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 Please do not quote without permission.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.

[†] Department of Economics, Universitat Jaume I, Avda. Sos Baynat s/n, 12071 Castellón, Spain.

[‡] Department of Economics and Institute of International Economics, Universitat Jaume I.

Corresponding author: Marta R. Casanova (mroig@eco.uji.es).

1. Introduction

Economic activity has an evident tendency towards the spatial concentration. The characterisation of the patterns of geographic concentration of firms and industries in space has been a subject much followed for many economics along the years, dating back to Marshall (1890), as well as the sources of these agglomeration economies[1].

The theoretical work in economic geography has been evolving over time and demanding the fulfilment of new requirements in the measurement of spatial concentration. As instance Duranton and Overman (2005) stressed that any test for measuring concentration should fulfil five essential requirements: (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.

Thus, the methods to perform the empirical work have also had to adapt to these new demands. Taking space into consideration and treating space as being continuous, avoiding like this the sensitivity of the results to the choice of a specific area of study[2], have been two of the most important changes in the methods of measurement of the spatial distribution of activity. In *Albert et al* (2011) we can see a brief overview of the literature on the empirical measurement of economic agglomeration. From this exhaustive summary, we conclude that this literature has been influenced by two very different traditions, 'economic geography' and 'spatial statistics', and finally have converged and the positions of the two approaches have gradually got closer to each other.

In this paper, we are going to focus on both of the distance-based methods that *treat space as continuous* and are the result of the evolution of the two approaches. So, we are going to apply Ripley's *K* function and *K*-density function simultaneously in the Spanish manufacturing establishments. In this way, we will be able to compare the

[1] For further details see Fujita et al. (1999), Krugman (1991) or Duranton and Puga (2004).

[2] It is known as Modifiable Areal Unit Problem (MAUP).

outcomes and the resulting location patterns of both methods in order to find the advantages and drawbacks of each of them.

2. Methodology

Ripley’s *K* function, as now known, was introduced by Ripley (1976) and named ‘*K* function’ in Ripley (1977). This methodology was not used at first with economic purposes and has been modified and improved over time by many authors. It was introduced into economics by Arbia and Espa (1996) and later popularised by Marcon and Puech (2003).

Ripley’s *K* function, $K(r)$, is a distance-based method that measures concentration by counting the average number of neighbours each firm has within a circle of a given radius, ‘neighbours’ being understood to mean all firms situated at a distance equal to or lower than the radius (r). From here on, firms will be treated as points. 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.

K-density function, introduced and popularised by Duranton and Overman (2005), introduces the treatment of space as something that is continuous in the perspective of the ‘economic geography’. Up to that moment, the indices used in this path had not taken space into consideration (Herfindahl or Gini) or had treated space as being discrete (Ellison and Glaeser, 1997). This measure computes the density of bilateral distances between all pairs of establishments in an industry. In this way, it is also unbiased with respect to scale and aggregation.

The most distinguishing feature between the two methods is the fact that Ripley’s *K* function is a cumulative measure, instead of being a density function of bilateral distances, as is the case of the *K*-density used by Duranton and Overman (2005). However, in this paper we are going to analyse in detail the contributions of both measures to the analysis of the spatial location patterns.

We use the ‘whole of manufacturing’ as a benchmark, thus we can compare the spatial distribution of each sector with the overall tendency of manufacturing industry to agglomerate. In order to construct the confidence intervals we will use the Monte Carlo method, which involves generating a large number of independent random simulations. We simulate random distributions with the same number of establishments as in each of the sectors under consideration, and the location of these hypothetical firms is restricted to the sites where we can currently find firms from the whole manufacturing sector.

3. Data

Our empirical analysis uses current establishment level data, for the year 2007, from the Analysis System of Iberian Balances database,[3] which contains detailed information about Spanish and Portuguese companies. We restrict our database to Spanish manufacturing establishments, using the National Classification of Economic Activities[4] and analysing sectors at the four-digit level. Furthermore, we add another two requirements to our database. First, we ensure that our database contains only Spanish manufacturing firms on the peninsula, without including firms from the Canary and Balearic Islands, Ceuta and Melilla. Second, we restrict our analysis just to firms employing at least ten workers. Finally, once these requirements have been applied, our database contains exactly 43,087 firms.

Spanish manufacturing activities are classified into 23 sectors according to ‘NACE 93 - Rev. 1’ and these are as follows: (15) Food products and beverages, (16) Tobacco products, (17) Textiles, (18) Wearing apparel and dressing, (19) Tanning and dressing of leather, (20) Wood and products of wood, (21) Pulp, paper and paper products, (22) Publishing, printing and recorded media, (23) Coke, refined petroleum products, (24) Chemical and chemical products, (25) Rubber and plastic products, (26) Other non-metallic mineral products, (27) Basic metals, (28) Fabricated metal products, (29) Other machinery and equipment, (30) Office machinery and computers, (31) Electrical

[3] SABI

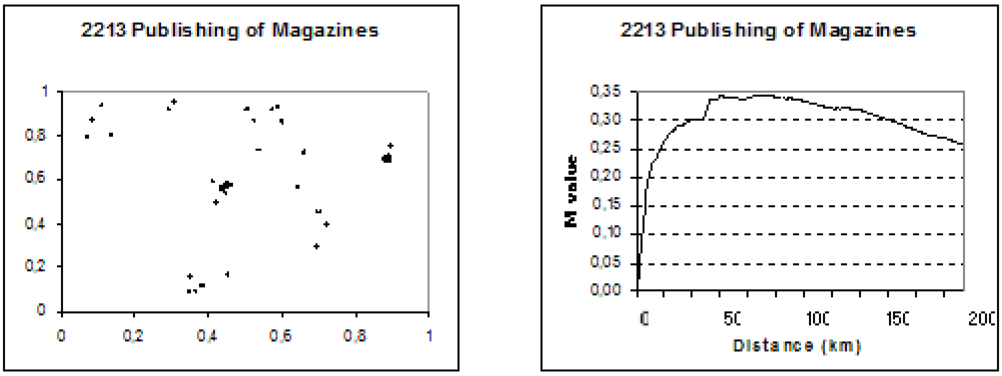
[4] NACE 93 - Rev. 1

machinery, (32) Radio, televisions and other appliances, (33) Instruments, (34) Motor vehicles and trailers, (35) Other transport equipment, (36) Furniture and other products, (37) Recycling.

4. Results and Discussion

The distance-based methods we are going to use to measure the spatial distribution of activity in Spain is Ripley's *K* function and *K*-density function, which offer important advantages over traditional concentration indices.

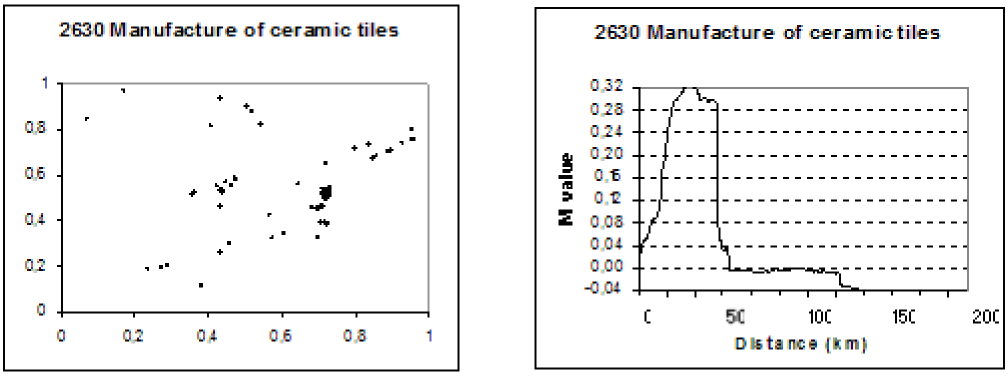
Here, we present two subsectors analysed by means of Ripley's *K* function.



(a) (b)

Figure 1. Relative location patterns of subsector 2213.

The subsector 2213 is very concentrated in space, its clusters are very reduced and its establishments are mostly located in Barcelona and Madrid. This information appears reflected in the M_{TM} curve. In fact, the values of M_{TM} increase very fast at a very low length of the radius (*r*). However, there is not a sudden drop of the values because there are two very distinct and separate clusters of one another; thus, the high concentration reached at a 'small' scale descends slowly.



(a) (b)

Figure 2. Relative location patterns of subsector 2630.

‘Manufacture of ceramic tiles’ (2630) is an industry heavily concentrated in the province of Castellón, where we can find the agglomeration of points. A particular location may specialize in a specific activity for two reasons. First, the location might have some underlying characteristic that gives a natural advantage to the activity. Second, some type of scale economy might be reached by concentrating production at that location. This second reason would be the main cause why the Spanish ceramics is, almost entirely, located in a radius lower than 50km surrounding Castellón.

If we look at the M_{TM} curve, it shows us that the increase of the M_{TM} value, and thus of the concentration, occurs at very small scales. However, unlike the previous case, this value increases very quickly and afterwards it decreases with the same speed. That is because there is a single cluster and owns the vast majority of the establishments analysed.

After this analysis by means of Ripley’s K function, we will do a further analysis with the K-density function.

References

Albert JM, Casanova MR, Orts V (2011) Spatial Location Patterns of Spanish Manufacturing Firms. *Mimeo*.

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

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

Duranton G, Puga D (2004) Micro-foundations of urban agglomeration economies. In: Henderson JV, Thisse JF (eds) *Handbook of Regional and Urban Economics*, volume 4: Cities and Geography. Elsevier, ch. 48, pp 2063-2117

Ellison G, Glaeser E (1997) Geographic concentration in U.S. manufacturing industries: a dartboard approach. *Journal of Political Economy* 105: 889-927

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

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

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

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

Morton A (2003) Workbook from Alan Morton. Electronic publication. Distribution mapping software (DMAP), http://www.dmap.co.uk

R Development Core Team (2007) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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.