Distance-Based Methods: Ripley's *K* function vs. *K* density function

José M. Albert[†], Marta R. Casanova[‡], Jorge Mateu[§] and Vicente Orts[**]

Preliminary Version. Please do not quote without permission.

Abstract

In this paper, we propose an analytical and methodological comparison between two of the most known distance-based methods in the evaluation of the geographic concentration of economic activity. These two methods are Ripley's *K* function, a cumulative function popularised by Marcon and Puech (2003) that counts the average number of neighbours of each point within a circle of a given radius, and *K* density function, a probability density function of point-pair distances introduced by Duranton and Overman (2005), which considers the distribution of bilateral distances between pairs of points. To carry out this comparison, we first apply both methodologies to an exhaustive database containing Spanish manufacturing establishments and we evaluate the spatial location patterns obtained from both analysis. After an initial analysis, we realise that although these functions have always been treated as substitutes they should be considered as complementary, as both cumulative function and probability density function provide relevant and necessary information about the distribution of activity in space. Therefore, our next step will be to assess what are the advantages and disadvantages of each methodology from a descriptive and analytical way.

Keywords: Distance-Based Methods, Spanish manufacturing establishments, spatial location patterns
JEL classification: C15, C40, C60, R12.

[†] Department of Economics, Universitat Jaume I, Av. Sos Baynat s/n, 12071 Castellón, Spain.
[‡] Department of Applied Economics, Universitat de València, Av. Tarongers s/n, 46022 Valencia, Spain.
[§] Department of Mathematics, Universitat Jaume I.
[**] Department of Economics and Institute of International Economics, Universitat Jaume I.
Corresponding author: Marta R. Casanova (marta.roig-casanova@uv.es)

1. Introduction

Spatial concentration of economic activity and the analysis of establishments' location have been subjects much followed for many economists along the years, dating back to Marshall (1890), and all of them have concluded that there are good reasons to expect that economic activity will be unevenly distributed across space.

More recently, theoretical research into the so called new economic geography, initially developed by Krugman (1991),[1] emphasized the role of reinforcing advantages and the interaction between the benefits associated with operating in a large market (large number of potential consumers and related industries) and the increased costs of competition (increasing number of companies operating in that market), as the main forces to explain the emergence of endogenous spatial disparities. Obviously, the intensity of these forces and the trade-off between them determine the location decision of firms and consumers (workers), and the location patterns of different industries in the territory, beyond the conditions established by physical geography. These theoretical developments have been accompanied, in recent years, by numerous empirical contributions, trying to establish the link between these forces and the location of economic activity in space, the intensity and the spatial scope of this agglomeration. Nevertheless, we must bear in mind that the intensity of these centripetal and centrifugal forces and the trade-off between them do not necessarily change monotonically with distance. Thus, measurement of economic concentration becomes again a field of renewed interest in economic geography. The current measures used into this issue must be capable of capturing this peculiarity in the formation of clusters in a better way and provide information about the value of the specific increase in the distance at which firms have less incentive to locate in a particular cluster.

The literature on the empirical measurement of economic concentration has been influenced by two different traditions, economic geography, culminated with the paper of Duranton and Overman (2005), and spatial statistics, which became more important with the publication of the paper of Marcon and Puech (2003) in the '*Journal of Economic Geography*'[2]. In these papers, Marcon and Puech followed the mathematical and statistical background of *Ripley's K function*, initiated years back by Ripley (1976, 1977), Diggle (1983) or Cressie (1993), while Duranton and Overman developed their

[1] This work has been surveyed in Fujita el al (1999).

[2] Previously, Arbia and Espa (1996) used already measures of spatial statistics.

own *K-density function* based on the paper of Silverman (1986) by using a Gaussian kernel function to estimate the density of bilateral distances.

These two traditions have been evolving with their own methodologies but, the economic literature has no many examples of discussions by the preference for one or the other methodology for assessing geographic concentration.

We can find the first comparison between Ripley's *K* function and *K* density function in the fifth section of Marcon and Puech (2003), where these authors pointed out some of the advantages and disadvantages of both measures, without opting expressly for one of them as the best measure for evaluating spatial location patterns.

Afterwards, Duranton and Overman (2005) stated in the conclusions of their paper that their methodology was more informative than that used in spatial statistics, by saying: '*We believe the k-densities are more informative than k-functions with respect to the scale of localization*' (p.1103).

Nevertheless, it was not until last year when Marcon and Puech (2010) tried to discuss the convenience of using a probability density function of point-pair distances or a cumulative function for evaluating spatial concentration, both the measures of economic geography and spatial statistics approaches, respectively; reaching the conclusion that the two measures provide us useful information about the location of activity in space and their results are complementary. Thus, they must be implemented simultaneously and should not be considered as substitutive measures.

Meanwhile, Albert, Casanova and Orts (2011) also got closer the positions of the two approaches by the use of a statistical function, *M function* (an extension of the *Ripley's K function*), and making this function meet the five requirements of Duranton and Overman (2005).[3] This last approach, as the *D function* of Diggle and Chetwynd (1991), has two advantages. First, it is defined to analyse the nature and physical scale of spatial clustering for inhomogeneous populations and, second, it has an easy interpretation in terms of expectations. That is, given the intensity of firms within an industry, our *M function* measures the expected number of excess firms within a predetermined distance of this industry in comparison with the expected number in absence of spatial clustering.

[3] (1) Be comparable across industries, (2) control for the overall agglomeration of manufacturing, (3) control for industrial concentration, (4) be unbiased with respect to scale and aggregation, and (5) give an indication of the significance of the results.

Now, in this paper, we go a step forward. Our approach still continues in the tradition of spatial statistics but, at the same time, continues getting/moving closer to the economic geography approach. At this point, we do not settle comparing a cumulative function, *M function*, with a probability density function, *K-density function*. We attempt to avoid that the function coming from the spatial statistics tradition accumulates spatial information on the distribution of points up to each distance. We introduce the possibility of considering the *M function* as a non cumulative function by means of the *M marginal function*, a modification of the *M function*. In this way, we slightly modify a function coming from the spatial statistics path, by converting it in a non cumulative function and by doing to provide us information more like that given by the *K-density function*, i.e. the points at each distance.

This *M marginal function* is an extension of the *M function*, proposed by Albert, Casanova and Orts (2011), that takes the trade-off between the centripetal and the centrifugal forces which form clusters, and its value informs us, at each distance, about the variation in the number of neighbours in each sector when *r* becomes higher as compared to the increase in neighbours of the overall manufacturing industry. Therefore, by means of the incorporation of the *M marginal function* we will avoid the most significant difference between the methodologies of the two paths, the way of calculating the neighbours: *K-density function* at each distance and *Ripley's K function* up to each distance.

The two measures, *M marginal function* and *K-density function*, satisfy the five essential requirements that any test for measuring concentration should fulfil and, in addition, they have other characteristics in common, they both (1) treat space as being continuous, (2) can detect the spatial location patterns at every scale, (3) let us know the distances at which significant concentration or dispersion occurs and, besides, (4) both measures of concentration test the significance of their results by capturing the deviation between the spatial distribution of establishments within a considered sector and the spatial distribution of establishments within a hypothetical sector generated with the same number of establishments than the sector considered randomly allocated across all locations where we can currently find a establishment from the whole manufacturing.

Finally, just say that the necessary dataset to implement both methodologies also coincides. The distance-based methods make use of micro economic data, treating each firm as a point on a continuous, rather than on a discrete space. For this reason we need a database that contains the geographic coordinates of every establishment, in order to know their precise location on a map. Therefore, we use establishment level data, for the year 2007, from the Analysis System of Iberian Balances database[4], which contains exhaustive information about Spanish manufacturing sectors at the four-digit level, classified using the National Classification of Economic Activities[5], and also latitude and longitude coordinates of every establishment.

2. Measures of Concentration

This section describes the properties of the measures of concentration that we will employ in our investigation, from different perspective.

2.1. Ripley's *K* function

Ripley's *K* function, *K*(*r*), is a tool to analyse completely mapped spatial point process data, i.e. data on the locations of establishments. Under some properties, the reduced second moment measure of *K function* is

$$K(r) = \frac{1}{\lambda} E[\text{Number of further events within distance r of an arbitrary event}]$$

where λ is the density, or mean number of events per unit area. This term is estimated by *N/A*, where *N* is the observed number of points in the region studied and *A* is the area of the study region. Additionally, the numerator of this function can be estimated by counting the average number of neighbours each establishment has within a circle of a given radius, 'neighbours' being understood to mean all establishments situated at a distance equal to or lower than the radius (*r*). This function is considered a cumulative function because accumulates spatial information on the distribution of points *up to each distance* (*r*).

The *K*(*r*) function describes characteristics of the point patterns at many and different scales simultaneously, depending on the value of '*r*' we take into account, that is,

[4] SABI

[5] NACE 93 - Rev. 1

$$K(r)=\frac{1}{\lambda N}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}w_{ij}I(d_{ij})$$

$$I(d_{ij})=\begin{cases}1, & d_{ij}\leq r\\ 0, & d_{ij}>r\end{cases}$$

where d_{ij} is the distance between the ith and jth establishments, $I(x)$ is the indicator function and w_{ij} is the weighting factor to correct for border effects.[6] The indicator function, $I(d_{ij})$, takes a value of 1 if the distance between the ith and jth establishments is lower than r, or 0 otherwise, and w_{ij} will be equal to the area of the circle divided by the intersection between the area of the circle and the area of study.

Finally, using the definition of λ, the $K(r)$ function can be rewritten as:

$$K(r)=\frac{A}{N^2}\sum_{i=1}^{N}\sum_{j=1,i\neq j}^{N}w_{ij}I(d_{ij})$$

Therefore, the $K(r)$ function is a distance-based method that measures concentration by showing the share of average number of neighbours in an area of radius (r), over the density of the whole study region (λ).

K-value will increase along different radius (r) when the average number of neighbours is higher than the density of the whole study region, and will decrease when the density of the whole study region is higher than the average number of neighbours in the different areas of radius (r). This density will always depend on the null hypothesis considered, i.e. the benchmark.

For many years, in the spatial statistics path, had been using a theoretical benchmark consisting of a kind of randomly distributed set of locations in the area of study, called C*omplete Spatial Randomness* (*CSR*). However, taking into account the economic point of view, this benchmark is not realistic because economic activity is not located in a random and independent way. Moreover, an appropriate benchmark must consider the inhomogeneity of the space, because of dissimilarities in such natural features as mountains, rivers or harbours, and the tendency of economic activity to agglomerate.

[6] These border-effect corrections should be incorporated to avoid artificial decreases in $K(r)$ when r increases, because the increase in the area of the circle under consideration is not followed by the increase of establishments (outside the study area there are no establishments).

2.2. M_{TM} function

In Albert et al (2011) was introduced the M_{TM} function. This function had the 'whole of manufacturing'[7] as benchmark and thus was able to compare the spatial distribution of each sector with the overall tendency of manufacturing industry to agglomerate.

$$M_{TM}(r) = K(r) - K_{TM}(r)$$

Here, $M_{TM}(r)$ is the difference between the K-value of each sector under consideration and the K-value of the total manufacturing at radius r. This difference allows identifying the existence and magnitude of spatial agglomeration of establishments in a considered sector over and above the level of spatial concentration of the whole of manufacturing attributable to natural inhomogeneity of countryside and to the general tendency of economic activity to agglomerate. In fact, with CSR as the benchmark employed we were not able to isolate the idiosyncratic tendency of each sector to locate itself in accordance with the general tendency of manufacturing establishments to agglomerate. Thus, by using this benchmark we take into account natural and economic factors that can condition the spatial distribution of activity.

Both K-values, the one of the sector and the other of the total manufacturing, are relative to the density of the whole study region, besides M_{TM} value will be relative to a benchmark, *TM* in this case.

On the one hand, M_{TM} value will increase as long as the average number of establishments at different (r) of a considered sector increases with the distance and this increase is higher than that occurred in the *TM*. On the other hand, M_{TM} value will decrease when the increase of the average number of establishments of a considered sector at different (r) is lower than the increase of establishments of the *TM*. Finally, M_{TM} value will remain constant for different values of (r) when the spatial location pattern of a considered sector coincides with the distribution of the benchmark, *TM*.

Thus, K-value will never take negative values, but M-value can take positive or negative values, depending on whether a subsector is concentrated or dispersed relative to the *TM*.

[7] Henceforth *TM*.

2.3. M_{TM} marginal function

By means of this function, we know at each distance the variation in the number of neighbours in each sector when r becomes higher as compared to the increase in neighbours of the overall manufacturing industry. It does not accumulate spatial information on the distribution of establishments up to each distance (r).

The agglomerative strengths that pull economic activity together and determine the differences in the shape and the size of the clusters do not act in a linear way/ monotonically with distance, so we need a function that reflects/reproduces this trade-off between centripetal and centrifugal forces that determines the location of the establishments and the different outlines of the clusters.

This function counts the variations of the neighbours when we change the radius, that is, $\Delta M_{TM}/\Delta r$ (the marginal M_{TM} value at each distance). Thus, stops being a cumulative function to make way for a non-cumulative function, as the K-density function, by getting closer to the economic geography tradition, i.e. counting the neighbours at each distance.

2.5 General methodological similarities and differences

The way of constructing the confidence interval, in order to test whether industrial location patterns significantly diverge from randomness, is the same in both cases, by generating 1000 simulations and rejecting the non-significant values. For each simulation we randomly reallocate as many establishments as the sector considered has across the sites where we can currently find establishments from the whole manufacturing. The sampling is done without replacement. Like this, by the specific Monte Carlo simulations drawn, we get/obtain/achieve that both measures of concentration applied in this paper share some similarity with the way of constructing the case-control counterfactual explained in the paper of Diggle and Chetwynd (1991), although, there is a difference. As we know the population density in advance, it is not necessary the use of a previously selected group of controls, or representative sample of the entire population, in order to construct the counterfactuals.

There are other differences between the two measures of concentration. The most important dissimilarity found is that M_{TM} function is a cumulative function and is relative to the total density of the whole region[8], whereas *K*-density function is a non-cumulative and absolute function, because has into account the frequency of distances at each (*r*) in order to know the number of neighbours next to an establishment, i.e. counts how many times is repeated each distance between pairs of establishments.

Regarding the necessity of an area of study, this is necessary in M_{TM} function because we need to know the density of the whole study region to calculate the M_{TM} value at every radius (*r*). In the case of the *K-density* function is unnecessary.

With reference to the total distance analysed by the two measures of concentration, we note it is different. The distance of Ripley's *K* function is shorter than that taken into account with *K*-density function. It is because a bias is found in Ripley's *K* function when we carry the analysis out at large distances. Thus, *K*-density function lets us detect clusters that are located at large distances between them, more than 250km, aspect that does not allow the Ripley's *K* function.

The way of reading the results also changes depending on the measure of concentration we are analysing. On the one hand, in the case of the M_{TM} function we say that relative localization appears within a particular sector when its *K*-value is higher than *K*-value of the total manufacturing. In such a case, our claim is that this sector is concentrated relative to the whole of the manufacturing industry. On the other hand, with the *K-density* function a sector is said to be globally localized if the observed density distribution of bilateral distances hits the upper band of the global confidence interval for at least one distance.

Finally, relative to the way of interpreting the results, we should ask ourselves why the initial value of M_{TM} function is always zero, while it is not when we implement *K*-density function. This is because the first distance of the radius considered in Ripley's *K* function is 0, thus, obviously, at this distance we have no neighbouring establishments and the first *M*-value is 0. In the case of *K*-density function, this initial distance taken into account should be also zero, but in this methodology we have a bandwidth (*h*), so, the first distance is not 0 strictly, depending always on the bandwidth taken into account. Therefore, if the initial distance considered is higher than zero, we may find neighbouring establishments and the first value of *K*-density function will not be 0.

[8] Porque es resultado de la diferencia de dos Ripley's *K* functions y esta función lo es.

3. Results. Empirical evidence

We will discuss the detailed exemplification coming from the empirical analysis, in order to test whether the resulting graphs of the two measures of concentration let us to verify the theoretical and methodological aspects explained in the previous section.

The number of subsectors concentrated or dispersed does not vary, regardless of the measure of concentration implemented. This confirms the robustness of both measures.

With reference to the *velocity* of the rise of M_{TM} function, we can say that this velocity is related to the size of the clusters. The smaller the size of the clusters, more rapidly will increase the *M*-value. This is because at very short distances of the radius (*r*) we will find an average number of establishments much higher than the density of the whole study region. Thus, by means of the Ripley's *K* function we have an accurate approximation of the closeness of the establishments within the cluster. In the case of the *K*-density function, this closeness will be reflected in its initial value, since higher values at short distances implies greater proximity of firms within the cluster.

The *intensity* reached by the subsectors does not always present a direct relationship between the values of both measures. This difference can be due to the intrinsic discrepancies between both indices and we will attempt to shed light on this issue, trying to find out to/at what features of the location pattern are sensitive both the Ripley's *K* function and the *K*-density function.

The statistical properties and the theoretical knowledge of each measure may help us to better interpret the results and to get the maximum benefit of them. Ripley's *K* function measures concentration by counting the average number of neighbours each establishment has within a circle of a given radius and *K*-density function is the estimator of the density of bilateral distances at each distance considered. Consequently, the maximum *intensity* reached by means of the two functions, i.e. the highest significant peak of M_{TM} and *K*-density value, will depend on the specific arrangement/location of points in space and on the way of calculating this maximum intensity each measure has.

On the one hand, the high intensity of M_{TM} value will be caused by the concentrated activity in space. Thus, being an average value, those establishments located in an isolated position will reduce the intensity of the function, because they will not find neighbours around them. On the other hand, the high intensity of *K*-density will also depend on the high concentration of activity in space, so the more establishments we find clustered at short distances, the more elevated will be the density of bilateral distances at these short distances. Therefore, the higher the percentage of establishments localised within clusters, the higher the intensity of both measures at short distances. However, can we detect the specific number of clusters responsible of the intensity of the two measures of concentration?

We should highlight that with none of the two measures we can ensure the precise number of clusters that actually exist, i.e. the total existing amount of spatial locations of establishments of each subsector. However, by means both measures, we can intuit whether most of the establishments of a specific subsector are localised in a single location, i.e. constitute a single cluster, or are concentrated in several locations, i.e. constitute several clusters.

According to the detection of local density at different spatial scales, Marcon and Puech (2010) tell us that *K*-density function detects more easily local clusters, while Ripley's *K* function is less precise to detect local clusters, being better detecting the interaction of different local clusters situated at larger distances. Nevertheless, the specific idiosyncrasy in location of Spanish establishments in space can be very varied and capricious and this fact may cause some variation in this evidence.

Based on our empirical results and when we speak about Ripley's *K* function, local clusters are clearly detected only in some specific situations. Only when the intensity of the *M*-value is very elevated at short distances and its increase and later decrease are very fast, we can clearly identify a local cluster. In the case of *K*-density function, we can interpret that a unique cluster exists when the intensity of the function at short distances is much more elevated than in other subsectors. In this way, the concentration of most of the establishments in a single location causes the rise of the density of bilateral distances at short distances. Taking into account this information obtained from our results, we must realize that the compliance of the evidence in relation to the detection of local density, referenced by Marcon and Puech, will always depend on two features of the point pattern: (1) the relative size of the cluster, depending on the companies in the subsector, and (2) the proximity of the companies within each cluster, taking this as a consequence the cluster size.

In figure 3 we observe a clear example about the specific situation previously discussed, a subsector with most of its establishments concentrated in a single location.

This figure shows the M_{TM} and the K-density curve of subsector 2630 and we realize that the point pattern of this subsector has a distinguishing feature, given that Ripley's K function detects more precisely the existence of the local cluster. This is because the subsector has a percentage very elevated of its establishments in that location and these establishments are very close to each other.

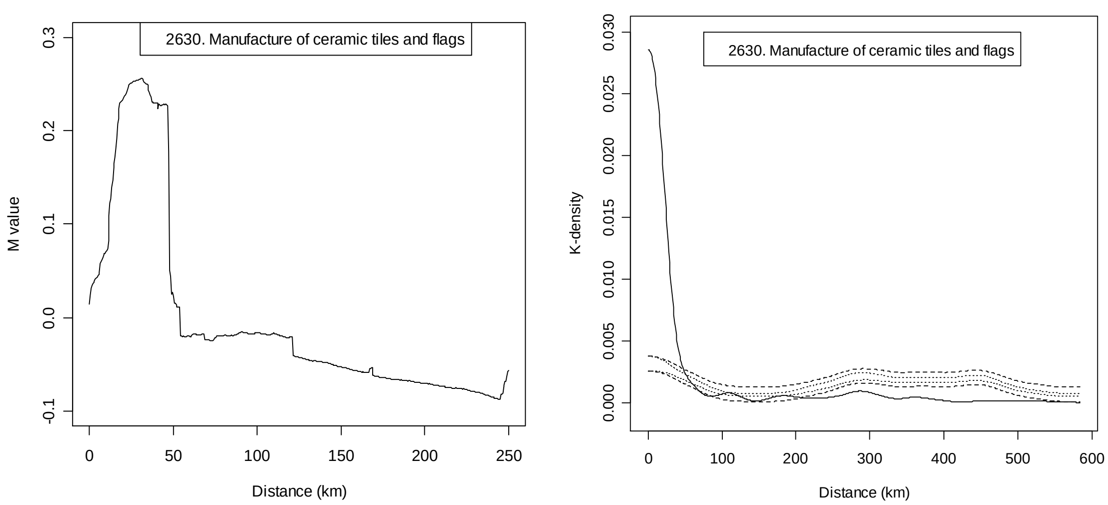

Figure 3. Comparison of M_{TM} function and K-density function of *subsector* 2630

If we compare the two subsectors that present the *highest intensity* according to the M_{TM} value, 2630 and 2213, we can observe that their spatial location patterns differ between them (figure 3 and 4 left). Both subsectors present high levels of concentration at short distances, and the increase of M_{TM} value is very fast, but the afterwards behaviour of this function is different. In subsector 2630, M-value decreases as quickly as it had risen while, in subsector 2213, M-value is maintained or its decline is very slow. This difference in the behaviour of M_{TM} value may be caused by a discrepancy in the number of clusters that constitute the two subsectors. Therefore, analysing in detail the location of the establishments of both subsectors we see that the subsector 2630 agglutinates the 78% of their establishments in a single location, in the province of Castellón, while the subsector 2213 agglutinates the 82% of their establishments in two locations, the 57% in Madrid and the 25% in Barcelona. So, we can conclude that in the case of M_{TM} function we detect the existence of more than one cluster by the slow drop/decrease of the M-value.

Although we know that exists more than one cluster in a specific subsector when the M-value decreases slowly, we cannot identify the exact quantity of clusters.

However, we can intuit this quantity by the intensity of the *M*-value. For instance, the subsectors 2213 and 3622 have the same percentage of establishments concentrated in specific locations (82%), although the quantity of locations differs. Subsector 2213 has two evident clusters, Madrid and Barcelona, and subsector 3622 has four evident clusters, Córdoba, Barcelona, Valencia and Madrid. As a result, the maximum intensity of subsector 3622 is much lower than the maximum intensity of subsector 2213, although the slow decrease of the *M*-value coincides in both subsectors.

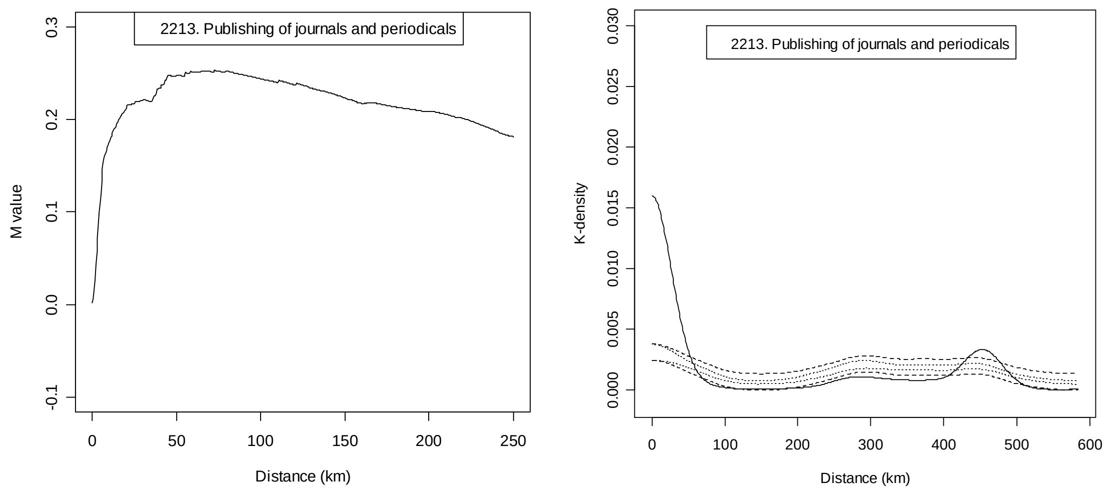

Figure 4. Comparison of M_{TM} function and *K*-density function of *subsector* 2213

The behaviour of the *K*-density function also changes depending on the existence of one cluster or more than one (figure 3 and 4 right). As we can see in these figures, *K*-density value of subsector 2630 is almost the double than *K*-density value of subsector 2213. In this way, the intensity of *K*-density value is related to the specific location of the establishments. As most of the establishments of subsector 2630 are located in a single location and, in the case of subsector 2213, they are located in two locations, the density of bilateral distances at short distances of sector 2630 will be higher that the density at short distances of sector 2213. Concluding, in the case of *K*-density function we detect the existence of more than one cluster by the reduction of its maximum value at initial distances.

Up to now, we have been speaking about subsectors shaped by establishments distributed in a concentrated way. In contrast, Figure 5 shows the spatial distribution of the establishments from subsector 3511, which are concentrated, but in a peculiar way.

They are distributed and clustered along the coastline. Here, each dot corresponds to an establishment.[9]

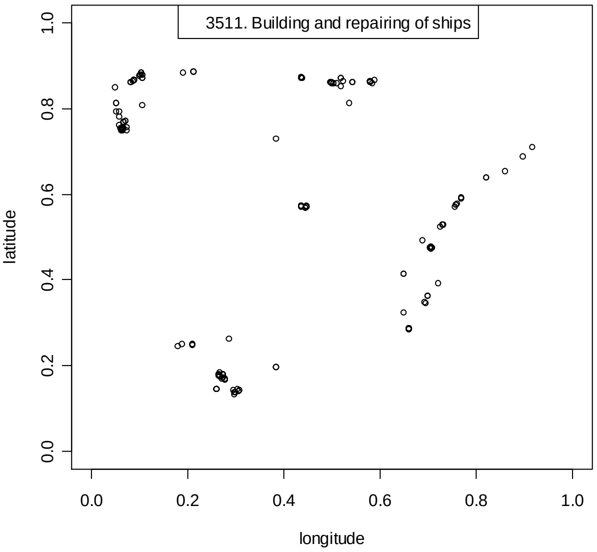

Figure 5. Spatial distribution of establishments of subsector 3511

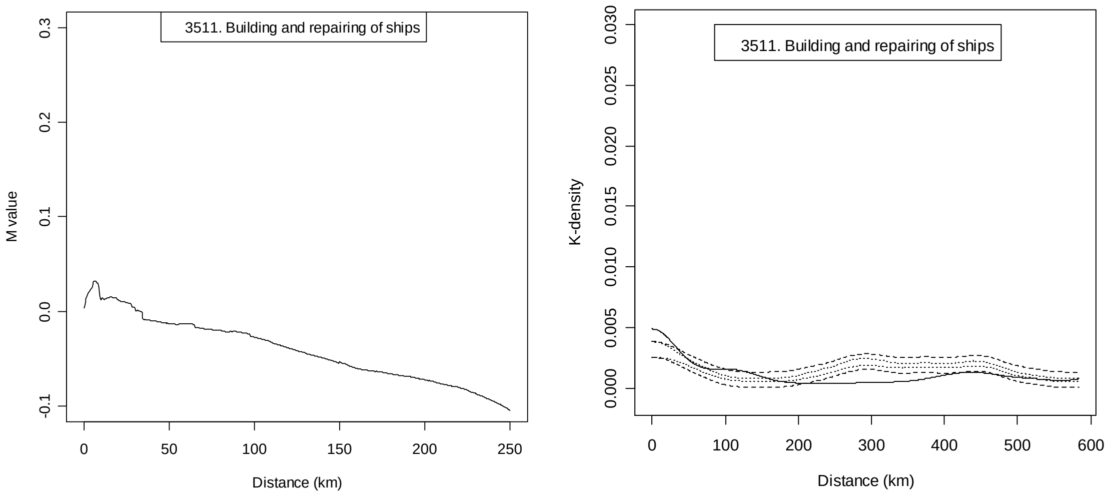

Figure 6. Comparison of M_{TM} function and *K*-density function of *subsector* 3511

[9] The area of study is normalised, (0, 1), but the real measures would be 1075 × 882 km.

References

Albert JM, Casanova MR and Orts V (2011) Spatial Location Patterns of Spanish Manufacturing Firms. *Papers in Regional Science*. Forthcoming.

Arbia G, Espa G (1996) *Statistica Economica Territoriale*, Cedam, Padua.

Cressie NA (1993) *Statistics for Spatial Data*. New York: Wiley

Diggle PJ (1983) *Statistical Analysis of Spatial Point Patterns*. London: Academic Press

Diggle PJ, Chetwynd AG (1991) Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics* 47: 1155-1163

Duranton G, Overman H (2005) Testing for Localization using Micro-Geographic Data. *Review of Economic Studies*, 72, 1077-1106.

Duranton G, Overman H (2008) Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data. *Journal of Regional Science*, 48 (1), 213-243.

Fujita M, Krugman P and Venables AJ (1999) The Spatial Economy: Cities, Regions and International Trade. MIT Press, Cambridge, MA.

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

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

Marcon E, Puech F (2010) Measures of the geographic concentration of industries: improving distance-based methods. *Journal of Economic Geography*, 10 (5), 745-762.

Marshall A (1890) Principles of Economics. MacMillan, London.

Ripley BD (1976) The second-order analysis of stationary point processes. *Journal of Applied Probability* 13: 255-266

Ripley BD (1977) Modelling Spatial Patterns. *Journal of the Royal Statistical Society - Series B (Methodological)* 39: 172-192

SABI. System of Iberian Balances Analysis. Bureau Van Dijk Electronic Publishing

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall