A typology of distance-based measures of spatial concentration

Eric Marcon, Florence Puech

A TYPOLOGY OF DISTANCE-BASED MEASURES OF SPATIAL CONCENTRATION

ERIC MARCON[1] AND FLORENCE PUECH[2]

UPDATED MARCH, 10TH 2012

[1] AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana. Corresponding author, e-mail: Eric.Marcon@ecofog.gf

[2] LET (Université de Lyon, CNRS, ENTPE), Institut des Sciences de l'Homme, 14 av. Berthelot, 69363 Lyon Cedex 07, France.

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 specific properties and various applications. However no typology of distance-based methods exists. This paper fills this gap. The proposed classification helps understanding all properties of distance-based methods and proves that they are variations on the same framework.

Running head: Distance-based measures of concentration

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

JEL Classification: C10, C60, R12

I. Introduction

In the "spatial economics" article of 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. For example, in the long-term interest of economists for the measurement of spatial concentration of activities, distance-based methods are the last tools added (Combes et al. 2008). Economists traditionally employ disproportionality methods (terminology by Bickenbach and Bode 2008) defined on a *discrete* definition of space[1] like Gini index (Gini 1912), Ellison and Glaeser index (1997) or the entropy index of overall localization (Cutrini 2009). In opposition, the newly introduced distance-based methods are *continuous* functions of space and provide information about concentration at *all* scales simultaneously. The seminal work by Ripley (1976, 1977), the *K* function, has been quickly transferred to field scientists in ecology (handbooks by Diggle 1983; Cressie 1993 for instance), but remained incidentally used in economics (Barff 1987; Feser and Sweeney 2000; Sweeney and Feser 1998; Arbia and Espa 1996; Arbia 1989) before the works of Marcon and Puech (2003, 2010) and Duranton and Overman (2002)[2] which introduced an alternative approach.

We propose in this paper a classification of distance-based methods. Two main reasons motivate our work. Firstly various distance-based methods are used today by economists. The richness of the toolbox provided by these measures may come with some confusion for the economist interested in testing a hypothesis rather than a methodology so a classification may be helpful. Secondly, we provide in this article a unified theoretical framework by proving that all of 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 testing the data against it and also solving remaining issues. As a result, if objects (for instance plants) attract each other, one will find on average more neighbors (the other plants) around them than if they were distributed randomly and independently. In conclusion, these methods are variations on the same framework to gauge the spatial concentration.

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

II. Basic principles

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

When studying the location of activities, economists are interested to precisely describe the spatial distribution of one kind of entities ("points"[3]), for example shops of a given activity. In that case, they aim at detecting phenomena of attraction ("agglomeration", "localization") or repulsion ("dispersion") between those entities on a territory. Another field of research rests on the relations between entities belonging not to one but two

[1] *e.g.* a continent is divided into country, divided in turn into regions etc.

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

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

different groups ("co-localization" phenomenon). All the tools we consider in this review identify the spatial structure of the distribution. The particularity of these tools is that they treat space as continuous and not as a collection of predefined zones. As a consequence, they are functions of distance between the analyzed entities. Results are presented as a plot of a function of distance, whose values are meaningful or not, and always compared to an envelope (two curves) representing the confidence interval of a null hypothesis to test. "Intertype functions" are used to characterize co-localization. They are built in the same way. The curve allows showing the potential attraction or repulsion of two types of points (between two types of shops for example).

In more technical terms, those distance-based methods investigate the spatial structure of point patterns. Their mathematical framework is that of point processes (see the first chapters of Møller and Waagepetersen 2004 for a rigourous introduction). It is clearly out of the scope of this paper to explore the point process theory but a 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 law, with parameters) but 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 whose law is unknown so its characterization must be non-parametric: some statistics will be used for tests and even quantify some properties, but identifying the point process will generally remain impossible. These statistics will be defined according to pertinent properties of the point process. Their value will be estimated from the data. The first property of interest is its 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 all over the space, it is denoted λ and the point process is said to be stationary. The probability to find a point around x is $\lambda(x)dx$. Intensity can be estimated by density $\hat{\lambda}(x)$, the number of observed points per unit of space. In simple cases, points are just counted. If density changes over 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.

III. Distance-based methods: a brief presentation

A. The g function

The second order property of point pattern characterizes the relation between points: attraction, repulsion or independence. It is defined as the ratio between the joint probability to find two points in two places x and y, denoted $\lambda(x,y)dxdy$ and the product of probabilities to find each of them. For practical purpose, this property is supposed 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 $g(r) = \frac{\lambda(x,y)}{\lambda(x)\lambda(y)}$ where r is the distance between two points x and y. If points are distributed independently, $g(r) = 1$ (no interaction between points). If $g(r) > 1$, the probability to find 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.

Writing the conditional probability to find a point around y when a point actually is at x was 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 deeply discuss hereinafter the importance of the reference choice.
- Neighbors of reference points are counted at distance *r*. Neighbors may be of a different type from reference points or the same. In the former case, intertype measures are defined

while in the second case intratype measures used.

- 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.

Estimating $g(r)$ from the data requires a technique to count neighbors at a given distance: kernel functions again are used, but in one dimension this time. A kernel function gives a weight to neighbors at distances around r. The weight is greater the closer to *r* the distance is, and the kernel function sums to 1 for all distances. Duranton and Overman (2005, Eq1) used a Gaussian kernel following Silverman (1986). The most efficient is also the simplest (Illian et al. 2008, chapter 4.3.3), the simple box kernel with bandwidth of parameter *h*:

$$k\left(\left\|x_i - x_j\right\|, r\right) = \begin{cases} \frac{1}{2h} \text{ if } r-h \le \left\|x_i - x_j\right\| \le r+h \\ 0 \text{ else} \end{cases} \quad (1)$$

x_i and x_j refer to the spatial position (exact location) of two points. In the remainder, x_i will designate reference points,x_j their neighbors.

$g(r)$ is estimated by:

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

$c(i,j)$ is an edge-effect correction depending on both points. 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 out of the area. Several corrections are conceivable for relatively simple shapes (Goreaud and Pélissier 1999) but are intractable for actual geographical units (like countries).[4]

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 area of the study area.

B. The K function

Ripley (1976, 1977) integrated 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 homogenous and independent (a homogenous Poisson process) the spatial pattern is called complete spatial randomness (CSR), and $K(r) = \pi r^2$. If $K(r)$ is greater than πr^2, then more points are found within a radius r apart from each point. The point process is said to be attractive: spatial concentration is detected. Values of $K(r)$ inferior to πr^2 indicate that points repulse each other up to distance r (dispersion). 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 $L(r) = \sqrt{K(r)/\pi} - r$ so that its reference value is 0.

$L(r)$ 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 were distributed independently from reference points. We believe that $K(r)/\pi r^2$ is a better normalization because it is 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:

[4] See Marcon and Puech (2003) for more insights and Law *et al.* (2009) for a recent review.

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

$\mathbf{1}\big(\|x_i - x_j\| \leq r\big)$ is the indicator function equal to 1 if the distance between x_i and x_j is less than r, 0 else. Like ĝ, $\widehat{K}$ suffers edge effects.

K is the cumulative function of g: it provides information *up to* a given distance while g gives them *at* a distance. Abusing language, in analogy with random variable terminology, the latter type of functions is called probability density functions below, even though 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 one. 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 (Goreaud and Pélissier 2003; Arbia et al. 2008):

- "*random labeling*" is appropriate when locations are given, types are chosen (think of shops in a city),
- while "*population independence*" is the good hypothesis when points can be set anywhere, but not independently from other points of the same type (think of exploring interactions between two types of sellers on a beach -ice creams and sun hats for instance, each of them having its own spatial structure).

C. The K_{mm} function

The K_{mm} function was introduced by Penttinen et al. (Penttinen et al. 1992; Penttinen 2006). It generalizes Ripley's *K* function by associating quantitative marks $w(x_i)$ to points, that can be used as weights (Espa et al. 2010)[5]. It can be understood as a *K* function computed on a data set where $w(x_i)$ points are superposed where a point with mark $w(x_i)$ is found.

Its estimator is:

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

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

$\widehat{K}_{mm}(r)$ 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 firms, Espa et al. (2010) divided $\widehat{K}_{mm}(r)$ by W^2 (we estimated it by $n(n-1)\overline{w}^2$) so that $\widehat{K}_{mm}(r)$ has the same properties as $\widehat{K}(r)$. We follow them here.

D. The D function

To deal with inhomogenous point patterns, Diggle and Chetwynd (1991) introduced the *D* function, equal to the difference of 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, it is limited to tests. We introduce here an alternate version of *D*, we will denote D_i:

[5] By considering the spatial distribution of firms, a classical weight for the entities is the number of employees. However, one can also consider other weights as the value-added per firm.

$$D_i = K_c - K_{c,0}$$

It also equals 0 under the null hypothesis and can be used exactly like *D*. Its advantage compared to Diggle and Chetwynd'*D* 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.

E. 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 g_{inhom}'s integral 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; Arbia et al. 2012) because it requires the estimation of the intensity of the point process by kernel methods. If the kernel's bandwidth is very small, intensity is highly variable and independency 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).

It is estimated by:

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

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

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

F. The O-ring function

Wiegand and Moloney (Wiegand et al. 1999; Wiegand and Moloney 2004) argue for the use of a probability density function (pdf) rather than a cumulative function to have more informative results at each distance. They defined the *O-ring* function: $O(r) = \lambda g(r)$ and developed a method to correct for edge effects by pixelization of the original map allowing any border shape. O-ring values are not directly interpreted but compared to the envelope of simulations of a null hypothesis to allow its rejection. As far as we know there is no application of the O-ring in economics. However, it may be a promising method if only discrete spatial data are available at a very thin level of observation.

G. The Kd function

Duranton and Overman's (2005) *Kd* is the pdf to find a point's neighbor at a given distance. It counts and averages the number point pairs at each distance, smoothes the results to obtain a continuous function that is normalized to sum to 1. *Kd*'s values are compared to the confidence interval of the null hypothesis that points are randomly placed on their actual location set. A variant of *Kd* named K^{emp} (also proposed by Duranton and Overman 2005) allows to weight points, counting employees in firms.

Duranton and Overman's *f* function in their definition of *Kd* is actually not our kernel function but $k(\|x_i - x_j\|, r)h$. We have:

$$K_d(r) = \frac{1}{n(n-1)}\sum_i \sum_{j,i\neq j} k(\|x_i - x_J\|, r)$$

$$K^{emp}(r) = \frac{1}{\sum_i \sum_{j,i\neq j} w(x_i)w(x_j)} \times \sum_i \sum_{j,i\neq j} w(x_i)w(x_j)k(\|x_i - x_J\|, r) \quad (8)$$

H. The M function

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 definition is:

$$M(r) = \frac{\sum_i \frac{\sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j^c\right\| \leq r\right) w\left(x_j^c\right)}{\sum_{j,i\neq j} \mathbf{1}\left(\left\|x_i - x_j\right\| \leq r\right) w\left(x_j\right)}}{\sum_i \frac{W_c - w(x_i)}{W - w(x_i)}} \quad (9)$$

W_c is the total weight of the points x_j^c. The denominator is slightly different in the inter-type function, $\sum_i \frac{W_c}{W - w(x_i)}$, avoiding a small bias.

IV. A typology of distance-based methods

In what follows, we see 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 of just counting. Second, an average number of neighbors $\bar{n}(r)$ is calculated. Third, $\bar{n}(r)$ is divided by a reference measure $m(r)$. In accordance with the

Table 1: Estimating the number of neighbors.

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

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 ($2\pi r dr$ 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, $\bar{n}(r)/m\,(r)$ is compared to the value it has on the whole domain, $\bar{n}_0/m_0$. Fifth and last, 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 steps are detailed below.

A. Step 1: a number of neighbors $\boldsymbol{n(x_i, r)}$

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 pdf functions (*g*, g_{inhom}, K_d and *O)*, the second one cumulative functions (*K*, K_{inhom}, and *M*). *M*, K^{emp} and K_{mm} attribute points a weight, such as a number of employees in firms. 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 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 over-weight.

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 inter-type 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. $\hat{\lambda}(x_j)$ is the density of points around x_j. It is the estimator of $\lambda(x_j)$, the intensity of the point process. Hats are used for estimators, to avoid confusion: *K*, *g*, *O* have a mathematical definition relying on the point process they are used to characterize and they are estimated from the data. In opposition, *Kd* and *M* are pure empirical functions. $k(\|x_i - x_j\|, r)$ 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.

Table 2: Average number of neighbors.

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

Table 3: Reference measure.

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

B. 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. $\bar{n}(r)$ is the average number of neighbors. n is the total number of reference points (the centers of circles).

C. Step 3: a reference measure

These numbers of neighbors are then divided by a reference measure $m(r)$ (Table 3).

Table 4: The reference value.

Function	Reference value	Observations
$\widehat{\boldsymbol{K}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$ $\widehat{\boldsymbol{g}}(\boldsymbol{r})$	$\frac{\bar{n}_0}{m_0} = \frac{n-1}{A}$	The reference value is the intensity of points, evaluated by the total number of points minus 1 (the circle center) divided by the area of the window.
$\widehat{\boldsymbol{K}}_{\boldsymbol{inhom}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$ $\widehat{\boldsymbol{g}}_{\boldsymbol{inhom}}(\boldsymbol{r})$ $\widehat{\boldsymbol{O}}(\boldsymbol{r})$	$\frac{\bar{n}_0}{m_0} = 1$	Local weights used in inhomogenous measures are such that the reference value is 1. $\hat{O}$ is a special case because its authors decided not to normalize it as a location quotient as *g* but have an expected value equal to the intensity in the case of CSR.
$\widehat{\boldsymbol{K}}_{\boldsymbol{mm}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$	$\frac{\bar{n}_0}{m_0} = \frac{(n-1)\bar{w}}{A}$	The reference value is the density multiplied by the average point weight.
$\boldsymbol{K_d}(\boldsymbol{r})$	$\frac{\bar{n}_0}{m_0} = n-1$	Absolute measure.
$\boldsymbol{K^{emp}}(\boldsymbol{r})$	$\frac{\bar{n}_0}{m_0} = \frac{\sum_i \sum_{j,i\neq j} w(x_i) w(x_j)}{\sum_i w(x_i)}$	Absolute measure.
$\boldsymbol{M}(\boldsymbol{r})$	$\frac{\bar{n}_0}{m_0} = \frac{1}{n}\sum_i \frac{W_c - w(x_i)}{W - w(x_i)}$	The reference value is calculated as $\bar{n}(r)/m\,(r)$ 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.

D. Step 4: the reference value

Normalization comes then (Table 4). $\bar{n}(r)/m\,(r)$ is divided by its reference value $\bar{n}_0/m_0$. The latter can be understood as the value of the former with r large enough for all points to be neighbors to each other. Then, possible spatial structure does not matter: $\bar{n}_0$ counts all points except the center of the circle, m_0 is the total area (topographic measures) or the total number of points (relative measures) or 1 (absolute measures).

All measures in the table are location quotients except for the O-ring (but $O(r)/\lambda$ is) and K_d which is an absolute measure.

E. Last step: null hypothesis

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

To decide whether these values reject the null hypothesis we want to test (generally independence between point locations), Monte-Carlo simulations summarized in Table 5 are drawn. They also allow solving various issues the analytical formulation of measures could not deal with.

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 allows giving the values a signification, while the values themselves are meaningless. Quite similarly, Kosfeld et al. (2011) use *K* in heterogenous space: to characterize the spatial structure of an industry sector, they draw 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.

M allows giving the points a weight, typically a number of employees. But it does not

Table 5: Simulation of the null hypothesis.

Function	Usual null hypothesis for intratype functions	Null hypothesis for intertype functions
$\widehat{\boldsymbol{K}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$ $\widehat{\boldsymbol{g}}(\boldsymbol{r})$	A homogenous Poisson process of intensity $\hat{\lambda}$ (estimated from the data).	Random labeling: point locations are kept, labels are redistributed. Population independence: point sets are shifted relatively to each other.
$\widehat{\boldsymbol{K}}_{\boldsymbol{inhom}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$ $\widehat{\boldsymbol{g}}_{\boldsymbol{inhom}}(\boldsymbol{r})$ $\widehat{\boldsymbol{O}}(\boldsymbol{r})$	An inhomogenous Poisson process of intensity $\hat{\lambda}(x)$ (estimated from the data).	As *K*.
$\widehat{\boldsymbol{K}}_{\boldsymbol{mm}}(\boldsymbol{r})/(\boldsymbol{\pi r^2})$	Generally, a homogenous Poisson process with random labeling.	No intertype function: marks are continuous.
$\boldsymbol{K_d}(\boldsymbol{r})$	Points are redistributed across actual locations.	Points are redistributed across actual locations of centers and neighbors.
$\boldsymbol{M}(\boldsymbol{r})$	Points are redistributed across actual locations.	Reference points are kept unchanged; all other points are redistributed randomly across actual locations. Reference and neighbor roles are exchanged and confidence intervals are computed again. The intertype function is significant if the null hypothesis is rejected in both cases.

take into account the possible influence of the structure of point weights, like Ellison and Glaeser index includes the value of the Herfindahl index *H* to correct for industrial concentration (see Ellison and Glaeser 1997). The simulations are done with the actual weights so weight structure is controlled for by the confidence interval of the null hypothesis, if not numerically.

The *O-ring* statistic does not take into account the heterogeneity of the point process. Its value is compared to that of an inhomogenous, independent simulated process to detect aggregation or repulsion.

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 very difficultly in other situations. Espa et al. (2010) integrate both point locations and mark structure in a single model of localization, whose parameters cannot be inferred yet.

Monte-Carlo simulations provide many values of the simulated function for each value of r. A proportion of them according to the accepted risk level is eliminated (often the greater and the smaller 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 observed value of the function of r are compared to. Duranton and Overman (2005) note that repeating the same local test for all values of r is not satisfactory as the resulting local confidence interval should be not too restrictive. They propose a way to build a global test followed by Marcon and Puech (2010). Loosmore and Ford (2006) proved the inadequacy of the local test.

Goodness-of-fit (GoF) tests are an alternative explored by Heinrich (1991) but not applied yet in the empirical literature.

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 distribition 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. An analytical (without simulations) GoF test of the *K* function against CSR is now available (Lang and Marcon 2010). Also, Jensen and Michel (2011) developed an analytical test for K_d and *M* based on a different approach.

V. Discussion

Various applications of distance-based methods can be found in the economic literature to gauge the spatial concentration of activities (see for instance Puech 2003 for a review). Unfortunately, the choice of a measure is far from being systematically justified by a rigorous comparison of the properties of existing distance-based methods. Generally, the only motivation given is that such an approach solves the Modifiable Areal Unit Problem (MAUP) introduced by Openshaw and Taylor, 1979 (1979). To put things clearer on the MAUP, if statistic tools rely on a discretization of space (like Gini, Herfindahl or Ellison and Glaeser indices) Morphet underlines that *"the result will be sensitive to the shape, size, and position of the areal units chosen"* (Morphet 1997, p. 1039). Thus, after introducing the well-known Ripley's *K* function in the field of economics (see Barff 1987; Arbia and Espa 1996; Marcon and Puech 2003), specific developments were done to adapt distance-based methods to the measurement of the spatial distribution of activity. The *Kd* function (Duranton and Overman 2005) or the *M* function (Marcon and Puech 2010) were proposed and now are widely used (Klier and McMillen 2008; Jensen and Michel 2011; Ellison et al. 2010). The main reason is that only these two functions respect a maximum of the good criteria for a concentration index in spatial economics (Combes and Overman 2004).

The statistical literature (Møller and Waagepetersen 2004) mainly deals with homogenous point processes, used for topographic measures. Many theoretical results exist, such as *g* and *K* expectancy for many processes (Diggle 1983; Illian et al. 2008). Topographic measures for inhomogenous space and relative measures still lack much mathematical support. K_d was built on empirical foundations: its values are not estimators of a theoretical statistic. From an economic point of view, it is now well admitted that measures of relative concentration should be preferred to gauge the geographic concentration of activities (see Duranton and Overman 2005; Combes and Overman 2004; Combes et al. 2008). In economic geography, the typical question to address is that of the spatial structure of employment, supposed to be aggregative because of externalities (Marshall 1890; Krugman 1991). Since the spatial distribution of activities is very heterogenous, homogenous topographic measures are of little use (Marcon and Puech 2003) but relative ones are able to detect aggregation scales (Combes and Overman 2004; Duranton and Overman 2005; Marcon and Puech 2010). Moreover, the use of relative measures allows testing the predictions of theoretical models because an industrial over-representativity is generally expected. Usually, studies assess deviations of the spatial distribution of a sector from that of the aggregate activity (see Krugman 1991; Brülhart and Traeger 2005 for example; Ellison and Glaeser 1997). In a more sophisticated way, comparing the distribution of the production of a given industry to that of its demand may be informative to test various theoretical predictions for example the Home Market Effect (Krugman 1980) or the tradability of services (Jensen and Kletzer 2006).

A recent debate rests on the respective advantages of pdf and cumulative measures (Marcon and Puech 2010; Wiegand and Moloney 2004; Law et al. 2009). Arguments are not repeated here (the reader should refer to the cited papers) because the choice depends on both on the question analyzed and

the availability of data. For instance it has been shown by Marcon and Puech (2010) that *M* or *Kd* are useful to evaluate the spatial concentration of activities and depending of the question analyzed, one could give clearer results than the other. As Marcon and Puech underline, M and Kd are more complements than substitutes. For instance, *Kd* provides more precise estimations than *M* for gauging the local density of activities. However, *M* better assesses the global effect of the superposition of spatial structures. As a consequence, if the question is "up to which distance do externalities matter?", then a cumulative function is more appropriate, while a pdf will answer "do externalities matter at a given distance?" better.

To go further, we provide in Table 6 the tools' properties so that a researcher can use the appropriate function. After the reference and neighbor types have been chosen, the basic question to answer is whether the reference is topographic (then, whether space is homogenous or not) or relative. Note that no distance-based methods gauge the absolute concentration. A benchmark is systematically retained: physical space (topographic measures) or another variable (relative measure). Last, some functions provide quantitative information about the point pattern structure: how many times more neighbors are found at (g, g_{inhom}) or up to ($K/\pi r^2$, *M*) the chosen distance? Others only allow a test to reject the null hypothesis of independence (*O-ring*, *Kd*).

K_{inhom} has been employed by Diggle et al. (2007) and Arbia et al. (2012) in a case-control design: some points, the controls (the whole employment location in Arbia et al.'s study of the high-tech industries in Milan), are used to estimate the point process intensity (assuming they are approximately independently distributed), while the different pattern of cases is attributed to dependence. This approach is quite similar to that of *M* where all points weight 1: estimating $\hat{\lambda}(x_j)$ in Table 1 for K_{inhom} with a simple box kernel with bandwidth r is not different from $m(r)$ 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).

The next step is clearly beyond descriptive statistics. Duranton and Overman (2008) emphasize that this way of research is very promising to explain the location strategies. However, these functions will be really useful if they can be related to some economic models they will allow to validate or not (as the one proposed by Picone et al. 2009). In a discrete space approach, Ellison and Glaeser (1997) built their index from a profit-maximization model. In opposition, distance-based methods were built from geometrical considerations (the point process theory for *g* and *K*). Very few works indeed relate concentration values to an economic model built around spillovers, labor pooling and input-output linkages. Feser and Sweeney (2002) use the *D* function to test for the existence of a factor's effect on ag-

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

Function choice	**Topographic, homogenous**	**Topographic, inhomogenous**	**Relative**
Probability Density Functions	*G*	g_{inhom} *O-ring (test only)*	*Kd (test only),* K^{emp}*(test only, with weights)*
Cumulative functions	*K* K_{mm} *(with weights)*	K_{inhom}	*M (with weights)* D_i *Case-control* K_{inhom}

glomeration. Models using explicitly a distance-based measure of concentration as the output of a combination of factors used as proxies of the expected causes of concentration (such as Rosenthal and Strange 2001 with Ellison and Glaeser's index) are just two as far as we know: Puech (2003, chapter 5) and Ellison et al. (2010). The first one is an attempt to model the value of *M* around each firm as a linear function of these proxies. With much better data and surely much more audience, Ellison et al. use *Kd* values as outputs and show the major role of input-output linkages.

VI. Conclusion

A decade ago, disproportionality methods such as Gini or Ellison and Glaeser index were economists' classical tools. It was proved by Briant et al. (2010) that they face serious issues known as the MAUP (Openshaw and Taylor 1979; Arbia 1989). Quite logically, methods were then developed to take advantage of the knowledge of the exact position of objects. The first ones were statistics based on the distance of the nearest neighbor of points, after Clark et 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. An exception is Leslie and Kronenfeld (2011) who develop a new statistic, the colocation quotient, based on the ratio of nearest neighbors of the type of interest.

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 (*M* is a location quotient, so it is easy to interpret). We explained in this article (Table 6) which tool to use according to the underlying framework (topographic or relative). Topographic measures are widely used and updated by ecologists in handbooks (Law et al. 2009; Illian et al. 2008; Fortin and Dale 2005) which widely ignore relative measures. Economists use mainly relative measures (Marcon and Puech 2010; Duranton and Overman 2005, 2008) because they better fit economic considerations. Several economists (Combes et al. 2008) clearly state that applications of distance-based methods should now be privileged 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). However, relating these descriptive tools to economic theory is the real challenge, following the way opened by Ellison et al. (2010).

References

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

Arbia G, Copetti M, Diggle P, Fratesi U, Senn L (2009) Modelling Individual Behaviour of Firms in the Study of Spatial Concentration. In: Fratesi U, Senn L (eds) Growth and Innovation of Competitive Regions. Advances in Spatial Science. Springer, Berlin, pp 297-327. doi:10.1007/978-3-540-70924-4_14

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. doi:10.1007/s00181-007-0154-1

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

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

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. doi:10.1177/0160017608319589

Briant A, Combes P-P, Lafourcade M (2010) Dots to boxes: Do the Size and Shape of Spatial Units Jeopardize Economic Geography Estimations? Journal of Urban Economics 67 (3):287-302

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 P-P, Mayer T, Thisse J-F (2008) Economic Geography, The Integration of Regions and Nations. Princeton University Press, Princeton

Combes P-P, Overman HG (2004) The spatial distribution of economic activities in the European Union. In: Henderson JV, Thisse J-F (eds) Handbook of Urban and Regional Economics, vol 4. 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. doi:10.1016/j.regsciurbeco.2008.08.005

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. doi:10.1111/j.1541-0420.2006.00683.x

Duranton G (2008) Spatial Economics. The New Palgrave Dictionary of Economics, Second, Palgrave Macmillan edn. Second Edition, S.N. Durlauf et L.E. Blume (Eds), Palgrave Macmillan,

Duranton G, Overman HG (2002) Testing for Localization Using Micro-Geographic Data. 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. doi:10.1257/aer.100.3.1195

Espa G, Giuliani D, Arbia G (2010) Weighting Ripley's K-function to account for the firm dimension in the analysis of spatial concentration. Università di Trento, Trento

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

Feser EJ, Sweeney SH (2002) Theory, methods, and a cross-metropolitan comparison of business clustering. In: McCann P (ed) Industrial Location Economics. Edward Elgar, Cheltenham,

Fortin M-J, Dale MRT (2005) Spatial Analysis. A guide for ecologists. Cambridge University Press, Cambridge

Gini C (1912) Variabilità e mutabilità. Studi Economico-Giuridici dell'Università di Cagliari, vol 3. Università di Cagliari,

Goreaud F, Pélissier R (1999) On explicit formulas of edge-effect correction for Ripley's K-function. J Veg Sci 10 (3):433-438

Goreaud F, Pélissier R (2003) Avoiding misinterpretation of biotic interactions

with the intertype K_{12} fonction: population independence *vs* random labelling hypotheses. J Veg Sci 14:681-692

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 BJ, Kletzer LG (2006) Tradable Services: Understanding the Scope and Impact of Services Offshoring. In: Brainard L, Collins SM (eds) Brookings Trade Forum: 2005. Offshoring White-Collar Work – The Issues and the Implications. Brookings Institution Press, Washington, pp 75-134

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. doi:10.1007/s00168-009-0342-3

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

Kosfeld R, Eckey H-F, Lauridsen J (2011) Spatial point pattern analysis and industry concentration. The Annals of Regional Science 47 (2):311-328. doi:10.1007/s00168-010-0385-5

Krugman P (1980) Scale Economies, Product Differentiation, and the Pattern of Trade. The American Economic Review 70 (5):950–959

Krugman P (1991) Geography and Trade. MIT Press, London

Lang G, Marcon E (2010) Testing randomness of spatial point patterns with the Ripley statistic. ArXiv e-prints 1006.1567

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. J Ecol 97 (4):616-628. doi:10.1111/j.1365-2745.2009.01510.x

Leslie TF, Kronenfeld BJ (2011) The Colocation Quotient: A New Measure of Spatial Association Between Categorical Subsets of Points. Geogr Anal 43 (3):306-326

Loosmore NB, Ford ED (2006) Statistical inference using the G or K point pattern spatial statistics. Ecology 87 (8):1925-1931. doi:10.1890/0012-9658(2006)87[1925:siutgo]2.0.co;2

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

Marshall A (1890) Principle of Economics. Macmillan, London

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

Morphet CS (1997) A statistical method for the identification of spatial clusters. Environment and Planning A 29 (6):1039-1055

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

Overman HG (2008) GIS data in economics. The New Palgrave Dictionary of Economics. Second Edition, S.N. Durlauf et L.E. Blume (Eds), Palgrave Macmillan,

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

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

Picone GA, Ridley DB, Zandbergen PA (2009) Distance decreases with differentiation: Strategic agglomeration by retailers. International Journal of Industrial Organization 27 (3):463-473

Puech F (2003) Concentration géographique des activités industrielles : Mesures et enjeux. PhD Thesis, Université de Paris I, Panthéon - Sorbonne, Paris

Ripley BD (1976) The Second-Order Analysis of Stationary Point Processes. Journal of Applied Probability 13:255-266

Ripley BD (1977) Modelling Spatial Patterns. Journal of the Royal Statistical Society B 39 (2):172-212

Rosenthal SS, Strange WC (2001) The Determinants of Agglomeration. Journal of Urban Economics 50 (2):191-229

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

Stoyan D, Stoyan H (2000) Improving ratio estimators of second order point process characteristics. Scand J Stat 27 (4):641-656

Sweeney SH, Feser EJ (1998) Plant Size and Clustering of Manufacturing Activity. Geogr Anal 30 (1):45-64

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. Am Nat 154 (6):605-627