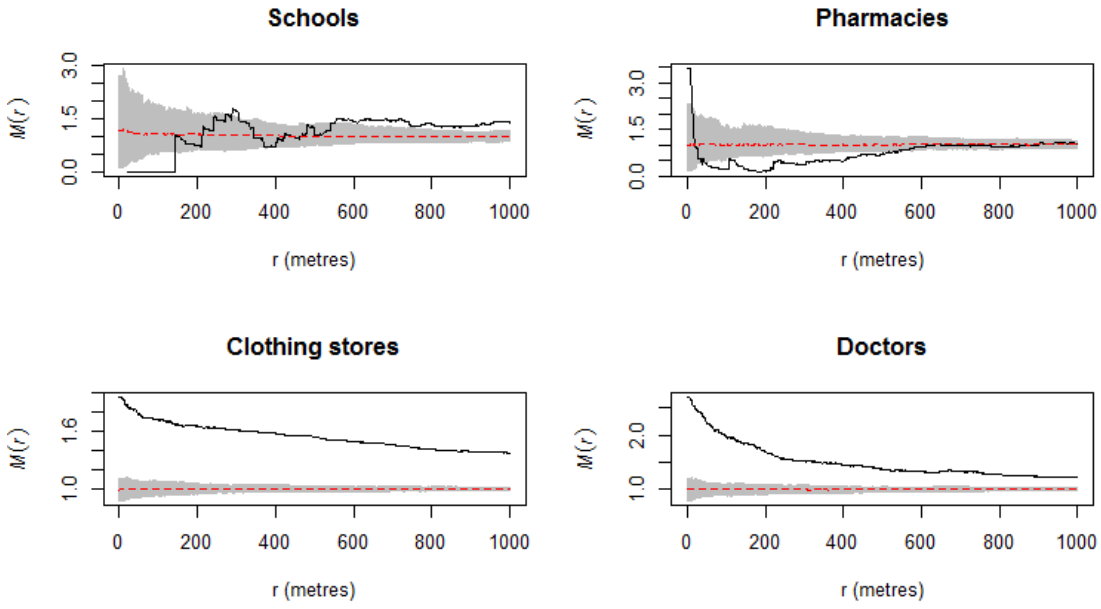


Figure 4.13 – Marcon and Puech functions for the four facilities
**Source:** *INSEE-BPE, packages spatstat and dbmss, authors' calculations*

```
library("dbmss")
# Set of marked points
bpe_equip<- bpe[bpe $TYPEQU %in%c ("C104","D301","B302","D201"),c (2,3,1)]
colnames(bpe_equip) <- c("X", "Y", "PointType")
bpe_equip_wmppp <- wmppp(bpe_equip)
r<- 0:1000
NumberOfSimulations<-  99
menv_sch<-  MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="C104")
menv_pha<-  MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="D301")
menv_clo<-  MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
    ReferenceType="B302")
menv_doc<-  MEnvelope(bpe_equip_wmppp, r, NumberOfSimulations,
```

```
    ReferenceType="D201")
par(mfrow=c(2, 2))
plot(menv_sch, legend=FALSE, main="Schools", xlab = "r (metres)")
plot(menv_pha, legend=FALSE, main="Pharmacies", xlab = "r (metres)")
plot(menv_clo, legend=FALSE, main="Clothing stores", xlab = "r (metres)")
plot(menv_doc, legend=FALSE, main="Doctors", xlab = "r (metres)")
par(mfrow=c(1, 1))
```

It is easy to see that levels of spatial concentration can be seen for all of the distances studied for doctors or clothing stores (both associated $M$ curves being located above their respective confidence interval up to 1 kilometre). As it is possible to compare the values obtained by the $M$ function, we can also conclude that the highest levels of aggregation appear at short distances. Thus, in the very first area of study, the proportion of clothing stores around clothing stores is approximately 2 times higher than the proportion of clothing stores observed in the city of Rennes. This result is quite close to the conclusions drawn by Marcon et al. 2015a in Lyon for this activity. With regard to schools or pharmacies, however, concentration or dispersion levels are detected according to the distances in question. Schools, for example, appear dispersed up to approximately 150 metres (the associated $M$ curve is located beneath the confidence interval of the null hypothesis up to this distance), then, beyond a distance of 500 metres, a phenomenon of spatial concentration is detected. At very short distances, pharmacies appear spatially aggregated, whereas their distribution is dispersed above approximately 50 metres. However, for schools and pharmacies, we note that the $M$ curves remain fairly close to their respective confidence intervals.

### 4.5.3 Other developments

This area of statistical literature is currently growing rapidly (Duranton 2008, Marcon et al. 2017). The contributions are varied: statisticians define the necessary theoretical framework and researchers develop tools applicable to their specific field. Among the work carried out recently, Bonneu et al. 2015 propose a family of indicators that have the merit of showing links between the Bonneu-Thomas (proposed in this article), Marcon-Puech and Duranton-Overman indicators. Not all indicators have yet been implemented in the usual software, even if efforts are made to take account of recent developments in the literature and make them freely available to interested users.

## 4.6 Multi-type processes

The introduction presented four maps relating to the respective locations of schools, pharmacies, general practitioners and clothing stores (Figure 4.1, p.73). All this information could have been gathered together, with each activity being a qualitative mark for the process. These marks make it possible to build **multi-type processes**, and to introduce new questions alongside those that have been developed previously: is there independence in location between types (marks)? If the answer is no, are there any phenomena of attraction or repulsion?

In order to provide answers to these questions, we must now consider processes that have specific characteristics: it is therefore possible for us to define indicators of the first order (intensity) and second order (neighbourhood relations), which we will do successively in the following two sub-sections.

### 4.6.1 Intensity functions

Analysis of variability in the intensity of processes that led to the observation of distribution of the analysed entities is interesting for an initial analysis.

In the field of ecology, one might wonder, for example, (*i*) if all tree species within a forest are located in the same way, (*ii*) if the dead trees are more agglomerated than the healthy trees (*iii*) if

the presence of young shrubs follows that of parent trees etc. The density study gives an initial indication of the observed spatial heterogeneity. In the example below, we have used the respective locations of the trees of a permanent experimental facility in Paracou, French Guiana, available in the `Paracou16` dataset in package `dbmss`. Three tree species are listed: *Vacapoua americana*, *Qualea rosea* and mixed tree species grouped under the term *Other*. The high number of trees present on the Paracou16 plot (2426 trees in total) makes it very difficult to identify any location trends for each species (see Figure 4.14).

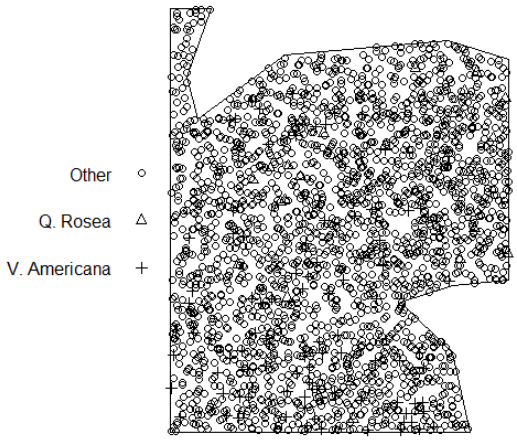


Figure 4.14 – Location of tree species *Vacapoua americana*, *Qualea rosea* or other (mix) in the Paracou16 forest system.
**Source:** *Paracou16 data from package dbmss, authors' calculations*

```
library("dbmss")
data(paracou16)
plot(paracou16, which.marks=2, main = "")
# the 2nd column makes it possible to differentiate the types of points (
    species)
```

On the other hand, a representation of the density by species is more informative and makes it possible to highlight differences in location according to the tree species in question (see Figure 4.15). A 2D representation of density is given in this example and obtained from the `density` function of package *spatstat*.

```
library("dbmss")
data(paracou16)
V.Americana<-  paracou16[paracou16$marks$PointType=="V. Americana"]
Q.Rosea<-  paracou16[paracou16$marks$PointType=="Q. Rosea"]
Other<-  paracou16[paracou16$marks$PointType=="Other"]
par(mfrow=c(1,3))
plot(density(V.Americana, 8), main="V. Americana")
plot(density(Q.Rosea, 8), main="Q. Rosea")
plot(density(Other, 8), main="Other")
par(mfrow=c(1,1))
```

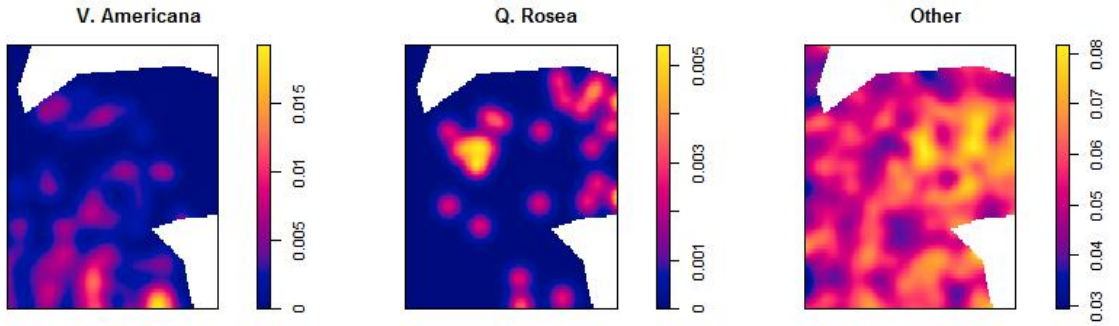


Figure 4.15 – Representation of the density of tree species *Vacapoua americana*, *Qualea rosea* or other (mix) in the Paracou16 forest system.
**Source:** *Paracou16 dataset in package* `dbmss`*, authors' calculations*

In the field of spatial economics, the study of multi-type processes could also be rich in information. We could, for example, question the possible interactions between the different types of facilities (general practitioners, schools, etc.). Using the extract from the permanent database of facilities in the city of Rennes, the four spatial sub-distributions were shown in Figure 4.1. In Figure 4.16, we mapped the densities of two facilities: pharmacies and doctors. Visually, quite similar implantation trends seem to be present, as confirmed by the 3D representation in Figure 4.16. The `persp` function in *spatstat* is used.

```
library("dbmss")
# BPE file on the INSEE.fr site: https://www.insee.fr
# Data for these examples:
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))

bpe_pha<- bpe[bpe $TYPEQU=="D301", ]
bpe_doc<- bpe[bpe $TYPEQU=="D201", ]

pharma <- as.ppp(bpe_pha[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    pha[,"lambert_x"]),max (bpe_pha[,"lambert_x"])),c (min(bpe_pha[,"
    lambert_y"]),max (bpe_pha[,"lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
doctors <- as.ppp(bpe_doc[ ,c ("lambert_x", "lambert_y")], owin(c(min(bpe_
    doc[,"lambert_x"]),max (bpe_doc[,"lambert_x"])),c (min(bpe_doc[,"
    lambert_y"]),max (bpe_doc[,"lambert_y"]))))
bpe_doctors_wmppp <- as.wmppp(doctors)

persp(density(doctors),col ="limegreen",
theta = -45,#Viewing angle
xlab = "Lambert X", ylab = "Lambert Y", zlab = "Density",
main = "Doctors")
persp(density(pharma),col ="limegreen", theta = -45,
xlab = "Lambert X", ylab = "Lambert Y", zlab = "Density",
main = "Pharmacies")
```

However, only the results of a second order process property analysis will allow us to reach a

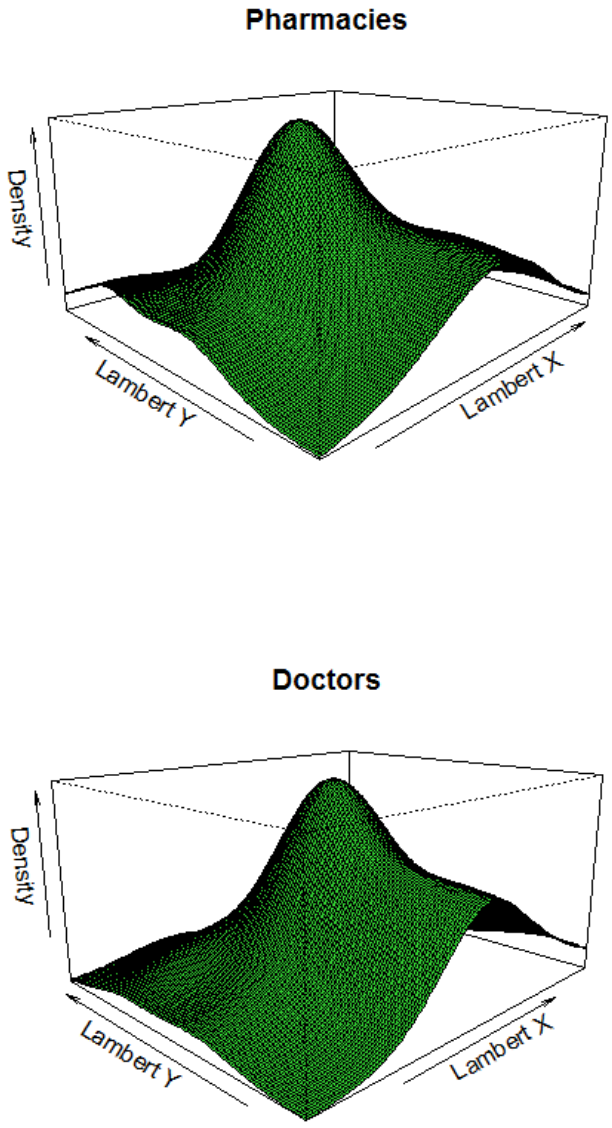


Figure 4.16 – Representation of the density of pharmacies and doctors in Rennes
**Source:** *INSEE-BPE, packages spatstat and dbmss, authors' calculations*

conclusion of any possible interaction (attraction or repulsion) between tree species or between facilities. This is why a first order property study is only a first step of analysis when studying spatial distribution.

### 4.6.2 Intertype functions

Various developments have been proposed to study the second order properties of multi-type processes. Indicators derived from Ripley's $K$ function (univariate) have been proposed to analyse relative locations of spatial sub-distributions related to different marks. These indicators are generally referred to as intertype or bivariate functions. We will look at two in more detail in the following sub-sections. From a practical point of view, it is possible to use R packages such as *spatstat* or *dbmss* to calculate the functions and represent the results graphically.

#### The $K$ intertype function

Consider the following case. We would like to study the spatial structure between two types of points, for example: $T$ type points located around $S$ type points. Using an intertype function then makes it possible to study the spatial structure of $T$ type points located at a distance of less than or equal to $r$ from $S$ type points.

An initial indicator can be used, the $K$ intertype function. This is written $\widehat{K}_{S,T}$ and is defined as follows:

$$\widehat{K}_{S,T}(r) = \frac{1}{\widehat{\lambda}_S n_S} \sum_{i \in S} \sum_{j \in T} \mathbf{1}\left\{\left\|x_i - x_j\right\| \leq r\right\}. \tag{4.16}$$

where $\widehat{\lambda}_S$ refers to the estimated intensity of the $S$ type sub-process. In the field of study, $n_S$ represents the total number of points $S$.

In the event that $S$ and $T$ are the same type, the definition of the univariate $K$ function is presented in the section 4.4.1 (p.83). Note, however, that the correction of edge effects is not included here in the definition of the intertype $K$ function for ease of presentation. The reference value is always $\pi r^2$, regardless of the radius $r$, since this is based on the null hypothesis of a completely random distribution of points (of types $S$ and $T$). If the $S$ type sub-process is independent of the $T$ type sub-process, then the number of $T$ type points within or equal to a distance of $r$ from an $S$ type point is the expected number of $T$ type points located in a disc or radius $r$, or $\lambda_T \pi r^2$. This null hypothesis corresponds to the independent distribution of two types of industrial establishments, for example. Another null hypothesis giving the same result is that the points are first distributed according to a homogeneous Poisson process and then receive their type in a second stage (for example, commercial spaces are created and then occupied by different types of shops). For all $r$ distances for which observed values of $\widehat{K}_{S,T}(r)$ are less than $\pi r^2$, a tendency to repulsion of $T$ points around $S$ points would be reported. Conversely, values of $\widehat{K}_{S,T}$ greater than $\pi r^2$ will tend to validate an attraction of $T$ points around $S$ points within a radius $r$. The simulation of a confidence interval using the Monte Carlo method will result in an attraction or a repulsion between the two types of points.

The $K$ intertype function can be implemented in package *spatstat* using the `Kcross` function. In application, let's look again at the example of the `Paracou16` data. Indeed, if we use the intertype $K$ function, we hypothesise that the space in question is homogeneous; however, this hypothesis is almost systematically used in empirical analyses in forest ecology (Goreaud 2000). In Figure 4.17, we have represented the intertype $K$ functions (or bivariates) for the species *Qualea rosea* or mixed *Other* with that of *Vacapoua americana*. The black curves represent observed $K$ intertype functions and red dotted lines represent reference intertype $K$ functions. As can be seen, there is a repulsive relationship between *Qualea rosea* and *Vacapoua americana* (observed $K$ intertypes are located below the reference value) whereas no association trend appears to exist between the *Vacapoua*

*americana* and other tree species (theoretical and observed $K$ intertype curves are mixed up for all distances).

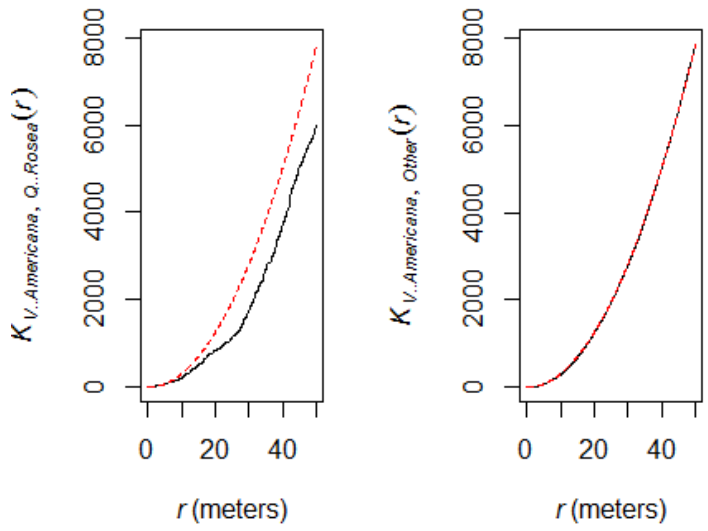


Figure 4.17 – Interactions of different tree species in the Paracou16 forest system
**Source:** *`Paracou16` dataset in package `dbmss`, authors' calculations*

```
library("dbmss")
# Simplification of marks
marks(paracou16)<-  paracou16$marks$PointType
par(mfrow=c(1,2))
# Calculation of K intertypes for the trees of species "Q.Rosea" around
    those of species "Q. Rosea"
plot(Kcross(paracou16, "V. Americana", "Q. Rosea", correction="isotropic"),
     legend=FALSE, main=NULL)
# calculation of K intertypes for trees of species "Q.Rosea" around those
    of species "Other"
plot(Kcross(paracou16, "V. Americana", "Other", correction="isotropic"),
    legend=FALSE, main=NULL)
par(mfrow=c(1,1))
```

#### The $M$ intertype function

Similarly, the previously presented $M$ function can be used as an intertype tool. The idea is always to compare a local proportion to a global proportion but in the case of the $M$ intertype function, the type of neighbouring points of interest is not the same as that of the centre points. For example, if we suspect an attraction of $T$ type points by $S$ type points, we will compare the local proportion of $T$ type neighbours around $S$ type points to the overall proportion observed throughout the territory in question. If the attraction between the $T$ type points around $S$ type points is real, the proportion of $T$ type points around $S$ type points should be locally higher than that observed across the entire study area. Conversely, if $T$ points are repulsed by $S$ type points, the relative proportion of $T$ type points around $S$ type points will be relatively lower than that observed for the whole

territory analysed. In this case, the unweighted empirical estimator of $M$ intertypes will be defined by:

$$\widehat{M}_{S,T}(r) = \sum_{i \in S} \frac{\sum_{j \in T} \mathbf{1}\left(\|x_i - x_j\| \leq r\right)}{\sum_{j \neq i} \mathbf{1}\left(\|x_i - x_j\| \leq r\right)} \Big/ \frac{n_T}{n-1}. \tag{4.17}$$

where $n$ means the total number of points across the entire study area, $n_S$ the $S$ type points. As for the intertype $K$ function, we will assume here that each point belongs to only one type that can be $S$, $T$ or other. For the intertype $M$ function, the reference value for all distances $r$ in question is always equal to 1. For more details on this function (taking into account the weighting, construction of the associated confidence interval, etc.), please refer to the article by Marcon et al. 2010. This intertype function can be calculated in R using the `Mhat` function of package *dbmss*. The `MEnvelope` function from the same package can be used to construct a confidence interval.

A concrete example of how to apply $M$ intertypes is given below, based on the Rennes facilities that were considered in the introduction. If we suspect relationships of attraction or repulsion between several facilities, it is then possible to analyse existing interactions using the intertype $M$ function. Remember that using the $M$ function makes it possible to reject the hypothesis of a homogeneous space that can be considered to be a strong hypothesis to characterise the location of economic activities (see, for example, Duranton et al. 2005, p. 1104). In this case, the use of $M$ intertypes would therefore seem more appropriate than $K$ intertypes. In Figure 4.18, based on the data extract from the facilities database, we have represented the links between the locations of doctors and pharmacies in Rennes. On the right-hand graphic, the locations of pharmacies in a neighbourhood of $r$ metres of doctors have been analysed. A repulsion would be detected at very short distances and then intertype aggregation would be observable up to 1 km. The left-hand graphic shows that doctors seem to be relatively agglomerated within a radius of 1 km around the locations of pharmacies in Rennes (the construction of a confidence interval with 100 simulations, for example, would allow us to conclude that the tendency towards dispersion at very short distances is not significant).

```
library("dbmss")

# BPE file on the INSEE.fr site: https://www.insee.fr
# Data for these examples:
load(url("https://zenodo.org/record/1308085/files/ConfPoints.gz"))

# Set of marked points
bpe_equip <- bpe[bpe$TYPEQU %in% c("C104","D301","B302","D201"), c(2,3,1)]
colnames(bpe_equip) <- c("X", "Y", "PointType")
bpe_equip_wmppp <- wmppp(bpe_equip)
bpe_pha <- bpe[bpe$TYPEQU=="D301", ]
bpe_doc <- bpe[bpe$TYPEQU=="D201", ]
pharma <- as.ppp(bpe_pha[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    pha[,"lambert_x"]), max(bpe_pha[,"lambert_x"])), c(min(bpe_pha[,"
    lambert_y"]), max(bpe_pha[,"lambert_y"]))))
bpe_pharma_wmppp <- as.wmppp(pharma)
doctors <- as.ppp(bpe_doc[ , c("lambert_x", "lambert_y")], owin(c(min(bpe_
    doc[,"lambert_x"]), max(bpe_med[,"lambert_x"])), c(min(bpe_med[,"
    lambert_y"]), max(bpe_med[,"lambert_y"]))))
bpe_doctors_wmppp <- as.wmppp(doctors)
```

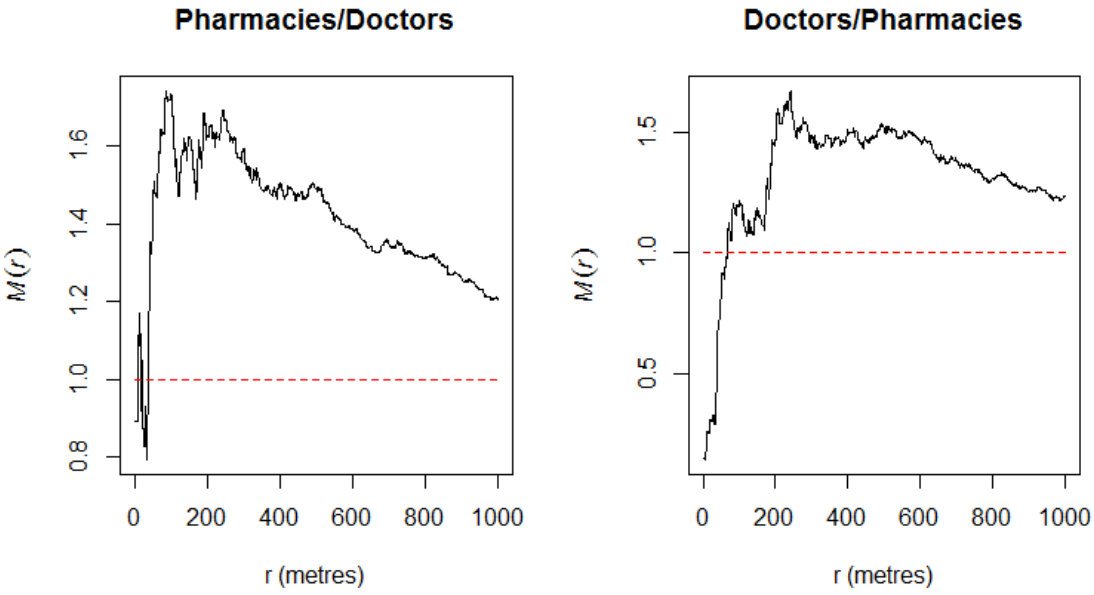


Figure 4.18 – Neighbouring relationships between doctors and pharmacies in Rennes
**Source:** *INSEE-BPE, packages spatstat and dbmss, authors' calculations*

```
# Set of marked points
r<- 0:1000

# M intertype: study of interactions between doctors' locations around
    pharmacies
M_pha_doc<- Mhat(bpe _equip_wmppp, r, ReferenceType="D301", NeighborType="
    D201")

# M intertype: study of interactions between pharmacy locations around
    doctors
M_doc_pha<- Mhat(bpe _equip_wmppp, r, ReferenceType="D201", NeighborType="
    D301")

par(mfrow=c(1, 2))
plot(M_pha_doc, legend=FALSE, main="Pharmacies/Doctors", xlab = "r (metres)
    ")
plot(M_doc_pha, legend=FALSE, main="Doctors/Pharmacies", xlab = "r (metres)
    ")
par(mfrow=c(1, 1))
```

Analysis of the neighbouring relations between Rennes facilities is not the only factor that can be explored. For example, we could suspect interactions between the locations of certain facilities and the population. To examine this relationship, the data in Figure 4.13 would have to be considered with the population data. The R code to establish the link between the population and the four types of facilities considered using the $M$ function is given below. Figure 4.19 clearly shows that the distribution of the four facilities in question does not appear to deviate significantly

from that of the population (the maximum distance reported was limited to 100 metres as no notable result is obtained beyond this radius).

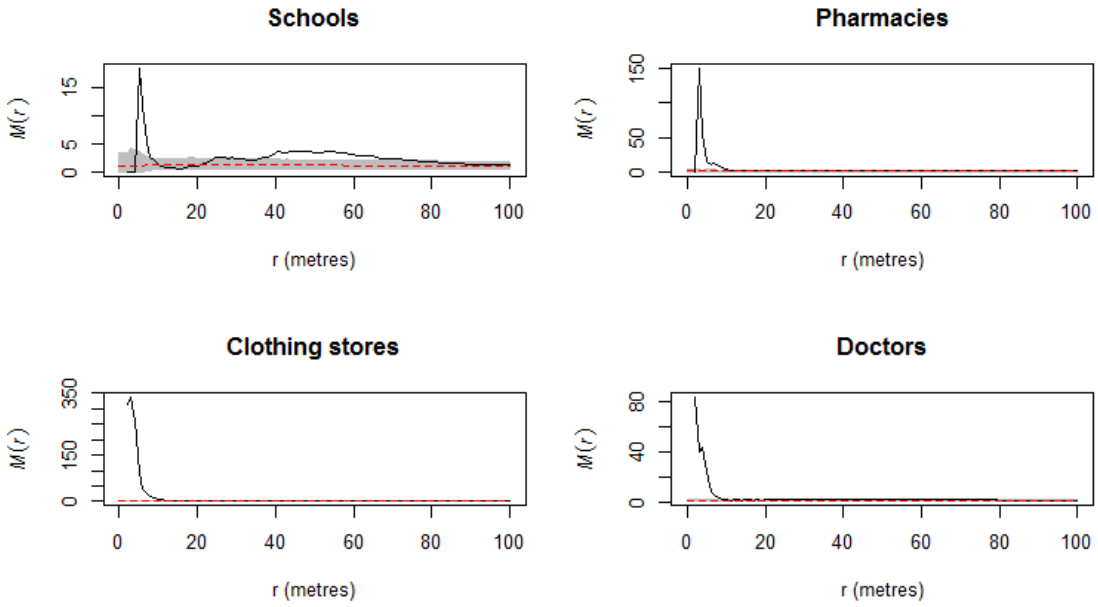


Figure 4.19 – Facilities/population interactions for the four facilities in Rennes
**Source:** *INSEE-BPE, packages spatstat and dbmss, authors' calculations*

```
library("dbmss")
colnames(popu) <- c("X", "Y", "PointWeight")
popu$PointType<-  "POPU"
popuwmppp<- wmppp(popu)

# Merger of point sets in the window  bpe_equip_dbmms
bpe_equip_popu<-  superimpose(popuwmppp, bpe_equip_wmppp, W=bpe_equip_wmppp
    $window)

# 100 simulations are selected by default
menv_popu_sch<-  MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="C104", SimulationType="RandomLabeling")
menv_popu_pha<-  MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="D301", SimulationType="RandomLabeling")
menv_popu_clo<-  MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="B302", SimulationType="RandomLabeling")
menv_popu_doc<-  MEnvelope(bpe_equip_popu, r, ReferenceType="POPU",
    NeighborType="D201", SimulationType="RandomLabeling")

par(mfrow=c(2, 2))
plot(menv_popu_sch,legend =FALSE, main="Schools", xlim=c(0,100), xlab = "r
    (metres)")
plot(menv_popu_pha,legend =FALSE, main="Pharmacies", xlim=c(0,100), xlab =
    "r (metres)")
```

```
plot(menv_popu_clo, legend=FALSE, main="Clothing stores", xlim=c(0,100),
    xlab = "r (metres)")
plot(menv_popu_doc, legend=FALSE, main="Doctors", xlim=c(0,100), xlab = "r
    (metres)")
par(mfrow=c(1, 1))
```

Lastly, note that the $M$ intertype function is not the only function available in heterogeneous space. Other univariate functions have a bivariate version such as $K_d$ or $K_{inhom}$ and can be implemented using package *package dbmss* in R.

## 4.7 Process modelling

The processes presented above, particularly the Poisson processes, are also used to build models. As in traditional statistical models, they are used to explain and predict. The aim is also to find the one with the best power of explanation among the competing models. To build these models, we use covariables. The flexibility of the R software allows the use of data that are associated with observation points, but also continuous data, images and grids.

### 4.7.1 General modelling framework

To adjust a Poisson point process to a spread of points, the shape of the intensity function can be specified $\lambda(\cdot)$ in order to look for the parameters that allow for the best adjustment. In the *spatstat*, package, the `ppm` function is an essential tool. If we call *trend* the intensity model and *mypp* the process analysed, the command is written:

```
ppm(mypp~trend)
# where "trend" refers generically to a trend and
#       "mypp" refers to the process analysed
```

The syntax of this point process modelling (PPM) command is similar to that of the standard command `lm` in R, which is used for linear regression models. There are many specifics in modelling: estimated models may result from a log-linear function of the explanatory variable, defined from several variables, etc. The choice and validation of the models must complete the analysis to provide a conclusive response. Among the solutions, the likelihood ratio test may be applied.

### 4.7.2 Application examples

To address such a question, the datasets analysed must be rich enough to satisfy theoretical models. Readers interested in this approach may refer to the two notable examples dealt with in detail in the work of Baddeley et al. 2005. The first is based on data (`Bei`) relating to trees of the species *Beischmiedia pendula* available in package *spatstat*. Indeed, in addition to the location of trees of this species in a tropical rainforest on the island of Barro Colorado, data on the altitude and slope of the land are also provided. The second dataset, named `Murchison` in package *spatstat*, relates to the location of gold deposits in Murchison in West Australia. This is used to model the intensity of gold deposits according to other spatial data: the distance to the nearest geological fault (the faults are described by lines) and the presence of a particular type of rock (described by polygons). Process intensity modelling can therefore be based on exogenous variables that are measured or calculated from geographical information.

Modelling progress is implemented regularly in the `ppm` function. The ability to model interactions between points (with the `interaction` argument in the function) in addition to density currently exists for only one particular type of processes, those of Gibbs, used for the modelling

of the spatial aggregation in industry by Sweeney et al. 2015. The `ppm` function can be used for updates.

## Conclusion

In this chapter, we have attempted to give an initial overview of the statistical methods that can be used to characterise point data. Our objective was to emphasise that the diversity of questions raised requires careful handling of statistical tools. Before any study, the question asked and its framework of analysis should, therefore, be clearly defined in order to select the most relevant statistical method. This theoretical warning is important because calculation routines are now widely accessible in the R software in particular and, in principle, pose few practical problems in use. These statistical methods may give rise to more advanced analyses in this field or additional studies, in particular in spatial econometrics for example (see Chapter 6: "Spatial econometrics: current models).

## References - Chapter 4


Arbia, Giuseppe (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*. Dordrecht: Kluwer.

Arbia, Giuseppe, Giuseppe Espa, and Danny Quah (2008). « A class of spatial econometric methods in the empirical analysis of clusters of firms in the space ». *Empirical Economics* 34.1, pp. 81–103.

Arbia, Giuseppe et al. (2012). « Clusters of firms in an inhomogeneous space: The high-tech industries in Milan ». *Economic Modelling* 29.1, pp. 3–11.

Baddeley, Adrian J., Jesper Møller, and Rasmus Plenge Waagepetersen (2000). « Non- and semi-parametric estimation of interaction in inhomogeneous point patterns ». *Statistica Neerlandica* 54.3, pp. 329–350.

Baddeley, Adrian J., Edge Rubak, and R Turner (2015b). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Interdisciplinary Statistics. 810 pages. Chapman and Hall/CRC.

Baddeley, Adrian J and Rolf Turner (2005). « Spatstat: an R package for analyzing spatial point patterns ». *Journal of Statistical Software* 12.6, pp. 1–42.

Barlet, Muriel, Anthony Briant, and Laure Crusson (2008). *Concentration géographique dans l´industrie manufacturière et dans les services en France : une approche par un indicateur en continu*. Série des documents de travail de la Direction des Études et Synthèses économiques G 2008 / 09. Institut National de la Statistique et des études économiques (Insee).

— (2013). « Location patterns of service industries in France: A distance-based approach ». *Regional Science and Urban Economics* 43.2, pp. 338–351.

Behrens, Kristian and Théophile Bougna (2015). « An anatomy of the geographical concentration of Canadian manufacturing industries ». *Regional Science and Urban Economics* 51, pp. 47–69.

Besag, Julian E. (1977). « Comments on Ripley's paper ». *Journal of the Royal Statistical Society* B 39.2, pp. 193–195.

Bonneu, Florent (2007). « Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process ». *Case Studies in Business, Industry and Government Statistics* 1.2, pp. 139–152.

Bonneu, Florent and Christine Thomas-Agnan (2015). « Measuring and Testing Spatial Mass Concentration with Micro-geographic Data ». *Spatial Economic Analysis* 10.3, pp. 289–316.

Briant, Anthony, Pierre-Philippe Combes, and Miren Lafourcade (2010). « Dots to boxes: Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations? » *Journal of Urban Economics* 67.3, pp. 287–302.

Brülhart, Marius and Rolf Traeger (2005). « An Account of Geographic Concentration Patterns in Europe ». *Regional Science and Urban Economics* 35.6, pp. 597–624.

Cole, Russel G. and Gregg Syms (1999). « Using spatial pattern analysis to distinguish causes of mortality: an example from kelp in north-eastern New Zealand ». *Journal of Ecology* 87.6, pp. 963–972.

Combes, Pierre-Philippe, Thierry Mayer, and Jacques-François Thisse (2008). *Economic Geography, The Integration of Regions and Nations*. Princeton: Princeton University Press.

Combes, Pierre-Philippe and Henry G Overman (2004). « The spatial distribution of economic activities in the European Union ». *Handbook of Urban and Regional Economics*. Ed. by J Vernon Henderson and Jacques-François Thisse. Vol. 4. Amsterdam: Elsevier. North Holland. Chap. 64, pp. 2845–2909.

Condit, Richard et al. (2000). « Spatial Patterns in the Distribution of Tropical Tree Species ». *Science* 288.5470, pp. 1414–1418.

Diggle, Peter J. (1983). *Statistical analysis of spatial point patterns*. London: Academic Press, 148 p.

Diggle, Peter J. and A. G. Chetwynd (1991). « Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations ». *Biometrics* 47.3, pp. 1155–1163.

Duranton, Gilles (2008). « Spatial Economics ». *The New Palgrave Dictionary of Economics*. Ed. by Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan.

Duranton, Gilles and Henry G. Overman (2005). « Testing for Localization Using Micro-Geographic Data ». *Review of Economic Studies* 72.4, pp. 1077–1106.

Ellison, Glenn and Edward L. Glaeser (1997). « Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach ». *Journal of Political Economy* 105.5, pp. 889–927.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr (2010). « What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns ». *The American Economic Review* 100.3, pp. 1195–1213.

Fehmi, Jeffrey S. and James W. Bartolome (2001). « A grid-based method for sampling and analysing spatially ambiguous plants. » *Journal of Vegetation Science* 12.4, pp. 467–472.

Goreaud, Françcois and Raphaël Pélissier (1999). « On explicit formulas of edge effect correction for Ripley's K-function ». *Journal of Vegetation Science* 10.3, pp. 433–438. ISSN: 1654-1103. DOI: 10.2307/3237072. URL: http://dx.doi.org/10.2307/3237072.

Goreaud, François (2000). « Apports de l'analyse de la structure spatiale en forêt tempérée à l'étude de la modélisation des peuplements complexes ». PhD Thesis. Nancy: ENGREF.

Heinrich, Lothar (1991). « Goodness-of-fit tests for the second moment funciton of a stationary multidimensional poisson process ». *Statistics: A Journal of Theoretical and Applied Statistics* 22.2, pp. 245–268. DOI: 10.1080/02331889108802308.

Holmes, Thomas J. and John J. Stevens (2004). « Spatial Distribution of Economic Activities in North America ». *Cities and Geography*. Ed. by J. Vernon Henderson and Jacques-François Thisse. Vol. 4. Handbook of Regional and Urban Economics Chapter 63 - Supplement C. Elsevier, pp. 2797–2843.

Illian, Janine et al. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Statistics in Practice. Chichester: Wiley-Interscience, p. 534.

Jensen, Pablo and Julien Michel (2011). « Measuring spatial dispersion: exact results on the variance of random spatial distributions ». *The Annals of Regional Science* 47.1, pp. 81–110.

Lagache, Thibault et al. (2013). « Analysis of the Spatial Organization of Molecules with Robust Statistics ». *Plos One* 8.12, e80914.

Lang, G., E. Marcon, and F. Puech (2015). « Distance-Based Measures of Spatial Concentration: Introducing a Relative Density Function ». *HAL* hal-01082178.version 2.

Lang, Gabriel and Eric Marcon (2013). « Testing randomness of spatial point patterns with the Ripley statistic ». *ESAIM: Probability and Statistics* 17, pp. 767–788.

Marcon, Eric and Florence Puech (2003). « Evaluating the Geographic Concentration of Industries Using Distance-Based Methods ». *Journal of Economic Geography* 3.4, pp. 409–428.

— (2010). « Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods ». *Journal of Economic Geography* 10.5, pp. 745–762.

— (2015a). « Mesures de la concentration spatiale en espace continu : théorie et applications ». *Économie et Statistique* 474, pp. 105–131.

— (2017). « A typology of distance-based measures of spatial concentration ». *Regional Science and Urban Economics* 62, pp. 56–67.

Marcon, Eric, Florence Puech, and Stéphane Traissac (2012). « Characterizing the relative spatial structure of point patterns ». *International Journal of Ecology* 2012.Article ID 619281, p. 11.

Marcon, Eric et al. (2015b). « Tools to Characterize Point Patterns: dbmss for R ». *Journal of Statistical Software* 67.3, pp. 1–15.

Maurel, Françoise and Béatrice SÉdillot (1999). « A measure of the geographic concentration in french manufacturing industries ». *Regional Science and Urban Economics* 29.5, pp. 575–604.

Møller, Jesper and Hakon Toftaker (2014). « Geometric Anisotropic Spatial Point Pattern Analysis and Cox Processes ». *Scandinavian Journal of Statistics*. Monographs on Statistics and Applies Probabilities 41.2, pp. 414–435.

Møller, Jesper and Rasmus Plenge Waagepetersen (2004). *Statistical Inference and Simulation for Spatial Point Processes*. Vol. 100. Monographs on Statistics and Applies Probabilities. Chapman and Hall, 300 p.

Openshaw, S. and P. J. Taylor (1979a). « A million or so correlation coefficients: three experiments on the modifiable areal unit problem ». *Statistical Applications in the Spatial Sciences*. Ed. by N. Wrigley. London: Pion, pp. 127–144.

Ripley, Brian D. (1976). « The Second-Order Analysis of Stationary Point Processes ». *Journal of Applied Probability* 13.2, pp. 255–266.

— (1977). « Modelling Spatial Patterns ». *Journal of the Royal Statistical Society* B 39.2, pp. 172–212.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall, 175 p.

Sweeney, Stuart H. and Edward J. Feser (1998). « Plant Size and Clustering of Manufacturing Activity ». *Geographical Analysis* 30.1, pp. 45–64.

Sweeney, Stuart H and Miguel Gómez-Antonio (2015). « Localization and Industry Clustering Econometrics: an Assessment of Gibbs Models for Spatial Point Processes ». *Journal of Regional Science* 56.2, pp. 257–287.

Szwagrzyk, Jerzy and Marek Czerwczak (1993). « Spatial patterns of trees in natural forests of East-Central Europe ». *Journal of Vegetation Science* 4.4, pp. 469–476.

Veen, Alejandro and Frederic Paik Schoenberg (2006). « Assessing Spatial Point Process Models Using Weighted K-functions: Analysis of California Earthquakes ». *Case Studies in Spatial Point Process Modeling*. Ed. by Adrian Baddeley et al. New York, NY: Springer New York, pp. 293–306.

Wiegand, T. and K. A. Moloney (2004). « Rings, circles, and null-models for point pattern analysis in ecology ». *Oikos* 104.2, pp. 209–229.