# Towards new identification strategies: point patterns

Ripley, B.D. (1977). Modelling spatial patterns, *Journal of the Royal Statistical Society, Series B*, 39, 172–192

PSE - Public economics research seminar

Fall 2022

# Introduction

- There is an old and growing literature about the location decisions of firms and FDI [Crozet, Mayer & Muchielli, 2004];

# Introduction

- There is an old and growing literature about the location decisions of firms and FDI [Crozet, Mayer & Muchielli, 2004];
- But due to data limitations, clustering behaviours of individuals and households depending of their income and social status largely remains inside the scope of segregation statistics;

# Introduction

- There is an old and growing literature about the location decisions of firms and FDI [Crozet, Mayer & Muchielli, 2004];
- But due to data limitations, clustering behaviours of individuals and households depending of their income and social status largely remains inside the scope of segregation statistics;
- Until recently, most of the segregation indexes relied on series computed at a relatively large geographical level ; the U.S. *Census track* is designed to cover approximately 4000 inhabitants; the INSEE's *IRIS*, slightly more than 2000.

# Introduction

- Yet we know, from the econometrics of peer effects, that the neighborhood interactions that are most powerful in predicting someone's wage and education, are very narrow ones. The vast U.S. literature on desegregation policies has long demonstrated that moving residents of a public housing unit located in a ghetto, to a similar unit, but located among an affluent suburb, makes very little difference for the future income and education of children raised there [Oreopoulos, 2001].

# Introduction

- Yet we know, from the econometrics of peer effects, that the neighborhood interactions that are most powerful in predicting someone's wage and education, are very narrow ones. The vast U.S. literature on desegregation policies has long demonstrated that moving residents of a public housing unit located in a ghetto, to a similar unit, but located among an affluent suburb, makes very little difference for the future income and education of children raised there [Oreopoulos, 2001].
- The relevant level to study individual location decisions is not the city, nor the district, but the neighbourhood. That's the rationale behind the *Enquête emploi* of the INSEE which relies on the study of very localised units (17 workers living near one another on average) [Goux & Maurin, 2007].

# Introduction

- Yet we know, from the econometrics of peer effects, that the neighborhood interactions that are most powerful in predicting someone's wage and education, are very narrow ones. The vast U.S. literature on desegregation policies has long demonstrated that moving residents of a public housing unit located in a ghetto, to a similar unit, but located among an affluent suburb, makes very little difference for the future income and education of children raised there [Oreopoulos, 2001].
- The relevant level to study individual location decisions is not the city, nor the district, but the neighbourhood. That's the rationale behind the *Enquête emploi* of the INSEE which relies on the study of very localised units (17 workers living near one another on average) [Goux & Maurin, 2007].
- Now that precisely geolocalized individual data are available in growing number, it might be interesting to use segregation indexes based, not on averages over a sub-region, but over point systems.

## Setting

We consider a configuration of $n$ points of $\mathbb{R}^2$ denoted $C = \{x_1, ..., x_n\}$

- Let $n(C)$ be the number of points in the configuration;

## Setting

We consider a configuration of $n$ points of $\mathbb{R}^2$ denoted $C = \{x_1, ..., x_n\}$

- Let $n(C)$ be the number of points in the configuration;
- Let $B$ be a subset of $\mathbb{R}^2$;

# Setting

We consider a configuration of $n$ points of $\mathbb{R}^2$ denoted $C = \{x_1, ..., x_n\}$

- Let $n(C)$ be the number of points in the configuration;
- Let $B$ be a subset of $\mathbb{R}^2$;
- $n(C \cap B)$ gives the number of points of the configuration located in region $B$;

## Setting

We consider a configuration of $n$ points of $\mathbb{R}^2$ denoted $C = \{x_1, ..., x_n\}$

- Let $n(C)$ be the number of points in the configuration;
- Let $B$ be a subset of $\mathbb{R}^2$;
- $n(C \cap B)$ gives the number of points of the configuration located in region $B$;

**Definition**

The process $X$ generating the configuration of points is said to be *defined* if, for any region $B$, we are able to provide the number $n(X \cap B)$.

# Complete Spatial Randomness

We would like to define a special type of process which generates points randomly over space.

- Let $W$ be the *window of observation*; $B$ designates any compact subset of $W$;

# Complete Spatial Randomness

We would like to define a special type of process which generates points randomly over space.

- Let $W$ be the *window of observation*; $B$ designates any compact subset of $W$;
- Let $u = (u_1, u_2)$ be a point. We have:

$$P(u \in B) = \frac{|B|}{|W|}$$

# Complete Spatial Randomness

We would like to define a special type of process which generates points randomly over space.

- Let $W$ be the *window of observation*; $B$ designates any compact subset of $W$;
- Let $u = (u_1, u_2)$ be a point. We have:

$$P(u \in B) = \frac{|B|}{|W|}$$

- Since the probability that a point be located in $B$ is fixed, equal to $P(u \in B)$, that we henceforth denote $p$, we can count the number of points using a binomial distribution:

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

# Complete Spatial Randomness

We would like to define a special type of process which generates points randomly over space.

- Let $W$ be the *window of observation*; $B$ designates any compact subset of $W$;
- Let $u = (u_1, u_2)$ be a point. We have:

$$P(u \in B) = \frac{|B|}{|W|}$$

- Since the probability that a point be located in $B$ is fixed, equal to $P(u \in B)$, that we henceforth denote $p$, we can count the number of points using a binomial distribution:

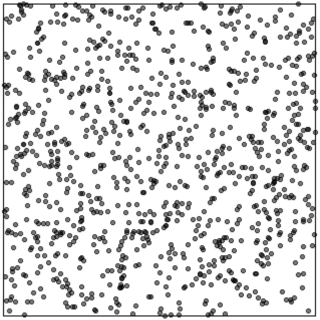
$$P(n(X \cap B) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Process $X$ is now defined.

# Complete Spatial Randomness (in R)

```
library(spatstat)
plot(runifpoint(100, win=owin(c(0, 10),c(0, 10))))
```

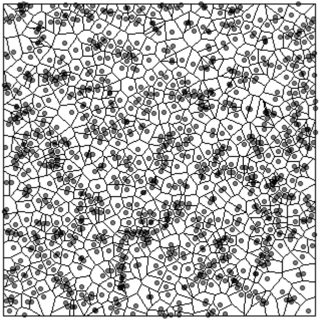
Figure 1: 1000 points (binomial process) over a $10 \times 10$ square



# Complete Spatial Randomness (in R)

```
library(dismo)
plot(voronoi(m, ext=d))
```

Figure 2: Corresponding Voronoi diagram



# Complete Spatial Randomness

The process $X$ is defined, but it cannot be said to be completely random.

Randomness is indeed defined in this context by two properties:

- Homogeneity - $\mathrm{E}(n(X \cap B))$ is determined by the area of $B$, not by its localisation, i.e. there is no preference for one region.

# Complete Spatial Randomness

The process $X$ is defined, but it cannot be said to be completely random.

Randomness is indeed defined in this context by two properties:

- Homogeneity - $\mathrm{E}(n(X \cap B))$ is determined by the area of $B$, not by its localisation, i.e. there is no preference for one region.
  - A binominal process is indeed homogeneous;

# Complete Spatial Randomness

The process $X$ is defined, but it cannot be said to be completely random.

Randomness is indeed defined in this context by two properties:

- Homogeneity - $\mathrm{E}(n(X \cap B))$ is determined by the area of $B$, not by its localisation, i.e. there is no preference for one region.
    - A binominal process is indeed homogeneous;
- Independence - $\forall (B, B')$, $n(X \cap B) \perp n(X \cap B')$

# Complete Spatial Randomness

The process $X$ is defined, but it cannot be said to be completely random.
Randomness is indeed defined in this context by two properties:

- Homogeneity - $\mathrm{E}(n(X \cap B))$ is determined by the area of $B$, not by its localisation, i.e. there is no preference for one region.
  - A binominal process is indeed homogeneous;
- Independence - $\forall(B, B')$, $n(X \cap B) \perp n(X \cap B')$
  - A binomial process is not independent: the number of points $k$ in $B$ determines the number of points $n-k$ in $W \setminus B$ ;

## Complete Spatial Randomness

The properties of the binomial distribution provide a way out. We have:

# Complete Spatial Randomness

The properties of the binomial distribution provide a way out. We have:

$$\mathrm{E}(n(X \cap B)) = np = n\frac{|B|}{|W|}$$

# Complete Spatial Randomness

The properties of the binomial distribution provide a way out. We have:

$$\mathrm{E}(n(X \cap B)) = np = n\frac{|B|}{|W|}$$

We'll now fix the number of points so that we might define a new constant $\lambda = \frac{n}{|W|}$. It can be interpreted as the average number of points in a unit of area 1: it is the *intensity* of the process.

# Complete Spatial Randomness

The properties of the binomial distribution provide a way out. We have:

$$\mathrm{E}(n(X \cap B)) = np = n\frac{|B|}{|W|}$$

We'll now fix the number of points so that we might define a new constant $\lambda = \frac{n}{|W|}$. It can be interpreted as the average number of points in a unit of area 1: it is the *intensity* of the process.
We now divide $W$ into $m$ equal regions. When these $m$ regions become infinitely small, the probability that a point be in $B$ is given by:

$$p_m \xrightarrow[m \to \infty]{} \frac{\lambda |B|}{m}$$

# Poisson homogeneous point process

Distribution $\mathcal{B}(m, p_m)$ provides a count of points located in region $B$ when it becomes infinitely small, and by the properties of the binomial distribution, it converges (for $m$ infinitely great) to a Poisson distribution of parameter $\lambda|B|$:

# Poisson homogeneous point process

Distribution $\mathcal{B}(m, p_m)$ provides a count of points located in region $B$ when it becomes infinitely small, and by the properties of the binomial distribution, it converges (for $m$ infinitely great) to a Poisson distribution of parameter $\lambda|B|$:

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

# Poisson homogeneous point process

Distribution $\mathcal{B}(m, p_m)$ provides a count of points located in region $B$ when it becomes infinitely small, and by the properties of the binomial distribution, it converges (for $m$ infinitely great) to a Poisson distribution of parameter $\lambda|B|$:

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

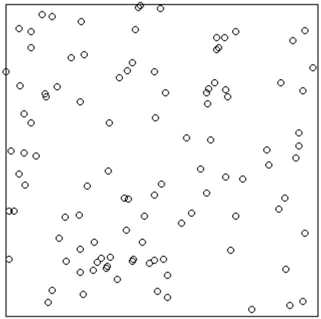
## Properties

- A Poisson process is homogeneous: $\mathrm{E}(n(X \cap B)) = \lambda|B|$;
- It has been defined in order to be independent (the probability of having one point in a region becoming negligible, it has no influence on the probability over the remaining regions).

# Poisson homogeneous point process (in R)

```
library(dismo)
plot(rpoispp(100))
```

Figure 3: Poisson process (100 points on a 1 $\times$ 1 square)



# Matern processes

We can derive from Poisson processes models akin to real phenomena:

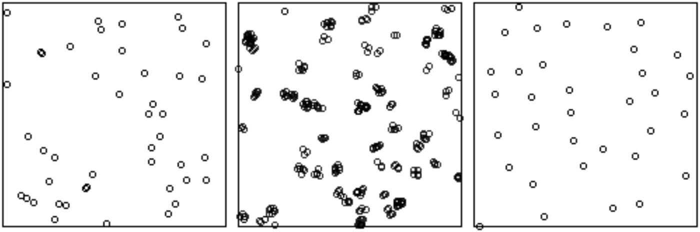
- Marten type II process draws points in order, new points being removed if they are within distance $r$ of a previously drawn point.

# Matern processes

We can derive from Poisson processes models akin to real phenomena:

- Marten type II process draws points in order, new points being removed if they are within distance $r$ of a previously drawn point.
- Marten cluster process defines a number of parent points, and draws points within a distance $r$ with a Poisson process.

Figure 4: Examples of Poisson, Marten cluster & Marten II processes



# Poisson inhomogeneous process

More formalized approaches exist to reach a realistic model. Indeed, if the independence property is genuine to any Poisson process, the homogeneity property can be relaxed.

# Poisson inhomogeneous process

More formalized approaches exist to reach a realistic model. Indeed, if the independence property is genuine to any Poisson process, the homogeneity property can be relaxed.
We can turn the constant $\lambda$ into a function of the coordinates of the points $\lambda(u_1, u_2)$. This will allow us to privilege some regions over others. Such a pattern defines an inhomogeneous Poisson process.

# Poisson inhomogeneous process

More formalized approaches exist to reach a realistic model. Indeed, if the independence property is genuine to any Poisson process, the homogeneity property can be relaxed.
We can turn the constant $\lambda$ into a function of the coordinates of the points $\lambda(u_1, u_2)$. This will allow us to privilege some regions over others. Such a pattern defines an inhomogeneous Poisson process. $\lambda()$ might be interpreted as the excepted number of points in an infinitely small subset of $W$.

### Density

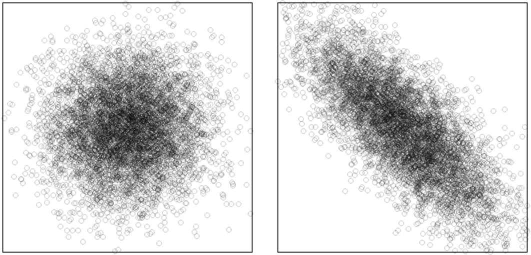
The probability density function of an inhomogeneous Poisson process is :

$$f(u) = \frac{\lambda(u)}{\int_B v dv}$$

# Poisson inhomogeneous process (in R)

```
plot(rpoispp(function(x, y) {...}))
```

Figure 5: Inhomogenous processes, $\lambda(x)$ being two different versions of $\mathcal{N}_2$



# Testing for randomness

How do test whether or not a real-world spatial configuration deviates from one of the models developed here?

# Ripley $K$ function

One idea is to define a measure of the number of points found within a certain radius $r$ around a center.

# Ripley $K$ function

One idea is to define a measure of the number of points found within a certain radius $r$ around a center.
This is the idea of B.D. Ripley's (1977) $K$ function.

# Ripley $K$ function

One idea is to define a measure of the number of points found within a certain radius $r$ around a center.
This is the idea of B.D. Ripley's (1977) $K$ function.

## Ripley's $K$ function

If we denote $\lambda$ the intensity of the Poisson process, $b(u,r)$ the disk of center $u$ and radius $r$, Ripley's $K$ is defined as:

$$K(r) = \frac{1}{\lambda}\mathbb{E}[n(X \cap b(u,r) \setminus u)|u \in X]$$

# Ripley $K$ function

One idea is to define a measure of the number of points found within a certain radius $r$ around a center.
This is the idea of B.D. Ripley's (1977) $K$ function.

## Ripley's $K$ function

If we denote $\lambda$ the intensity of the Poisson process, $b(u, r)$ the disk of center $u$ and radius $r$, Ripley's $K$ is defined as:

$$K(r) = \frac{1}{\lambda}\mathbb{E}[n(X \cap b(u, r) \setminus u)|u \in X]$$

The function counts the neighboring points, normalizing that count through a division by the intensity of the process.

# Ripley $K$ function

The most common estimator of the $K$ function relies on the most simple estimator of the intensity, $\hat{\lambda} = (n/|W|)$.

## Ripley's $K$ function

If we denote $W$ the window of observation, $n$ the number of points of the configuration, and $r$ the radius of analysis, the estimated $K$ function writes:

$$\hat{K}(r) = \frac{1}{\hat{\lambda}^2|W|} \sum_i \sum_{i \neq j} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}} e(x_i, x_j; r)$$

# Ripley $K$ function

The most common estimator of the $K$ function relies on the most simple estimator of the intensity, $\hat{\lambda} = (n/|W|)$.

## Ripley's $K$ function

If we denote $W$ the window of observation, $n$ the number of points of the configuration, and $r$ the radius of analysis, the estimated $K$ function writes:

$$\hat{K}(r) = \frac{1}{\hat{\lambda}^2 |W|} \sum_i \sum_{i \neq j} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}} e(x_i, x_j; r)$$

The parameter $e$ is an edge factor ; it is applied when there exists over $W$ certain obstacles (administrative frontiers, geographical limits) that might bias the randomness of the distribution (for instance, a certain firm might operate only over a certain region).

# Tests of randomness with the $K$ function

If the process generating the points configuration was perfectly random, the theoretical value of the function should be:

# Tests of randomness with the $K$ function

If the process generating the points configuration was perfectly random, the theoretical value of the function should be:

$$K_{poisson}(r) = \pi r^2$$

# Tests of randomness with the $K$ function

If the process generating the points configuration was perfectly random, the theoretical value of the function should be:

$$K_{poisson}(r) = \pi r^2$$

That value can be used as the null in a test of deviation from randomness. The idea is to compare $K_{poisson}$ and the $\hat{K}$ estimated from the data to check for a significant deviation.

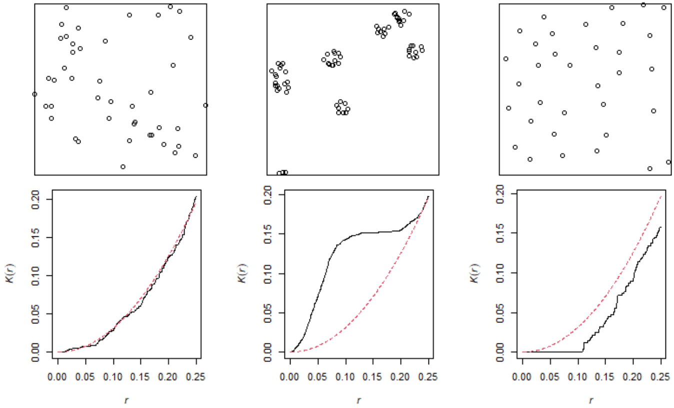
# Tests of randomness with the $K$ function

```
plot(Kest(...,correction="isotropic"))
```

# Tests of randomness with the $K$ function

```
plot(Kest(...,correction="isotropic"))
```

Figure 6: $\hat{K}(r)$ for examples of a Poisson, Matern cluster & Martin II processes



# Tests of randomness with the $K$ function

Ripley derived from the $K$ a very intuitive test based on Monte Carlo simulations.

# Tests of randomness with the $K$ function

Ripley derived from the $K$ a very intuitive test based on Monte Carlo simulations.
The idea is to generated $D$ spatial simulations over $W$ based on $\mathcal{H}_0$ (here, it shall be a Poisson process, but it might be a different one).

- For each simulation we compute the function $\hat{K}^{(d)}$.
- We define an upper bound $U$ and a lower bound $L$ so as to generate an envelope:

$$U(r) = \max_d \hat{K}^{(d)}(r)$$

$$L(r) = \min_d \hat{K}^{(d)}(r)$$

# Tests of randomness with the $K$ function

If the generating process is a Poisson one, $\hat{K}(r)$ and the $D$ Monte Carlo simulations are perfectly independent, equivalent, processes.

# Tests of randomness with the $K$ function

If the generating process is a Poisson one, $\hat{K}(r)$ and the $D$ Monte Carlo simulations are perfectly independent, equivalent, processes. The probability that $\hat{K}(r)$ be above $U$ (i.e. that it is the maximum of the $D+1$ drawings) is $\frac{1}{D+1}$ and by symmetry, the probability that $\hat{K}(r)$ lies outside the envelope $[L(r), U(r)]$, which is also the risk of our test, shall be:

# Tests of randomness with the $K$ function

If the generating process is a Poisson one, $\hat{K}(r)$ and the $D$ Monte Carlo simulations are perfectly independent, equivalent, processes. The probability that $\hat{K}(r)$ be above $U$ (i.e. that it is the maximum of the $D+1$ drawings) is $\frac{1}{D+1}$ and by symmetry, the probability that $\hat{K}(r)$ lies outside the envelope $[L(r), U(r)]$, which is also the risk of our test, shall be:

$$\alpha = \frac{2}{D+1}$$

# Tests of randomness with the $K$ function

If the generating process is a Poisson one, $\hat{K}(r)$ and the $D$ Monte Carlo simulations are perfectly independent, equivalent, processes. The probability that $\hat{K}(r)$ be above $U$ (i.e. that it is the maximum of the $D+1$ drawings) is $\frac{1}{D+1}$ and by symmetry, the probability that $\hat{K}(r)$ lies outside the envelope $[L(r), U(r)]$, which is also the risk of our test, shall be:

$$\alpha = \frac{2}{D+1}$$

I.e., for a 5% risk, we have to generate $D = 39$ simulations.

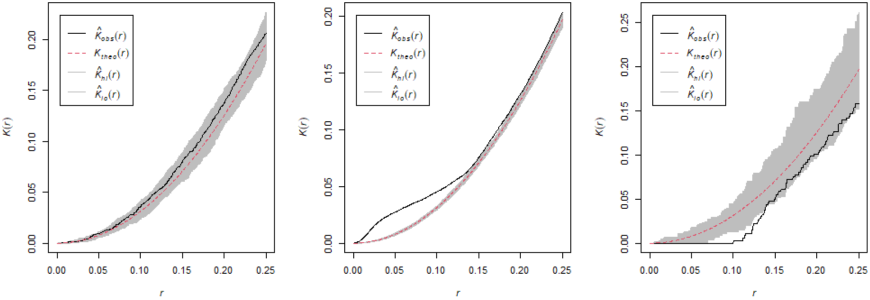
# Tests of randomness with the $K$ function (in R)

```
p <- rpoispp(100,win=owin(c(0, 1),c(0, 1)))
plot(envelope(p, Kest, nsim = 39, rank = 1))
```

# Tests of randomness with the $K$ function (in R)

```
p <- rpoispp(100,win=owin(c(0, 1),c(0, 1)))
plot(envelope(p, Kest, nsim = 39, rank = 1))
```

Figure 7: Envelopes for examples of a Poisson, Matern cluster & Martin II processes



# Alternatives

Issue :

# Alternatives

Issue : Perfect randomness in space as $\mathcal{H}_0$ makes a very undemanding null hypothesis. To check for non-randomness or clustering effects of people and firms, we must rely on a more realistic null, which fits better the economic intuition.

# Alternatives

Issue : Perfect randomness in space as $\mathcal{H}_0$ makes a very undemanding null hypothesis. To check for non-randomness or clustering effects of people and firms, we must rely on a more realistic null, which fits better the economic intuition.
The same Monte Carlo based tests can be done with a non-Poisson model (for instance, a clustered one). But the choice of such a model contains an element of arbitrariness.

# Alternatives

Issue : Perfect randomness in space as $\mathcal{H}_0$ makes a very undemanding null hypothesis. To check for non-randomness or clustering effects of people and firms, we must rely on a more realistic null, which fits better the economic intuition.

The same Monte Carlo based tests can be done with a non-Poisson model (for instance, a clustered one). But the choice of such a model contains an element of arbitrariness.

An alternative is to use modified $K$ functions...

# The $M$ function

Many developments of Ripley's $K$ have been introduced in the literature, especially in natural sciences. One measure specific to economists is Marcon & Puech (2010)'s $M$ function.

# The $M$ function

Many developments of Ripley's $K$ have been introduced in the literature, especially in natural sciences. One measure specific to economists is Marcon & Puech (2010)'s $M$ function.
We consider one subtype $S$ of points over the configuration $X$, and we take the ratio between the frequency of that type of points over increasingly large disks of radius $r$ drawn around each point of $S$, and their frequency over the whole window.

# The $M$ function

Many developments of Ripley's $K$ have been introduced in the literature, especially in natural sciences. One measure specific to economists is Marcon & Puech (2010)'s $M$ function.
We consider one subtype $S$ of points over the configuration $X$, and we take the ratio between the frequency of that type of points over increasingly large disks of radius $r$ drawn around each point of $S$, and their frequency over the whole window.

### $M$ function

Let $n_S$ be the number of points of type $S$. Then Marcon & Puech's $M()$ function is defined as:

$$M(r) = \sum_{i \in S} \frac{\sum_{j \neq i, j \in S} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}}}{\sum_{i \neq j} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}}} \Big/ \frac{n_S - 1}{n - 1}$$

# The $M$ function (in R)

One example of application to economics provided by Marcon & Puech is the analysis of the structure of the BPE (*Base publique des équipements*) of the INSEE, which provides Lambert93 coordinates for a whole set of amenities (public buildings, but also different types of shops and firms).

# The $M$ function (in R)

One example of application to economics provided by Marcon & Puech is the analysis of the structure of the BPE (*Base publique des équipements*) of the INSEE, which provides Lambert93 coordinates for a whole set of amenities (public buildings, but also different types of shops and firms).
We might select, over these, one type of equipment as the $S$ reference category.
A test similar to the one previously presented can be built with Monte Carlo simulations.

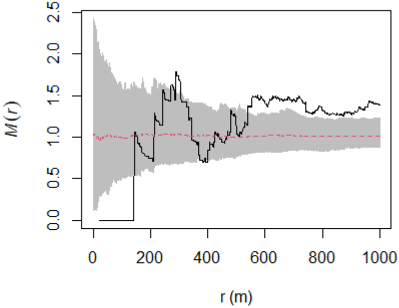
# The $M$ function (in R)

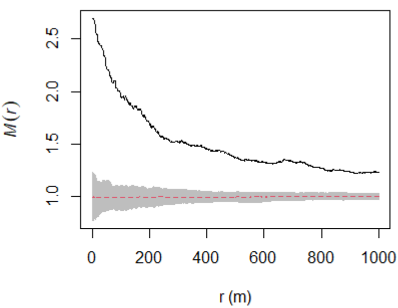
```
bpe_equip_wmppp <- wmppp(bpe_equip)
r<- 0:1000
nb <- 99
menv_eco <- MEnvelope(bpe_equip_wmppp, r, nb,
                      ReferenceType="C104")
```

# The $M$ function (in R)

```
bpe_equip_wmppp <- wmppp(bpe_equip)
r<- 0:1000
nb <- 99
menv_eco <- MEnvelope(bpe_equip_wmppp, r, nb,
                      ReferenceType="C104")
```

Figure 8: Estimated $M$ over Rennes, ref. schools (left) and doctors (right)



# Marked points

In most economic data, points are *marked*. To $X$ is added a variable $M$ providing further information about the points.

# Marked points

In most economic data, points are *marked*. To $X$ is added a variable $M$ providing further information about the points.
$M$ can be categorical (if points are equipments, $M$ can give their type) or continuous (if points are persons, $M$ can provide their taxable income).

# Marked points

In most economic data, points are *marked*. To $X$ is added a variable $M$ providing further information about the points.
$M$ can be categorical (if points are equipments, $M$ can give their type) or continuous (if points are persons, $M$ can provide their taxable income).

### Null hypotheses

- $[M][X|M]$; there is one process by submark. $\mathcal{H}_0$ postulates the independence between these different processes (*Independence of components*).
- $[X][M|X]$; locations are drawn first, and marks assigned in a second time. $\mathcal{H}_0$ postulates that, given the locations, the marks are conditionally independently distributed (*Random labelling*)
- $[X, M]$ combines the two preceding nulls (*CSR and Independence*)

# Marked points (measures)

(Baddeley, Rubak & Turner, 2015) provide a way to test each null. For $K$ and $M$ (or any other type of measure), we can define a cross function, i.e. a function, the cumulative count of which is not based any more on points of the same type, but over points of another type:

# Marked points (measures)

(Baddeley, Rubak & Turner, 2015) provide a way to test each null. For $K$ and $M$ (or any other type of measure), we can define a cross function, i.e. a function, the cumulative count of which is not based any more on points of the same type, but over points of another type:

$K_{cross}$ and $M_{cross}$

For two categories $S$ and $T$, the cross functions write:

$$\hat{K}_{S,T}(r) = \frac{1}{\hat{\lambda}_S n_S} \sum_{i \in S} \sum_{j \in T} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}}$$

$$\hat{M}_{S,T}(r) = \sum_{i \in S} \frac{\sum_{j \in T} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}}}{\sum_{i \neq j} \mathbb{I}_{\{\|x_i - x_j\| \leq r\}}} / \frac{n_T}{n-1}$$

# Marked points tests (*CSR and independence* case)

If we want to test the combined $\mathcal{H}_0$, we can use the very same Monte Carlo based test over $K_{cross}$.

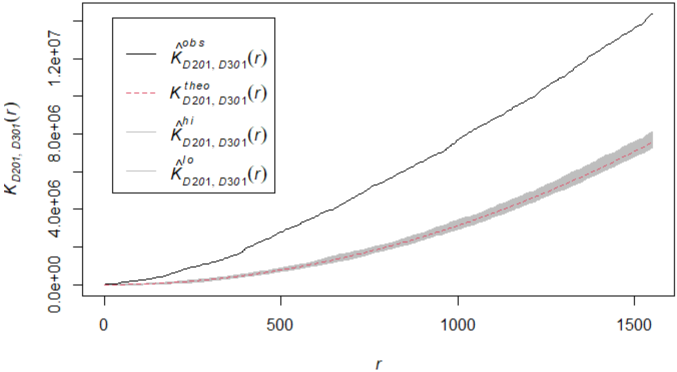
# Marked points tests (*CSR and independence* case)

If we want to test the combined $\mathcal{H}_0$, we can use the very same Monte Carlo based test over $K_{cross}$.

```
plot(envelope(bpe_ppp, Kcross, nsim = 39,
                                i = "D201", j = "D301"))
```

# Marked points tests (*CSR and independence* case)

Figure 9: $K_{cross}$ CSRI-test ; doctors (D201) to pharmacies (D301) in Rennes



# Marked points tests (*Independence of components* case)

The null model is generated step-wise: 1. $X$ is generated as a Poisson process ; 2. Marks are assigned randomly to create categories; 3. To each category is applied a random vector translation.

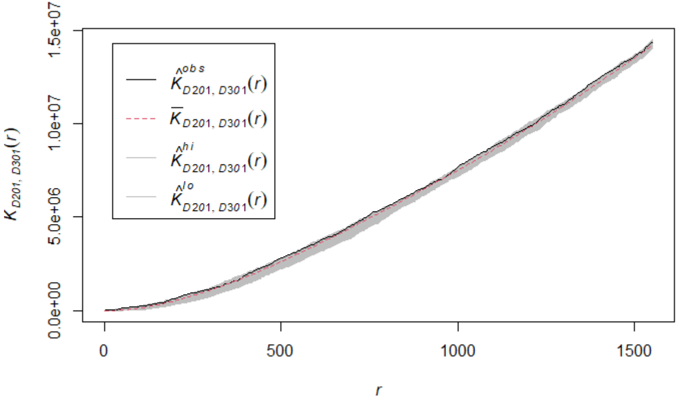
# Marked points tests (*Independence of components* case)

The null model is generated step-wise: 1. $X$ is generated as a Poisson process ; 2. Marks are assigned randomly to create categories; 3. To each category is applied a random vector translation.

```
plot(envelope(bpe_ppp, Kcross, nsim = 39, i = "D201",
              j = "D301", simulate = expression(rshift
              (bpe_ppp, radius = 200))))
```

# Marked points tests (*Independence of components* case)

Figure 10: $K_{cross}$ test (*Independence of components*)



# Marked points tests (*Random labelling* case)

We define a new *one-type to any-type* measure, denoted $K_{\bullet i}(r)$. It's Ripley's $K$, with $1/\lambda$ as initial factor, and as count, the number of points of any type within distance $r$ of a certain type of center.

# Marked points tests (*Random labelling* case)

We define a new *one-type to any-type* measure, denoted $K_{\bullet i}(r)$. It's Ripley's $K$, with $1/\lambda$ as initial factor, and as count, the number of points of any type within distance $r$ of a certain type of center.
Here, the null model is built by holding the points' coordinates of the real set fixed, and randomly reshuffling the marks.

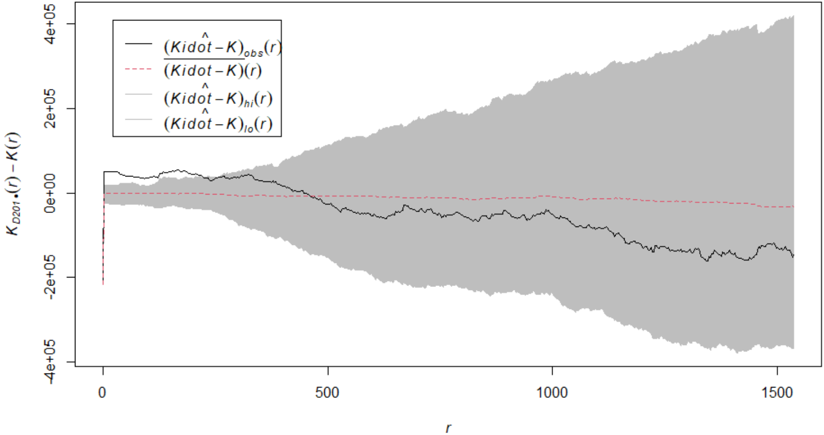
# Marked points tests (*Random labelling* case)

We define a new *one-type to any-type* measure, denoted $K_{\bullet i}(r)$. It's Ripley's $K$, with $1/\lambda$ as initial factor, and as count, the number of points of any type within distance $r$ of a certain type of center.
Here, the null model is built by holding the points' coordinates of the real set fixed, and randomly reshuffling the marks.
Under the null, we should have $K_{\bullet i}(r) = K(r)$. We then create a function $K_{\bullet i}(r) - K(r)$ and use Monte Carlo simulations to create a corresponding envelope.

```
Kdif <- function(X, ..., i){
  Kidot <- Kdot(X,...,i=i)
  K <- Kest(X,...)
  dif <- eval.fv(Kidot - K)
  return (dif)
}
E <- envelope(bpe_ppp, Kdif, nsim = 39, i = "D201",
              simulate = expression(rlabel(bpe_ppp)))
```

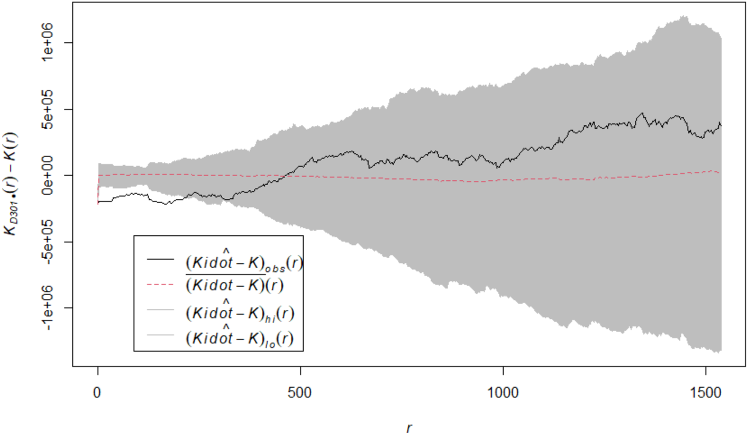
# Marked points tests (*Random labelling* case)

Figure 11: $K_{\bullet} - K$ test (D201 as center)



# Marked points tests (*Random labelling* case)

Figure 12: $K_{\bullet} - K$ test (D301 as center)



# Marked points tests (*Random labelling* case)

- When we start the computation from D201 (doctors), $K_\bullet(r) > K(r)$ for the first 200 m, before its comes inside the confidence envelope.

# Marked points tests (*Random labelling* case)

- When we start the computation from D201 (doctors), $K_\bullet(r) > K(r)$ for the first 200 m, before its comes inside the confidence envelope.
- Conversely, when we start from D301 (doctors), $K_\bullet(r) < K(r)$ for the first 200 m.

# Marked points tests (*Random labelling* case)

- When we start the computation from D201 (doctors), $K_\bullet(r) > K(r)$ for the first 200 m, before its comes inside the confidence envelope.
- Conversely, when we start from D301 (doctors), $K_\bullet(r) < K(r)$ for the first 200 m.
- I.e., when we start from doctors, we are set to meet more health services in the immediate neighbourhood (than we would, if we started from a random point).

# Marked points tests

- $\mathcal{H}_{0,\text{Independence of components}}$ holds with a 5% risk.

# Marked points tests

- $\mathcal{H}_{0,\text{Independence of components}}$ holds with a 5% risk.
- But $\mathcal{H}_{0,\text{Random labelling}}$ does not.

# Marked points tests

- $\mathcal{H}_{0,\text{Independence of components}}$ holds with a 5% risk.
- But $\mathcal{H}_{0,\text{Random labelling}}$ does not.
- I.e. $\mathcal{H}_{0,\text{CSRI}}$ does not hold because of the former one. $X$ is CSR conditional on $M$, but $M$ is not randomly assigned conditional on $X$. I.e. doctors and pharmacists do not purposely cluster near one another. But among every single equipment location, not every one is allocated to each activity (doctors benefit from more central positions). That's why we empirically observe a deviation from theoretical randomness.

# Road-map for application

- This can be used directly on localised data on household disposable income (INSEE/DGFiP-*Filocom*);

# Road-map for application

- This can be used directly on localised data on household disposable income (INSEE/DGFiP-*Filocom*);
- Could also be used on data covering limited geographical units ; bracket decomposition of taxable income by commune (DGFiP-*Revenus communaux*), information on income at the IRIS level (INSEE-*FiLoSoFi*)...

# References (Empirics)


- CROZET, Matthieu, Thierry MAYER & Jean-Louis MUCHIELLI (2004), How do firms agglomerate? A study of FDI in France, *Regional Science and Urban Economics*, 34, issue 1, p. 27-54
- GOUX, Dominique & Eric MAURIN (2007). Close Neighbours Matter: Neighbourhood Effects on Early Performance at School, *Economic Journal*, Wiley, 117 (523), p. 1193-1215
- OREOPOULOS, Philip (2003). The Long-Run Consequences of Living in a Poor Neighborhood, *The Quarterly Journal of Economics*, Volume 118, Issue 4, November, p. 1533–1575

# References (Statistics)


- BADDELEY, Adrian J., Rolf TURNER (2005). Spatstat : an R package for analyzing spatial point patterns. *Journal of Statistical Software*. 12.
- BADDELEY, Adrian J., Edge RUBAK & Rolf TURNER (2015). *Spatial Point Patterns : Methodology and Applications with R*. Chapman Hall/CRC Interdisciplinary Statistics. 810 p.
- MARCON, Eric & Florence PUECH (2010). Measures of the Geographic Concentration of Industries : Improving Distance-Based Methods. *Journal of Economic Geography*, 10.5, p. 745–762


*Thanks!*