A Typology of Distance-Based Measures of Spatial Concentration

Eric Marcon[1], Florence Puech[2]*

Abstract
Over the last decade, distance-based methods have been introduced and then improved in the field of spatial economics to gauge the geographic concentration of activities. There is a growing literature on this theme including new tools, discussions on their specific properties and various applications. However, there is currently no typology of distance-based methods. This paper fills that gap. The proposed classification helps understand all the properties of distance-based methods and proves that they are variations on the same framework.

JEL Classification: C10, C60, R12

Keywords
Spatial concentration, Aggregation, Point patterns, Agglomeration, Economic geography

[1] *AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana.*
[2] *Université de Paris-Sud, RITM, 54 Boulevard Desgranges, F-92330 Sceaux, France*
***Corresponding author**: florence.puech@u-psud.fr

Contents

Introduction 1
1 Basic principles 2
2 Distance-based methods: a brief presentation 3
2.1 The *g* function . . . 3
2.2 The *K* function . . . 3
2.3 The K_{mm} function . . . 4
2.4 The *D* function . . . 5
2.5 The g_{inhom} and K_{inhom} functions . . . 5
2.6 The K_d function . . . 5
2.7 The *M* and *m* functions . . . 7
3 A typology of distance-based methods 8
3.1 Step 1: a number of neighbors $n(x_i, r)$ . . . 8
3.2 Step 2: computing an average number of neighbors . 8
3.3 Step 3: a reference measure . . . 9
3.4 Step 4: the reference value . . . 9
3.5 Last step: null hypothesis . . . 9
4 Discussion 11
4.1 Is there a leading function in economics? . . . 11
4.2 Ways of research . . . 11
4.3 Computation of distance-based methods . . . 12
5 Conclusion 13

Introduction

In the article on "spatial economics" in the New Palgrave Dictionary of Economics, Gilles Duranton wrote *"On the empirical front, a first key challenge is to develop new tools for spatial analysis. With very detailed data becoming available, new tools are needed. Ideally, all the data work should be done in continuous space to avoid border biases and arbitrary spatial units."* (Duranton, 2008). In recent years, economists have made every effort in that direction. Measurement of the spatial concentration of activities is certainly one of the most striking examples and has been considerably renewed in the last decade with the development of distance-based methods (Combes *et al.*, 2008). To present the motivation for the use of distance-based methods briefly, let us say that economists traditionally employ disproportionality methods (terminology used by Bickenbach and Bode, 2008) defined on a *discrete* definition of space. In the latter, the territory being analyzed is divided in several exclusive zones (*e.g.* a country is divided in turn into regions) and the spatial concentration of activities is evaluated at a given level of observation with the Gini (Gini, 1912), the Ellison and Glaeser (Ellison and Glaeser, 1997) or the entropy indices of overall localization (Cutrini, 2009), for example. However, the issues arising from discrete spaces are now well known and linked to the Modifiable Areal Unit Problem – MAUP (Arbia, 1989; Openshaw and Taylor, 1979): the position of the zoning boundaries and level of observation have an impact.[1] This seems sufficiently problematic (Marcon and Puech, 2003) to justify development of spatial concentration indices in order to take more effective account of geography. This encourages the use of distance-based methods which are *continuous* functions of space. Distance-based measures provide information about concentration at *all* scales simultaneously and do not rely on

[1] Some authors underline another problem of indices based on a discrete space: any permutation of zones does not affect the spatial concentration results (see Arbia, 2001, for an illustrative example). However, some investigations have been proposed recently to integrate the spatial position of zones in the discrete-based measures (see Guimarães *et al.*, 2011). The criticism of aspatial discrete space measures does not hold anymore.

zoning. The seminal work by Ripley (1976, 1977) introduced the best known of the existing distance-based methods: the K function. The latter was taken up quickly by field scientists in ecology (see handbooks by Diggle, 1983; Cressie, 1993, for instance) but its use remained incidental in economics (Arbia, 1989; Arbia and Espa, 1996; Barff, 1987; Feser and Sweeney, 2000; Sweeney and Feser, 1998) until the works of Marcon and Puech (2003, 2010) and Duranton and Overman (2002)[2] who introduced an alternative approach.

In this paper, we propose a typology of distance-based methods. There are two main reasons behind our work. First, a great variety of distance-based methods are used by economists today. The varied toolbox provided by these measures may bring some confusion for economists interested in testing a hypothesis rather than a methodology, so a state of the art may be helpful. Second, in this article we provide a unified theoretical framework by showing that all distance-based methods rely on counting the number of neighbors of points, normalizing this number by space or another number of neighbors, averaging the results in the appropriate way and finally normalizing the result. Monte-Carlo simulations of the null hypothesis allow the data to be tested against it and can also solve remaining issues. As a result, if objects (for example plants) attract each other, more neighbors (other plants) will be found around them on average than if they were distributed randomly and independently. In conclusion, these methods are variations on the same framework to gauge spatial concentration. This being the case, this typology can be useful for readers to choose the appropriate distance-based tool to answer their question.

The paper is organized as follows. In the first part, we give a quick presentation of the common framework and basic vocabulary. Then, all the available distance-based measures are introduced. The third part builds a typology of these methods, showing that they follow the same pattern but vary because they assume different theoretical choices. The last part is a discussion of each tool's properties and their relevance to address economic questions.

1. Basic principles

Before presenting distance-based measures in detail, we shall propose a general overview of the framework of these functions.

When studying the location of activities, economists document the spatial distribution of one kind of entity (*points*[3]), for example shops with a given activity. Their aim is to detect phenomena of attraction (also called *aggregation*, *agglomeration*, *localization*), repulsion (*dispersion*) or independence between those entities in a territory. Industries that are spatially concentrated or dispersed are sometimes referred as *diverging industries* (Barlet *et al.*, 2013). Another field of research analyzes on the relations between entities belonging not to one, but to two different groups (*co-localization* phenomenon). All the tools considered in this review identify the spatial structure of the point distribution. Their particularity resides in their analysis of space: they treat space as continuous and not as a collection of predefined zones. They are based on the distance separating pairs of entities, which is why they are called *distance-based methods*. Results are presented as a plot of a function of distance, whose values are meaningful or not, and always compared to an envelope representing the confidence interval of a null hypothesis to be tested. Intratype (or univariate) functions consider a single type of points to address their localization. Intertype (or bivariate) functions are used to characterize co-localization. They are built in the same way. The curve shows the potential attraction or repulsion of two types of points (between two types of shops for example).

In more technical terms, these distance-based methods investigate the spatial structure of point patterns. Their mathematical framework is that of point processes (see the first chapter of Møller and Waagepetersen, 2004, for a rigorous introduction). It is clearly outside the scope of this paper to explore the point process theory, but basic knowledge is required for a good understanding of what follows. Point processes are similar to random variables (the best known are described by their distribution, with parameters) however their output is not a number but a point set in space (often called a *point pattern*). Practical interest is limited to two-dimensional finite spaces, that is to say points on a map. The observed point set is a realization of an underlying point process for which the law is unknown, so its characterization must be non-parametric: some statistics will be used for tests and even to quantify some properties, but identifying the point process will generally remain impossible. These statistics will be defined according to relevant properties of the point process. Their value will be estimated from the data. The first property of interest is intensity, the number of points per unit of space observed in a small area. Intensity is denoted $\lambda(x)$, where x is a point of the sample. If intensity is the same throughout the space, it is denoted λ and the point process is said to be stationary. The probability of finding a point around x is $\lambda(x)\,\mathrm{d}x$. Intensity can be estimated by density $\widehat{\lambda}(x)$, the number of observed points per unit of space. In simple cases, points are just counted. If density changes over the space, *kernel methods* (Diggle, 1985) must be used. They rely on counting points around x, up to a chosen distance called bandwidth, while giving a decreasing weight to the further points. Several types of smoothing exist (Silverman, 1986) but Gaussian kernel smoothing is generally employed in spatial economics (see Duranton and Overman, 2005, among others).

[2] Published as Duranton and Overman (2005).

[3] In the entire article, the term "points" refers to the studied entities (shops, plants etc.) of the sample.

2. Distance-based methods: a brief presentation

2.1 The *g* function

The second-order property of a point pattern characterizes the relation between points: *attraction*, *repulsion* or *independence*. It is defined as the ratio between the joint probability of finding two points in two places x and y, denoted $\lambda(x,y)\,dx\,dy$ and the product of the probabilities of finding each of them. For practical purposes, this property is assumed to depend only on the distance between the points (as it does not change with direction, the point process is said to be *isotropic*). A stationary and isotropic process is called *homogenous*. The second-order property is denoted $g(r) = \lambda(x,y)/[\lambda(x)\,\lambda(y)]$ where r is the distance between two points x and y. If points are distributed independently, $g(r)$ =1. This value corresponds to the benchmark where no interaction between points is detected. If $g(r) > 1$, the probability of finding two points r apart is greater than if they were independent and inversely for $g(r) < 1$. The first case corresponds to attraction, the second to dispersion.

The following idea, "writing the conditional probability of finding a point around y when a point is actually at x", is at the origin of all the measures presented here. It results in the following method:

- Some points are chosen as the reference, for example the general distribution of economic activities. We shall discuss the importance of the reference choice in depth hereinafter.
- Neighbors of reference points are counted at distance r. Neighbors may be of a different type to the reference points or the same. In the former case, intertype (or bivariate) measures are defined, and intratype (univariate) in the latter.
- The ratio of the average density of neighbors at distance r from reference points to the density of this type of points anywhere is the estimator of g, noted $\hat{g}(r)$. It is a location quotient (Florence, 1972).

Estimating $g(r)$ from the data requires a technique to count neighbors at a given distance: kernel functions are used again, but in one dimension this time. A kernel function gives a weight to neighbors at distances around r. The smaller the distance to r, the greater the weight, and the kernel function sums to 1 for all distances. Duranton and Overman (2005, Eq.1) used a Gaussian kernel following Silverman (1986). Its estimator is $k(||x_i - x_j||, r)$ where x_i is the exact location of point i and x_j the spatial position of point j.[4]

The maximum value of $k(||x_i - x_j||, r)$ is reached at the exact distance r between points i and j and decreases according to a Gaussian distribution of standard error h if the distance between points deviates from r. We can define:

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

$g(r)$ is estimated by:[5]

$$\hat{g}(r) = \frac{1}{2\pi r \widehat{\lambda^2}} \sum_i \sum_{j, i \neq j} k\left(\left\|x_i - x_j\right\|, r\right) c(i,j) \quad (2)$$

where $c(i,j)$ is an edge-effect correction depending on both points.[6] When a point is close to the boundary of the area under study, some of its neighbors are not observed because no data is available outside the area. Since the number of neighbors is underestimated for points located near the border of the domain, a statistical bias appears if no edge-effect correction is applied. For example, Marcon and Puech (2003), working on French data, showed that a severe bias existed in the distribution of manufacturing plants. A variety of edge-effect corrections are conceivable and have been proposed for relatively simple shapes (see Goreaud and Pélissier, 1999). However, these corrections are intractable for actual geographical units, such as countries or regions (Law *et al.*, 2009). This has compelled authors to work on simple shapes to analyze the geographic distribution of economic activities (see Rowlingson and Diggle, 1993; Feser and Sweeney, 2000, among others). Ohser (1983) developed complex corrections for any polygonal window. They are implemented in the **spatstat** package for R (Baddeley and Turner, 2005; R Development Core Team, 2014), including the possibility to approximate any window by a raster image, as introduced by Wiegand *et al.* (1999).

2.2 The *K* function

Ripley (1976, 1977) summed g on a range of distances from 0 to r to define the K function: $K(r) = \int_0^r g(\rho) 2\pi\rho d\rho$.

If the point process is homogeneous and independent (*i.e.* it is a homogeneous Poisson process) the spatial pattern is called complete spatial randomness (CSR). In concrete terms, the CSR hypothesis means that all the points of the distribution have the same probability of being located anywhere on the territory: the density is constant everywhere in the domain. A complete spatial random distribution thus corresponds to a random distribution of independent points (no attraction or repulsion between points of the distribution under study). In the case of CSR, Ripley proved that the K function reaches its reference value: $K(r) = \pi r^2$. If $K(r)$ is greater than πr^2, then more points are found within a radius of r from each point. The point process is said to be attractive and spatial concentration is detected. Values of $K(r)$ below πr^2 indicate that points repulse each other up to distance r (dispersion).

[4] In the remainder, x_i will designate reference points, x_j their neighbors.

[5] To avoid any confusion in the paper, estimators will systematically be written with hats.

[6] Note that the unbiased estimator of λ^2 is $n(n-1)/A^2$ (Stoyan and Stoyan, 2000) where n is the number of points and A the surface of the study area.

As $K(r)$ is not easy to plot and πr^2 is not an easy-to-compare reference, Besag (1977) proposed to transform it into $L(r) = \sqrt{K(r)/\pi}$ so that its reference value is r, *i.e.* a straight line. It has often been plotted as $L(r) - r$ in the literature (Pélissier, 1998, for example) making its reference value 0. The notation has gradually shifted and has become $L(r) = \sqrt{K(r)/\pi} - r$ (Goreaud and Pélissier, 1999, for example). $L(r)$ can be interpreted as a distance (Marcon and Puech, 2003): $L(r) = l$ means that as many neighbors are found around reference points up to distance r as would be expected at distance $r + l$ if neighbor points are distributed independently of reference points. We believe that $K(r)/\pi r^2$ is a better normalization because it becomes a location quotient: the density of neighbors around reference points divided by the density of neighbors anywhere. Note that the reference value is 1.

Estimation of K is done by counting neighbors up to r and is defined as:

$$\hat{K}(r) = \frac{1}{A\widehat{\lambda^2}} \sum_i \sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) c(i,j) \tag{3}$$

$\mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right)$ is the indicator function equal to 1 if the distance between x_i and x_j is less than r, 0 else. A is the surface area under study. Like $\hat{g}$, $\hat{K}$ suffers edge effects.

K is the cumulative function of g: it provides information *up to* a given distance while g provides information *at* a distance. Misusing a term from random variable terminology, the latter type of functions are called probability density functions below, although few of them are actually normalized so that they sum to 1.

With K intertype functions, neighbors of a particular type (Lotwick and Silverman, 1982) are counted around points of another type. The null hypothesis may be that points are labeled randomly or that point locations are independent. It must be chosen with care to avoid erroneous results (Arbia *et al.*, 2008; Goreaud and Pélissier, 2003):

- *random labeling* is appropriate when locations are given, types are chosen. More precisely, in this case we assume that locations of points are given and that type-labeling is independent of the position of points. The hypothesis of *random labeling* is accurate for the location of shops in a city.

- *population independence* is the right hypothesis when points can be set anywhere, but not independently of other points of the same type. A good example of this hypothesis is trickier to find. One possible example could be interactions between two types of sellers on a beach (ice creams and sun hats, for instance) where each of them has its own spatial structure.

Empirical applications of Ripley's K and Besag's L functions to assess the geographical concentration of economic activities are limited. Some authors aim to detect the location patterns of subsectors of manufacturing industries (Marcon and Puech, 2003) or services (Ó hUallacháin and Leslie, 2007). Others depict specific location patterns of industrial production by focusing on one characteristic: plants size (Arbia, 1989) or the technology used (Barff, 1987), for example. However, these functions have two main limitations in the field of spatial economics preventing them from becoming wide used. The first one is related to the CSR hypothesis. The constant density benchmark is very strong regarding for evaluating of the spatial distribution of activities and considerably limits the usefulness of the K and L functions in the field of spatial economics. To give some intuitive examples for the location of plants (we will come back to that point in the discussion section), this implies that there are no lakes or other locations where no buildings are permitted etc. The second important limitation is that the number of employees in the plants can not be taken into account: points can not be weighted. From an economic point of view it is hardly convincing if we aim to explain the agglomeration forces at work. This second limitation can be solved, however, by introducing a new function presented in the next subsection.

2.3 The K_{mm} function

The K_{mm} function was introduced by Penttinen *et al.* (Penttinen, 2006; Penttinen *et al.*, 1992). It generalizes Ripley's K function by associating quantitative marks $w(x_i)$ to points that can be used as weights. By considering the spatial distribution of firms, a classical weight for the entities is the number of employees. However, other weights are possible, such as value-added per establishment.

The K_{mm} function can be understood as a K function computed on a data set where every point is associated with a mark (its weight in our case). The mark of the point i is denoted $w(x_i)$ and that of point j is $w(x_j)$. Its estimator is:

$$\begin{aligned}
&\hat{K}_{mm}(r) \\
&= \frac{1}{A\widehat{\lambda^2}\overline{w}^2} \sum_i \sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) w(x_i)\, w(x_j)\, c(i,j) \\
&= \frac{A}{n(n-1)\,\overline{w}^2} \sum_i \sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) w(x_i)\, w(x_j)\, c(i,j)
\end{aligned} \tag{4}$$

$\overline{w}$ is the average point weight.

$\hat{K}_{mm}(r)$ was not normalized by $n(n-1)\overline{w}^2$ by Penttinen *et al.* (1992) who transformed it into $L_{mm}(r) = \sqrt{\frac{K_{mm}(r)}{\pi W^2}}$ where W is the total weight of points. Studying the spatial concentration of plants, Giuliani *et al.* (2014) divided $\hat{K}_{mm}(r)$ by W^2 (we estimated it by $n(n-1)\overline{w}^2$) so that $\hat{K}_{mm}(r)$ has the same properties as $\hat{K}(r)$. We follow them here. The reference value is reached if the points and marks are independent: $K_{mm} = \pi r^2$ and $L_{mm} = 0$.

Note that the derivative of K_{mm}, that is to say the weighted equivalent of g denoted g_{mm}, was first introduced by Stoyan and Ohser (1984, 1985) but not used in empirical applications.

In spatial economics, only one application of the K_{mm} function can yet be found to date in Giuliani *et al.* (2014). In a few words, over the same period of observation (1996-2004), the K_{mm} function detected agglomeration of high and high-medium technology manufacturing firms in the Milan's area (Italy), whereas no significant results appeared for these industries in the area of Turin. The authors advanced theoretical economic arguments to explain the differences. However, no additional statistical test was provided to clear up the results.

2.4 The *D* function

Ripley's K, Besag's L and Penttinen *et al.*'s K_{mm} functions consider space as homogeneous (as defined in section 2.1). To deal with inhomogenous point patterns, Diggle and Chetwynd (1991) introduced the D function, equal to the difference between two K functions: that of the points of interest, called cases, and that of other points, called controls: $D = K_c - K_0$. The authors show that under the null hypothesis $K_c = K_0$. Both also equal the intertype function of cases and controls $K_{c,0}$. When not zero, D cannot be interpreted and is limited to tests. We introduce here an alternate version of D (previously advocated by Arbia *et al.*, 2008), which we will denote D_i:

$$D_i = K_c - K_{c,0} \tag{5}$$

It also equals to 0 under the null hypothesis and can be used exactly like D. Its advantage compared to the D of Diggle and Chetwynd is that it compares two K functions computed around the same points (the cases). Thus, $D_i/\pi r^2$ is the difference between two location quotients: that of the cases around themselves and that of the cases around the controls. Finally, note that the D function is also called the KD function by Waller (2010), certainly due to the use of Ripley's K function in the definition of the D function.

From a statistical point of view, three important limitations of the D function can be given. Firstly, values are not comparable on the same plot of the D function. The excess number of points does not have the same signification at small distances as it does at greater ones (the expected number of points is greater at large distances). Secondly, over the same distances, values of two D plots are not comparable if the controls are not the same. Changing the reference points (the controls) implies a change in the benchmark distribution: comparisons are then unfounded whatever the distance. Thirdly, the D function results from a difference between two K functions and therefore points can not be weighted.

This function was originally used by Diggle and Chetwynd to detect the spatial concentration of rare diseases. Despite the three limitations cited above, numerous applications of the D function can be found in the field of spatial economics. In a pioneering empirical study, using the D function, Sweeney and Feser (1998) showed that plants size matters in the measuring agglomeration of manufacturing firms in North Carolina (medium-sized plants show greater levels of spatial concentration). Marcon and Puech (2003) evaluated the spatial distribution of manufacturing firms in France and provided some comparisons with the results obtained using Ripley's original K function on the same data. Other studies have investigated more sophisticated questions, such as a possible greater degree of spatial concentration for firms that are linked (Feser and Sweeney, 2000) or the existence of differences in the geographic concentration of patents according to the sector considered (Arbia *et al.*, 2008).

2.5 The g_{inhom} and K_{inhom} functions

The K function cannot be estimated from data if the point process is not stationary. Baddeley *et al.* (2000) derived the inhomogenous version of K called K_{inhom}, equal to the integral of g_{inhom} and centered on πr^2 under the assumption of independence of points. It has been little used in economics (but see Arbia *et al.*, 2009, 2012) because it requires the intensity of the point process to be estimated by kernel methods. If the kernel's bandwidth is very small, intensity is highly variable and independence is found, while a wide kernel results in more stationarity and dependence. In other words, the results are highly dependent on the arbitrary choice of the estimation kernel bandwidth (Diggle *et al.*, 2007). If it is not guided by additional knowledge supporting it, results may be arbitrary.

It is estimated by:

$$\hat{K}_{inhom}(r) = \frac{1}{A}\sum_i \sum_{j,i\neq j} \frac{\mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) c(i,j)}{\widehat{\lambda}(x_i)\,\widehat{\lambda}(x_j)} \tag{6}$$

g_{inhom}'s estimator can be found in Law *et al.* (2009):

$$\hat{g}_{inhom}(r) = \frac{1}{2\pi r}\sum_i \sum_{j,i\neq j} \frac{k\left(\left\|x_i - x_j\right\|, r\right) c(i,j)}{\widehat{\lambda}(x_i)\,\widehat{\lambda}(x_j)} \tag{7}$$

In the spirit of Besag's L function, the L_{inhom} function is also proposed Arbia *et al.* (2012) and is estimated by:

$$\hat{L}_{inhom}(r) = \sqrt{\frac{\hat{K}_{inhom}}{\pi}} \tag{8}$$

Despite the great qualities of the K_{inhom} function, its applications are still scarce in economics. To the best of our knowledge, the first paper that introduced this function in our field was by Bonneu (2007). In this paper, he analyzed the distribution of fire department emergencies in the area of Toulouse (France) in 2004. Another application was recently proposed by Arbia *et al.* (2012). They provide an evaluation of the spatial distribution of five sectors of the high-tech industry in 2001 in Milan (Italy). The authors justified the use of the L_{inhom} function for its potential in a dynamic context, although even if they did not deal with space-time analysis in their paper.

2.6 The K_d function

Duranton and Overman (2005)'s K_d is the probability density function to find a point's neighbor at a given distance. It counts and averages the number of point pairs at each distance, and smooths the results to obtain a continuous function that

Table 1. Estimating the number of neighbors

Function	Neighbors around x_i	Observations
$\widehat{K}(r)/(\pi r^2)$	$n(x_i,r)=\sum_{j,i\neq j}\mathbf{1}\left(\left\|x_i-x_j\right\|\leq r\right)c(i,j)$	The number of neighbors is counted and corrected from edge effects.
$\widehat{K}_{mm}(r)/(\pi r^2)$	$n(x_i,r)=\sum_{j,i\neq j}\mathbf{1}\left(\left\|x_i-x_j\right\|\leq r\right)w(x_j)c(i,j)$	As K above, but each neighbor counts for its weight.
$\widehat{K}_{inhom}(r)/(\pi r^2)$	$n(x_i,r)=\sum_{j,i\neq j}\frac{\mathbf{1}\left(\left\|x_i-x_j\right\|\leq r\right)c(i,j)}{\widehat{\lambda}(x_j)}$	As K above, but each neighbor counts for the inverse of the density around it.
$\widehat{g}(r)$	$n(x_i,r)=\sum_{j,i\neq j}k\left(\left\|x_i-x_j\right\|,r\right)c(i,j)$	As K above, but the neighbors are counted at distance r.
$\widehat{g}_{inhom}(r)$	$n(x_i,r)=\sum_{j,i\neq j}\frac{k\left(\left\|x_i-x_j\right\|,r\right)c(i,j)}{\widehat{\lambda}(x_j)}$	As g above, but each neighbor counts for the inverse of the density around it.
$\widehat{K}_d(r)$	$n(x_i,r)=\sum_{j,i\neq j}k\left(\left\|x_i-x_j\right\|,r\right)$	As g above, without edge-effect correction.
$\widehat{K}^{emp}(r)$	$n(x_i,r)=\sum_{j,i\neq j}k\left(\left\|x_i-x_j\right\|,r\right)w(x_j)$	As K_d above, but each neighbor counts for its weight.
$\widehat{M}(r)$	$n(x_i,r)=\sum_{j,i\neq j}\mathbf{1}\left(\left\|x_i-x_j^c\right\|\leq r\right)w\left(x_j^c\right)$	Each neighbor of the type of interest counts for its weight.
$\widehat{m}(r)$	$n(x_i,r)=\sum_{j,i\neq j}k\left(\left\|x_i-x_j^c\right\|,r\right)w\left(x_j^c\right)$	As M above, but the neighbors are counted at distance r .

is normalized to sum to 1. Values of K_d are compared to the confidence interval of the null hypothesis that points are randomly placed on their actual location set. A variant of K_d named K^{emp} (also proposed by Duranton and Overman, 2005) allows points to be weighted, counting employees in firms. Actually, K_d and K^{emp} are the densities of neighbor points and neighbor employees around reference points, according to the definition of their estimators by their authors. Duranton and Overman's function in their definition of $\hat{K}_d(r)$ is actually not the previously defined kernel function but $k\left(\left\|x_i - x_j\right\|, r\right)h$. We have:

$$\hat{K}_d(r) = \frac{1}{n(n-1)} \sum_i \sum_{j,i\neq j} k\left(\left\|x_i - x_j\right\|, r\right) \tag{9}$$

$$\hat{K}^{emp}(r) = \frac{1}{\sum_i \sum_{j,i\neq j} w(x_i)\, w(x_j)} \sum_i \sum_{j,i\neq j} w(x_i)\, w(x_j)\, k\left(\left\|x_i - x_j\right\|, r\right) \tag{10}$$

There is not one single reference value for K_d and K^{emp} under the null hypothesis of independence between point locations. For each distance, the benchmark is obtained by simulations of the null hypothesis: it is the center of the confidence interval.

The K_d function is now considered one of the leading functions in spatial economics. In consequence, Duranton and Overman's methodology has been widely applied in our field since their seminal paper. In this paper, they broadly depicted the advantages of the K_d function on an exhaustive dataset of UK manufacturing plants in 1996 at the four-digit sectoral level. In a later study, Duranton and Overman (2008) paved the way for future research by studying various pertinent economic questions on location patterns: plant entries and exits, affiliated and non-affiliated plants, domestic and foreign plants etc. Then their methodology has been applied to various countries (see Klier and McMillen, 2008, among others) and extended to empirical studies in services (Nakajima *et al.*, 2012; Barlet *et al.*, 2013; Koh and Riedel, 2014). Behrens and Bougna (2015) preferred to use the K_d function to depict change in the spatial distribution of manufacturing activities in Canada from 2001 to 2009. They also tested the importance of the level of sectoral aggregation as Fratesi (2008) did for the distribution of the pharmaceutical and optical-photographic sectors in Great-Britain. Finally, Marcon and Puech (2010, 2015) provided theoretical and empirical comparisons of the K_d results with other distance-based methods.

Behrens and Bougna (2015) introduced the cumulative versions of the weighted and unweighted K_d function (*i.e.* $\int_0^r K_d(r)\,\mathrm{d}r$). This development seems to be motivated by the difficulties of interpreting K_d results. These new but unnamed cumulative functions of K_d provide the proportion of plant pairs located less than a distance r apart. Some comparisons with the K_d function or Ellison and Glaeser's index (1997) were given. The authors did not provide the confidence interval of the null hypothesis, which could be calculated the same way as that of K_d.

2.7 The *M* and *m* functions

Marcon and Puech's (2010) M function is a cumulative function that gives the relative frequency of neighbors of a chosen type (denoted x_j^c) up to each distance, compared to the same ratio in the whole area under study. Its estimator is:

$$\hat{M}(r) = \sum_i \frac{\sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j^c\right\| \leq r\right) w\left(x_j^c\right)}{\sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) w(x_j)} \Big/ \sum_i \frac{W_c - w(x_i)}{W - w(x_i)} \tag{11}$$

W_c is the total weight of the points x_j^c. The denominator is slightly different in the intertype function, $\sum_i \{W_c/[W - w(x_i)]\}$, avoiding a small bias. For any distance, the reference value is thus 1.[7]

Some empirical applications of the M function can be found in economics. For instance Jensen and Michel (2011) developed the unweighted version of the M function to gauge the spatial pattern of shops in the Lyon area in France. Marcon and Puech (2015) compare the empirical results obtained with the M function to other distance-based methods (D and K_d) to show the limitation of each measure. Finally, although Marcon and Puech (2010) showed that M and K_d are complements rather than substitutes, applications using the M function in economics are undoubtedly less numerous than those using K_d. However, it is interesting to note that the M function has been rapidly transferred to other scientific fields, such as geography (Deurloo and De Vos, 2008), ecology (Marcon *et al.*, 2012), biology (Fernandez-Gonzalez *et al.*, 2005) or seismology (Nissi *et al.*, 2013).

In a recent paper Lang *et al.* (2015) proposed the m function: a density function version of M. By keeping the same notations, the estimator of m is:

$$\hat{m}(r) = \sum_i \frac{\sum_{j,i\neq j} k\left(\left\|x_i - x_j^c\right\|, r\right) w\left(x_j^c\right)}{\sum_{j,i\neq j} k\left(\left\|x_i - x_j\right\|, r\right) w(x_j)} \Big/ \sum_i \frac{W_c - w(x_i)}{W - w(x_i)} \tag{12}$$

As in Duranton and Overman (2005), a Gaussian kernel of optimal bandwidth was applied according to Silverman (1986). The numerator is the *local ratio*: the relative weight of neighbors of interest at distance r from all reference points and averaged over all reference points. The denominator corresponds to the *global ratio*: it is defined in the same way as the local ratio, but over the whole data set. 1 is the reference value for any distance r and the generation of a confidence interval of the null hypothesis with Monte-Carlo simulations provide the significance of the m estimates.

[7] In extreme cases, if the industrial concentration (in the sense of Ellison and Glaeser, 1997) is too high, for each radius r, the reference value would be the center of the confidence interval.

Table 2. Average number of neighbors

Function	Average number of neighbors	Observations
$\widehat{K}(r)/(\pi r^2)$ $\widehat{g}(r)$ $\widehat{K}_d(r)$ $\widehat{M}(r)$ $\widehat{m}(r)$	$\overline{n}(r) = \frac{1}{n}\sum_i n(x_i, r)$	The number of neighbors around each point is not weighted.
$\widehat{K}_{inhom}(r)/(\pi r^2)$ $\widehat{g}_{inhom}(r)$	$\overline{n}(r) = \sum_i \frac{n(x_i, r)}{\widehat{\lambda}(x_i)}$	The weight of each reference point is the inverse of the intensity of the point process around it.
$\widehat{K}_{mm}(r)/(\pi r^2)$ $\widehat{K}^{emp}(r)$	$\overline{n}(r) = \frac{1}{n\overline{w}}\sum_i w(x_i)\, n(x_i, r)$	The average is weighted ($\overline{w}$ is the average weight).

Note the similarities between m and the cumulative M function, but that the local ratio is defined *at* distance r and not *up to* it for the m function.

3. A typology of distance-based methods

In what follows, we shall prove that all these functions can be built empirically following the same five steps. First, neighbors are counted around each point *at* or *within* a distance r, sometimes weights are summed instead. Second, an average number of neighbors $\overline{n}(r)$ is calculated. Third, $\overline{n}(r)$ is divided by a reference measure $z(r)$. In accordance with the typology of Brülhart and Traeger (2005) we shall use the following vocabulary:

- *Topographic measures* use space as their reference: the number of neighbors is divided by the area of a ring or a disk ($2\pi r\,\mathrm{d}r$ or πr^2).
- *Relative measures* divide the number of neighbors of interest by that of all neighbors.
- *Absolute measures* do not have any reference values.

Fourth, $\overline{n}(r)/z(r)$ is compared to the value it has on the whole domain, $\overline{n}_0/z_0$. Fifth and last, the significance of the values of the functions at several distances is generally tested against a null hypothesis by Monte-Carlo simulations of the appropriate counterfactual. These five steps are detailed below.

3.1 Step 1: a number of neighbors $n(x_i, r)$

The first step consists in counting neighbors of each point, at distance r (*on* the circle of radius r) or up to distance r (*in* the circle of radius r). The first option defines density functions: g, g_{inhom}, K^{emp}, K_d and m. The second defines cumulative functions: K, K_{inhom}, K_{mm}, D_i, the cumulative of K^{emp}, the cumulative of K_d, and M. M, K^{emp} and K_{mm} attribute a weight to points, such as the plants' number of employees. This raw number of neighbors of a point x_i at a distance r or up to r is denoted $n(x_i, r)$. g_{inhom} and K_{inhom} do not just count 1 for each point but give them a weight that is inversely proportional to the local density of points: more neighbors are expected where more points are located and these portions of space must not be overweighted.

Table 1 summarizes the way neighbors are counted around reference points for each function. Reference points (circle centers) are denoted x_i, their neighbors are x_j. The point types may be identical or not, defining intertype functions in the latter case. By construction, M focuses on one special type of neighbor points denoted x_j^c and compares their distribution to that of all neighbors denoted x_j. $w(x_j)$ is the weight of point x_j. $\widehat{\lambda}(x_j)$ is the density of points around x_j. It is the estimator of $\lambda(x_j)$, the intensity of the point process. As we noted, we use hats in equations for estimators to avoid any confusion: K, g have a mathematical definition relying on the point process they are used to characterize and are estimated from the data. We also define K_d and M in this way in this paper. $k\left(\left\|x_i - x_j\right\|, r\right)$ is some kernel function able to evaluate the number of neighbors at distance r. $c(i, j)$ is some edge-effect correction depending on both points.

3.2 Step 2: computing an average number of neighbors

The value obtained around each point following (table 1) is then averaged for all reference points. In topographic, inhomogenous measures, the weight of each point is inversely proportional to the intensity of the process around it so that space is sampled uniformly. All points have the same weight in K_d and M.

Table 2 summarizes the way the average number of neighbors is calculated. $\overline{n}(r)$ is the average number of neighbors.

n is the total number of reference points (the centers of circles).

3.3 Step 3: a reference measure

These numbers of neighbors are then divided by a reference measure $z(r)$, table 3. This step determines the nature of the measure:

- if $z(r)$ is a measure of space, the function is topographic,
- if it is a number of neighbors, the function is relative,
- if there is no reference, the function is absolute.

3.4 Step 4: the reference value

Normalization comes next (table 4). $\overline{n}(r)/z(r)$ is divided by its reference value $\overline{n}_0/z_0$. The latter can be understood as the value of the former with r large enough for all points to be neighbors of each other. Then, possible spatial structure does not matter: $\overline{n}_0$ counts all points except the center of the circle, z_0 is the total area (topographic measures) or the total number of points (relative measures) or 1 (absolute measures).

All measures in table 4 are location quotients except for K_d which is an absolute measure.

3.5 Last step: null hypothesis

We have obtained a value for each function at or within a distance r. Computation can be repeated for several values of r to get each concentration measure as a function of distance.

How can the significance of the results be tested? In the vast majority of empirical studies, authors use Monte-Carlo simulations to decide whether or not values obtained by the previous functions reject the null hypothesis tested.

Practically, thanks to Monte-Carlo simulations, a confidence interval of the results is proposed. In a general manner, Monte-Carlo simulations provide many values of the simulated function for each value of r. A proportion of them is eliminated according to the accepted risk level (often the greatest and the smallest 2.5% in order to obtain a 95% confidence interval). The remaining values constitute the local (*i.e.* at r) confidence interval of the null hypothesis. To test the significance of the results, for every value of r the observed value of the function is compared to the local confidence interval. Lagache *et al.* (2013) developed a method to calculate the quantiles of $K(r)$ under the null hypotheses of CSR analytically, without any simulation.

Duranton and Overman (2005) noted that repeating the same local test for all values of r is not satisfactory as the resulting local confidence interval would be restrictive enough. This is confirmed by the findings of Loosmore and Ford (2006) who proved the inadequacy of the local test. Duranton and Overman (2005) proposed a way to build a global (more conservative) test, followed by Marcon and Puech (2010). This global test has little mathematical support. Barlet *et al.* (2013) showed empirically that it is still not restrictive enough when the number of simulations is too small relatively to the number of points: its confidence level is overestimated so K_d detects localization where there may not be any. The solution is to increase the number of simulations.

For every intratype or intertype function, technical explanations of the construction of the associated confidence interval are summarized in table 5.

It can be noted that K_d does not contain any reference to the overall distribution of points, as a relative measure should. Comparing K_d to its simulated values (null hypothesis) provides a concentration test since the departure from randomness is detected from all occupied sites observed in the real distribution. Quite similarly, Kosfeld *et al.* (2011) used K in inhomogenous space: to characterize the spatial structure of an industry sector, they drew point sets of the same size among the actual locations of all industries to build a confidence envelope of the null hypothesis that the sector under study is distributed like all industries. The expectation of the value K_d of the simulated distributions is simply the value of K_d calculated on the whole point set since points of the sector under study are redistributed on all actual locations with the same probability.

M and m allow the points to be given a weight, typically a number of employees. But it does not take into account the possible influence of the structure of point weights. However, the necessity of controlling for industrial concentration was highlighted by Ellison and Glaeser (1997). The answer is the following. The simulations are done with the actual weights, so the weight structure is controlled for by the confidence interval of the null hypothesis, if not numerically.

The null hypothesis of K_{mm} varies. Penttinen *et al.* (1992) inferred the unmarked point process model and used it to simulate points with a random permutation of marks. The null hypothesis was the independence of marks, given the point set structure. Practically, K_{mm} can be used when the unmarked point process is a homogenous Poisson, but with great difficultly in other situations. Giuliani *et al.* (2014) integrated both point locations and mark structure in a single model of localization, whose parameters cannot be inferred yet.

K_{inhom} was employed by Diggle *et al.* (2007) and Arbia *et al.* (2012) in a case-control design: a number of points, the controls, were used to estimate the point process intensity (assuming they are approximately independently distributed), while the different pattern of cases was attributed to dependence. This approach is quite similar to that of M where all points weight 1: estimating $\widehat{\lambda}(x_j)$ in table 1 for K_{inhom} with a simple box kernel with bandwidth r is not different from $z(r)$ in table 3 for M. The main difference (normalization apart) is that all reference points have the same weight in M, while each piece of space has in K_{inhom} (table 2).

Goodness-of-fit (GoF) tests are an alternative explored by Heinrich (1991). They consist in calculating the discrepancy between an estimated function for the real data and its expectation under the null hypothesis: the value obtained is compared to its distribution under the null hypothesis. In the general case, neither the expectation nor the quantiles of the

Table 3. Reference measure

Function	Reference measure	Observations
$\widehat{K}(r)/(\pi r^2)$ $\widehat{K}_{inhom}(r)/(\pi r^2)$ $\widehat{K}_{mm}(r)/(\pi r^2)$	$z(r) = \pi r^2$	The reference measure is the area of a circle
$\widehat{g}(r)$ $\widehat{g}_{inhom}(r)$	$z(r) = 2\pi r$	As K above, but the measure is the length of a ring.
$\widehat{K}_d(r)$ $\widehat{K}^{emp}(r)$	$z(r) = 1$	K_d is not compared to anything. It is an absolute measure.
$\widehat{M}(r)$	$z(r) = \sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) w(x_j)$	The number of neighbors of the type of interest is compared to the number of neighbors of all types.
$\widehat{m}(r)$	$z(r) = \sum_{j,i\neq j} k\left(\left\|x_i - x_j\right\|, r\right) w(x_j)$	As m above but the surface of reference is a ring.

Table 4. Reference value

Function	Reference value	Observations
$\widehat{K}(r)/(\pi r^2)$ $\widehat{g}(r)$	$\frac{\overline{n}_0}{m_0} = \frac{n-1}{A}$	The reference value is the intensity of the point process, evaluated by the total number of points minus 1 (the circle center) divided by the area of the window.
$\widehat{K}_{inhom}(r)/(\pi r^2)$ $\widehat{g}_{inhom}(r)$	$\frac{\overline{n}_0}{m_0} = 1$	Local weights used in inhomogenous measures are such that the reference value is 1.
$\widehat{K}_{mm}(r)/(\pi r^2)$	$\frac{\overline{n}_0}{m_0} = \frac{(n-1)\overline{w}}{A}$	The reference value is the density multiplied by the average point weight.
$\widehat{K}_d(r)$	$\frac{\overline{n}_0}{m_0} = n-1$	Absolute measure. $z_0 = 1$. $\overline{n}_0$ is what $\overline{n}(r)$ would be if the indicator function were always true.
$\widehat{K}^{emp}(r)$	$\frac{\overline{n}_0}{m_0} = \frac{\sum_i \sum_{j,i\neq j} w(x_i) w(x_j)}{\sum_i w(x_i)}$	As K_d above.
$\widehat{M}(r)$ $\widehat{m}(r)$	$\frac{\overline{n}_0}{m_0} = \frac{1}{n}\sum_i \frac{W_c - w(x_i)}{W - w(x_i)}$	The reference value is calculated as $\overline{n}(r)/z(r)$ with r large enough for all points to be neighbors to each other. W_c is is the total weight of points belonging to the neighbor type, W the total weight of all points.

Table 5. Simulation of the null hypothesis

Function	Null Hypothesis
$\widehat{K}(r)/(\pi r^2)$ $\widehat{g}(r)$	A homogenous Poisson process of intensity $\widehat{\lambda}$ (estimated from the data).
$\widehat{K}_{inhom}(r)/(\pi r^2)$ $\widehat{g}_{inhom}(r)$	An inhomogenous Poisson process of intensity $\widehat{\lambda}(x)$ (estimated from the data).
$\widehat{K}_{mm}(r)/(\pi r^2)$	Generally, a homogenous Poisson process with random labeling.
$\widehat{K}_d(r)$ $\widehat{K}^{emp}(r)$	Points are redistributed across actual locations.
$\widehat{M}(r)$ $\widehat{m}(r)$	Points are redistributed across actual locations.

distribution are known so they must be obtained by Monte-Carlo simulations. Loosmore and Ford (2006) published a GoF test for the K function in ecology. Marcon *et al.* (2012) proposed an application of the GoF test for the M function in ecology too. Barlet *et al.* (2013) developed a GoF test for the K_d function and prove it to be unbiased.

Global, analytical tests (providing a p-value to reject the null hypothesis erroneously, calculated from the data without simulations) are scarce. Jensen and Michel (2011) provided one for K_d based on the exact calculation of the variance. Lang and Marcon (2013) developed a test for K against CSR in a square domain, Marcon *et al.* (2013) extended it to a rectangle domain and applied it to ecological data. A non-trivial advantage of analytical tests is saving the time involved in calculating of confidence intervals, which is quite long for large datasets.

4. Discussion

The aim of the previous section was to propose a common framework for understanding the statistic construction of the most popular distance-based methods. In this section, we shall provide a discussion of those functions with the objective of addressing economic questions. The nature of the spatial concentration (topographic, relative or absolute) and the type of the function (cumulative or density functions) are essential.

4.1 Is there a leading function in economics?

To begin with, as we underlined in sub-section 2.6, Duranton and Overman's K_d function is certainly now the most employed distance-based method in the economic literature. The reason seems simple: K_d respects a maximum of the properties required for any statistical measure that aiming to detect spatial structures in economics. In particular in their seminal paper Duranton and Overman listed five important criteria (Duranton and Overman, 2005, p.1079): (1) results should be comparable across industries, (2) the global agglomeration pattern should be taken into account, (3) the measure should control for industrial concentration in the sense of Ellison and Glaeser (1997), (4) results should be unbiased across geographic scales (this is related to the MAUP issues) and lastly (5) a test of the significance of the results should be provided.

K_d respects all of them. An overview of the typology provided in this paper shows that only three other functions K_{inhom}, M and m respect all five criteria above. They can therefore also be good candidates to analyze spatial concentration in spatial economics. However, researchers may be interested not only in detecting the spatial structures but also in quantifying agglomeration or dispersion results. Up to now, only a small number of the distance-based methods listed previously can be mobilized to reach that objective. If the question is how many times more neighbors are found *at* the chosen distance then only g, g_{inhom} and m are of interest. If the question is how many times more neighbors are found *up to* the chosen distance then only $K/\pi r^2$, M, K_{inhom} and K_{mm} can be used.

The debate is still open as to the "best" candidate for detecting spatial structures in economics as noticed by Gibbons *et al.* (2015). This dilemma will certainly be resolved by providing a strong theoretical background to spatial concentration measures in continuous space. However, the literature is not yet stabilized on the criteria that any distance-based method should respect (Combes and Overman, 2004; Bonneu and Thomas-Agnan, 2014).

4.2 Ways of research

Broadly speaking, methods treating space as continuous are still in their "infancy" in economics in that they have only been developed over the past decade. As Combes and Overman (2004) stressed, improvements in that direction are needed because no distance-based methods include a theoretical model to explain agglomeration.

We believe that two points are essential in this debate and shall guide further developments: the nature of the spatial concentration (topographic, relative or absolute) and the type of function (cumulative or density functions). Let us explain these two key-points.

First, concerning the type of spatial concentration under study, it is rare that the authors motivate the nature of the spatial concentration that they are trying to grasp, although, some authors, such as Brülhart and Traeger (2005), have already shown that this is of a great importance. As we underlined, distance-based methods that integrate physical space as their benchmark are *topographic measures*. Those which refer to another variable are called *relative measures*. If no benchmark is used, distance-based measures are called *absolute measures*. Researchers should carefully motivate the type of concentration that they want to use the appropriate function. For example, if the basic question is "how many manufactur-

Table 6. Choice of the appropriate function to describe a point pattern structure

Function choice	Topographic, homogenous	Topographic, inhomogenous	Absolute	Relative
Density functions	g	g_{inhom}	K_d	m
			K^{emp}	
Cumulative functions	K	K_{inhom}	Cumulative of K_d	M
	K_{mm}	D_i	Cumulative of K^{emp}	Case-control K_{inhom}

ing plants are there?" then it refers to absolute concentration. It the question is "what is the proportion of the manufacturing plants of sector of interest in the neighborhood?" then relative concentration methods are appropriate. Finally if the question is "how many manufacturing plants are located per unit area?" this refers to topographic concentration. To go further, in table 6 we provide the tools' properties so that researchers can use the appropriate function for their analysis. In each case, a significant departure from randomness can be detected but the analysis is not the same, so results on the underlying spatial structure obviously will not converge. For example, K_d is an absolute measure: comparing it to its null hypothesis confidence envelope allows the independence of locations to be tested. This is very similar to the way the D function works, comparing a number of cases and a number of controls around each point at the same scale. D and K_d actually have very similar properties but their results could be very different from the estimates provided by the M or m functions which evaluate relative concentration. Let us consider a concrete example to make things clearer. Marcon and Puech (2015) showed that in the area of Lyon (France), gas stations are agglomerated in relative terms (because they are generally located in low density shopping areas) but these retail shops exhibit an absolute and a topographic dispersion (because under the null hypothesis of a random distribution, gas stations are located in areas with a higher shopping density than they are in the observed distribution). As a result, improving the theoretical background of the function should make the choice of distance-based measure easier.

Second, a recent open debate concerns the respective advantages of probability density functions and cumulative measures (Wiegand and Moloney, 2004; Law *et al.*, 2009; Marcon and Puech, 2010). Arguments are not repeated here (the reader should refer to the cited papers) because the choice depends on both the question analyzed and the availability of data. Marcon and Puech (2010) have shown that K_d or M are useful to evaluate the spatial distribution of activities and depending on the question raised, one may give clearer results than the other. K_d provides more precise estimations than M for gauging the local density of activities. Thus, K_d should be preferred if the objective is the evaluation of local densities. M assesses the global effect of the superposition of spatial structures better. As a consequence, if the question is "up to which distance do externalities matter?" then a cumulative function is more appropriate. A probability density function will answer "do externalities matter at a given distance?" in a more satisfactory way. As a consequence, M and K_d seem to be complements rather than substitutes.

To conclude, the next step is to provide a complete theoretical background for distance-based methods. The statistical literature (Møller and Waagepetersen, 2004) mainly deals with homogeneous point processes, used for topographic measures. Many theoretical results exist, such as g and K expectations for many processes (Diggle, 1983; Illian *et al.*, 2008). For the moment, topographic measures for inhomogeneous space and relative measures still lack statistical support. This could be explained by the long tradition in various scientific domains of employing (and thus developing) functions based on homogeneous point processes.[8] Since the use of distance-based measures is more recent in economics, theoretical developments are just beginning. To give some examples in spatial economics, note that the widely-employed K_d was built on empirical foundations: its values are not estimators of a theoretical statistic. Further theoretical investigations are needed to improve the statistical background of the distance-based measures in a spatial economic framework. First studies are being done in that direction. The growing interest in the K_d measure is very promising (Ellison *et al.*, 2010; Alfaro and Chen, 2014) and theoretical developments are now proposed (Barlet *et al.*, 2013). It is also the case of the M function: it has been proved recently that M is the generalization of Ripley's function with inhomogeneous point processes (Marcon *et al.*, 2012).

4.3 Computation of distance-based methods

One important point concerns the computation of all these functions and their respective confidence intervals, which can be very time-consuming (Scholl and Brenner, 2013). An easy-to-use package called **dbmss** for "distance-based measures of spatial structures" (Marcon *et al.*, 2014) is now developed,

[8] This is the case for example in forestry where relative measures have not been employed before Marcon *et al.* (2012).

based on the R **spatstat** package.[9] This package allows the simple computation of all distance-based measures presented in this article.

5. Conclusion

A decade ago, disproportionality methods such as the Gini or Ellison and Glaeser indices were classical tools for economists. Quite logically, methods were then developed to take advantage of the knowledge of the exact position of objects and solve issues linked to the Modifiable Areal Unit Problem (Openshaw and Taylor, 1979). The first were statistics based on the distance of the nearest neighbor of points, after Clark and Evans (1954). They have been outdated by the distance-based measures of concentration reviewed in this paper because the latter use the information provided by all points less than r apart from each reference point instead of just one.

When geo-referenced data are available, distance-based measures of concentration are a complete set of tools to test data against null hypotheses of independence (to show aggregation or repulsion) and, for some of them, to quantify the phenomena. We have explained in this article (table 6) which tool to use according to the underlying framework (topographic, absolute or relative). Topographic measures are widely used and updated by ecologists in handbooks (Fortin and Dale, 2005; Illian *et al.*, 2008) which ignore relative measures. Economists mainly use absolute and relative measures to take into account of the overall distribution of economic activities. Several economists (among others Combes *et al.*, 2008) clearly state that applications of distance-based methods should now be given preference by researchers. The problem of the availability of geo-referenced economic data or easy-to-use programs to implement these functions are short-term issues (Overman, 2008), although they have not been solved yet (Marcon *et al.*, 2014). However, relating these descriptive tools to economic theory is the real challenge (Alfaro and Chen, 2014), following the way opened by Ellison *et al.* (2010).

Acknowledgments

We thank participants at the 61st Congress of the French Economic Association (Paris), Hotelling Seminar (Université de Paris Sud / ENS Cachan) and the 12th International Workshop Spatial Econometrics and Statistics (Orléans, France). The second author gratefully acknowledges financial support from the LET (Université de Lyon, CNRS, ENTPE), IUT de Sceaux and AAP Attractivité 2014 (Université de Paris-Sud). This work has benefited from an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref.ANR-10-LABX-0025).

[9] Available at the following address: http://cran.r-project.org/web/packages/dbmss/

References

Alfaro L, Chen MX (2014). "The global agglomeration of multinational firms." *Journal of International Economics*, **94**(2), 263–276.

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

Arbia G (2001). "The Role of Spatial Effects in the Empirical Analysis of Regional Concentration." *Journal of Geographical Systems*, **3**(3), 271–281.

Arbia G, Copetti M, Diggle P, Fratesi U, Senn L (2009). "Modelling Individual Behaviour of Firms in the Study of Spatial Concentration.", In U Fratesi, L Senn (eds.), *Growth and Innovation of Competitive Regions*, pp. 297–327. Advances in Spatial Science, second edition. Springer, Berlin.

Arbia G, Espa G (1996). *Statistica economica territoriale*. Cedam, Padua.

Arbia G, Espa G, Giuliani D, Mazzitelli A (2012). "Clusters of firms in an inhomogeneous space: The high-tech industries in Milan." *Economic Modelling*, **29**(1), 3–11.

Arbia G, Espa G, Quah D (2008). "A class of spatial econometric methods in the empirical analysis of clusters of firms in the space." *Empirical Economics*, **34**(1), 81–103.

Baddeley AJ, Møller J, Waagepetersen RP (2000). "Non- and semi-parametric estimation of interaction in inhomogeneous point patterns." *Statistica Neerlandica*, **54**(3), 329–350.

Baddeley AJ, Turner R (2005). "Spatstat: an R package for analyzing spatial point patterns." *Journal of Statistical Software*, **12**(6), 1–42.

Barff RA (1987). "Industrial Clustering and the Organization of Production: A Point Pattern Analysis of Manufacturing in Cincinnati, Ohio." *Annals of the Association of American Geographers*, **77**(1), 89–103.

Barlet M, Briant A, Crusson L (2013). "Location patterns of service industries in France: A distance-based approach." *Regional Science and Urban Economics*, **43**(2), 338–351.

Behrens K, Bougna T (2015). "An anatomy of the geographical concentration of Canadian manufacturing industries." *Regional Science and Urban Economics*, **51**, 47–69.

Besag JE (1977). "Comments on Ripley's paper." *Journal of the Royal Statistical Society*, **B 39**(2), 193–195.

Bickenbach F, Bode E (2008). "Disproportionality Measures of Concentration, Specialization, and Localization." *International Regional Science Review*, **31**(4), 359–388.

Bonneu F (2007). "Exploring and Modeling Fire Department Emergencies with a Spatio-Temporal Marked Point Process." *Case Studies in Business, Industry and Government Statistics*, **1**(2), 139–152.

Bonneu F, Thomas-Agnan C (2014). "Measuring and testing spatial mass concentration with micrio-geographic data." In *TSE Working Paper*, No 14-474.

Brülhart M, Traeger R (2005). "An Account of Geographic Concentration Patterns in Europe." *Regional Science and Urban Economics*, **35**(6), 597–624.

Clark PJ, Evans FC (1954). "Distance to nearest neighbor as a measure of spatial relationships in populations." *Ecology*, **35**(4), 445–453.

Combes PP, Mayer T, Thisse JF (2008). *Economic Geography, The Integration of Regions and Nations*. Princeton University Press, Princeton.

Combes PP, Overman HG (2004). "The spatial distribution of economic activities in the European Union." In JV Henderson, JF Thisse (eds.), *Handbook of Urban and Regional Economics*, volume 4, chapter 64, pp. 2845–2909. Elsevier. North Holland, Amsterdam.

Cressie NA (1993). *Statistics for spatial data*. John Wiley & Sons, New York.

Cutrini E (2009). "Using entropy measures to disentangle regional from national localization patterns." *Regional Science and Urban Economics*, **39**(2), 243–250.

Deurloo MC, De Vos S (2008). "Measuring segregation at the micro level: an application of the M measure to multi-ethnic residential neighbourhoods in Amsterdam." *Tijdschrift voor economische en sociale geografie*, **99**(3), 329–347.

Diggle PJ (1983). *Statistical analysis of spatial point patterns*. Academic Press, London.

Diggle PJ (1985). "A Kernel Method for Smoothing Point Process Data." *Applied Statistics*, **34**(2), 138–147.

Diggle PJ, Chetwynd AG (1991). "Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations." *Biometrics*, **47**(3), 1155–1163.

Diggle PJ, Gomez-Rubio V, Brown PE, Chetwynd AG, Gooding S (2007). "Second-order analysis of inhomogeneous spatial point processes using case-control data." *Biometrics*, **63**(2), 550–557.

Duranton G (2008). "Spatial Economics.", In SN Durlauf, LE Blume (eds.), *The New Palgrave Dictionary of Economics*. Palgrave Macmillan.

Duranton G, Overman HG (2002). "Testing for Localization Using Micro-Geographic Data." *Discussion Paper 3379*, CEPR.

Duranton G, Overman HG (2005). "Testing for Localisation Using Micro-Geographic Data." *Review of Economic Studies*, **72**(4), 1077–1106.

Duranton G, Overman HG (2008). "Exploring the Detailed Location Patterns of UK Manufacturing Industries using Microgeographic Data." *Journal of Regional Science*, **48**(1), 213–243.

Ellison G, Glaeser EL (1997). "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach." *Journal of Political Economy*, **105**(5), 889–927.

Ellison G, Glaeser EL, Kerr WR (2010). "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns." *The American Economic Review*, **100**(3), 1195–1213.

Fernandez-Gonzalez R, Barcellos-Hoff MH, de Solorzano CO (2005). "A Tool for the Quantitative Spatial Analysis of Complex Cellular Systems." *IEEE Transactions on Image Processing*, **14**(9), 1300–1313.

Feser EJ, Sweeney SH (2000). "A test for the coincident economic and spatial clustering of business enterprises." *Journal of Geographical Systems*, **2**(4), 349–373.

Florence PS (1972). *The Logic of British and American Industry: A Realistic Analysis of Economic Structure and Government*. 3rd edition. Routledge & Kegan Paul, London.

Fortin MJ, Dale MRT (2005). *Spatial Analysis. A guide for ecologists.* Cambridge University Press, Cambridge.

Fratesi U (2008). "Issues in the measurement of localization." *Environment and Planning A*, **40**(3), 733–758.

Gibbons S, Overman HG, Pattachini E (2015). "Spatial methods." In G Duranton, J Henderson, WC Strange (eds.), *Handbook of Urban and Regional Economics*, volume 5A, chapter 3, pp. 115–168. Elsevier. North Holland, Amsterdam.

Gini C (1912). *Variabilità e mutabilità*, volume 3. Università di Cagliari.

Giuliani D, Arbia G, Espa G (2014). "Weighting Ripley's K-Function to Account for the Firm Dimension in the Analysis of Spatial Concentration." *International Regional Science Review*, **37**(3), 251–272.

Goreaud F, Pélissier R (1999). "On explicit formulas of edge-effect correction for Ripley's K-function." *Journal of Vegetation Science*, **10**(3), 433–438.

Goreaud F, Pélissier R (2003). "Avoiding misinterpretation of biotic interactions with the intertype K_{12} function: population independence vs random labelling hypotheses." *Journal of Vegetation Science*, **14**(5), 681–692.

Guimarães P, Figueiredo O, Woodward D (2011). "Accounting for neighboring effects in measures of spatial concentration." *Journal of Regional Science*, **51**(4), 678–693.

Heinrich L (1991). "Goodness-of-fit tests for the second moment function of a stationary multidimensional poisson process." *Statistics: A Journal of Theoretical and Applied Statistics*, **22**(2), 245 – 268.

Illian J, Penttinen A, Stoyan H, Stoyan D (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Statistics in Practice. Wiley-Interscience, Chichester.

Jensen P, Michel J (2011). "Measuring spatial dispersion: exact results on the variance of random spatial distributions." *The Annals of Regional Science*, **47**(1), 81–110.

Klier T, McMillen DP (2008). "Evolving agglomeration in the U.S. auto supplier industry." *Journal of Regional Science*, **48**(1), 245–267.

Koh HJ, Riedel N (2014). "Assessing the Localization Pattern of German Manufacturing and Service Industries: A Distance-based Approach." *Regional Studies*, **48**(5), 823–843.

Kosfeld R, Eckey HF, Lauridsen J (2011). "Spatial point pattern analysis and industry concentration." *The Annals of Regional Science*, **47**(2), 311–328.

Lagache T, Lang G, Sauvonnet N, Olivo-Marin JC (2013). "Analysis of the Spatial Organization of Molecules with Robust Statistics." *Plos One*, **8**(12), e80914.

Lang G, Marcon E (2013). "Testing randomness of spatial point patterns with the Ripley statistic." *ESAIM: Probability and Statistics*, **17**, 767–788.

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

Law R, Illian J, Burslem D, Gratzer G, Gunatilleke CVS, Gunatilleke I (2009). "Ecological information from spatial patterns of plants: insights from point process theory." *Journal of Ecology*, **97**(4), 616–628.

Loosmore NB, Ford ED (2006). "Statistical inference using the G or K point pattern spatial statistics." *Ecology*, **87**(8), 1925–1931.

Lotwick HW, Silverman BW (1982). "Methods for Analysing Spatial Processes of Several Types of Points." *Journal of the Royal Statistical Society*, **44**(3), 406–413.

Marcon E, Puech F (2003). "Evaluating the Geographic Concentration of Industries Using Distance-Based Methods." *Journal of Economic Geography*, **3**(4), 409–428.

Marcon E, Puech F (2010). "Measures of the Geographic Concentration of Industries: Improving Distance-Based Methods." *Journal of Economic Geography*, **10**(5), 745–762.

Marcon E, Puech F (2015). "Mesures de la concentration spatiale en espace continu : théorie et applications." *Economie et Statistique*, **474**, 105–131.

Marcon E, Puech F, Traissac S (2012). "Characterizing the relative spatial structure of point patterns." *International Journal of Ecology*, **2012**(Article ID 619281), 11.

Marcon E, Traissac S, Lang G (2013). "A Statistical Test for Ripley's Function Rejection of Poisson Null Hypothesis." *ISRN Ecology*, **2013**(Article ID 753475), 9.

Marcon E, Traissac S, Puech F, Lang G (2014). "Tools to Characterize Point Patterns: dbmss for R." *Journal of Statistical Software*, **in revision**.

Møller J, Waagepetersen RP (2004). *Statistical Inference and Simulation for Spatial Point Processes*, volume 100 of *Monographs on Statistics and Applies Probabilities*. Chapman and Hall.

Nakajima K, Saito YU, Uesugi I (2012). "Measuring economic localization: Evidence from Japanese firm-level data." *Journal of the Japanese and International Economies*, **26**(2), 201–220.

Nissi E, Sarra A, Palermi S, Luca G (2013). "The Application of M-Function Analysis to the Geographical Distribution of Earthquake Sequence.", In A Giusti, G Ritter, M Vichi (eds.), *Classification and Data Mining*, chapter 32, pp. 271–278. Studies in Classification, Data Analysis, and Knowledge Organization. Springer Berlin Heidelberg.

Ó hUallacháin B, Leslie TF (2007). "Producer Services in the Urban Core and Suburbs of Phoenix, Arizona." *Urban Studies*, **44**(8), 1581–1601.

Ohser J (1983). "On estimators for the reduced second moment measure of point processes." *Series Statistics*, **14**(1), 63–71.

Openshaw S, Taylor PJ (1979). "A million or so correlation coefficients: three experiments on the modifiable areal unit problem.", In N Wrigley (ed.), *Statistical Applications in the Spatial Sciences*, pp. 127–144. Pion, London.

Overman HG (2008). "GIS data in economics." *The New Palgrave Dictionary of Economics*, (Second Edition).

Penttinen A (2006). "Statistics for Marked Point Patterns.", In *The Yearbook of the Finnish Statistical Society*, pp. 70–91. The Finnish Statistical Society, Helsinki.

Penttinen A, Stoyan D, Henttonen HM (1992). "Marked Point Processes in Forest Statistics." *Forest Science*, **38**(4), 806–824.

Pélissier R (1998). "Tree spatial patterns in three contrasting plots of a southern Indian tropical moist evergreen forest." *Journal of Tropical Ecology*, **14**(1), 1–16.

R Development Core Team (2014). "R: A Language and Environment for Statistical Computing."

Ripley BD (1976). "The Second-Order Analysis of Stationary Point Processes." *Journal of Applied Probability*, **13**(2), 255–266.

Ripley BD (1977). "Modelling Spatial Patterns." *Journal of the Royal Statistical Society*, **B 39**(2), 172–212.

Rowlingson B, Diggle (1993). "SPLANCS: Spatial Point Pattern Analysis Code in S-Plus." *Computers & Geosciences*, **19**(5), 627–655.

Scholl T, Brenner T (2013). "Optimizing Distance-Based Methods for Big Data Analysis." In *Working Papers on Innovation and Space. Philipps University Marburg*, 2013-09.

Silverman BW (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Stoyan D, Ohser J (1984). "Cross-correlation measures for weighted random measures." *Teoriya Veroyatnostei i ee Primeneniya*, **29**(2), 338–347.

Stoyan D, Ohser J (1985). "Cross-Correlation Measures of Weighted Random Measures and Their Estimation." *Theory of Probability and its Applications*, **29**(2), 345–355.

Stoyan D, Stoyan H (2000). "Improving ratio estimators of second order point process characteristics." *Scandinavian Journal of Statistics*, **27**(4), 641–656.

Sweeney SH, Feser EJ (1998). "Plant Size and Clustering of Manufacturing Activity." *Geographical Analysis*, **30**(1), 45–64.

Waller L (2010). "Point Process Models and Methods in Spatial Epidemiology.", In A Gelfand, P Diggle, P Guttorp, M Fuentes (eds.), *Handbook in Spatial Statistics*, chapter 22, pp. 403–423. CRC Handbooks of Modern Statistical Methods Series. Chapman & Hall.

Wiegand T, Moloney KA (2004). "Rings, circles, and null-models for point pattern analysis in ecology." *Oikos*, **104**(2), 209–229.

Wiegand T, Moloney KA, Naves J, Knauer F (1999). "Finding the Missing Link between Landscape Structure and Population Dynamics: A Spatially Explicit Perspective." *The American Naturalist*, **154**(6), 605–627.