

# Spatial distribution of economic activities: an empirical approach using self-organizing maps\*

Federico Pablo-Martí (♠): [federico.pablo@uah.es](mailto:federico.pablo@uah.es)  
Josep-Maria Arauzo-Carod (♣, ♦): [josepmaria.arauzo@urv.cat](mailto:josepmaria.arauzo@urv.cat)

## Abstract

The aim of this paper is to analyse the co-location patterns of industries and firms. According to spatial distribution of firms at a microgeographic level, we identify how firms from different industries use to be close to each other and then which are the main reasons for this locational behaviour. The empirical application uses data from Mercantil Registers of Spanish firms (manufactures and services). Intersectorial linkages are shown using self-organizing maps technique.

(♠) Facultad de CC. Económicas y Empresariales  
(Universidad de Alcalá)  
Pl. de la Victoria, 2; 28802 - Alcalá de Henares  
Phone: + 34 918 854 231, Fax + 34 918 854 201

(♣) Quantitative Urban and Regional Economics (QUIRE)  
Department of Economics (Universitat Rovira i Virgili)  
Av. Universitat, 1; 43204 - Reus  
Phone: + 34 977 758 902, Fax + 34 977 759 810

(♦) Institut d'Economia de Barcelona (IEB)

Key words: clusters, microgeographic data, self-organizing maps, firm location  
JEL classification: R10, R12, R34

\*This research was partially funded by SEJ2007-64605/ECON, SEJ2007-65086/ECON, the "Xarxa de Referència d'R+D+I en Economia i Polítiques Públiques" of the Catalan Government and the PGIR program N-2008PGIR/05 of the Rovira i Virgili University. Any errors are, of course, our own.

## 1. Introduction

Analysis of spatial distribution of economic activity has plenty of implications in several areas like urban planning, infrastructures, firm supporting policies and land use, among others, and is receiving an increasing attention by researchers. Traditionally, scholars have approached this issue using extant administrative units (e.g. counties, regions, etc.) and then analysing how economic activities were spatially distributed. Unfortunately, those analyses suffer from some shortcomings, as administrative units are not ever coincident with real economic areas and are sometimes arbitrary. Additionally, administrative units present large differences in terms of size and shape, for instance, and those spatial specificities could make analysis more difficult.

In order to face such constraints, recent developments have shifted to microgeographic data, trying to overcome previous shortcomings. Concretely, smaller spatial units are being used, while such units do not match exactly with any extant administrative unit as they are created as a result of equally dividing space into homogeneous cells.<sup>1</sup> The two most usual cell shapes are squares and hexagons. The main advantage of a hexagonal map over a square map is that the distance between the centre of every hexagonal cell (or hexadecimal) and the centre of the six adjacent hexagons is constant, while for a square map the distance varies depending whether we consider the four cells adjacent to each cell (Rooks Case contiguity) or the four cells that are at the diagonal (Bishops Case contiguity). But a disadvantage of a hexadecimal map is that the adjacent cells are only in six directions instead of eight, as in a square map. Besides, no hexagonal cell has another adjacent extended directly towards the east or towards the west.

[INSERT FIGURE 1]

---

<sup>1</sup> There are also other approaches, like those that use the stochastic methodology of Point Pattern or those that use Neuronal Networks for pattern recognition. Unfortunately, previous approaches are not able to deal with multisectorial analyses, which are the goal of this paper.

There is currently an important shift of production systems worldwide that has extending the idea of a deeper public intervention on economic activity. This implies that policy makers must previously identify and selected key industries where public intervention is expected to be intensified. Such intervention implies knowing most dynamic industries as well as their spatial distribution patterns in terms of geographical location and clustering. Accordingly, mapping spatial distribution of economic activity appears to be of key importance, but there is not an agreement about with approach fits better with policy design. Currently there are two man perspectives: such of Industrial Districts and such of Clusters. While the former is so popular mainly due to the methodology of Sforzi-ISTAT, the later is potentially easier to implement thanks to the lower data requirements. Therefore, in this paper we are following cluster approach both due to data availability and to shortcoming of Sforzi-ISTAT methodology.<sup>2</sup> Having said that, the methodology proposed in this paper aims to overcome previous limitations in order to get more exact results and, therefore, to better design public policies.

According to this framework, in this paper we try to identify manufacturing and service (all sectors) clusters in Spain, using data from Mercantil Registers of 2006. Additionally we classify those clusters according to the reasons that explain the clusterization processes. This is, whether firms tend to locate together because they look for the same type of sites (no matter the industry to which they belong to), or whether firms look to be located close to their suppliers / customers in order to optimise the commercial exchanges among them.

This paper is organised as follows. In the next section we review main contributions about the spatial distribution of economic activity and about the spatial units used in empirical analysis. In the third section we explain the data set, we describe and analyse the spatial distribution of firms in Spain, we define

---

<sup>2</sup> Boix and Galletto (2008) identify some shortcomings of Sforzi-ISTAT. Among them, there are the lack of precision when defining boundaries of local labour markets, the use of national input-output matrices, the existence of polispecialised districts, the lack of local data about social capital and some general drawbacks of the methodology about how to capture socioeconomic characteristics of local communities.

the methodology used for identifying clusters and we explain the use of GIS (Geographical Information Systems) techniques for location analysis. In the fourth section we present and discuss our main empirical results. In the final section we present our conclusions.

## **2. Spatial distribution of economic activity**

Spatial distribution of economic activity has been a major topic since seminal contributions of scholars like Johann Heinrich Von Thünen (land use model), Alfred Marshall (agglomeration economies), Alfred Weber (the impact of transportation costs on location decisions), Walter Christaller (Central Place Theory) or William Alonso (Central Business District), to name just a few.

From a dynamic approach, researchers have analysed how specific characteristics of sites (cities, counties or regions, among others, but usually administrative units) affect their probability of being chosen by new firms, while from a static approach efforts have focused on the estimation of the degree of spatial concentration of firms, jobs or individuals. This paper shares both approaches since we are interested in which is the current spatial distribution of incumbent firms but, at the same time, we want to understand which are the reasons that drive this distribution. So, we do not aim just to measure degrees of concentration (dispersion) of economic activity, but also to explain location determinants of selected sites.

If we review empirical literature on spatial distribution of economic activity most researches agree on that there is a high level of concentration (specially in most developed countries), no matter the way in which such concentration is measured, as Duranton and Overman (2005), Devereux et al. (2004), Maurel and Sédillot (1999) or Ellison and Glaeser (1997) show, among others. The Spanish case is roughly the same, as scholars like Paluzie et al. (2004) and Viladecans (2004) have widely demonstrated. Additionally, data obtained from our data set also points out into the same direction, which means that land sites

considered by firms are only a very small part of absolute available land. So, firms tend to cluster in a few number of sites while most of available land remains empty. Our data also shows that firms and individuals compete for the same areas, since most of those “economic sites” are so close to big urban areas.

In any case, there is plenty of worldwide empirical evidence about this concentration pattern, and usually scholars explain the geographical concentration of production in terms of the existence of some increasing returns (Krugman, 1991) or because there are some kind of external scale economies at the industry level. In this sense, Karlsson et al. (2005, p. 10) argue that *“(w)hen external economies of scale of this type are present in a functional region, the unit costs of each firm in the industry decreases as the number of firms in the industry in the region increases. With decreasing costs, co-located firms can increase their productivity and their factor rewards. Hence wages and profits can rise”*. Usually, this location behaviour is explained in terms of agglomeration economics (such benefits that firms obtain from getting close to other firms), but additional knowledge is needed about what is hiding behind agglomeration economies. Since Marshall (1890) findings, agglomeration economies have been identified as main drivers of firms’ concentration due to three important reasons: a specialised labour market (explained by the presence of a pool of skilled workers), suppliers’ availability (given the size of the market) and knowledge spillovers (due to knowledge transfers among firms). Later, Hoover (1936) tried to better measure this phenomenon and classified agglomeration economies into urbanisation economies (as a consequence of the concentration of diverse activities) and localisation economies (as a consequence of the concentration of similar activities)<sup>3</sup>.

Nevertheless, empirical evidence shows that there are mixed examples of firm agglomerations: specialized areas and diversified areas. In any case, more agglomerated areas tend to be more diverse. This map pictures a Herfindahl-Hirschman Index (HHI) for services and manufacturing activities in Spain, where

---

<sup>3</sup> See Parr (2002) for a review of agglomeration economies’ classification.

blue areas imply higher diversity and red areas imply lower diversity. A close look to the map will show that bigger urban areas are more diverse (blue) than smaller urban areas or rural ones (red).

So, firms look both for being located close to similar firms (e.g., firms from the same industry) but also to different firms (e.g., firms from another industry). Therefore, firms look for neighbours, but not all neighbours could be equally useful, and even some of them could be useless and harmful. This is why sometimes firms look to be located close to other firms to which they are vertically integrated, because they need to have close linkages with those firms that are providers / suppliers of them. So, spatial proximity<sup>4</sup> appears to be an argument good enough for sharing the same location. Additionally, there are other reasons that explain why certain types of firms tend to be located in the same areas than other certain (different) types of firms. This is why even if they belong to different industries and have different characteristics, they share the need for specific territorial inputs that push them to the sites where those inputs are available (e.g., skilled human capital, energy supply, specific transport infrastructures, access to main markets, etc.).

It is at this point that *Modifiable Area Unit Problem* (MAUP) appears<sup>5</sup>, since the scope of the area used in the empirical analysis will strongly determine results obtained by the researchers and, of course, will make comparisons difficult (Duranton and Overman, 2005). In this sense, Arbia (2001) provides an excellent example of such problems. Concretely, he portrays a hypothetical distribution of firms' location (Figure 2), in which there are four firms inside the spatial area to be analysed (Figure 2a). Arbia (2001) shows that, depending on how spatial borders are designed, this location could result in a minimum concentration pattern (Figure 2b), in a maximum concentration pattern (Figure 2c) or in an Intermediate concentration pattern (Figure 2d).

[INSERT FIGURE 2]

---

<sup>4</sup> At this paper "spatial proximity" means to be located into the same cell. An extension of this approach could be to use XClusters for the spatial delimitation of such proximity.

<sup>5</sup> See Openshaw and Taylor (1979) for a detailed analysis and Wrigley (1995) for a further review.

Figure 2 shows that spatial aggregation really matters, so researchers should be aware of this circumstance and carefully select the most appropriate areas. Unfortunately, this has not been a major concern in empirical analysis, mainly due to lack of sufficiently disaggregated data<sup>6</sup>, but recently, researchers have get accessibility to dramatically improved datasets with extended spatial disaggregation. This is, for instance, the case of our data set which comprises accurate individual information about the location of firms, which allows us to technically address previous shortcomings and to freely decide way in which space is disaggregated, no matter where administrative (usually arbitrary) boundaries are. This is of special importance since “(...) *any statistical measure based on spatial aggregates is sensitive to the scale and aggregation problems*” (Arbia, 2001, p. 414). So, as Duranton and Overman (2005, p. 1079) point out, “(...) *any good measure of localization must avoid these aggregation problems*”.

According to previous considerations, our goal is to empirically asses location patterns of both manufacturing and service firms in Spain and try to determine if those firms tend to be close to firms of the same industry, to firms with close industry linkages (e.g., providers and suppliers) or to firms that share location requirements (e.g., accessibility to inputs, labour and infrastructures). In this sense, there are previous contributions that have faced similar approaches. Concretely, Duranton and Overman (2008, 2005) analyse manufactures using microgeographic (postcode level) data coming from the Annual Census of Production in the United Kingdom. They compute Euclidean distances between every pair of entering establishments and compare those results with extant distances between incumbent establishments in order to check if location patterns of entrants and incumbent establishments are similar o not.

Duranton and Overman (2008) try to identify two specific situations: first one occurs when firms from different industries use to locate in the same areas (*joint-localization*); second one occurs when firms from different industries also

---

<sup>6</sup> A prior insight into the influence of spatial units when analysing location of firms has been done in Arauzo-Carod and Manjón-Antolín (2004) and in Arauzo-Carod (2008). See Olsen (2002) for a discussion about the units to be used in geographical economics.

use to locate in the same areas but because of some kind of interindustry linkages among them (*colocalization*). This distinction is of extreme importance because allows to better understand location process and, therefore, to help firms providing the type of environment (e.g., spatial characteristics, firms, specialised services, interindustry linkages, and so on) that they do need. Following previous distinction, the *joint-localization* means that there are some firms (from different industries) that share the same spatial requirements (e.g., they do need to access the same type of inputs, services, infrastructures, etc.), so they tend to locate in the same areas. But the *colocalization* is strongly different and implies that firms need to be close to their suppliers / clients, which means that firms of different industries will cluster together.

### **3. Data and methodology**

#### **3.1 Data**

Our data set is referred to 2006 and comprises firms<sup>7</sup> from manufacturing, services and agriculture.

The source of this data base is SABI (*Sistema de Análisis de Balances Ibéricos*), which uses data from Mercantil Register including balance sheets and income and expenditure accounts. For each firm we also know the number of employees, the industry to which the firm belongs (four digit NACE code), and the amount of sales and assets, among other variables and, a relevant information for the purposes of this paper, as detailed geographical location of the firm. Nevertheless, SABI dataset has also two important shortcomings. The first one is referred to the sample. Even if the number of firms is so large (e.g., 581,712 service firms for the 2007 edition), microfirms and self-employed individuals are not considered, but it is reasonable to assume that spatial distribution of such activities is so close than those of included firms. The second one is about the nature of the units, since SABI covers only firms, not

---

<sup>7</sup> It is worth to note that the data set is about firms (not establishments) and that each firm could have more than one establishment, although in most of the registers the firm has only one establishment.



establishments,<sup>8</sup> being the later more appropriate for the analysis of the spatial distribution of economic activity. In any case, since SABI covers most of economic activity carried out in Spain previous disadvantages are easily overtaken.<sup>9</sup>

[INSERT MAP 1]

Map 1 shows spatial distribution of firms included at the data set. Red (blue) points mean higher (lower) number of firms. It is important to notice that number of firms varies strongly across industries, that more populated areas concentrate higher number of firms and that some industries tend to cluster into specific areas.

In order to discriminate if closeness of firms belonging to different industries is explained by interindustry linkages (i.e., supply chains across industries) or by sharing accessibility to similar spatial characteristics we use Spanish Input Output Tables to check first option.<sup>10</sup>

### **3.2 Methodology of cluster identification**

Proposed methodology partially follows contributions of Duranton and Overman (2005), Brenner (2003 and 2004) and Ellison and Glaser (1997), but departing from previous contributions we improve such approaches by several ways.

Firstly, we divide space into homogeneous cells of different sizes. This is quite different from strategies followed by other scholars, like administrative units (Brenner, 2003 and 2004; and Ellison and Glaser, 1997) or distance among firms (Duranton and Overman, 2005). Previous strategies have several shortcomings as López-Bazo (2006) points out: not taking into account precise

---

<sup>8</sup> Other alternative statistical sources as Censo de Locales (INE) are not currently updated while having as observation units firms instead of establishments also provides useful information since it highlights role played by municipalities chosen by firms as headquarters' sites.

<sup>9</sup> There are alternative datasets like DIRCE (INE) but, unfortunately, data is presented only at 2-digit level and geographical location of the firms is also highly spatially aggregated.

<sup>10</sup> Spanish Input Output matrix is from 2000 and covers all economic industries at 2 digits of NACE classification (*Instituto Nacional de Estadística*).

location of firms, limitations due to administrative special aggregation levels in each country, the difficulty to compare the results obtained for different levels of administrative aggregation, the non-economic nature of such administrative units, the size differences across administrative units, MAUP problem that could create spurious correlation among variables and the fact that such administrative divisions do not take into account neighbour effects across units.

Secondly, we create industry specific maps departing from firms' georeferenced data. Even if this approach is similar to the one used by Duranton and Overman (2005), they consider distances among firms while we focus at the areas occupied by firms. Nevertheless our dataset (SABI) offers data at 3-digit level, we have decided to use a selected version at 2-digit level since the reliability is higher at this level and there are also some computational constraints when working with a high number of industries. By this way we can analyse both big areas like countries and smaller areas like cities. Nevertheless, this approach has some shortcomings, being the most import of them is the fact that we are considering only those areas where there are firms located, without taking into account size neither number of firms located there.<sup>11</sup> In any case we could partially solve this disadvantage by reducing cell's size to a certain extent,<sup>12</sup> but if size is so small that there is only one firm, then it is not possible to identify the existence of any agglomeration pattern. Our approach allows us to compare the observed spatial distribution of firms with random simulations of such distribution and to check if there is some kind of concentration compared to the random distribution.

Thirdly, we create multiple random industry specific maps under two conditions: *i*) total number of firms at each industry remains constant and *ii*) total number of

---

<sup>11</sup> This is a (simple) starting point that could be easily improved by taking into account intensity of land use by the way of consider some indicators like number of jobs, production value or sales' levels, among others. This should allow comparing expected results in terms of, for instance, number of jobs, with real results, but has also some (potential) limitations regarding accuracy of data.

<sup>12</sup> Nevertheless, main problems are about heterogeneity of firm size, so it seems that a better solution should be to use size of firms (e.g., employment) instead than just the number (or the existence) of firms.

firms at each cell remains constant.<sup>13</sup> By this way we are comparing the same number of firms but with different industry distributions (for each cell we expect to find the same industry distribution that for the whole sample). Thus, if according to real data there is a cell with only one firm, in our simulations this cell will have also one firm, but the industry will result as a random variable depending on industry distribution.

Fourthly, we compare the number of cells where there are firms (according to real data) with the expected number of cells with firms, and we obtain a concentration index similar to the one by Ellison and Glaeser (1997), but with some important differences. Concretely, our methodology does not focus on agglomeration issues which allows us to analyse industry distribution, and while our index is centered at 1 (values below 1 indicate concentration and values over 1 indicates dispersion), Ellison and Glaeser's (1997) index ranges between zero and infinite, so they arbitrary define a concentration threshold.

Fifthly, we generalize our approach to several industries (X-Clustering). Methodologically, this is quite similar to using only one sector but here we are analysing if a group of industries tend to locate together (colocalization).

Sixthly, we make a cluster mapping using raster data in the following way: we compare real spatial distribution of firms with several computational simulations; if number of firms of an industry is significantly higher than the one obtained by simulation procedures we assume that there is a cluster.

Seventhly, we make a cluster mapping using vectorial data in the following way: once we have determined the cells – clusters we evaluate the economic activity

---

<sup>13</sup> This later requirement implies that firms localise randomly inside "occupied" cells (i.e., areas where real firms are located) as Duranton and Overman (2008) do. This approach means that firms are expected to be located only in such places available for economic activity (as real data shows). Unfortunately, a major shortcoming that arises from this point of view is that it assumes that firms could be located elsewhere where are other firms, no matter their industry, which is not so much realistic (especially at a 2/3 digit level). An extension of this work (and, additionally, a possible solution for this shortcoming) could be to consider that manufacturing, services and agriculture firms could be located where are other firms from, respectively, manufacturing, services and agriculture.

(firms, jobs and production) in each one of clusters, both in absolute and relative terms.

Eighthly, the self-organizing maps allow to show the local microstructure of industries.

According to previous comments, our methodology can complement previous approaches based on distribution's comparisons (Brenner, 2003 and 2004; Ellison and Glaser, 1997) and on distance's distributions (Duranton and Overman, 2005).

#### **4. Main results**

Our main results show that location decisions of firms (and, therefore, their concentration / dispersion patterns) are driven by several industry-specific determinants (i.e., belonging to a manufacturing or services activity or to a specific industry inside them) and also by their technological level. In some vertically integrated industries reducing distance to providers / suppliers is a key issue, while other types of industries do not need such spatial proximity. Additionally, there are industries in which there are no clear location patterns and show a homogeneous distribution of firms.

[INSERT TABLE 1]

Table 1 illustrates which are the expected spatial distributions of firms across regular cells<sup>14</sup> (according to the number of firms inside each industry) and which is the real (observed) spatial distribution of such firms. Concretely, it is shown in how many cells ( $X$ ) there are firms from industry  $y$  (i.e., this is the "real" spatial distribution of firms"); the expected number of cells (Mean) where firms of industry  $y$  should appear (according to the total number of firms at each industry) if they were randomly spatially distributed; and a collocation index

---

<sup>14</sup> Those regular cells have an area of 100 km<sup>2</sup> (10 km \* 10 km).

(Index) that relate previous measures (i.e.,  $\text{Index} = X / \text{Mean}$ ). This index can be understood in the following way: if  $\text{Index} < 1$ , this means that the industry  $y$  appears in less cells times than expected (i.e., this industry is spatially concentrated in a smaller number of cells); and if  $\text{Index} > 1$ , this means that the industry  $y$  appears in more cells than expected (according to a random distribution), so this industry is spatially dispersed. So, there is some kind of location behaviour that deserves to be analyzed, since it could be a cluster (firms from industry  $y$  tend to locate together) or not.

If we care on technological level, it seems that lower the technological level of the industry higher the spatial dispersion (Table 1). So, high-tech firms tend to be more spatially concentrated than low-tech firms<sup>15</sup>. This appears to be logical since markets and resources of such firms tend to be concentrated in few areas, so there is no logical reason for a dispersion pattern.

Regarding differences between manufacturing and services, our results (Table 1) are even clearer than previous ones, and show that while most of services activities show high concentration levels (e.g., Financial intermediation, Education, Business services, etc.), manufacturing activities are more dispersed (Agriculture and fishing, Food, beverages and tobacco, etc.). These results have to do with spatial distribution of population and economic activity and with production and distribution requirements of manufacturing and services. Concretely, while most of services need face-to-face interactions, they strongly depend on where customers (both firms and individuals) are located, but given that manufacturing goods can be easily transported, such interactions are not essential, and firms can locate elsewhere and later transport final products to their markets.

Until this point we have analysed spatial distribution of firms at a single industry level and we have shown that looking at some industry specificities (i.e.,

---

<sup>15</sup> As an example, Index of high-tech industries like Office machinery, computers and medical, precision and optical instruments (0,644) or Electrical machinery and apparatus (0,664) are clearly lower than those of some low-tech industries like Food, beverages and tobacco (1,452) or Agriculture and fishing (1,424).

manufacturing vs. services and high-tech vs. low-tech) helps us to understand such location patterns.

[INSERT TABLE 2]

But this situation gets more complicated if we take into account location of more than one industry. So, next step is to check for the existence and extent of clusters by checking if pairs of industries (or groups of three or four) tend to be located close to each other. For instance, if we consider what's happening for pairs of industries, Table 2 summarises main findings and shows a selection of all the possible combinations of pairs of industries<sup>16</sup>. Now, previous indicators are slightly different and include: the codes of industries  $y$  and  $i$ , respectively; the number of times ( $X$ ) that firms of industry  $y$  and industry  $i$  appear together inside the same cell; the expected number of times (Mean) that firms of industry  $y$  and industry  $i$  should appear together inside the same cell (according to the total number of firms for both industries) if they were randomly spatially distributed; and a collocation index (Index) that relate previous measures (i.e.,  $\text{Index} = X / \text{Mean}$ ). The Index can be understood in the following way: if  $\text{Index} < 1$ , this means that this industry combination ( $y$  and  $i$ ) appears less times than expected and (for same reasons that we will analyze later) this pair of firms tend to do not be located in the same areas; and if  $\text{Index} > 1$ , this means that this industry combination appears more times than expected, so this pair of industries use to be located in the same areas (they use to cluster together). Therefore, if  $\text{Index} > 1$  it could be a cluster (both industries locate together because they have strong interindustry linkages) or it could be an example of collocation (both industries locate together because they do need the same type of economic environment, but without having any kind of interindustry relationship between them). The procedure to be followed is, firstly, to identify such similar location patterns and, secondly, to discriminate between previously mentioned proximity explanations.

---

<sup>16</sup> Since all the possible combinations of pairs of industries consists on 378 pairs, here we only show results for the top-10 pairs with the lower values of the Index and for the top-10 pairs with the higher values of the Index.

According to the industrial classification of 28 industries, there are 378 possible pairs of industries to be found. Most of them (324) show a collocation index  $< 1$ , which means that this pair of industries appears fewer times than expected, while only in 54 pairs results of collocation index are  $> 1$ , which means a cluster or a collocation.

[INSERT TABLE 3]

Table 3 shows that among the pairs of industries with higher values of the collocation index there are the following: Agriculture and Fishing / Food, beverages and Tobacco, and Extraction activities / Food, beverages and Tobacco. A close analysis to such possible clusters or collocated activities shows that there is a small number of industries involved (most of them usually appear in the industry pairs with higher levels of the collocation index): Agriculture and Fishing; Extraction activities; Food, beverages and Tobacco; Wood, Furniture and other manufacturing activities; Non-metallic Mineral Products; Construction; Retail and Repair of personal and household goods.

As in Table 2, it is not feasible to explain in a detailed way all the 378 combinations so we have selected again the “top 10” and the “bottom 10” pairs of industries. Once we have identified them, next step is to try to explain those results in terms of interindustry relationships between pairs of industries according to Input-Output tables.

At Table 3 we show interindustry linkages in terms of intermediate consumption between pairs of industries. We assume that if two pairs of industries are linked due to such interindustry intermediate consumption they can be identified as a part of a cluster, while if there is no such relationship their location patterns can be explained in terms of collocation.

Our results show that there is not a clear pattern in terms of interindustry linkages. So, looking at this data it is not obvious to explain firm collocation behaviour (or, alternatively, absence of firm collocation behaviour) in terms of such linkages. Therefore, a cluster explanation could not be found. Concretely,

looking at the bottom of the table, will show that some pairs of industries have important linkages (e.g., 34,51% of the intermediate consumption of industry 1 comes from industry 3, while 15,84% of the intermediate goods sold by industry 3 goes to industry 1), while others do not have such linkages or they are much weaker (e.g., industries 2 and 3, industries 2 and 9, industries 3 and 17, etc.). Finally, an overview of the top of the table will show a similar portrait: while some of the pairs of industries reach important interindustry linkages (e.g., industries 22 and 24, industries 19 and 22, etc.) others are less linked (e.g., industries 12 and 26, industries 14 and 26, etc.).

[INSERT FIGURE 3]

[INSERT FIGURE 4]

Only 15 out of 28 industries show a collocation index higher than 1, while the remaining 13 industries are identified as a concentrated (blue), according to our concentration index. But, surprisingly, industries with higher collocation relationships are defined as disperse (red) or intermediate (green) in terms of industry concentration. If we focus on disperse industries only 4 of them (all from services) show collocation's relationships but only with a few industries. In order to better illustrate such interindustry relationships we present some self-organizing maps. Those maps show that there is a dual situation regarding collocation relationships: firstly, most of industries do not have such relationships and, secondly, a smaller number of industries that tend to collocate together frequently.

## **5. Conclusions**

We have contributed to extant literature on cluster identification by designing a procedure to identify groups of industries that tend to cluster together and, then, to show up if this behaviour is explained in terms of vertical integration or if those industries share some common location determinants. This distinction



allows going further in the analysis of firm location determinants since our results shown that diversified clusters are not casual and are strongly determined by industry characteristics. Concretely, it means that firms do not need only neighbours but they do need “specific” neighbours in order to maximise their performance.

The methodology proposed in this paper allows to better explain main reasons driving cluster formation but much more work needs to be done into this direction, concretely about identifying cluster’s size in order to better capture cluster’s borders. This methodology implies dividing space into homogeneous cells of equal size but, obviously, cell’s size influences the number and characteristics of the identified clusters. Concretely, bigger cells means higher probability of finding a cluster inside them, while smaller cells means that probability of interindustrial cluster diminishes as the number of firms in each cell will be smaller. Therefore, since in this paper we have assumed equality of sizes for all the clusters, it appears that the use of flexible sizes fits better with real distribution of economic activity and is a promising line for future research.

## References

- Arauzo-Carod, J.M. (2008): "Industrial Location at a Local Level: Comments on the Territorial Level of the Analysis", *Tijdschrift voor Economische en Sociale Geografie - Journal of Economic & Social Geography* **99**: 193-208.
- Arauzo-Carod, J.M. and Manjón-Antolín, M. (2004): "Firm Size and Geographical Aggregation: An Empirical Appraisal in Industrial Location", *Small Business Economics* **22**: 299-312
- Arbia, G. (2001): "Modeling the Geography of Economic Activities on a Continuous Space", *Papers in Regional Science* **80**: 411-424.
- Boix, R. (2008): "Los distritos industriales en la Europa Mediterránea. Los mapas de Italia y España", in V. Soler (ed.), *Mediterráneo Económico*, Fundación Cajamar: Almería.
- Boix, R. and Galletto, V. (2008): "Marshallian industrial districts in Spain", *Scienze Regionali / Italian Journal of Regional Science* **7 (3)**: 29-52.
- Brenner, T. (2003): "An identification of local industrial clusters in Germany", *Papers on Economics & Evolution* # 0304, Max Plack Institute, Jena.
- Brenner, T. (2004): *Local industrial clusters: existence, emergence and evolution*, Routledge: London.
- Devereaux, M.; Griffith, R. and Simpson, H. (2004), "The geographic distribution of production activity in the UK", *Regional Science and Urban Economics* **34 (5)**: 533-564.
- Duranton, G. and Overman, H.G. (2008): "Exploring the Detailed Location Patterns of U.K. manufacturing Industries using Microgeographic Data", *Journal of Regional Science* **48 (1)**: 213-243.
- Duranton, G. and Overman, H.G. (2005): "Testing for Localization Using Microgeographic Data", *Review of Economic Studies* **72**: 1077-1106.
- Duranton, G. and Puga, D. (2004): "Micro-foundations of urban agglomeration economies". In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. IV. North-Holland.
- Ellison, G. and Glaeser, E.L. (1997): "Geographic concentration in US manufacturing industries: A dartboard approach", *Journal of Political Economy* **195**: 889-927.
- Hoover (1936): "The measurement of industrial location", *The Review of Economics and Statistics* **18**: 162-171.

Karlsson, C.; Johansson, B. and Stough, R.R. (2005): "Industrial Clusters and Inter-Firm Networks: An Introduction". In: Karlsson, C.; Johansson, B. and Stough, R.R. (Eds.), *Industrial Clusters and Inter-Firm Networks*, Edward Elgar: Cheltenham.

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

Lambert, D.M.; McNamara, K.T. and Garrett, M.I. (2006): "An Application of Spatial Poisson Models to Manufacturing Investment Location Analysis", *Journal of Agricultural and Applied Economics* **38**: 105-121.

López-Bazo, E. (ed.) (2006): *Definición de la metodología de detección e identificación de clusters industriales en España*, Dirección General de la Pequeña y Mediana Empresa (DGPYME): Madrid.

Marshall, A. (1890): *Principles of Economics*, MacMillan: New York.

Maurel, F. and Sédillot, B. (1999), "A measure of the geographic concentration in French manufacturing industries", *Regional Science and Urban Economics* **29**: 575-604.

NWB Team (2006), Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan, <http://nwb.slis.indiana.edu>

Olsen, J. (2002): "On the Units of Geographical Economics", *Geoforum* **33**: 153-164.

Openshaw, S. and Taylor, P.J. (1979): "A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem". In N. Wrigley, *Statistical Applications in the Spatial Sciences*, London, Pion: 127-144.

Pablo-Martí, F. and Muñoz-Yebra, C. (2009): "Localización empresarial y economías de aglomeración: el debate en torno a la agregación espacial", *Investigaciones Regionales* **15**: 139-166.

Paluzie, E; Pons, J. and Tirado, D. (2004): "The geographical concentration of industry across Spanish regions, 1856-1995", *Review of Regional Research* **24 (2)**: 143-160.

Parr, J.B. (2002): "Missing Elements in the Analysis of Agglomeration Economies", *International Regional Science Review* **25 (2)**: 151-168.

Porter, M. (1998): "Clusters and the new economics of competition", *Harvard Business Review* **76 (6)**: 77-90.

Sonis, M.; Hewings, G.J.D. and Guo, D. (2008): "Industrial clusters in the input-output economic system". In C. Karlsson, *Handbook of Research on Cluster Theory*, Cheltenham, Edward Elgar: 153-168.

Viladecans, E. (2004): "Agglomeration economies and industrial location: city-level evidence", *Journal of Economic Geography* **4/5**: 565-582.

Wrigley, N. (1995): "Revisiting the Modifiable Areal Unit Problem and the Ecological Fallacy". In A.D. Cliff, P.R. Gould, A.G. Hoare and N.J. Thrift (eds.), *Diffusing Geography*, Oxford, Blackwell: 49-71.

## Tables

**Table 1: Concentration patterns of firms at a single industry level**

Code	Industry	X	Mean	STD	Index	X-2S	X+2S	Concentrated	Dispersed
22	Financial intermediation	882	1480,11	17,6811804	0,59590166	1444,74764	1515,47236	TRUE	FALSE
6	Paper and publishing	947	1494,58	17,9619013	0,63362282	1458,6562	1530,5038	TRUE	FALSE
13	Office machinery, computers and medical, precision and optical instruments	324	502,86	13,3553001	0,64431452	476,1494	529,5706	TRUE	FALSE
26	Education	790	1209,17	17,5580164	0,65334072	1174,05397	1244,28603	TRUE	FALSE
14	Electrical machinery and apparatus	520	782,36	14,6890463	0,66465566	752,981907	811,738093	TRUE	FALSE
24	Business services	1360	1979,03	21,1557261	0,68720535	1936,71855	2021,34145	TRUE	FALSE
23	Real estate activities	1957	2803,29	18,8970069	0,69810829	2765,49599	2841,08401	TRUE	FALSE
28	Other services	1375	1819,52	21,5638493	0,75569381	1776,3923	1862,6477	TRUE	FALSE
12	Machinery and equipment	820	1076	17,2533118	0,76208178	1041,49338	1110,50662	TRUE	FALSE
4	Textiles, leather clothes and shoes	1169	1523,26	17,384319	0,76743301	1488,49136	1558,02864	TRUE	FALSE
27	Health and veterinary activities, social services	1122	1458,21	20,5029168	0,7694365	1417,20417	1499,21583	TRUE	FALSE
8	Rubber and plastic products	698	903,5	19,1498609	0,77255119	865,200278	941,799722	TRUE	FALSE
25	Public administration	141	179,3	7,24812759	0,78639152	164,803745	193,796255	TRUE	FALSE
7	Chemical products	734	837,17	14,8691634	0,87676338	807,431673	866,908327	TRUE	FALSE
10	Basic metals	567	629,55	16,7460986	0,90064332	596,057803	663,042197	TRUE	FALSE
15	Transport and communications	668	726,47	16,8111645	0,91951491	692,847671	760,092329	TRUE	FALSE
19	Trade and repair	2888	3035,78	16,6336521	0,95132058	3002,5127	3069,0473	TRUE	FALSE
16	Recycling	349	359,69	9,90020406	0,97027996	339,889592	379,490408	FALSE	FALSE
11	Fabricated metal products	1682	1701,7	19,8267751	0,98842334	1662,04645	1741,35355	FALSE	FALSE
21	Transport and communications	2090	2034,14	19,9479221	1,02746124	1994,24416	2074,03584	FALSE	TRUE
17	Construction	2706	2585,57	21,9674944	1,04657774	2541,63501	2629,50499	FALSE	TRUE
20	Hotels and restaurants	2238	2136,5	20,4181045	1,04750761	2095,66379	2177,33621	FALSE	TRUE
18	Electricity and water distribution	795	739,43	15,2674838	1,07515248	708,895032	769,964968	FALSE	TRUE
5	Wood, furniture and other manufactures	1734	1610,89	20,5956232	1,07642359	1569,69875	1652,08125	FALSE	TRUE
9	Non-metallic mineral products	1297	1125,88	18,1566027	1,15198778	1089,56679	1162,19321	FALSE	TRUE
2	Extractive activities	1152	823,16	15,7015858	1,39948491	791,756828	854,563172	FALSE	TRUE
1	Agriculture and fishing	2409	1691,54	20,5354682	1,42414604	1650,46906	1732,61094	FALSE	TRUE
3	Food, beverages and tobacco	2236	1540,31	20,5001577	1,45165584	1499,30968	1581,31032	FALSE	TRUE

Note: X-2S equals X minus 2 standard deviations and X+2S equals X plus 2 standard deviations.

Source: own calculations.

**Table 2:** Concentration patterns of firms for pairs of industries

<i>Top-10 industries with the lower values of the collocation index</i>										
<b>Code industry y</b>	<b>Code industry i</b>	<b>X</b>	<b>Mean</b>	<b>STD</b>	<b>Index</b>	<b>X-2S</b>	<b>X+2S</b>	<b>Concentrated</b>	<b>Dispersed</b>	
4	22	639	1092,83	12,749	0,585	1067,333	1118,327	TRUE	FALSE	
22	23	835	1424,44	16,580	0,586	1391,280	1457,600	TRUE	FALSE	
14	22	391	662,3	12,630	0,590	637,039	687,561	TRUE	FALSE	
22	24	748	1259,35	15,338	0,594	1228,675	1290,025	TRUE	FALSE	
14	26	361	606,94	10,773	0,595	585,394	628,486	TRUE	FALSE	
6	22	651	1080,17	14,169	0,603	1051,833	1108,507	TRUE	FALSE	
12	26	464	769,81	13,134	0,603	743,542	796,078	TRUE	FALSE	
4	26	574	948,17	14,318	0,605	919,534	976,806	TRUE	FALSE	
19	22	878	1449,2	17,590	0,606	1414,021	1484,379	TRUE	FALSE	
22	26	569	934,68	13,485	0,609	907,709	961,651	TRUE	FALSE	

<i>Top-10 industries with the higher values of the collocation index</i>										
<b>Code industry y</b>	<b>Code industry i</b>	<b>X</b>	<b>Mean</b>	<b>STD</b>	<b>Index</b>	<b>X-2S</b>	<b>X+2S</b>	<b>Concentrated</b>	<b>Dispersed</b>	
1	17	2013	1572,54	18,303	1,280	1535,934	1609,146	FALSE	TRUE	
1	19	2120	1648,27	20,232	1,286	1607,805	1688,735	FALSE	TRUE	
2	9	785	609,29	11,853	1,288	585,584	632,996	FALSE	TRUE	
2	17	1059	799,41	15,180	1,325	769,049	829,771	FALSE	TRUE	
2	19	1100	815,4	15,300	1,349	784,801	845,999	FALSE	TRUE	
3	17	1957	1446,37	19,204	1,353	1407,963	1484,777	FALSE	TRUE	
3	19	2042	1506,4	19,019	1,356	1468,361	1544,439	FALSE	TRUE	
1	2	990	723,86	13,682	1,368	696,496	751,224	FALSE	TRUE	
2	3	985	700,7	13,457	1,406	673,787	727,613	FALSE	TRUE	
1	3	1773	1193,15	15,684	1,486	1161,782	1224,518	FALSE	TRUE	

Source: own calculations.

**Table 3: Interindustry linkages according to the collocation index**

<i>Top-10 industries with the lower values of the collocation index</i>								
Industry x buys (a)	Industry y sells (b)	Purchases x to y (c)	Total purchases x (d)	Total sells y (e)	(c / d) (%)	(c / e) (%)	Index	
4	22	258,10	12.305,50	16.868,80	2,10	1,53	0,584	
22	23	635,90	8.520,00	18.743,00	7,46	3,39	0,586	
14	22	135,50	8.143,10	16.868,80	1,66	0,80	0,590	
22	24	3.869,60	8.520,00	59.803,40	45,42	6,47	0,593	
14	26	20,60	8.143,10	1.597,20	0,25	1,29	0,594	
6	22	220,40	11.111,00	16.868,80	1,98	1,31	0,602	
12	26	17,20	8.694,10	1.597,20	0,20	1,08	0,602	
4	26	58,80	12.305,50	1.597,20	0,48	3,68	0,605	
19	22	1.838,10	40.752,90	16.868,80	4,51	10,90	0,605	
22	26	35,90	8.520,00	1.597,20	0,42	2,25	0,608	
<i>Top-10 industries with the higher values of the collocation index</i>								
Industry x buys (a)	Industry y sells (b)	Purchases x to y (c)	Total purchases x (d)	Total sells y (e)	(c / d) (%)	(c / e) (%)	Index	
1	17	212,40	13.773,00	43.515,60	1,54	0,49	1,280	
1	19	1.675,00	13.773,00	34.413,70	12,16	4,87	1,286	
	9	29,30	6.282,90	16.546,10	0,47	0,18	1,288	
2	17	112,90	6.282,90	43.515,60	1,80	0,26	1,324	
2	19	208,40	6.282,90	34.413,70	3,32	0,61	1,349	
3	17	225,90	45.829,90	43.515,60	0,49	0,52	1,353	
3	19	2.504,90	45.829,90	34.413,70	5,47	7,28	1,355	
1	2	505,60	13.773,00	13.841,30	3,67	3,65	1,367	
2	3	0,40	6.282,90	30.001,60	0,01	0,00	1,405	
1	3	4.752,60	13.773,00	30.001,60	34,51	15,84	1,485	

Notes: (a) and (b) are industry codes and (c), (d) and (e) are millions euros.

Source: Spanish Input – Output Table (INE) and own calculations.

**Table 4:** Times in which the sector appears co-located (P-A Index)

<b>Code</b>	<b>Industry</b>	<b>All</b>	<b>0.787</b>	<b>Mean</b>	<b>1</b>
1	Agriculture and fishing	5	2	9	11
2	Extractive activities	2	2	11	12
3	Food, beverages and tobacco	4	2	9	12
4	Textiles, leather clothes and shoes	18	3	5	0
5	Wood, furniture and other manufactures	5	6	6	10
6	Paper and publishing	21	4	2	0
7	Chemical products	9	4	14	0
8	Rubber and plastic products	17	5	5	0
9	Non-metallic mineral products	2	4	9	12
10	Basic metals	6	4	17	0
11	Fabricated metal products	7	6	7	7
12	Machinery and equipment	16	4	7	0
13	Office machinery, computers and medical, precision and optical instruments	26	1	0	0
14	Electrical machinery and apparatus	26	0	1	0
15	Transport and communications	5	3	18	1
16	Recycling	1	0	24	2
17	Construction	12	1	5	9
18	Electricity and water distribution	5	4	9	9
19	Trade and repair	12	1	8	6
20	Hotels and restaurants	10	3	8	6
21	Transport and communications	10	3	7	7
22	Financial intermediation	23	2	1	0
23	Real estate activities	14	4	6	3
24	Business services	18	2	7	0
25	Public administration	12	12	3	0
26	Education	24	2	1	0
27	Health and veterinary activities, social	17	4	6	0
28	Other services	15	4	7	1

Source: own elaboration and NOMBRE SOFTWARE.

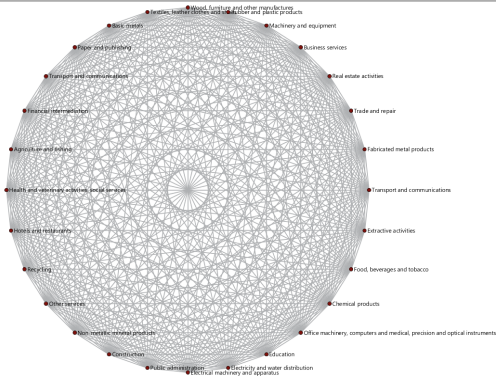
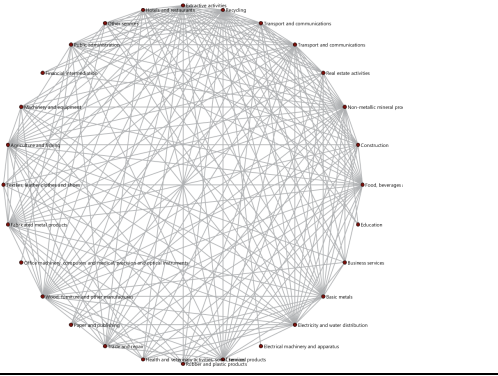
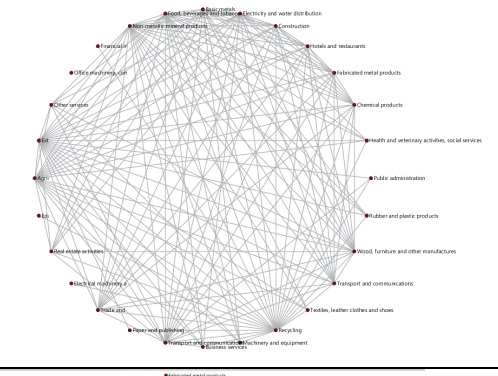
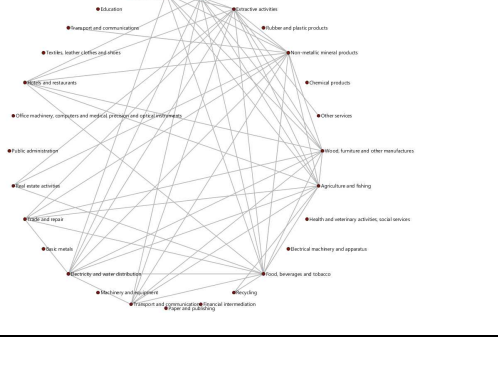


**Table 6:** Number of samples in which the co-location appears (P-A Index)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
1		4	4	3	4	1	3	2	4	3	4	3	1	1	3	4	4	4	4	4	4	1	3	3	2	1	3	3	
2			4	3	4	3	3	3	4	3	4	3	1	1	3	3	4	4	4	4	4	2	4	3	3	2	3	4	
3				3	4	2	3	3	4	3	4	3	1	1	3	4	4	4	4	4	4	1	4	3	2	1	3	3	
4					2	1	1	1	3	1	2	1	1	1	1	3	1	2	1	1	1	0	1	1	1	1	1	1	
5						1	3	2	4	3	4	2	1	1	3	3	4	4	4	4	4	1	3	2	2	1	2	3	
6							1	1	2	1	1	1	1	1	2	3	1	2	1	1	1	1	1	1	1	1	1	1	
7								2	3	3	3	2	1	1	3	3	3	3	3	3	3	1	2	1	2	1	1	1	
8									3	3	2	1	1	1	2	3	1	1	1	1	1	1	1	1	1	1	1	1	
9										3	4	3	1	1	4	3	4	4	4	4	4	2	4	3	2	2	3	3	
10											3	3	1	1	3	3	3	3	3	3	3	1	3	2	2	1	2	2	
11												2	1	1	3	3	4	4	3	3	3	1	2	1	1	1	2	2	
12													1	1	3	3	1	2	1	1	1	1	1	1	1	1	1	1	
13														1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	
14															1	3	1	1	1	1	1	1	1	1	1	1	1	1	
15																3	3	3	3	3	3	1	3	3	3	1	2	3	
16																	3	3	3	3	3	3	3	3	3	3	3	3	
17																		4	3	4	4	1	1	1	2	1	1	1	
18																			4	3	4	1	3	3	2	1	3	3	
19																				3	3	1	1	1	2	1	1	1	
20																					3	1	2	1	2	1	1	2	
21																						1	2	1	2	1	1	2	
22																								1	1	1	1	1	
23																									1	1	1	1	
24																										1	1	1	
25																											1	1	
26																												1	
27																													1

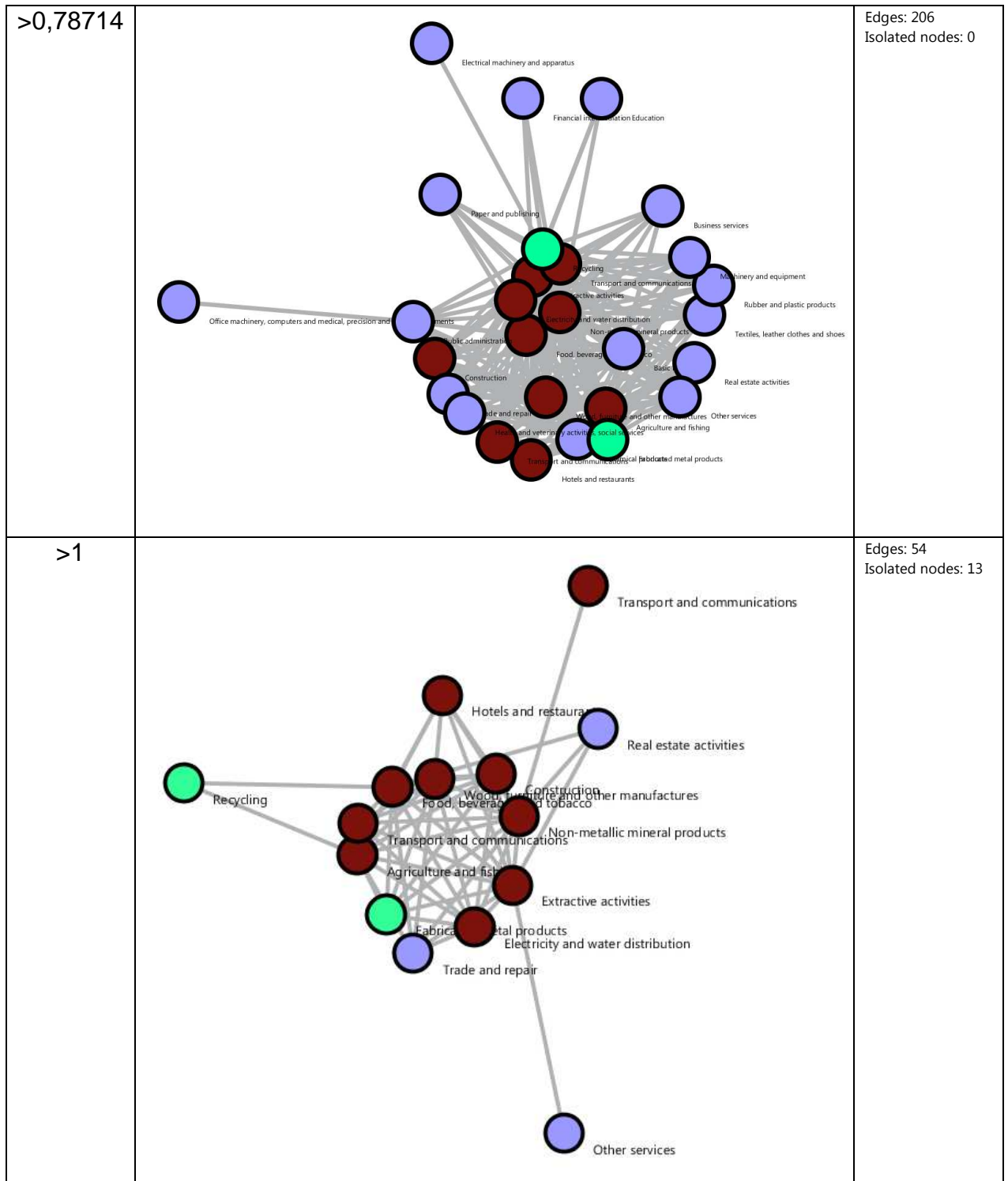
Source: Own elaboration

**Figure 3: Sectorial conectiveness. Radial Tree Graphs**

	Radial Tree/Graph	Nodes: 28
All		Edges: 378 Isolated nodes: 0 Average degree: 27.000000000000007 This graph is weakly connected. There are 1 weakly connected components. (0 isolates) The largest connected component consists of 28 nodes. Density (disregarding weights): 1 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 0,83905 densities (weighted against observed max) weight: 0,56465
>0,78714		Edges: 206 Isolated nodes: 0 Average degree: 14.71428571428572 This graph is weakly connected. There are 1 weakly connected components. (0 isolates) The largest connected component consists of 28 nodes. Density (disregarding weights): 0,54497 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 0,52023 densities (weighted against observed max) weight: 0,35009
>Mean 0.84070		Edges: 160 Isolated nodes: 1 Average degree: 11.428571428571429 This graph is not weakly connected. There are 2 weakly connected components. (1 isolates) The largest connected component consists of 27 nodes. Density (disregarding weights): 0,42328 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 0,42136 densities (weighted against observed max) weight: 0,28355
>1		Edges: 54 Isolated nodes: 13 Average degree: 3.857142857142857 This graph is not weakly connected. There are 14 weakly connected components. (13 isolates) The largest connected component consists of 15 nodes. Density (disregarding weights): 0,14286 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 0,16448 densities (weighted against observed max) weight: 0,11069

Source: Own elaboration using NWB.

**Figure 4: Sectorial proximity. Spring Graphs**



Source: Own elaboration using NWB.