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[1] **Eric Marcon**[2] **Florence Puech**[3]

ABSTRACT

For a decade, distance-based methods have been largely employed and improved in the field of spatial economics. Such tools are very powerful to evaluate accurately the spatial distribution of plants or retail stores for example (Duranton and Overman, 2008; Jensen and Michel, 2011). In the present paper, we introduce a new statistic measure based on distances to evaluate the spatial concentration of economic activities. As far as we know, the *m* function is the first relative density function proposed in the economic literature. This tool completes the typology of distance-based methods recently drawn up by Marcon and Puech (2014). By working on several theoretical and empirical examples, we prove the advantages and the limits of the *m* function to gauge the spatial structures in spatial economics.

Keywords: Spatial concentration, Aggregation, Point patterns, Agglomeration, Economic geography.
JEL Classification: C10, C60, R12

Acknowledgments:
We are grateful to the Lyon Chamber of Commerce and Industry and Christophe Baume for providing us with data on stores. We thank the participants to the *13th International Workshop Spatial Econometrics and Statistics* (Toulon) and to the *Workshop on spatial statistics: Measurement of spatial concentration and its applications* (Sceaux) for their valuable comments and discussions. We gratefully acknowledge funding from the AAP Attractivité 2014 (Université de Paris-Sud).

[1] AgroParisTech, UMR 518 Mia - Mathématique et Informatique appliquées, 19 avenue du Maine, 75732 Paris Cedex 15, France. Email: Gabriel.Lang@agroparistech.fr

[2] AgroParisTech, UMR EcoFog - Ecologie des Forêts de Guyane, Campus agronomique BP 316, 97379 Kourou Cedex, France. Email: Eric.Marcon@ecofog.gf

[3] Corresponding author, Université de Paris-Sud RITM, 54 Boulevard Desgranges, 92331 Sceaux Cedex, France. Email: Florence.Puech@u-psud.fr

1. INTRODUCTION

"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).

This sentence of Gilles Duranton in the *"Spatial Economics"* article of The New Palgrave Dictionary of Economics underlines all the importance for proposing new statistic tools that allow an accurate description of the spatial distribution of economic activities. Industrial agglomerations are certainly the main feature of the present economic geography (Krugman, 1991; Henderson and Thisse, 2004). Thus, it is not surprising that a lot of research has been recently devoted to improve the measurement of the spatial concentration of activities (see Duranton and Overman, 2005; Marcon and Puech, 2003, 2010; Arbia *et al.*, 2012; Mori and Smith, in press, among others).

In the present paper we propose a new tool for evaluating the spatial structures in economics: a relative density function called *m*. This new measure belongs to the distance-based methods that are the latest statistical measures proposed in the field of spatial economics (Combes *et al.*, 2008). Such measures are now privileged in our field because they do not rely on any zoning contrarily to the Gini or Ellison and Glaeser (1997) indices for example. It has been proved (Arbia, 2001; Briant *et al.*, 2011) that such indices are sensitive to the zoning chosen as described by the Modifiable Areal Unit Problem – MAUP (Openshaw and Taylor, 1979; Arbia, 1989). Thus it should be better to treat space as "continuous". Working directly on the position of entities (shops, plants…) allows an exact and unbiased analysis of the spatial structure of the distribution at all scales simultaneously (and not at only one level of observation as the zoning-space approach does). Multiple patterns can be detected: for example aggregation or repulsion between entities according to the distance considered. These functions are thus very powerful and a lot of studies now use distance-based methods 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, 2013; Koh and Riedel, 2014, Giuliani *et al.*, in press).[4]

The introduction of the *m* function is important for at least two reasons. First, our paper fills a gap in the economic literature because as far as we know no relative density function was yet proposed

[4] Marcon and Puech (2014) provide a survey of studies employing a distance-based method for evaluating spatial patterns.

to gauge the spatial concentration in continuous space as Marcon and Puech (2014) noticed. A great attention was devoted in defining the proposed *m* function to respect a maximum of the good criteria for the measurement of spatial concentration in economics (Duranton and Overman, 2005; Combes and Overman, 2004). Second, various theoretical and empirical examples are provided to show that the results brought by the *m* function are complementary to the widely used Duranton and Overman's *Kd* function. The latter distance-based method is also a density function but not a relative one. As a result, the *m* function is a relevant statistic tool to detect spatial structures in economics.

In the following sections, at first our motivation is explained (II). Then, the *m* function is presented (III) and some simple illustrative examples are given (IV). In the last section of our paper (V) comparisons with the *Kd* function are proposed. We provide simple theoretical examples and also an analysis of the spatial distribution of pharmacies in the Lyon area is done to illustrate the advantages and the limits of the use of the *m* function in our field

2. MOTIVATION

As we underlined in the introduction, a lot of efforts have been devoted during the last decade to introduce or improve distance-based methods in the field of economics (see Duranton and Overman, 2005; Arbia *et al.*, 2009; Marcon and Puech, 2010; Arbia *et al.*, 2012). These methods are particularly attractive because they provide a complete analysis of the location patterns of industries.

The basic idea of the distance-based methods is simple. Consider the case of the textile industry. The evaluation of the spatial distribution of plants in this sector rests on an assessment of the surroundings of textile plants for all distances considered. If there are locally more textile plants around textile plants than in general on the whole territory analyzed then a phenomenon of *concentration* is detected (textile plants attract textile plants). On the opposite if in the surroundings of the textile plants there are less textile plants than we find on average on the whole territory, then a phenomenon of *dispersion* will be observed (in that case textile plants repulse textile plants). At last, if there is no relation between entities, a phenomenon of *independence* is identified (textile plants are randomly and independently distributed). The significance of the results is provided by the generation of 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 notions require additional explanations: the definition of the surroundings of plants and the nature of the spatial concentration (topographic, relative or absolute). Let us develop these two important features.

Firstly, the **notion of the surroundings of plants** is central because it defines the type of the function retained *i.e.* cumulative or density functions. Practically, 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 of *at* a given distance. If the first option is chosen, it calls for a cumulative function. On the opposite, if the second one is retained, a density function is appropriate. The choice between the one or the other type of functions depends on question under study (Marcon and Puech, 2010).

The second clarification concerns **the nature of the spatial concentration.** The evaluation of the spatial concentration of economic activities implies to choose a reference value to which the observed distribution of activities will be compared to (Brülhart and Traeger, 2005).

- The first possibility is to retain a *topographic reference*. Here, the physical space is chosen as a benchmark (space may be homogenous or not). To give an example, it is the case of the number of neighboring plants located per unit of space (that is on a disk of radius r for a cumulative function or on the ring for a density function).
- The second possibility is a *relative reference*. In that case, another variable is taken as a benchmark. For instance, if we evaluate the spatial distribution of textile plants, we can detect in the close environment of the plant the deviations of this distribution of plants to another distribution. The benchmark plants could be all plants at the aggregate industrial level.
- The last possibility is an *absolute reference*. In that case, there is no reference value. For example, the absolute number of neighboring textile plants located at a given distance from a textile plant.

The growing number of measures in continuous space has incited us to recently provide a typology of those functions (Marcon and Puech, 2014). We proved that a classification of statistical measures can be drawn by considering only the nature of the concentration and the definition of the type of the function. Table 1 gives an overview of all the distance-based measures that have been used to gauge the spatial distribution of the 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 (unamed) cumulative function of K_d proposed by Berhens and Bougna (2013).

We can easily see in table 1 that one cell is empty: no relative density function has yet been proposed in the field of spatial economics. The present paper fills this gap. In the next section we will complete the table 1 by proposing a new density function, named the *m* function, which detects the relative spatial concentration.

Table 1: Typology of distance-based measures (Marcon and Puech, 2014)

Function choice	Topographic, homogenous	Topographic, inhomogenous	Absolute	Relative
Density functions	g	g_{inhom}	K_d	
Cumulative functions	K K_{mm}	K_{inhom} D	*Cumulative* K_d	M

Finally, it is important to note that the field of applications of distance-based methods is not confined to spatial economics. They have been developed and applied at first in other sciences. For example numerous empirical researches have been made with these tools in ecology (Law *et al.*, 2009) or in epidemiology (Waller, 2010).

3. Presentation of the *m* function

As a relative measure of concentration of dispersion, *m* considers the ratio between the number of neighbors of interest in the neighborhood of reference points and the number of all neighbors. This ratio is estimated from the observed data, and normalized by the same ratio measured on the whole data set.

The estimator of *m* is:

$$\hat{m}(r) = \sum_i \frac{\sum_{j,i\neq j} k\left(\left\|x_i - x_j\right\|, r\right) w(x_j^c)}{\sum_{j,i\neq j} k\left(\left\|x_i - x_j\right\|, r\right) w(x_j)} \Bigg/ \sum_i \frac{W_c - w(x_i)}{W - w(x_i)} \quad (1)$$

x_i designates reference points, x_j neighbors. $k\left(\left\|x_i - x_j\right\|, r\right)$ is a kernel estimator function whose sum returns the density of neighbors of point x_i at distance *r*. Following Illian *et al.* (2008, chapter 4.3.3), we used the simple and efficient box kernel with bandwidth parameter *h*:

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

Loosely speaking, the kernel estimator counts the number of points in a ring of width 2*h* at distance *r* apart from the reference point x_i and returns a number of points per unit of ring width. $w(x_j)$ is the weight of point x_j. $w(x_j^c)$ is the weight of neighbors of interest, *i.e.* it equals 0 if point x_j does not belong to the chosen point type. W_c is the total weight of the points x_j^c and W the total weight of all points.

We focus on the intratype *m* function in this paper, but the intertype function is defined the same way, taking reference points x_i and neighbor points x_j^c in different point types.

The equation reads as follows. The numerator is the *local ratio*: the relative weight of neighbors of interest at distance *r* from all reference points. It is summed over all reference points (actually, it is summed because the number of reference points simplifies with the denominator). The denominator is the same ratio over the whole data set, the *global ratio*. It is not just W_c/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 neighbor to each other. For this reason, the denominator is slightly different in the intertype function: $\sum_i \frac{W_c}{W - w(x_i)}$.

The choice of the bandwidth *h* is arbitrary. The wider it is, the smoother the estimator is.

The reference value is 1 for any distance *r*. This value is reached in case of independence of points. *m* values greater than 1 indicate spatial concentration of points while *m* values lower than 1 detect dispersion. *m* values can be interpreted. For example if at a distance *r*, the *m* function reaches 1.5, it means that the proportion of the neighbor points of interest at this distance is 50% higher than on the whole territory.

The significance of the *m* estimates is given by the generation of a confidence interval of the null hypothesis (Monte Carlo simulations). This technique is widely developed for distance-based methods (see Marcon and Puech, 2014). Practically, random distributions of points are generated by permuting the marks (type and weight couples) of the actual points on the actual spatial positions of points (coordinates). We only retain a global confidence interval, following Duranton and Overman (2005).

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 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 fulfills the criteria of Duranton and Overman mentioned in the introduction: (i) it compares the geographic concentration results across industries, (ii) controls for industrial concentration (indirectly, comparing its values to the confidence envelope of the appropriate null hypothesis), (iii) controls for the overall aggregation patterns of industries, (iv) tests the significance of the results (thanks to the confidence interval) and, (v) keeps the empirical results unbiased across geographic scales.

4. THEORETICAL EXAMPLES

Simple examples are now provided on three theoretical cases. In every example we retain a 1-by-1 window and a maximal distance for the *m* function equal to the half of the diagonal of the window (~0.707). We systematically use 64 regular intervals to calculate the *m* function. All point weights equal 1. A global confidence interval (CI) at the 1% risk level is generated with 10,000 simulations. All simulations are made with the help of the `R` (R Development Core Team, 2014) package **`spatstat`** (Baddeley et Turner, 2005) for the realizations of the point processes and the `R` package **`dbmss`** (Marcon *et al.*, 2014) for the calculation of the *m* function.

a) Random distribution

Figure 1 shows two distributions: one for the cases (black points) and the other for the controls (grey points). Cases and controls are simulated from a homogeneous Poisson process of parameter respectively equal to 25 and 100. 29 points are simulated for cases and 103 for the controls. Figure 2 depicts the *m* function results associated to this case.

Figure 1: Complete spatial random distributions for cases and controls

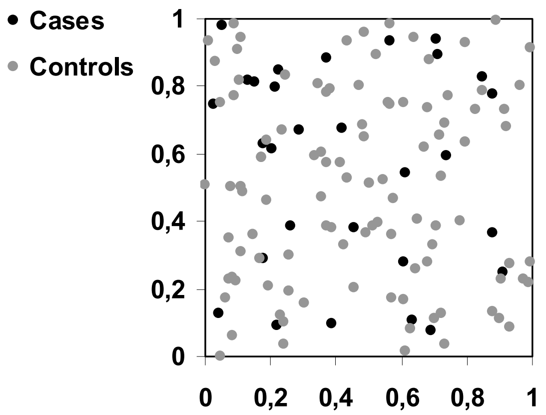
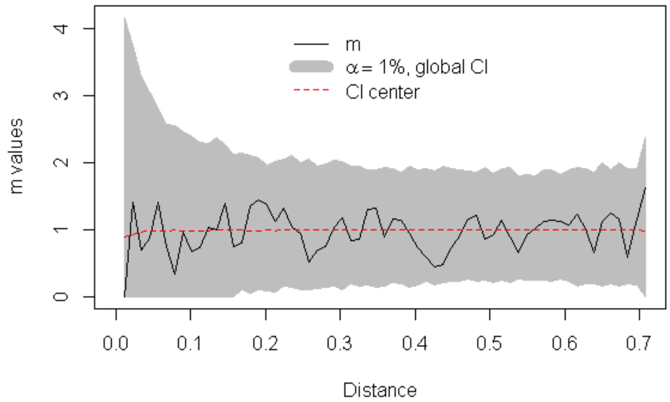


Figure 2: *m* function results



No significant result is observed: for all distance ranges *m* fluctuates but stays inside the confidence interval of the null hypothesis. As expected for random distributions of cases and controls, there is no evidence of any attraction or repulsion between cases on Figure 2.

Two minor comments can be done at small distances. First, there is a little number of neighbors at very small radii: the global confidence interval is quite large. Second, the lower band of the confidence interval is equal to zero up to 0.15. For more than 0.5% of the simulated distributions, cases are separated with gaps larger than 0.15; for these distribution, *m* is equal to zero up to the distance 0.15, so that the lower band of the confidence interval is zero.

b) <u>Aggregate distribution</u>

Figure 3 shows a multiple pattern: an aggregate distribution of cases (black points) and a complete spatial random distribution of controls (grey points). For the clusters of cases, we generate simulations from a Matérn process of 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 are simulated from a homogeneous Poisson process of density equal to 100. On Figure 3, 203 points are present: 102 cases shared out among two clusters and 101 controls are randomly distributed on the entire domain. Figure 4 depicts the results of the *m* function.

Figure 3: Aggregate distribution of cases, complete spatial randomness for controls

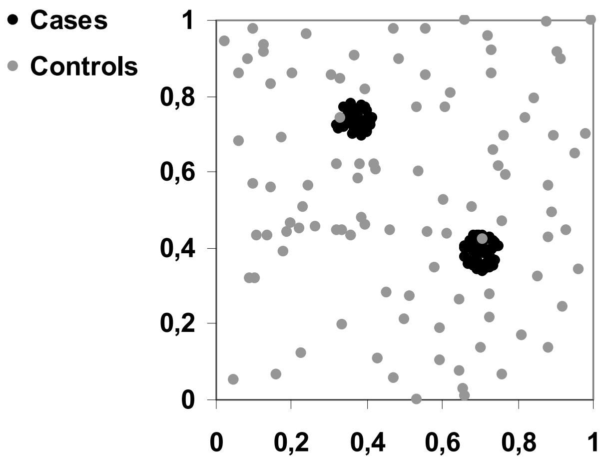
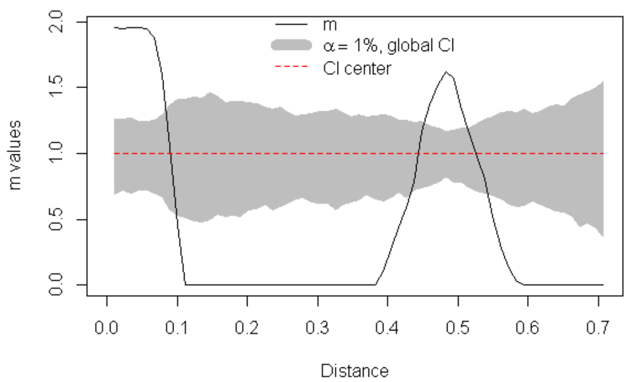


Figure 4: *m* function results



On Figure 4, two significant peaks of concentration are detected. They occur exactly at distances at which the relative local density of cases is the greatest. The first distance at which the peak is observed corresponds to the radius of the clusters (0.05). The second peak identifies the distance between clusters (approximately 0.5). This peak reaches a lower value. This is due to the presence of controls in the ring at this distance. 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 remains anecdotic.

Three additional comments have to be made. First, by construction there is no case between the aggregates. For these distances, the maximal dispersion is detected: between the clusters the *m* function reaches its lowest possible value (zero). The rapid decreasing of the *m* slope is a feature of the density functions. Contrarily to cumulative functions, values are very sensitive and large ranges of results may be observed in small intervals of distances. Second, one can observe that the *m* plot is constant for the first peak and then decreases. In the first radii, the local relative density is the greatest because around cases at these distances a maximum of cases is observed. This is not visible for the second peak: the *m* plot increases and then decreases. The explanation is simple. Around a distance of 0.4, the *m* function detects the first cases locating at the periphery of the (other) cluster: as a result, the *m* function rises. Then the local relative density continues to grow rapidly because cases are numerous inside the cluster. The *m* function decreases when the ring is greater than the inter-distance between both clusters. Third and last, it is interesting to note that the confidence interval of the null hypothesis is thinner at the inter-distance clusters: more points are observed at theses distances.

c) **Regular distribution**

Figure 5 shows another multiple patterns. A regular distribution of cases (black points) appears clearly. 100 cases are positioned on a square grid of 0.1x0.1. The complete spatial random distribution of controls (grey points) is a realization of a homogeneous Poisson process of parameter equal to 200. On Figure 6, 209 controls are present. The *m* function estimates are given in Figure 6.

Figure 5: Regular distribution of cases, complete spatial randomness for controls

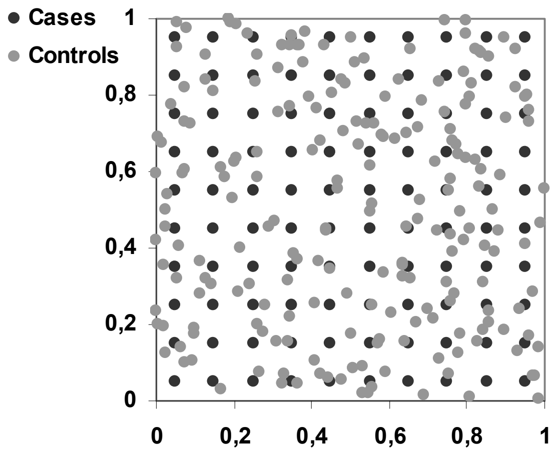
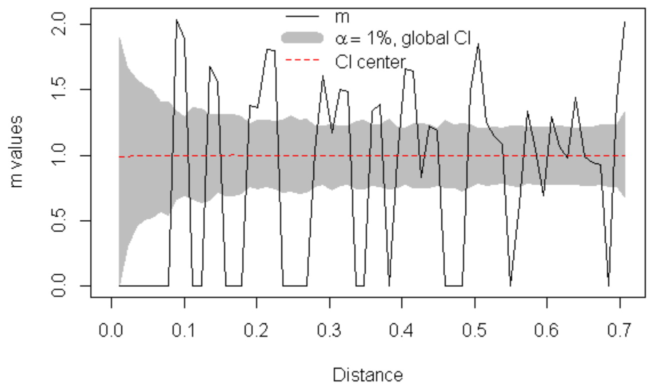


Figure 6: *m* function results



Numerous peaks of concentration and dispersion are observable. Up to the size of the square grid (0.1), every case has no case-neighbor: the local relative density of case-neighbors is equal to zero. Then at a distance equal to the size of the grid, any case has four neighbors: a positive peak is observed, detecting a phenomenon of spatial concentration of cases at this distance. Then, the *m* plot plummets when the radius increases: no cases are located in the close environment of cases, the *m* value returns rapidly to its minimum value (zero), detecting dispersion. Then, at a distance equal to the diagonal the grid (around 0.141) a second positive peak is observed. Four cases are detected: this is assimilated to a relative spatial concentration of cases. After the diagonal of the grid, the *m* plot rapidly falls to zero because the absence of case-neighbors leads to a maximum of dispersion. At larger distances, the observed positive and negative peaks have the same explanations. It is interesting to note that the *m* function is very sensitive: numerous peaks are detected, large and rapid variations in the results are visible. Two main reasons can explain it. First as we previously underlined, the density function results may vary a lot between two steps of computation. Thus they can identify precisely the local pattern and much more precisely as a cumulative function can do (Marcon and Puech, 2010). Second, in our example, a box kernel estimator is chosen: this also accentuates irregularities of the plot.

Finally, at very small distances, the large confidence interval of the null hypothesis is due to the small number of controls located near cases.

5. DISCUSSION

This section provides some comparisons with the density function the most used in spatial economics: the *Kd* function of Duranton and Overman (2005). By keeping the same previous notations of equation (1), the *Kd* function is defined by :

$$\hat{K}_d(r) = \frac{1}{n(n-1)} \sum_i \sum_{j,i \neq j} k\left(\left\| x_i - x_j \right\|, r\right) \tag{3}$$

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

$$\hat{K}^{emp}(r) = \frac{1}{\sum_i \sum_{j,i \neq j} w(x_i) w(x_j)} \sum_i \sum_{j,i \neq j} w(x_i) w(x_j) k\left(\left\| x_i - x_j \right\|, r\right) \tag{4}$$

The *Kd* function is very popular in spatial economics (see Marcon and Puech, 2014, for details). Thus comparing the results of the *m* function to those provided by the *Kd* function is useful to understand the main differences between those statistic measures. In what follows, we use the `R` package **`dbmss`** to estimate the *Kd* and *m* functions.

a) Comparisons of *Kd* and *m* results on the three previous theoretical cases

The *Kd* function was estimated for the three previous theoretical cases studied. *Kd* plot is shown on Figure 7 for the random distribution of cases, on the Figure 8 for the aggregated one and on the Figure 9 for the regular distribution of cases. The half of the maximum distance between points is systematically retained as it is suggested by Duranton and Overman (2005). As a result, the maximum value varies from one example to another.

To begin with, it is striking that the *Kd* results are less erratic than the *m* results for all examples. The kernel smoothing differs for *m* and *Kd*. As we previously underlined a narrow box kernel has been chosen for *m* while a Gaussian kernel is preferred for *Kd* (see Duranton and Overman, 2005). Future improvements of the *m* function will allow more smoothing.

For the random distributions of cases and controls (Figure 7), the results of *m* and *Kd* are identical. No level of significant dispersion or concentration is detected.

Figure 7: *Kd* results for complete spatial random distributions for cases and controls (map on Figure 1)

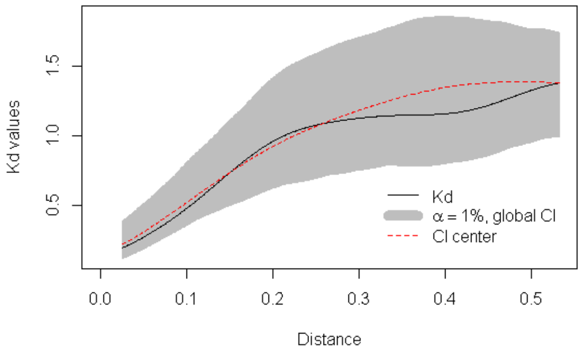


Figure 8: *Kd* results for aggregate distribution of cases (map on Figure 3)

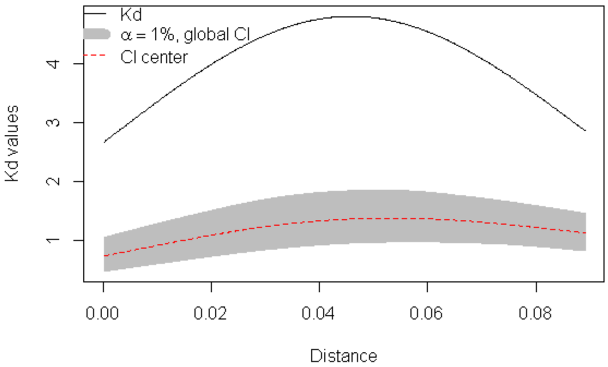
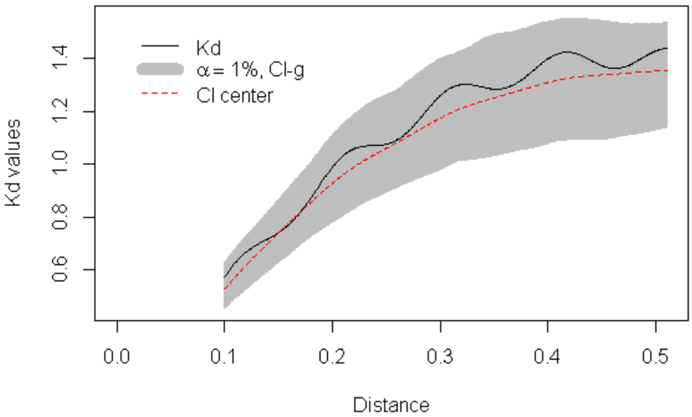


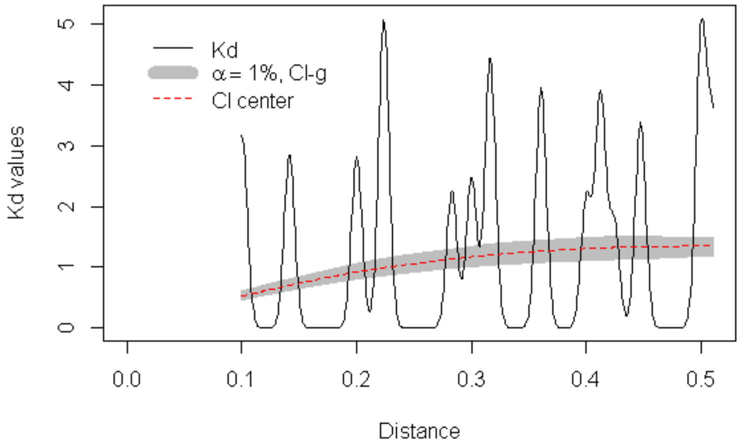
Figure 9: *Kd* results for regular distribution of cases (map on Figure 5) with the original Duranton and Overman's smoothing



For the aggregate distribution (Figure 8), as the *m* function *Kd* detects the first peak of concentration occurring at a distance approximately equal to 0.05. The main difference relies on the shape of the concentration peak. At very small distances *Kd* values increase up to a distance of 0.05 which corresponds to the radius of the cluster. Then, the value of *Kd* is declining. The increasing of *Kd* contrasts with the form *m* function at short distances. The explanation relies on the definition of *Kd* which evaluates the probability of finding one case-neighbor at a given distance. At short distances this probability increases because there are more and more case-neighbors observed. This probability reaches it maximum at the radius of the circle. Then, it progressively decreases since the radius of the circle increases.

The regular theoretical example pattern is more surprising. *Kd* detects some irregularities in the distribution but no significant results are observed on Figure 9. Here, results suffer from a too important smoothing of the results if we retain the original Duranton and Overman's technique. If we modify the smoothing by choosing a thinner bandwidth, then positive and negative peaks appear and results corroborate the *m* plot findings. Figure 10 provides such *Kd* plot with a thinner bandwidth (the risk level, the number of simulations, the type of the confidence interval remain unchanged). It is shown that the smoothing technique is of great importance for density functions. The latest are very sensitive to the type of the kernel estimator used as well as the bandwidth. Cumulative functions are undoubtedly less sensitive to the number of intervals on which they are computed. Finally, note that the estimates of the *Kd* function begins at a distance of $r = 0.1$ while *m* values can be estimated at lower distances. This is explained by the definition of *Kd*: no case-neighbor are located at a distance lower than 0.1 thus *Kd* is not defined. This is not the case for *m*: if there are no case-neighbors but control-neighbors are present, then the *m* function is defined and takes it lowest value (zero) as it is shown on Figure 6.

Figure 10: *Kd* results for regular distribution of cases (map on Figure 5) with a more detailed smoothing



b) Comparisons of *Kd* and *m* results for a retail sector in the city of Lyon (France)

In the three previous theoretical cases, *Kd* and *m* results do not differ greatly. However, it is not always the case. In real life, the distribution of activities could be indeed more complex. In that sub-section, we want to draw attention on the type of concentration that the function can identify. To make the things clearer, if we come back to the Table 1 we can see that the *m* function evaluates the relative concentration while *Kd* appraises the absolute one. This distinction should be crucial to understand spatial structures.

To give a concrete example, consider the spatial distribution of the non-eating retail stores in the area of Lyon (France). We exploit a database from the Chamber of Commerce and Industry of Lyon. It registers the exact geographic position of 3,124 non-eating shops in April 2012. Activities are classified in 26 sectors from 47.30Z to 47.79Z of the French NAF rev.2 classification of activities. Hereinafter, we focus on the "dispensing chemist in specialised stores" (47.73Z) which is composed of 156 stores. Hereinafter, this activity is renamed "pharmacies".

First of all, a comparison of the density of all non-eating stores in Lyon (Figure 11) and the spatial distribution of the pharmacies over the same area (back points on Figure 12) is worthwhile. We can see that the central area of Lyon (downtown) and the left bank of Rhône river (on east of the central Lyon) locate a lot of non-eating shops in the city (Figure 11). However, the distribution of pharmacies is undoubtedly more regularly located (Figure 12) than the global trend of non eating-stores. One can easily observe the presence of these activities all over the Lyon area (black points on Figure 12).

Figure 11: Density of the non-eating retail stores in the area of Lyon (France)

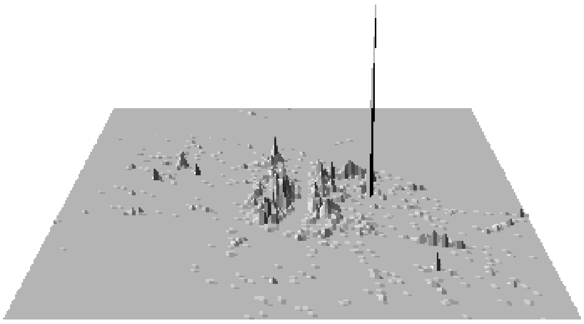
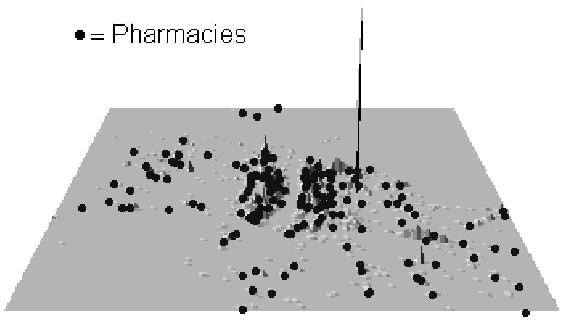


Figure 12: Spatial distribution of pharmacies in Lyon



The consequences on the *Kd* and *m* results will be of interest. They are given on Figures 13 and 14. Their respective global confidence intervals (CI) are computed at the 5% risk level with 10,000 simulations. All simulations are made with the help of the R package **dbmss** (R Development Core Team, 2014; Marcon *et al.*, 2014) for the calculation of the *Kd* and *m* functions. All weight of the pharmacies is equal to 1 and the maximum distance analyzed is around 2.5 kilometers.

The spatial structures detected by *Kd* and *m* differ. Up to a distance approximately of 2 kilometers, the *Kd* plot indicates that pharmacies are dispersed while *m* argues a certain degree of spatial concentration. These results are in line with the type of concentration they are looking for.

Remember that *Kd* evaluates the absolute concentration: it has no reference value for assessing the spatial concentration because it evaluates the probability of finding one case-neighbor at a distance *r* (in our example we are evaluating from any pharmacy the probability to find another pharmacy). What are the implications for our case-study? As we shown, this activity is more regularly distributed than global non-eating retail activities in the Lyon area. Thus, under the null hypothesis pharmacies are more concentrated than in the observed distribution (points are randomly labelled under the null hypothesis). As a consequence, *Kd* detects logically an observed dispersion of these stores in that sector.

Figure 13: *m* results for pharmacies in Lyon

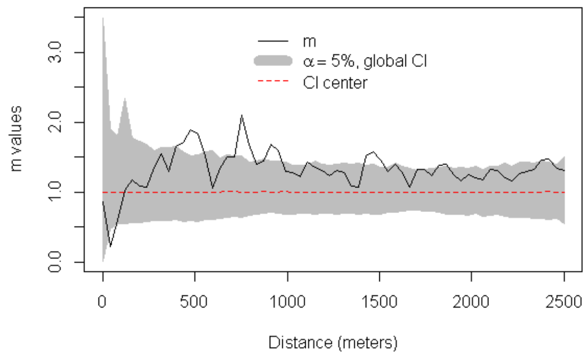
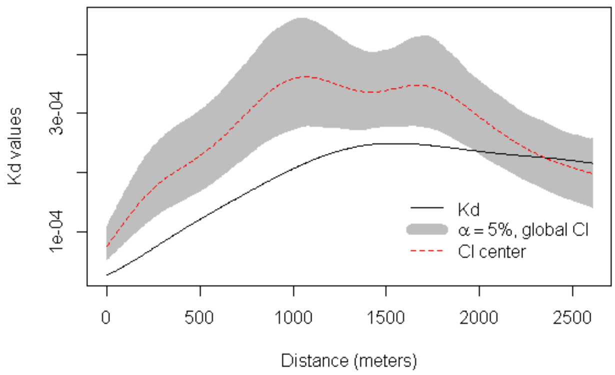


Figure 14: *Kd* results for pharmacies in Lyon



If we now turn to the *m* results, a spatial level of concentration is now detected. The reason is that *m* identifies the relative spatial concentration (and not the absolute one). Pharmacies are located in high density business areas (central area of Lyon, the left bank of Rhône river…) but particularly in low density business areas. In the latter, pharmacies are over-represented in the surroundings of these stores than under the null hypothesis. When we simulate distributions, these activities will be located in areas where the number of non-retailed shops is greater so the relative concentration of these stores will be lower under the null hypothesis. As a result, the observed *m* plot is over the confidence interval of the null hypothesis and indicates a relative spatial concentration of pharmacies around 500m or 750m for example. Note that a negative peak of the *m* function is observed at very short distances. This is explained by administrative constraints of the location of pharmacies: they must be regularly distributed at short distances (thus a significant relative dispersion is detected by the *m* function).

6. CONCLUSION

The first objective of this article was to introduce a new distance-based function called *m* which is a relative density-function. *m* is the first relative density function proposed in our field. In that sense, the *m* function will be undoubtedly useful for economists for detecting the relative spatial structures of any distribution. As any density function, *m* depicts local patterns more precisely than any existing cumulative function such as the *M* function of Marcon and Puech (2010). The second objective of the paper was to prove that *m* is not equivalent to the leading density function *Kd*. The reason is that *m* evaluates the relative concentration and *Kd* the absolute one. We take a real example to show the complementarity of *m* and *Kd*. More precisely, we illustrate that point with the distribution of pharmacies in the Lyon area in France. Both functions detect irregularities in the spatial distribution of this activity but *Kd* assimilate that spatial structure to dispersion and *m* to aggregation. As a conclusion, a great attention should have been given to the statistic measure to gauge the spatial structure of activities.

REFERENCES

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, 271-281.

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, 81-103.

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., and Senn, L., (Eds.), Growth and Innovation of Competitive Regions, Springer, Berlin, 297-327.

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, 3-11.

Baddeley, A.J., Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns, Journal of Statistical Software 12, 1-42.

Baddeley, A.J., Møller, J., Waagepetersen, R.P., 2000. Non- and semi-parametric estimation of interaction in inhomogeneous point patterns, Statistica Neerlandica 54, 329-350.

Barlet, M., Briant, A., Crusson, L., 2013. Location patterns of service industries in France: A distance-based approach, Regional Science and Urban Economics 43, 338-551.

Behrens, K., Bougna, T., 2013. An Anatomy of the Geographical Concentration of Canadian Manufacturing Industries, Cahier de recherche/Working paper CIRPEE 13-27.

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, 139-152.

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, 287-302.

Brülhart, M., Traeger, R., 2005. An Account of Geographic Concentration Patterns in Europe, Regional Science and Urban Economics 35, 597-624.

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, H.G., 2004. The spatial distribution of economic activities in the European Union. *in*: J.V. Henderson, et J.-F. Thisse, (Eds), Handbook of Urban and Regional Economics, North Holland, Amsterdam, Elsevier, pp. 2845-2909.

Diggle, P.J., Chetwynd, A.G., 1991. Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations, Biometrics 47, 1155-1163.

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

Duranton, G., Overman, H.G., 2005. Testing for Localisation Using Micro-Geographic Data, Review of Economic Studies 72, 1077-1106.

Duranton, G., Overman, H.G., 2008. Exploring the Detailed Location Patterns of UK Manufacturing Industries using Microgeographic Data. Journal of Regional Science 48, 213-243.

Ellison, G., Glaeser, E.L., 1997. Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach, Journal of Political Economy 105, 889-927.

Giuliani, D., Arbia G., Espa, G., in press. Weighting Ripley's K Function to Account for the Firm Dimension in the Analysis of Spatial Concentration. International Regional Science Review.

Henderson, J.V., Thisse, J.-F., 2004, Handbook of Urban and Regional Economics, North Holland, Amsterdam, Elsevier.

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, 81-110.

Koh, H.J., Riedel, N., 2014. Assessing the Localization Pattern of German Manufacturing and Service Industries: A Distance-based Approach. Regional Studies 48, 823-843.

Krugman, P., 1991. Geography and Trade. MIT Press, 156p.

Law, R., Illian, J., Burslem, D., Gratzer, G., Gunatilleke, C. V. S., Gunatilleke, I., 2009. Ecological information from spatial patterns of plants: insights from point process theory, Journal of Ecology, 97, 616–628.

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

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

Marcon, E., Puech, F., 2014. A typology of distance-based measures of spatial concentration. *HAL-SHS Working Paper*, no. halshs-00679993, version 2 (4 February 2014).

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

Mori, T.,Smith, T.E., in press, A Probabilistic Modeling Approach to the Detection of Industrial Agglomerations, Journal of Economic Geography.

Nakajima, K., Saito, Y.U., Uesugi, I., 2012. Measuring economic localization: Evidence from Japanese firmlevel data, Journal of the Japanese and International Economies 26, 201-220.

Ó hUallacháin B., Leslie, T.F., 2007. Producer Services in the Urban Core and Suburbs of Phoenix, Arizona, Urban Studies, 44, 1581-1601.

Openshaw, S., Taylor, P.J., 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, 127-144.

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

Penttinen, A., Stoyan, D., Henttonen, H.M., 1992. Marked Point Processes in Forest Statistics, Forest Science 38, 806-824.

R Development Core Team, 2014. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Ripley, B.D., 1976. The Second-Order Analysis of Stationary Point Processes, Journal of Applied Probability 13, 255-266.

Ripley, B.D., 1977. Modelling Spatial Patterns, Journal of the Royal Statistical Society B 39, 172-212.

Sweeney, S.H., Feser, E.J., 1998. Plant Size and Clustering of Manufacturing Activity, Geographical Analysis 30, 45-64.

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