DISTANCE-BASED MEASURES OF SPATIAL CONCENTRATION: INTRODUCING A RELATIVE DENSITY FUNCTION

Gabriel Lang, Eric Marcon, Florence Puech

Distance-Based Measures of Spatial Concentration: Introducing a Relative Density Function

Gabriel Lang[a], Eric Marcon[b], Florence Puech[c*]

Abstract
For a decade, distance-based methods have been widely employed and constantly improved in spatial economics. These methods are a very useful tool for accurately evaluating the spatial distribution of economic activity. We introduce a new distance-based statistical measure for evaluating the spatial concentration of industries. The m function is the first relative density function to be proposed in economics. This tool supplements the typology of distance-based methods recently drawn up by Marcon and Puech (2012). By considering several theoretical and empirical examples, we show the advantages and the limits of the m function for detecting spatial structures in economics.

JEL Classification: C10, C60, R12

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

[a] *AgroParisTech, UMR 518 Mia, 19 avenue du Maine, F-75732 Paris Cedex 15, France.*
[b] *AgroParisTech, UMR EcoFoG, BP 709, F-97310 Kourou, French Guiana.*
[c] *RITM, Univ. Paris-Sud, Université Paris-Saclay & EXCESS, 92330, Sceaux, France.*
***Corresponding author**: Florence.Puech@u-psud.fr. Authors are in alphabetical order.

Introduction

Industrial agglomerations are doubtless the main feature of today's economic geography (Krugman, 1991; Henderson and Thisse, 2004). Thus, it is not surprising that much recent research has attempted to improve the measurement of the spatial concentration of activities.[1] Distance-based methods are the latest statistical measures to be proposed in the field of spatial economics for detecting spatial structures (geographic concentration or dispersion). By treating space as continuous, distance-based methods provide a detailed analysis with robust results.[2] Consequently, many authors consider them to be very promising techniques (Combes and Overman, 2004; Combes *et al.*, 2008; Duranton, 2008) and that they open the way for new explanations of the spatial concentration of activities (Ellison *et al.*, 2010; Alfaro and Chen, 2014; Kerr and Kominers, 2015). Today, the Duranton and Overman's K_d function (Duranton and Overman, 2005) is the most used distance-based method in economics. It is a density function that evaluates absolute concentration (Marcon and Puech, 2012, 2015). In this article for the first time we shall introduce a relative density function in spatial economics, in a similar vein to the well-known location quotient (Florence, 1972). We have developed this for two main reasons. First, this new function, called m, supplements the typology of distance-based methods recently drawn up by Marcon and Puech (2012). Second, as Brülhart and Traeger (2005) have stressed, the nature of the spatial concentration (absolute, relative) matters. We shall prove that both K_d and m are useful for evaluating the spatial distribution of activities because of the complementary results they provide. For this, we shall give various comparisons of K_d and m results obtained with theoretical and empirical examples to understand the advantages and the limits of both distance-based measures.

There is growing evidence that distance-based measures are now preferred in spatial economics since Duranton and Overman's seminal paper (Duranton and Overman, 2005). One of the main reasons for this is that they preserve the richness of individual data. Unlike the Gini (1912) or the Ellison and Glaeser (1997) indices, distance-based methods do not rely on any predefined zoning (regions, counties…). Distance-based methods are implemented directly using the position of entities (stores, plants…). The analysis of the spatial distribution of entities is based on the distance between them. In contrast, it has been proven that indices which aggregate data at a zonal level are sensitive to the zoning chosen (Arbia, 2001; Briant *et al.*, 2010) as described by the Modifiable Areal Unit Problem – MAUP (Openshaw and Taylor, 1979; Arbia, 1989). One way of solving MAUP issues is to treat space as "continuous" as suggested by Duranton (2008).[3] This is the main feature of distance-based measures. Consequently, MAUP issues vanish. These techniques therefore allow an exact and unbiased analysis of the spatial structure

[1] See Duranton and Overman (2005); Marcon and Puech (2003, 2010); Arbia *et al.* (2012); Mori and Smith (2013); Jensen and Michel (2011); Howard *et al.* (forthcoming), among others.

[2] Duranton and Overman (2008) provide many concrete examples of the problems such functions can solve.

[3] Gilles Duranton wrote in the article on "Spatial Economics" in The New Palgrave Dictionary of Economics: "*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.*"

of the distribution at all scales simultaneously (and not at only one level of observation as is the case with spatial zoning). Multiple patterns can be detected: for example aggregation or repulsion between entities according to the distance considered. These distance-based measures are now deemed to be very powerful and we can easily understand why many studies now employ them to evaluate spatial patterns (Arbia and Espa, 1996; Sweeney and Feser, 1998; Ó hUallacháin and Leslie, 2007; Bonneu, 2007; Arbia *et al.*, 2008; Nakajima *et al.*, 2012; Barlet *et al.*, 2013; Behrens and Bougna, 2015; Koh and Riedel, 2014; Giuliani *et al.*, 2014). Alternative approaches exist: Billings and Johnson (2012) for instance developed a test to detect and localize concentration based on the local density of a chosen sector, compared to that of the whole economic activity. We will rather focus in this paper on the distance-based methods, which rely on the second-order property of the distribution of points, *i.e.* the excess or lack of neighbors, because they belong to a common, coherent framework (Marcon and Puech, 2012).

In what follows, we shall introduce a new distance-based method: the m function. We will devote a great deal of attention to defining it in order to respect a maximum number of the good criteria for measuring spatial concentration in economics (Combes and Overman, 2004). In particular, the m function satisfies Duranton and Overman's five important criteria (Duranton and Overman, 2005): (1) the results are comparable across industries, (2) it controls for the overall agglomeration of manufacturing, (3) it controls for industrial concentration in the sense of Ellison and Glaeser (1997), (4) it is unbiased across geographic scales (this is related to the MAUP issues) and lastly (5) it gives an indication of the significance of the results. There are several ways to control for the overall agglomeration. Duranton and Overman's widely used K_d function ignores it (this is why it is classified as an absolute measure) but its value is compared to a confidence interval of its possible values under a counterfactual null hypothesis. The m function relies on the local share of employment (or whatever measure of size, including the number of establishments) of the sector under study to directly control for the distribution of the whole activity: it is a relative measure. As K_d, it considers neighbors at a given distance rather than up to it: both are density functions. No relative density function has yet been proposed to gauge the spatial concentration in continuous space, as Marcon and Puech (2012) pointed out.

Various theoretical and empirical examples are provided to show that the results provided by the m function are complementary to those of K_d: their results do not converge systematically, so we recommend that the nature of the spatial concentration analyzed should be studied carefully to avoid any erroneous conclusions. Moreover, relative measures of concentration are formally location quotients: the m function can be interpreted as the location quotient of a sector of activity in the neighborhood of a reference sector. This property opens the way for the development of locational choice models following Guimarães *et al.* (2009) where the strength of externalities or natural advantages can be directly linked to the value of the function. As a result, we believe the m function is an appropriate statistical tool for detecting spatial structures in economics.

In the following sections, first of all, the background is explained (Section 1). Then, the m function is presented (Section 2) and some simple illustrative examples are given (Section 3). In the last section of the paper (Section 4) we propose some comparisons with the K_d function. We also provide a simple theoretical example and an analysis of the spatial distribution of pharmacies in the Lyon area (France) to illustrate the advantages and limits of the use of the m function in our field.

1. Background

As we have already mentioned, distance-based methods are particularly attractive for economists because they provide a complete and unbiased analysis of the location patterns of industries. The basic idea of distance-based methods is simple. Let us consider the case of the textile industry. Evaluating the spatial distribution of plants in this sector depends on an assessment of the surroundings of textile plants for all the distances that are considered. On average, if there are locally more textile plants around textile plants than in the whole of the area under investigation, concentration is detected ("textile plants attract textile plants"). On the other hand, if there are fewer textile plants in the surroundings of the textile plants than there are in the area as a whole, we talk of a phenomenon of dispersion (in which case "textile plants repel textile plants"). Alternatively, if there is no relationship between the entities, independence is identified ("textile plants are randomly and independently distributed"). The significance of the results is provided by the confidence interval of the null hypothesis. More technically, all distance-based methods explore the spatial structure of point patterns. Their mathematical framework is that of point processes (Møller and Waagepetersen, 2004). Two concepts require additional explanations: the definition of the surroundings of plants and the nature of the spatial concentration (topographic, relative or absolute). Let us now examine these two important factors in depth.

Firstly, the notion of the **surroundings of plants** is central because it defines the type of function applied *i.e.* a cumulative or a density function. In practice, the evaluation of the neighboring plants is done for all distances, for example every 100 meters up to the median distance between all pairs of plants. The spatial distribution can be estimated up to a given distance or at a given distance. If the first option is chosen, it calls for a cumulative function. If the second option is selected, a density function is appropriate. The choice between one type of function and the other depends on the issue under study (Marcon and Puech, 2010).

The second clarification concerns **the nature of the spatial concentration**. In order to evaluate the spatial concentration of economic activities, it is necessary to choose a

benchmark value with which to compare the observed distribution of activities (see Brülhart and Traeger, 2005, among others):

- The first possibility is to use a topographic reference. In this case, physical space is chosen as the benchmark. One possible example is the number of neighboring plants per unit of space (that is on a disk of radius r for a cumulative function or on the ring at distance r for a density function). Space may be homogeneous or not. The homogeneity of space implies a constant density all over the study area (in our previous case this means that all the plants in the distribution have the same probability of being located anywhere in the study area). Some authors (Duranton and Overman, 2005; Marcon and Puech, 2003) consider this hypothesis to be generally irrelevant in the field of spatial economics and that a non-homogeneous space framework is needed.

- The second possibility is a relative reference. In this case, another variable is taken as a benchmark. Any variable can be used except space (if it is, the concentration is topographic). For instance, if we evaluate the spatial distribution of textile plants, we can detect in the plant's near environment the deviations of this distribution of plants from another distribution. The benchmark plants can be all plants at the aggregate industrial level.

- The last possibility is to have no reference. In this case, an absolute measure is defined. For example, the number of neighboring textile plants located at a given distance from a textile plant.

The growing number of measures in continuous space recently prompted Marcon and Puech (2012) to provide a typology of such functions. A classification of statistical measures can be drawn up by considering the nature of the geographic concentration and the definition of the type of function. Table 1 gives an overview of all the distance-based measures that have been used to evaluate the spatial distribution of economic activities:

- the K function of Ripley (1976, 1977),
- the g function of Ripley (1976, 1977),
- the K_{mm} function introduced by Penttinen (2006) and Penttinen *et al.* (1992),
- the D function of Diggle and Chetwynd (1991),
- the K_{inhom} function of Baddeley *et al.* (2000),
- the g_{inhom} function of Baddeley *et al.* (2000),
- the K_d function of Duranton and Overman (2005),
- the M function of Marcon and Puech (2010),
- the (unnamed) cumulative function of K_d proposed by Behrens and Bougna (2015).

One cell in the table is empty: no relative density function has yet been proposed for the field of spatial economics (the K_d function does not control explicitly for the distribution of the economic activity). The present paper fills this gap. In the next section we shall complete Table 1 by proposing a new density function, named the m function, which expresses relative spatial concentration.

Finally, it should be noted that the application of distance-based methods is not confined to spatial economics. They were first developed and applied in other disciplines. Much empirical research has thus been conducted in the fields of ecology (Law *et al.*, 2009) and epidemiology (Waller, 2010), for example.

2. Presentation of the *m* function

2.1 An intuitive presentation

The idea of the m function is as follows. Consider an area in which various plants belonging to several industrial sectors are located. The m function is a relative measure that compares the proportion of plants of interest in the neighborhood of the reference plants to the proportion of neighbors of interest in the area as a whole. If plants are agglomerated, the proportion of neighbors of interest in the neighborhood of reference plants is greater than in the area as a whole. On the contrary, if plants are dispersed then the proportion of plants of interest in the neighborhood of the reference plants is lower than in the area as a whole. These proportions (ratios) are estimated from observed data. If the neighbors of interest belong to the same sector as the reference plants, the m function helps to detect agglomeration phenomena. If the neighbors of interest do not belong to the same sector as the reference plants, the m function identifies co-agglomeration.

2.2 Definition of the m function

Let us now turn to the mathematical definition of the m function. Plants are defined as points. All points belong to a point pattern denoted by $\mathscr{X}$. Two subsets are considered: that of the reference points $\mathscr{R}$ (*e.g.* a given sector of activity) and that of the neighboring points of interest $\mathscr{N}$.

The estimator of m is:

$$\hat{m}(r) = \frac{\sum_{x_i \in \mathscr{R}} \frac{\sum_{x_j \neq x_i, x_j \in \mathscr{N}} k\left(\left\|x_i - x_j\right\|, r\right) w(x_j)}{\sum_{x_j \neq x_i, x_j \in \mathscr{X}} k\left(\left\|x_i - x_j\right\|, r\right) w(x_j)}}{\sum_{x_i \in \mathscr{R}} \frac{W_{\mathscr{N}} - w(x_i)}{W - w(x_i)}} \tag{1}$$

where x_i denotes the reference points, and x_j the neighbors. $w(x_i)$ is the weight of point x_i. $W_{\mathscr{N}}$ is the total weight of the neighboring points of interest and W is the total weight of all the points. If points represent industrial establishments or shops, the weight can be the number of employees working in those entities. $k(\cdot)$ is a kernel estimator whose sum

Table 1. Choice of the appropriate function to describe a point pattern structure

Function choice	Topographic, homogeneous	Topographic, inhomogeneous	Absolute	Relative
Probability density functions	g	g_{inhom}	K_d K^{emp}	
Cumulative functions	K	K_{inhom}	Cumulative of K_d	M
	K_{mm}	D_i	Cumulative of K^{emp}	Case-control K_{inhom}

can be used to estimate the number of neighbors of point x_i at distance r. We have followed Duranton and Overman (2005) and used a Gaussian kernel of optimal bandwidth as described by Silverman (1986). The kernel estimator considers the neighbors of the reference point and gives them a maximum weighting if they are exactly r apart. Their weighting decreases according to a Gaussian distribution with standard deviation h. The choice of the bandwidth h is arbitrary but important (Illian *et al.*, 2008). The wider it is, the smoother the estimator. In this paper, we shall only analyze the spatial distribution of one sector, so we shall focus on the intratype (or univariate) m function taking $\mathscr{R} = \mathscr{N}$. But we can extend the m function for the analysis of inter-industrial spatial distributions (co-agglomeration): the intertype (or bivariate) function can be defined in the same way, choosing different point types as the reference and neighbors.[4]

The equation of the m function reads as follows. The numerator is the sum of the local ratios *i.e.* the relative weight of the neighbors of interest at distance r from all the reference points. This is averaged over all the reference points (actually, it is simply summed because the number of reference points is simplified with the denominator). The denominator is the same ratio over the whole data set, *i.e.* the global ratio. It is not just $W_{\mathscr{N}}/W$ because the reference points are never counted as neighbors. An unbiased estimator of the global ratio is thus the average local ratio considering all points are neighbors to each other. For this reason, the denominator is slightly different in the intertype function: $\sum_{x_i \in \mathscr{R}} \frac{W_{\mathscr{N}}}{W - w(x_i)}$.

The benchmark value of the m function is 1 for any distance r. This value is obtained when points are independent. m values greater than 1 indicate the spatial concentration of points while m values lower than 1 express dispersion. m values can be interpreted. For example, if the m function is 1.5, at distance r, the proportion of the neighbor points of interest at this distance is 50% higher than in the area as a whole.

The significance of the estimates of m is given by the confidence interval of the null hypothesis (Monte-Carlo simulations). This technique is widely employed in the case of distance-based methods. In practice, random distributions of points are generated by permuting the marks (type and weighting pairs) of the actual points on the actual spatial positions of points (coordinates). We generate only a global confidence interval, following Duranton and Overman (2005).

2.3 Discussion

The m function fulfills all of Duranton and Overman's criteria mentioned in the introduction: (i) it compares the geographic concentration results across industries, (ii) it controls for industrial concentration (indirectly, comparing its values to the confidence envelope of the appropriate null hypothesis), (iii) it controls for the overall aggregation patterns of industries, (iv) it enables the significance of the results to be tested (using the confidence interval) and, (v) it keeps the empirical results unbiased across geographic scales. Only a few continuous-space based methods respect all of these criteria (Marcon and Puech, 2012).

In continuous space, the definition of m is similar to that of the cumulative M function (Marcon and Puech, 2010) except that the local ratio is defined at distance r and not up to it. In contrast with the topographic functions g and K, the cumulative function is not the integral of the density function over r (Ripley, 1977) because relative functions are not derived from a measure of space.

The m function can be interpreted as an extension to continuous space of the location quotient (Florence, 1972). It is not a smoothed Ellison and Glaeser's index: the latter relies on the squared difference between the local share of the sector of interest and that of the whole activity, not on their ratio.

3. Theoretical examples

We shall now provide simple examples for three theoretical cases. In every example we have considered a 1-by-1 window

[4] To give an example, if the aim is to evaluate the spatial distribution of the textile industry, the analysis of the distribution of textile plants around textile plants is relevant. In that case of intra-industrial analysis, the intratype function should be used. If the focus is now on the co-agglomeration of the textile and clothing sectors, the intertype functions will deal with the distribution of textile plants around clothing plants or the distribution of clothing plants around textile plants.

and a maximum distance for the *m* function equal to one-third of the diagonal of the window (≈ 0.471). We have systematically used 512 regular intervals to calculate the *m* function. In all theoretical examples, we simulate two distributions of points: the cases and the controls.[5] The cases are the points of interest. The controls are the points that constitute the benchmark. For simplicity, all points have a weighting of 1. A global confidence interval (CI) at the 1% risk level was generated after 10,000 simulations. All simulations are made with the help of the R package (R Development Core Team, 2014). **spatstat** (Baddeley and Turner, 2005) was used to realize the point processes and the **dbmss** package (Marcon *et al.*, 2014) was used to compute the *m* function.

3.1 Random distribution

Figure 1a shows two distributions: one for the cases (diamonds) and the other for the controls (crosses). The cases and controls were simulated under the hypothesis of complete spatial randomness (CSR), under which points are distributed randomly and independently from each other. To achieve this, we generated two distributions from a homogeneous Poisson process with a parameter respectively equal to 25 and 100. The parameter of the Poisson process is the expectation of the number of points for each distribution.[6] In figure 1a, 29 points were simulated for the cases and 103 for the controls.

Figure 1b depicts the *m* function results for this case. No significant result is observed: *m* fluctuates for all distance ranges but stays inside the confidence interval of the null hypothesis. As expected for random distributions of cases and controls, figure 1b provides no evidence of any attraction or repulsion between cases. Two additional minor comments should be made. First, the global confidence interval is quite large at small radii: there is a small number of neighbors at very small radii. Second, the *m* function is not defined for very small distances: this indicates that cases are separated by gaps of more than 0.01.

3.2 Aggregate distribution

Figure 2a shows a multiple pattern: an aggregate distribution of cases (diamonds) and a completely random distribution of controls (crosses). For the clusters of cases, we generated simulations from a Matérn process with the following parameters: 2 for the density of the Poisson process that generates cluster centers, 0.05 for the radius of clusters and 50 for the average number of points per cluster. Controls were simulated from a homogeneous Poisson process with a density of 100. 203 points were plotted on figure 2a: 102 cases shared between two clusters and 101 controls were randomly distributed over the entire domain. Figure 2b shows the results for the *m* function.

On figure 2b, two significant concentration peaks are apparent. They occur at distances for which the relative local density of cases is the greatest. The first distance at which a peak is observed corresponds approximately to the radius of the clusters (around 0.05). This is due to the presence of controls in the ring at this distance, the peak is not reached exactly at a distance of 0.05 but only approximately. The second peak identifies the distance between clusters (approximately 0.5) and has a lower value. The local relative density of cases over controls is greater for the first peak because the presence of controls in the cluster is possible but rare.

Three additional comments have to be made. First, by construction there are no cases between the aggregates. For these distances, the maximum dispersion is detected: between the clusters the *m* function attains its lowest possible value (zero). The rapid decrease in the gradient of *m* is a feature of density functions. In contrast to cumulative functions, the values are very sensitive and large ranges of results may be observed over small intervals of distance. Second, one can observe that the *m* plot takes on its highest values in the case of small distances (first concentration peak) and then decreases. The explanation for this is simple. In the first radii, the local relative density is the greatest because the maximum number of cases is observed around these distances. Around a distance of 0.3, the *m* function detects the first cases located at the periphery of the (other) cluster: as a result, the *m* function raises. Then the local relative density continues to increase rapidly because of the large number of cases inside the cluster. Third and last, it is interesting to note that the confidence interval of the null hypothesis is narrower in the inter-distance clusters, because more points are observed at these distances.

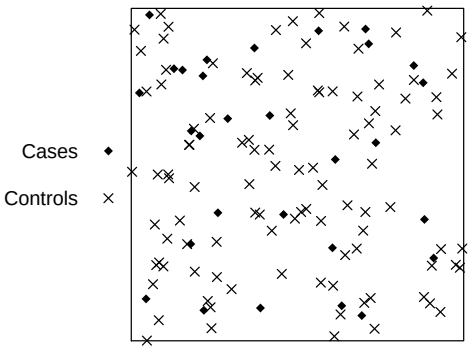
3.3 Regular distribution

Figure 3a shows another multiple pattern. A regular distribution of cases (diamonds) is clearly visible. 100 cases are positioned on a square grid measuring 0.1×0.1. The completely random distribution of controls (crosses) is a realization of a homogeneous Poisson process whose parameter is equal to 200. Figure 3a shows 209 controls. The *m* function estimates are given in figure 3b.

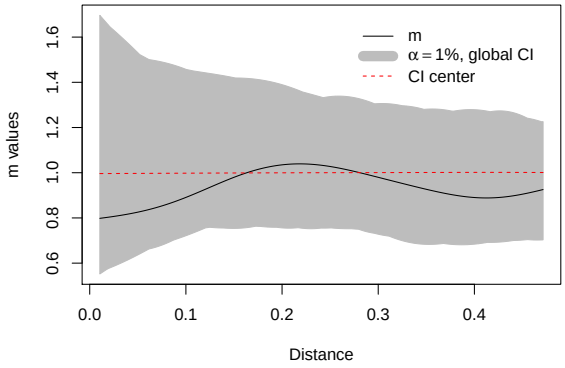
Up to the size of the square grid (0.1), the cases have no case neighbor: for small distances there is a significant amount of dispersion. Then, a large number of peaks can be observed but results are not significant. The reason is simple. In this example as we previously said we retained the optimal bandwidth as described by Silverman (1986). A thinner bandwidth would have shown significant positive and negative peaks. The choice of the bandwidth is important, unfortunately "*in general, however, no simple recipe for the choice of the bandwidth exists*"(Illian *et al.*, 2008, p.115). Let us take half of the previous bandwidth to better explain the spatial structure under study. Results of the *m* function are

[5] The vocabulary "cases" and "controls" is well established in the literature of point processes (Diggle, 1983; Arbia *et al.*, 2012) but some authors as Billings and Johnson (2012) prefer employing respectively "samples" and "counterfactuals".

[6] The Poisson process is commonly used for simulating CSR distributions. As Diggle (1983) wrote the Poisson process "*is the cornerstone on which the theory of spatial point processes is built. It represents the simplest possible stochastic mechanism for the generation of spatial point patterns, and in applications is used as an idealized standard of complete spatial randomness* (...)" (p.50).

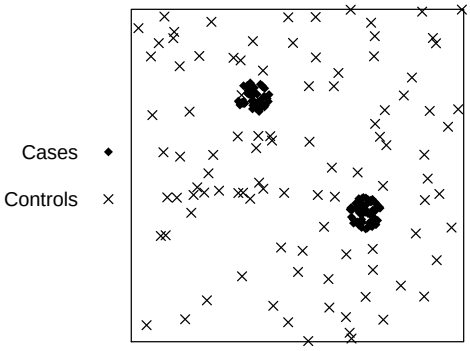


(a) Completely random distributions for cases and controls

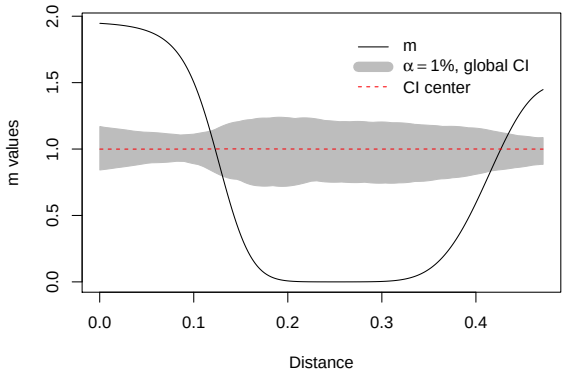


(b) *m* function results

Figure 1. Random distribution

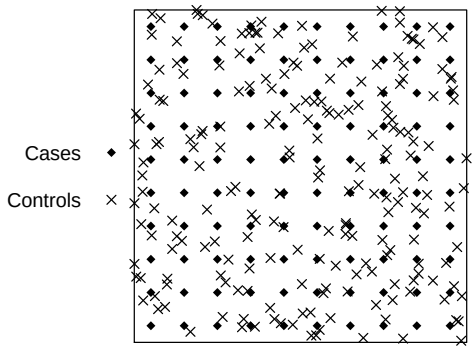


(a) Aggregate distribution of cases, complete spatial randomness for controls

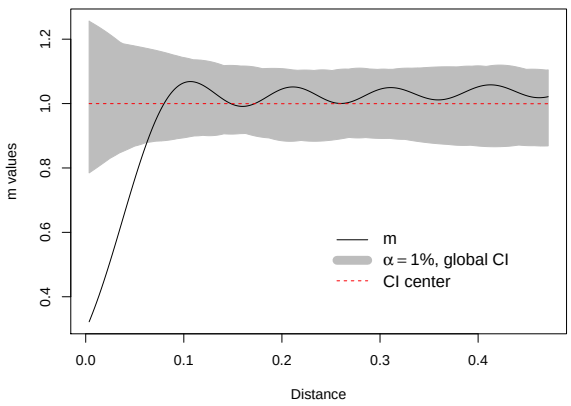


(b) *m* function results

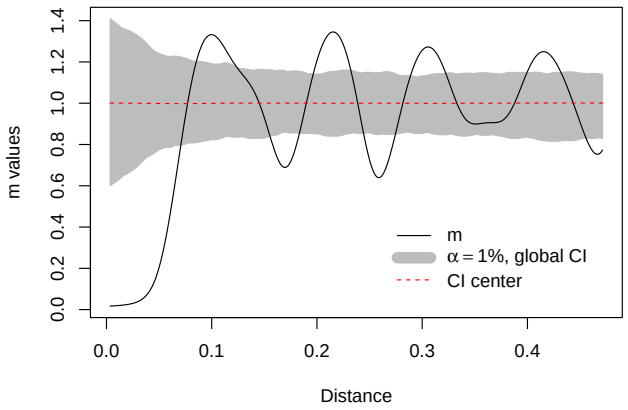
Figure 2. Aggregate distribution



(a) Regular distribution of cases, complete spatial randomness for controls



(b) *m* function results, original bandwidth of Silverman (1986)



(c) *m* function results, half of the original bandwidth of Silverman (1986)

Figure 3. Regular distribution

given in figure 3c. Up to the size of the square grid (0.1), m plots are the same whatever the definition of the bandwidth. At a distance equal to the size of the grid, the cases have four neighbors: a significant positive peak is observed in figure 3c, indicating the spatial concentration of cases at this distance. At this distance, a positive peak is also detected in figure 3b but the m plot stays within the confidence interval. Due to smoothing, significant values of m can also be observed just below the grid size (0.1). The m plot then plummets when the radius increases: no cases are located in the close environment of cases, the m value returns rapidly to below the confidence interval, indicating dispersion. On figure 3c, the irregularity in the gradient of m between a distance of 0.10 and 0.15 is interesting. At a distance equal to the diagonal of the grid (around 0.141) new neighbors are present. The irregularity in the gradient of m shows the existence of these new neighboring points. However, there is no positive peak because the smoothing we applied was too strong, but weaker smoothing would have generated positive m values. The original bandwidth appears too weak in that case: the irregularity in the gradient of m is not visible between a distance of 0.10 and 0.15 in figure 3b. Note at larger distances, the observed positive peaks appear at around twice the grid size (0.2), three times the grid size (0.3) etc. Between these peaks, m indicates the absence of neighboring cases: depending on the distance, significant dispersion may (as in the case for a distance of 0.25) or may not (as in the case for a distance around 0.35) be observed.

4. Discussion

This section provides some comparisons with the most used density function in spatial economics, Duranton and Overman's K_d function (Duranton and Overman, 2005). Keeping the notations previously used in equation 1 and using n to denote the total number of points, the K_d function is defined by:

$$\hat{K}_d(r) = \frac{1}{n(n-1)} \sum_{x_i \in \mathscr{R}} \sum_{x_j \neq x_i, x_j \in \mathscr{N}} k\left(\left\|x_i - x_j\right\|, r\right) \quad (2)$$

The weighted version of the K_d function, called the K^{emp} function (Duranton and Overman, 2005), is given by:

$$\hat{K}^{emp}(r) = \frac{\sum_{x_i \in \mathscr{R}} \sum_{x_j \neq x_i, x_j \in \mathscr{N}} w(x_i) w(x_j) k\left(\left\|x_i - x_j\right\|, r\right)}{\sum_{x_i \in \mathscr{R}} \sum_{x_j \neq x_i, x_j \in \mathscr{N}} w(x_i) w(x_j)} \quad (3)$$

The K_d function is very popular in spatial economics (see Marcon and Puech, 2012, for a review). It is therefore interesting to compare their properties in order to understand the main differences between the two statistical measures. In what follows, we have used the R package **dbmss** to estimate the K_d and m functions. As suggested by Duranton and Overman (2005), we use the reflection technique to estimate density close to the lowest distance for the K_d function. As a result, K_d plots start systematically for $r = 0$.

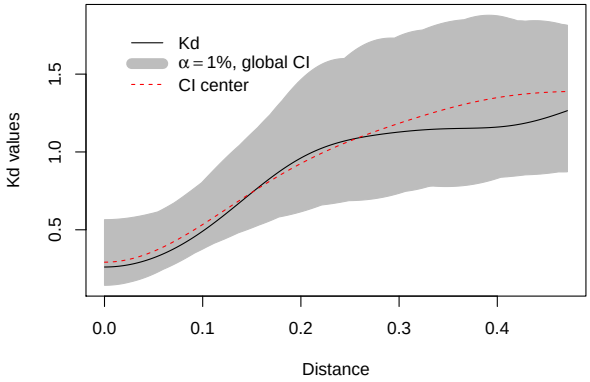
4.1 Comparisons of K_d and m results on the three previous theoretical cases

The K_d function was estimated for the three theoretical cases considered above. K_d plots are shown on figure 4a for the random distribution of cases, on figure 4b for the aggregate distribution and on figures 4c and 4d for the regular distribution. For all the cases, we have used the same maximum distance, *i.e.* one-third of the diagonal of the window (≈ 0.471).

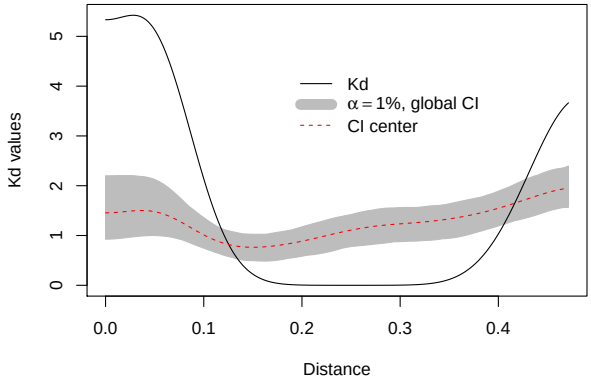
For the random distributions of cases and controls (figure 4a), m and K_d give identical results. No significant level of dispersion or concentration was detected. The value of m is 1 for all distances, subject only to stochastic fluctuations. Relative distance-based measures do not suffer edge-effects: points close to the domain borders have less neighbors but this issue cancels out when the ratio of their numbers is calculated. In contrast, K_d increases with distance for geometrical reasons. K_d evaluates the probability of finding a case neighbor at a given distance, *i.e.* on the circle of radius r around each point of interest: it first increases linearly with respect to r, then increases less because of edge effects (parts of the circles lay outside the domain when r is large enough) and finally drops to 0 when r gets larger than the diameter of the domain (not shown on the figure). Finally note a very minor difference in the results provided by K_d and m: the K_d plot is defined for $r = 0$ contrarily to the m plot (figure 1b). This is due to the reflection method used for the K_d function as we previously explained. No case-neighbor is located at a distance less than 0.01 thus m is not defined.

For the aggregate distribution (figure 4b), like the m function, K_d detects the first peak of concentration occurring at a distance of approximately 0.05. The main difference relates to the shape of the concentration peak. At very small distances the K_d values increase up to a distance of 0.05 which corresponds to the radius of the cluster. After this, the value of K_d starts to fall. The increase in K_d contrasts with the shape of the m function at short distances. The explanation is geometric again: at short distances this probability increases proportionally to the perimeter of the circle around reference points until r is too large and the circle partly leaves the cluster. Then, it progressively decreases.

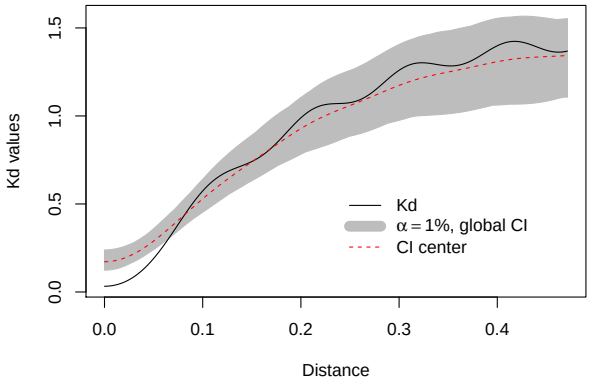
Let us now turn to the regular theoretical example. With the same original smoothing of Silverman (1986) also chosen by Duranton and Overman (2005), the results for this spatial pattern with the m (figure 3b) or the K_d (figure 4c) function are totally in accordance. If we modify the smoothing by choosing a narrower bandwidth, that is half of the original bandwidth of Silverman (1986), the results for the K_d function are given in figure 4d and should be compared with the m results given in figure 3c. As one would expect, there are a large number of positive and negative significant peaks of the K_d plot in comparison to weaker smoothing of the results.



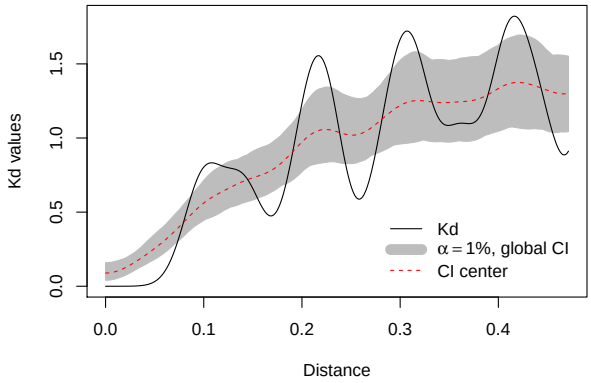
(a) K_d results for completely random distributions of cases and controls (map on figure 1a)



(b) K_d results for aggregate distribution of cases (map on figure 2a)



(c) K_d results for regular distribution of cases (map on figure 3a) with the original Duranton and Overman's smoothing



(d) K_d results for regular distribution of cases (map on figure 3a) with a more detailed smoothing

Figure 4. K_d results for the three theoretical cases

The m results and the K_d results are, again in that example, totally in accordance.

4.2 Comparisons between the results from the K_d and m functions on a more complex example

In the three above theoretical cases, there is not a great deal of difference between the results with the K_d and m functions. However, this is not always the case. In the real world, the distribution of activities is more complex. In this sub-section, we shall draw attention to the type of concentration that the m function can identify. To make things clearer, if we return to table 1 we can see that the m function evaluates the relative concentration while K_d appraises the absolute concentration. This distinction may be crucial for a comprehensive understanding of spatial structures.

4.2.1 Divergence in the K_d and m results on a theoretical case

Consider the following theoretical example. A city is delimited by a 1-by-1 window. For the sake of simplicity "cases" and "controls" are the only two types of shops located in the city. Figure 5 shows the distribution of cases (diamonds) and controls (crosses). A multiple pattern is observed: a completely random distribution of controls (crosses) and cases (diamonds) and also an aggregate distribution of controls (crosses). More technically, for the cluster of controls we generated simulations from a Matérn process with the following parameters: 1 for the density of the Poisson process that generates cluster centers, 0.1 for the radius of clusters and 75 for the average number of points per cluster. Controls were also simulated from a homogeneous Poisson process with a density of 50. Cases were simulated from a homogeneous Poisson process with a density of 25. 134 points were plotted on figure 5. The cluster is composed of 66 controls, 42 controls are randomly distributed on the area and 26 cases are randomly distributed over the entire domain.

Figures 6a and 6b show the results for the K_d function and the m function. In line with our previous examples, all points have a weighting of 1. The maximum distance for the m function equals one-third of the diagonal of the window (≈ 0.471). A global confidence interval (CI) at the 1% risk level was generated after 10,000 simulations. All simulations were conducted with the help of the R package the **dbmss** package for computing the K_d and m functions. The divergence in the results between K_d and m plots show all the importance of the definition of the nature of the spatial concentration studied.

As we underlined, the m function detects the relative spatial concentration. Cases are more regularly distributed than the controls in the example. As a consequence, around the cases the presence of cases is *relatively* more important than the one of controls: relative concentration is detected by the m function up to a distance approximately equal to 0.23 (figure 6b). Now, consider the K_d results (figure 6a): dispersion of cases is detected. By construction of the example, cases do not take place in the clusters. Under the null hypothesis of random location of cases among all observed points, some cases will be located in the cluster: the probability to find neighbors at short distances is thus higher than in the real data set: in other words, the points of the dataset are less concentrated than under the null hypothesis. The concentration characterized by K_d is called absolute because it just counts neighbors without comparing their number to a benchmark. In the dataset, cases are located in low-density areas so they are far from each other: a dispersion of cases is detected by K_d. However, they are relatively abundant and close to each other (in comparison with controls), cases are agglomerated in relative terms, as detected by the m function.

4.2.2 Confirmation of the previous results for a retail sector in the city of Lyon (France)

To give a concrete example of the previous theoretical case, we shall consider the spatial distribution of the non-food retail stores in the Lyon area (France). We have exploited a database provided by the Chamber of Commerce and Industry of Lyon. This contains the exact geographic position of 3,124 non-food stores in April 2012. The different types of store have been classified into 26 sectors from 47.30Z to 47.79Z of the French NAF rev. 2 classification of activities. We shall focus on "dispensing chemists in specialized stores" (47.73Z), of which there are 156 in the Lyon area. We shall refer to these stores as pharmacies in what follows.

First of all, a comparison of the density of all non-food stores in Lyon (figure 7a) and the spatial distribution of the pharmacies over the same area (black points on figure 7b) is worthwhile. We can see that many of the city's non-food stores are located in central Lyon and the left bank of Rhône river (to the east of central Lyon, figure 7a). However, pharmacies are undoubtedly more regularly distributed than non-food stores as a whole. One can easily observe the presence of these activities over the entire Lyon area (black points on figure 7b).

The impacts on the results for m and K_d will be of interest. These are given on figures 8a and 8b (distances are reported in meters on the horizontal axis). Their respective global confidence intervals (CI) were computed at the 1% risk level after 10,000 simulations. All the pharmacies were assigned a weight of 1 and the maximum distance analyzed was around 2,500 meters. The spatial structures detected by K_d and m differ. Up to a distance of approximately of 2 kilometers, the plot of K_d indicates that pharmacies are dispersed while that of m indicates a degree of spatial concentration up to 1.5 kilometers. As we underlined, pharmacies more regularly distributed than non-food retail activities as a whole. As a result, pharmacies are more concentrated under the null hypothesis than in the real distribution: a dispersion of pharmacies is detected by K_d. Moreover, even though there are pharmacies in high density business areas (central Lyon, the left bank of the River Rhône…), they are over-represented in low-density business areas, relatively to other shops. When we simulate distributions, pharmacies are located in areas where the number of non-retail stores is greater so the relative concentration of these stores will be lower under the null hypothesis. As a

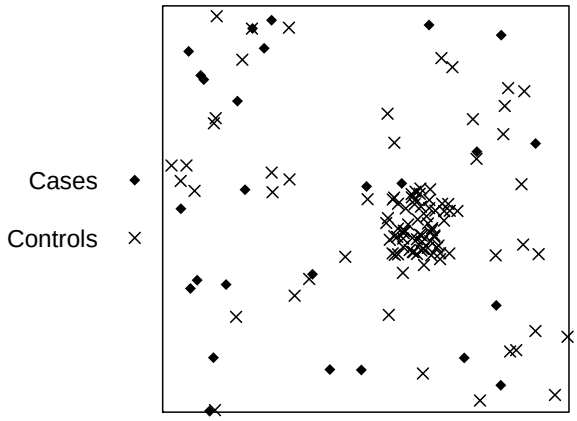
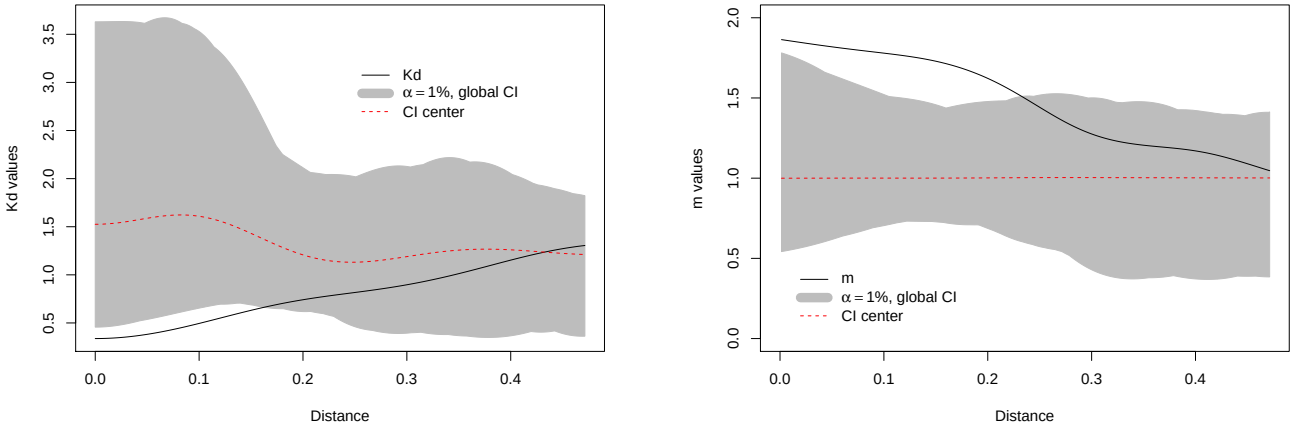


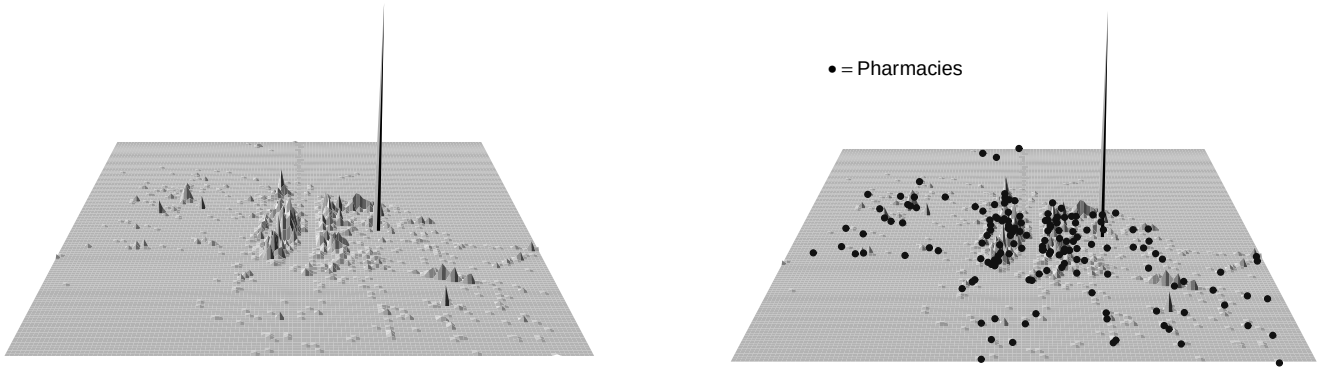
Figure 5. A more complex theoretical case



(a) K_d function results

(b) m function results

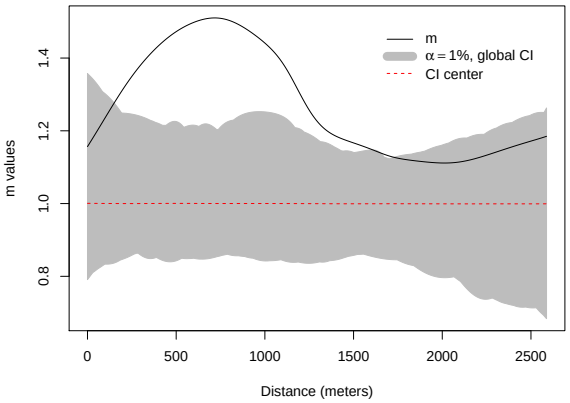
Figure 6. K_d and m results for a more complex theoretical case



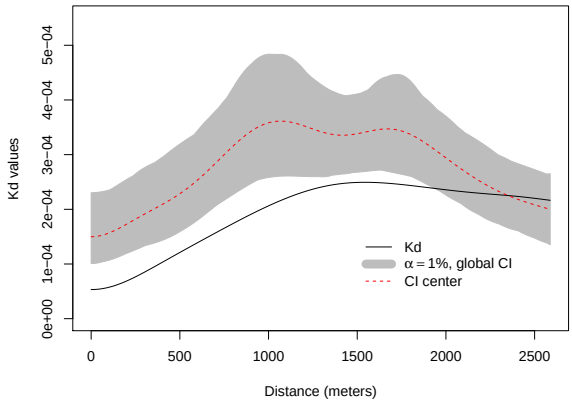
(a) Density of the non-food retail stores

(b) Spatial distribution of pharmacies in Lyon

Figure 7. Non-food retail stores in the area of Lyon (France)



(a) m results for pharmacies in Lyon



(b) K_d results for pharmacies in Lyon

Figure 8. Spatial structure of the pharmacies

result, the observed m plot is over the confidence interval of the null hypothesis and indicates a relative spatial concentration of pharmacies between approximatively 250 meters and 1,500 meters. This comparison of results emphasizes that the nature of the spatial concentration should be studied with care. Our conclusion is in line with Brülhart and Traeger (2005) or Marcon and Puech (2015) findings.

Conclusion

In this article, we introduced a new distance-based function called m. The m function is the first relative density function to be proposed in the field of spatial economics. It respects all of the good criteria of Duranton and Overman (2005) for evaluating the spatial distribution of economic activities. So, m will certainly be useful for economists for detecting the relative spatial concentration or dispersion of activities. For example, any density function detects local patterns more precisely than cumulative functions so m can be preferred in that case to relative cumulative distance-based measures. Moreover, we showed on theoretical and empirical data that m and the leading density function K_d of Duranton and Overman (2005) may be used conjointly for having a comprehensive approach of the distribution of activities. The main reason is that K_d evaluates the spatial absolute concentration in continuous space while m evaluates the relative one. At the end of the article, the analysis of the distribution of the pharmacies in Lyon provides a good example of the complementary of the results of K_d and m.

Acknowledgments

We are grateful to the Lyon Chamber of Commerce and Industry and Christophe Baume for providing us with data on stores. We thank those who took part in the *13th International Workshop on Spatial Econometrics and Statistics* (Toulon, France), the *Workshop on spatial statistics: Measurement of spatial concentration and its applications* (Sceaux, France) and the *61st Annual North American Meetings of the Regional Science Association International* (Washington D.C., USA) for their valuable comments and discussions. We gratefully acknowledge funding from the AAP Attractivité 2014 (Université de Paris-Sud). This work was supported by the Paris-Saclay Center for Data Science funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02 and an "Investissement d'Avenir" grant managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01).

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, 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.

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.

Billings SB, Johnson EB (2012). "A non-parametric test for industrial specialization." *Journal of Urban Economics*, **71**(3), 312–331.

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.

Briant A, Combes PP, 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.

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.

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

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

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

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.

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

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.

Guimarães P, Figueiredo O, Woodward D (2009). "Dartboard tests for the location quotient." *Regional Science and Urban Economics*, **39**(3), 360–364.

Henderson JV, Thisse JF (2004). *Handbook of Urban and Regional Economics*. Elsevier. North Holland, Amsterdam.

Howard E, Newman C, Tarp F (forthcoming). "Measuring industry coagglomeration and identifying the driving forces." *Journal of Economic Geography*, pp. 1–24, doi:10.1093/jeg/lbv037.

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.

Kerr WR, Kominers SD (2015). "Agglomerative Forces and Cluster Shapes." *The Review of Economics and Statistics*, **97**(4), 877–899.

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.

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

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.

Marcon E, Lang G, Traissac S, Puech F (2014). "dbmss: Distance-based measures of spatial structures." URL http://cran.r-project.org/web/packages/dbmss/.

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 (2012). "A typology of distance-based measures of spatial concentration." *HAL SHS*, **00679993**(version 1).

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

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.

Mori T, Smith TE (2013). "A probabilistic modeling approach to the detection of industrial agglomerations." *Journal of Economic Geography*, **14**(3), 547–588.

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.

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

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.

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.

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.

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

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*, CRC Handbooks of Modern Statistical Methods Series, chapter 22, pp. 403–423. Chapman & Hall.