UNIVERSITAT ROVIRA I VIRGILI

WORKING PAPERS

Col·lecció "DOCUMENTS DE TREBALL DEL DEPARTAMENT D'ECONOMIA"

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

Federico Pablo-Martí
Josep-Maria Arauzo-Carod

Document de treball nº - 6 - 2010

DEPARTAMENT D'ECONOMIA
Facultat de Ciències Econòmiques i Empresarials

Edita:

Departament d'Economia
http://www.fcee.urv.es/departaments/economia/public_html/index.html
Universitat Rovira i Virgili
Facultat de Ciències Econòmiques i Empresarials
Avgda. de la Universitat, 1
432004 Reus
Tel. +34 977 759 811
Fax +34 977 300 661

Dirigir comentaris al Departament d'Economia.

Dipòsit Legal: T - 1835 - 2010

ISSN 1988 - 0812

DEPARTAMENT D'ECONOMIA
Facultat de Ciències Econòmiques i Empresarials

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

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

Abstract

The aim of this paper is to analyse the colocation patterns of industries and firms. We study the spatial distribution of firms from different industries at a microgeographic level and from this identify the main reasons for this locational behaviour. The empirical application uses data from Mercantile Registers of Spanish firms (manufacturers and services). Inter-sectorial linkages are shown using self-organizing maps.

(♠) 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 (QURE)
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)
Av. Diagonal, 690; 08034 - Barcelona

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

*This research was partially funded by the grants SEJ2007-64605/ECON and SEJ2007-65086/ECON, the "Xarxa de Referència d'R+D+I en Economia i Polítiques Públiques" and the SGR Program (2009-SGR-322) of the Catalan Government and by the PGIR program N-2008PGIR/05 of the Rovira i Virgili University. We are grateful to seminar participants at the 12th EUNIP Conference (Universitat Rovira i Virgili). Any errors are, of course, our own.

1. Introduction

The analysis of the spatial distribution of economic activity can be applied in various areas such as urban planning, infrastructures, firm supporting policies and land use, among others, and is receiving increasing attention from researchers. Traditionally, scholars have analyzed how economic activities were spatially distributed in extant administrative units (e.g. counties, regions, etc.). Unfortunately, these analyses suffer from the shortcoming that administrative units do not always coincide with real economic areas and are sometimes arbitrary. Additionally, administrative units vary greatly in size and shape, for instance, and these spatial specificities can make analysis more difficult.

To deal with such constraints, recent research has started to investigate microgeographic data. In particular, smaller spatial units are being used. These units are created by equally dividing a space into homogeneous cells and, therefore, do not exactly match any extant administrative unit.[1] The two commonest 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, whereas for a square map the distance varies depending on whether we consider the four cells adjacent to each cell (rook contiguity) or the four cells that are at the diagonal (bishop contiguity). However, a disadvantage of a hexadecimal map is that the adjacent cells are placed only in six directions rather than eight, as is the case in a square map. Furthermore, a hexagonal cell can have no other adjacent cell directly to the east or to the west.

[INSERT FIGURE 1]

[1] There are also other approaches such as those that use the stochastic methodology of Point Pattern or those that use Neuronal Networks for pattern recognition. However, these approaches are not able to do the multisectorial analyses that are the goal of this paper.

In addition to the issues regarding the spatial distribution of economic activity, there are other issues regarding whether public authorities should intervene in economic activity in order to increase the productivity of firms and territories. Regardless of the important academic debate surrounding such interventions, it should be remembered that prior to carrying out any public intervention, policy makers must first identify and select which key industries where they think this is necessary. Such analyses involve identifying the most dynamic industries and their spatial distribution patterns in terms of geographical location and clustering. Accordingly, mapping the spatial distribution of economic activity is of key importance, but there is no agreement as to which technical approach is best for designing policy. Currently, there are two main approaches: Industrial Districts and Clusters. While the former is popular mainly due to the Sforzi-ISTAT methodology, the latter is potentially easier to use because of its lower data requirements. Therefore, in this paper we will use the cluster approach due to both data availability and the shortcomings of Sforzi-ISTAT methodology.[2]

The methodology proposed in this paper aims to overcome previous constraints, to obtain more precise results and, as a result, to improve public policy design. Accordingly, in this paper we try to identify manufacturing and service clusters (from all sectors) in Spain, using data from the Mercantile Registers of 2006. Additionally, we classify these clusters according to the reasons behind the clusterization processes; that is, whether firms tend to locate together because they look for the same types of site (regardless of the industry to which they belong), or whether firms look to be located close to their suppliers / customers in order to optimise commercial exchanges.

This paper is organised as follows. In the next section we review the main literature on the spatial distribution of economic activity and 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 the

[2] Boix and Galletto (2008) identify some of the shortcomings of Sforzi-ISTAT. These include the lack of precision when defining the boundaries of local labour markets, the use of national input-output matrices, the existence of polispecialised districts, the lack of local data regarding social capital and certain general drawbacks when trying to capture the socioeconomic characteristics of local communities.

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. The spatial distribution of economic activity

The spatial distribution of economic activity has been a major topic since the seminal contributions of scholars such as 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) and William Alonso (Central Business District), to name just a few.

Researchers have used a dynamic approach to analyse how the specific characteristics of sites (usually administrative units, but also cities, counties and regions, among others) affect the location choices of new firms, and have used a static approach to estimate the spatial concentration of firms, jobs or individuals. This paper combines both approaches because we are interested in the current spatial distribution of incumbent firms and we also wish to understand the reasons for this distribution. Consequently, we do not solely aim to measure degrees of concentration (dispersion) of economic activity, but also to explain the location determinants of selected sites.

If we review empirical literature on the spatial distribution of economic activity, most researchers agree that there is a high level of concentration (especially in most developed countries), regardless of how such a concentration is measured, as is shown by Duranton and Overman (2005), Devereux et al. (2004), Maurel and Sédillot (1999) and Ellison and Glaeser (1997), among others. The Spanish case is roughly the same, as scholars such as Paluzie et al. (2004) and Viladecans (2004) have demonstrated. Our data set also points in the same direction, which means that the land sites considered by firms are

only a very small part of the absolute available land. Thus, firms tend to cluster in a few sites while most of the available land remains empty. Our data also show that firms and individuals compete for the same areas because most of those "economic sites" are close to big urban areas.

In any case, this concentration pattern is supported by plenty of empirical evidence from across the world, and scholars usually explain it in terms of increasing returns (Krugman, 1991) or external scale economies at 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 (e.g. the benefits that firms obtain by being close to other firms), but additional knowledge is needed to determine the motives behind agglomeration economies. Since the findings of Marshall (1890), agglomeration economies have been identified as the main drivers of firm concentration and three important reasons have been given for this: the presence of a specialised labour market (the presence of a pool of skilled workers), supplier availability (depending on the size of the market) and knowledge spillovers (resulting from knowledge transfers between firms). Hoover (1936) tried to measure this phenomenon more accurately and classified agglomeration economies into urbanisation economies (which result from the concentration of diverse activities) and localisation economies (which result from the concentration of similar activities)[3].

Thus, firms look to be located both close to similar firms (e.g. firms from the same industry) and to different firms (e.g. firms from another industry). However, although firms look for neighbours, not all neighbours are equally useful, and some could even be useless and harmful. This is why firms sometimes look to be located close to other firms with which they are vertically

[3] See Parr (2002) for a review of the classification of agglomeration economies.

integrated, because they need to have close linkages with their providers / suppliers. Spatial proximity[4] therefore appears to be a good enough argument for sharing the same location. There are additional reasons that explain why certain types of firms are more likely to be located in the same area than other types of firms. 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 the *Modifiable Area Unit Problem* (MAUP) appears[5], since the size of the area used in the empirical analysis will strongly determine the results obtained by the researchers and, of course, will make comparisons difficult (Duranton and Overman, 2005). Arbia (2001) provides an excellent example of such problems. In particular, he portrays a hypothetical distribution of firm location (Figure 2) in which there are four firms inside the spatial area under analysis (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 an intermediate concentration pattern (Figure 2d).

[INSERT FIGURE 2]

Figure 2 shows that spatial aggregation really matters, so researchers should be aware of this circumstance and carefully select the most appropriate areas. Unfortunately, in the past this has not been a major concern in empirical analysis mainly due to the lack of sufficiently disaggregated data[6]; recently, however, researchers have started to get access to dramatically improved datasets with extended spatial disaggregation. This is the case with our data set which contains accurate individual information about the location of firms, thus

[4] In this paper "spatial proximity" means to be located in the same cell. An extension of this approach would be to use XClusters for the spatial delimitation of such proximity.
[5] See Openshaw and Taylor (1979) for a detailed analysis and Wrigley (1995) for a further review.
[6] The influence of spatial units on the analysis of firm location has been studied in Arauzo-Carod and Manjón-Antolín (2004) and in Arauzo-Carod (2008). See Olsen (2002) for a discussion of the units to be used in geographical economics.

allowing us to technically address previous shortcomings and to freely decide the way in which a space is disaggregated, regardless of where the administrative (and usually arbitrary) boundaries are. This is particularly important because "(…) *any statistical measure based on spatial aggregates is sensitive to the scale and aggregation problems*" (Arbia, 2001, p. 414). As Duranton and Overman (2005, p. 1079) point out, "(…) *any good measure of localization must avoid these aggregation problems*".

Given these considerations, our goal is to empirically asses the location patterns of both manufacturing and services firms in Spain and try to determine if these firms tend to locate close to other firms from the same industry, to firms with close industry linkages (e.g. providers and suppliers) or to firms that share the same location requirements (e.g. accessibility to inputs, labour and infrastructures). Previous contributions have taken a similar approach. Specifically, Duranton and Overman (2008, 2005) used microgeographic (postcode level) data from the Annual Census of Production in the United Kingdom to analyse manufacturers. They computed Euclidean distances between every pair of entering establishments and compared those results with the extant distances between incumbent establishments in order to check for any possible similarities between the location patterns of entrants and incumbent establishments.

In their 2008 study, Duranton and Overman tried to identify two specific situations: the first occurs when firms from different industries locate in the same areas (*joint-localization*); the second occurs when firms from different industries also locate in the same areas because there are certain inter-industry linkages between them (*co-localization*). This distinction is extremely important because allows us to better understand location process and, therefore, to advise firms as to which is the best type of environment (e.g. in terms of spatial characteristics, firms, specialised services, inter-industry linkages, and so on). Following on from this distinction, the term *joint-localization* means that there are some firms (from different industries) that share the same spatial

requirements (i.e. they need access to the same type of inputs, services, infrastructures, etc.), which means they tend to locate in the same areas. However, *co-localization* is very different as it implies that firms need to be close to their suppliers / clients, with the result that firms from different industries will cluster together.

3. Data and methodology

3.1 Data

Our data set refers to 2006 and comprises Spanish firms[7] from manufacturing, services and agriculture.

The source of this data base is SABI (*Sistema de Análisis de Balances Ibéricos*), which uses data from the Mercantile Register including balance sheets and income and expenditure accounts. For each firm we also know the number of employees, the industry to which it belongs (the four digit NACE code), and its sales and assets, among other variables. We also have detailed information about the firm's geographical location; that is, information which is particularly relevant for the purposes of this paper. Nevertheless, the SABI dataset also has two important shortcomings. The first concerns the sample. If the number of firms is very high (e.g. 581,712 service firms for the 2007 edition), then microfirms and self-employed individuals are not taken into account, despite that fact that it is reasonable to assume that the spatial distribution of such activities is similar to that of the firms included. The second concerns the nature of the units; that is, SABI only covers firms, not establishments,[8] the latter being more appropriate for analyzing the spatial distribution of economic activity. In any case, since SABI covers most of the

[7] It is worth noting that the data set refers to firms (not establishments) and that each firm could have more than one establishment, although in most registers each firm has only one establishment.

[8] Other alternative statistical sources such as *Censo de Locales* (INE) are not currently updated, although having firms as observation units instead of establishments also provides useful information since it highlights the role of municipalities when firms are choosing where to locate their headquarters.

economic activity carried out in Spain, these disadvantages are easily overcome.[9]

[INSERT MAP 1]

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

We use the Spanish Input Output Tables to determine whether the geographical proximity of firms belonging to different industries can be explained by inter-industry linkages (i.e. supply chains across industries) or by the need to access similar areas.[10]

3.2 Methodology of cluster identification

The proposed methodology partially follows the contributions of Duranton and Overman (2005), Brenner (2006 and 2004) and Ellison and Glaser (1997), but also improves on these approaches in several ways.

First, we divide the space into homogeneous cells of different sizes. This is quite different from the strategies followed by other researchers, who have used administrative units (Brenner, 2006 and 2004; and Ellison and Glaser, 1997) or the distance between firms (Duranton and Overman, 2005). As López-Bazo (2006) points out, these strategies have several shortcomings, these being: the inability to take into account the precise location of firms, the limitations resulting from the special administrative aggregation levels in each country, the difficulties in comparing the results obtained for different levels of administrative aggregation, the non-economic nature of such administrative units, the size differences across administrative units, the modifiable areal unit problem

[9] There are alternative datasets such as DIRCE (INE) but their data is presented only at 2-digit level and geographical location of the firms is also highly spatially aggregated.
[10] Spanish Input Output matrix is from 2000 and covers all economic industries at 2 digits of NACE classification (*Instituto Nacional de Estadística*).

(MAUP), which can create spurious correlations between variables, and the fact that such administrative divisions do not take into account neighbour effects across units.

Second, we create industry specific maps that depart from the firms' georeferenced data. This approach is similar to those used by Duranton and Overman (2005); however, their maps consider distances between firms whereas we focus on the areas occupied by firms. Although our dataset (SABI) provides data at 3-digit level, we have decided to use data at 2-digit level because it is more reliable and because there are certain computational constraints when working with a high number of industries. In this way we can analyse both large areas such as countries and smaller areas such as cities. Nevertheless, this approach has some shortcomings, the most import being the fact that we are only considering those areas where firms are located without taking into account either the size or the number of firms located there.[11] We could partially solve this disadvantage by reducing the cell size to a certain extent,[12] although if the cell were so small that it contains only one firm, it would not be possible to identify the existence of any agglomeration pattern. Our approach allows us to compare the spatial distribution of firms with random simulations of such distributions to check whether there are any concentrations in the former. There are other alternatives such as kernel-smoothing (see Barlet et al., 2009; Duranton and Overman, 2005 and Silverman, 1986), but using kernels is not feasible when using bigger units such as 10 km * 10 km cells, as are used in this paper; that is, kernels could be a good strategy for urban areas because they allow smooth contiguous areas (i.e. between cells), but smoothing has already been done with the cells for larger areas, and additional smoothing could homogenise non-adjacent and heterogeneous areas. Consequently, we have decided not to use kernel-smoothing.

[11] This is a (simple) starting point that could be easily improved by taking into account the intensity of land use by considering certain indicators such as the number of jobs, the production value and sales' levels, among others. This should allow the expected results to be compared with real results to determine, for example, the number of jobs. However, there are also some (potential) limitations regarding data accuracy.

[12] Nevertheless, the main problems concern the heterogeneity of firm size, so it seems that a better solution would be to use size of firms (e.g. the number of employees) rather than just the number (or the existence) of firms.

Third, we create multiple random industry specific maps with two conditions: *i*) the total number of firms in each industry remains constant and *ii*) the total number of firms in each cell remains constant.[13] In this way, we compare the same number of firms but with different industry distributions (for each cell we expect to find the same industry distribution as that of the whole sample). Thus, if the real data shows a cell with only one firm, our simulations will also show this cell with one firm, although the industry will appear as a random variable depending on industry distribution.

Fourth, we compare the actual number of cells with firms (according to real data) with the expected number of cells with firms, and obtain a concentration index similar to that of Ellison and Glaeser (1997), but with some important differences. In particular, our methodology does not focus on agglomeration issues, which allows us to analyse industry distribution. Furthermore, whereas our index is centred at 1 (values below 1 indicate concentration and values over 1 indicate dispersion), Ellison and Glaeser's (1997) index ranges between zero and infinite, which means that they arbitrarily define the concentration threshold.

Fifth, we generalize our approach to several industries (X-Clustering). Methodologically, this is quite similar to an approach that uses only one sector but here we are analysing whether or not a group of industries tends to locate together (co-localization).

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

[13] This latter requirement implies that firms localise randomly inside "occupied" cells (i.e. areas where real firms are located) as stated by Duranton and Overman (2008). This approach means that firms are expected to be located only in those places that are available for economic activity (as the real data shows). Unfortunately, a major shortcoming of this approach is that it assumes that firms could be located elsewhere with other firms, regardless of the industry they are involved in, which is not as realistic (especially at a 2/3 digit level). An extension of this work (and a possible solution for this shortcoming) would be to regard manufacturing, services and agricultural firms as being located with other firms from the fields of manufacturing, services and agriculture respectively.

Seventh, we make a cluster map using vectorial data in the following way: once we have determined the cells – clusters we evaluate the economic activity (firms, jobs and production) in each of the clusters, both in absolute and relative terms.

Eighth, the self-organizing maps are used to show the local microstructure of industries.

Our methodology can complement previous approaches based on distribution comparisons (Brenner, 2006 and 2004; Ellison and Glaser, 1997) and on distance distributions (Duranton and Overman, 2005) and thus enable industry to better understand the determinants behind the spatial distribution of firms.

4. Main results

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

[INSERT TABLE 1]

Table 1 illustrates the expected spatial distributions of firms across regular cells[14] (according to the number of firms in each industry) and the real (observed) spatial distribution of such firms. In particular, it shows how many cells (*X*) contain firms from industry *y* (i.e. this is the "real" spatial distribution of

[14] These regular cells have an area of 100 km^2 (10 km * 10 km).

firms); the expected number of cells (Mean) where firms from industry *y* should appear if they were randomly spatially distributed (according to the total number of firms in each industry); and a co-location index (Index) that relates these measurements to each other (i.e. Index = *X* / Mean). This index can be understood in the following way: if Index < 1, this means that the industry *y* appears in fewer cells than expected (i.e. this industry is spatially concentrated in a smaller number of cells); and if Index > 1, this means that the industry *y* appears in more cells than expected (according to a random distribution), which means that this industry is spatially dispersed. This indicates that there is a certain location behaviour taking place that should be analyzed to determine whether or not it is a cluster (i.e. whether or not firms from industry *y* tend to locate together).

On a technological level, it seems that the lower the technological level of the industry, the higher the spatial dispersion (Table 1). Thus, high-tech firms tend to be more spatially concentrated than low-tech firms[15]. This appears to be logical since the markets and resources of such firms tend to be concentrated in a few areas, which means there is no logical reason for a dispersion pattern.

Our results regarding the differences between manufacturing and services, (Table 1) are even clearer than those of previous studies and show that whereas most 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 reflect the spatial distribution of population and economic activity and the production and distribution requirements of manufacturing and services. Specifically, most services need face-to-face interactions and thus their location decisions are strongly motivated by the locations of their customers (both firms and individuals). In contrast, manufacturers can transport their goods easily, which means that such interactions are not essential and that these firms can locate elsewhere.

[15] As an example, indices of high-tech industries such as office machinery, computers and medical equipment, precision and optical instruments (0.644) and electrical machinery and apparatus (0.664) are clearly lower than those of some low-tech industries such as food, beverages and tobacco (1.452) and agriculture and fishing (1.424).

So far we have analysed the spatial distribution of firms at single industry level and have shown that looking at certain industry specificities (i.e. manufacturing vs. services and high-tech vs. low-tech) helps us to understand such location patterns.

[INSERT TABLE 2]

However, this situation gets more complicated if we take into account the location patterns of more than one industry. Therefore, the next step is to check for the existence and extent of clusters by checking if pairs of industries (or groups of three or four industries) tend to be located close to each other. Table 2 summarises the main findings and shows a selection of all the possible combinations of pairs of industries[16]. The previous indicators are slightly different and include: the codes of industries *y* and *i* respectively; the number of times (*X*) that firms from industry *y* and industry *i* appear together inside the same cell; the expected number of times (Mean) that firms from industry *y* and industry *i* should appear together inside the same cell if they were randomly spatially distributed (according to the total number of firms for both industries); and a co-location index (Index) that relates these measurements to one another (i.e. Index = *X* / Mean). The Index can be understood in the following way: if Index < 1, this means that this industry combination (*y* and *i*) appears fewer times than expected and (for reasons that we will analyze later) this pair of firms tends not to be located in the same areas; and if Index > 1, this means that this industry combination appears more times than expected, so this pair of industries locates in the same areas (they cluster together). Therefore, if Index > 1 it could indicate a cluster (both industries locate together because they have strong inter-industry linkages) or it could be an example of co-location (both industries locate together because they need the same type of economic environment but do not have any kind of inter-industry relationship). The procedure to be followed is first to identify such location patterns and second to distinguish between the previously mentioned proximity explanations.

[16] Given that there are 378 possible combinations of industry pairs, here we only show results for the 10 pairs with the lowest index values and for the 10 pairs with the highest index values.

The 28 industrially classified industries can be put into 378 possible pairs. Most of these (324) show a co-location index < 1, which means that these pairs of industries appear fewer times than expected. In contrast, only in 54 pairs show a co-location index > 1, which indicates a cluster or a co-location.

[INSERT TABLE 3]

Table 3 shows that the pairs of industries with higher co-location index values are: agriculture and fishing / food, beverages and tobacco, and extraction activities / food, beverages and tobacco. Close analysis of these possible clusters or co-located activities shows that they involve a small number of industries and that most of them usually appear in the industry pairs with higher co-location index levels, these being: agriculture and fishing; extraction activities; food, beverages and tobacco; wood, furniture and other manufacturing activities; non-metallic mineral products; construction; and retail and repair of personal and household goods.

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

Table 3 shows inter-industry linkages in terms of intermediate consumption between pairs of industries. We assume that if two pairs of industries are linked by such inter-industry intermediate consumption they can be identified as part of a cluster, whereas if there is no such relationship, their location patterns can be explained in terms of co-location.

Our results show no clear pattern in terms of inter-industry linkages. Therefore, this data cannot be used to explain firm co-location behaviour (or the absence thereof) in terms of such linkages; that is, it seems that there is no intermediate consumption pattern to the way in which firms locate close to other firms.

Therefore, a cluster explanation cannot be ascertained. In particular, the bottom of the table shows that some pairs of industries have important linkages (e.g. 34.51% of the intermediate consumption of industry 1 comes from industry 3, and 15.84% of the intermediate goods sold by industry 3 goes to industry 1) whereas others do not have such linkages or that any linkages are much weaker (e.g. industries 2 and 3, industries 2 and 9, industries 3 and 17, etc.). Finally, the top of the table gives a similar picture: while some of the pairs of industries reach important inter-industry linkages (e.g. industries 22 and 24, industries 19 and 22, etc.) others are less well linked (e.g. industries 12 and 26, industries 14 and 26, etc.).

Tables 4 and 5 summarise the main results regarding co-location behaviour. Specifically, Table 4 shows linkages of each industry in terms of the co-location index. Thus, the first three rows (1, 2 and 3) have P-A indexes lower than 1 and show (for each industry) the number of industries that do not tend to co-locate. This behaviour ranges from aversion (row 1) to neutrality (row 3). Finally, row 4 indicates the number of industries that do tend to co-locate.

[INSERT TABLE 4]

An industry-specific analysis shows that the main industries that strongly tend not to co-locate are: office machinery, computers and medical equipment, precision and optical instruments (code 13), electrical machinery and apparatus (14), financial intermediation (22) and education (26). There are other industries that also tend not to co-locate but their effect is weaker here.[17] Among the industries with stronger levels of co-location are: extractive activities (code 2), food, beverages and tobacco (3) and non-metallic mineral products (9).Finally, there are 13 industries (i.e. those with 0 value in row 4) that do not tend to co-locate with any industry.

[17] Specifically, textiles, leather clothes and shoes (code 4), paper and publishing (6), rubber and plastic products (8), machinery and equipment (12), business services (24) and health and veterinary activities (27).

Generally speaking, our results show four types of industrial co-location relationship:

- The first consists of 6 quite disperse industries (P-A Index = 1.02) that tend to co-locate with a high number of industries (7,33) which are also mainly disperse: (11), (17), (18), (19), (20) and (21).
- The second consists of 2 industries which are quite concentrated (P-A Index = 0.83) and are co-located with a small number of industries (2,50): (16) and (23).
- The third consists of 5 highly dispersed industries (P-A Index = 1,30) that tend to co-locate with an important number of industries (11.40): (1), (2), (3), (5) and (9).
- The fourth consists of 15 highly concentrated industries (P-A Index = 0,75) that rarely co-locate with other industries (0.13): (4), (6), (7), (8), (10), (12), (13), (14), (15), (22), (24), (25), (26), (27) and (28).

[INSERT TABLE 5]

Table 5 illustrates the types of industry co-locations: 1, 2 and 3 if industries tend not to be together and 4 if industries are co-located.

The complex structure of intersectorial relations is difficult to understand using indexes and other quantitative measures, especially when relationships greater than two levels are analysed. “Direct relationships” (e.g. vis-à-vis) can be easily checked, but situations such as “the friends of my friends are also my friends” (i.e. “indirect relationships”) are much more complex.

Self-organizing maps (SOM) (also known as self-organizing feature maps or SOFM) are types of neural network that follow an unsupervised learning process in order to create a low-dimensional (typically two-dimensional) discretised representation of the initial structures.

Like other neural networks, SOM follow a recursive process of learning and mapping. During the learning stage, input examples are used to build maps, and during the mapping stage a new input vector is classified automatically.

An SOM is made up of a group of spatially located nodes with the same weight as the input vectors. Nodes are usually placed in a hexagonal or rectangular grid from which SOM creates a map from a higher dimensional input vector to a lower dimensional map space. The input vector is positioned on the map by finding the node with the most similar weight vector to the data space vector and giving the map coordinates of this node to initial input vector. SOM differs from other neural networks because it uses a neighbourhood function to preserve topological properties.

[INSERT FIGURES 3a, and 3b]

Figures 3a and 3b show circular graphs for the co-location index and for the Input-Output relationships according to Spanish Input-Output Table. These circular graphs are algorithm-based networks that place nodes around a circle according to the internal network structure of the node connections. It is a simple algorithm that illustrates the number of nodes and edges in a network. Unfortunately, it is not suitable for bigger networks or for checking upper-level relationships.

The circular graph (Figure 3a) shows different thresholds according to the PA index, these being all the connections over the median, over the mean and over 1 (this value indicates the existence of co-location). Figure 3b reproduces the same scheme but takes into account Input-Output linkages to try and obtain a similar number of connections and thus allow the two graphs to be compared. These figures provide two important results: *i*) when the intensity of the connections is not taken into account, all industries are interrelated and *ii*) when the intensity of connections is taken into account (i.e. when only the strongest connections are considered) only some industries are shown to be interrelated (the other industries remain isolated and do not belong to alternative networks).

We have used radial tree graphs to illustrate the top-level relationships of each of the industries. Due to space constraints (and also because their dynamic

structure) we do not present these relationships here, but they can be found at http://gandalf.fcee.urv.es/professors/JosepMariaArauzo/WEB%20CLUSTERS/P&A.html.

The radial tree graphs place the analysed node in the centre and arrange around it those nodes which are primarily connected to it. The second-order connections are placed around the first-order connections so that the hierarchic relations between the different nodes can be clearly seen. The central node is known as the concentration node and the remaining nodes are placed around it in concentric circles. Each node is placed at the ring corresponding to its shortest distance to the network departing from the focus.

The radial tree graphs are helpful for analysing the interactions between a specific node and the other nodes, but they are not suitable for conducting a global analysis. In order to carry out such an analysis, it is better to use more complex algorithms that can manage all the network interactions simultaneously. The spring algorithm is the simplest algorithm among those that rely on the inyensity of relations. This algorithm shows the network connections in terms of wharves, where the length of these connections depends on the intensity of the interactions.

If the wharf is compressed (i.e. if its current length is shorter than its natural length) then it tends to spread and push its edge nodes out. But if the wharf is stretched (i.e. if its length is longer than its natural length), then it tends to contract, pulling its edge nodes in. The strength used by the spring is proportional to the difference between its current length and its natural length. The linked nodes tend to form clusters against a repulsive force that tends to separate them. Lengths change until an iterative result is obtained using the minimum amount of energy. The results are shown in graphs 4a and 4b.

[INSERT FIGURE 4a]

Only 15 out of 28 industries show a co-location index higher than 1; the remaining 13 industries are identified as concentrated (blue), according to our

concentration index[18]. Surprisingly, however, those industries with higher co-location relationships are defined as disperse (red) or intermediate (green) in terms of industry concentration. This behaviour can be explained in terms of scale economies and market characteristics because the competitive strategy of firms belonging to these industries is to disperse their production units. Despite this, these firms still need to be close to firms from other specific industries, as Figure 4a shows.

[INSERT FIGURE 4b]

However, if Figure 4a shows strong connectivity between groups of industries, the results change completely when we introduce inter-industry relationships from Input-Output Tables, because this means that connectivity now exists only for groups of 2 and (sometimes) 3 industries, but not as strongly as in the previous figure. It is also important to notice that concentrated industries are now more involved in such relationships.

Figures 4a and 4b are particularly interesting when they are compared; that is, when the proximity of industries in terms of Input-Output Tables is compared with the real data regarding the spatial distribution of firms belonging to these industries.[19] Such a comparison provides very interesting information about the nature of inter-industry relationships. One finding is that Input-Output linkages seem to be more homogeneous (i.e. the Input-Output linkages show that a higher number of industries have a lower number of connections) than the real collocation data, which seems to show a core of (mainly) dispersed but strongly connected industries[20]. Another important difference is that the real data show a large number of isolated industries (13) whereas this is not shown at all by the Input-Output data. Thus, it seems that in addition to the inter-industry linkages among firms, there are other determinants that explain the spatial proximity

[18] If we relax the collocation requirements and use the mean index (0.78714), we get all the industries (see the top of Figure 4a).

[19] Specifically, we will analyse the bottom area of both figures (>1 for 4a and >1500 for 4b) because the number of edges is quite similar (54 and 59).

[20] This core includes the following industries: wood, furniture and other manufactures; construction; non-metallic mineral products; extractive activities; electricity and water distribution; trade and repair; fabricated metal products; agriculture and fishing; transport and communications and food, beverages and tobacco.

between them; that is, there are some intense inter-industry linkages (in terms of intermediate consumption) that do not require spatial proximity. This is a key finding because it implies that traditional interpretations of clusterization processes could be biased.

[INSERT FIGURE 4c]

Network scaling algorithms take into account only the most significant nodes, thus simplifying the node analysis process. One of these algorithms is called the MST-Pathfinder Network Scaling algorithm, which is a variation of the traditional Pathfinder Network Scaling algorithm (PNS) and which has the advantage of noticeably reducing the calculation time. PNS is a structural assessment technique (Schvaneveldt et al., 1998) that significantly reduces the number of links and, therefore, provides a concise representation of clarified proximity patterns (IVC, 2005). PNS also not only provides "(…) a fuller representation of the salient semantic structures than minimal spanning trees, but also a more accurate representation of local structures than multidimensional scaling techniques" (Chen, 1999, p. 408).

PNS relies on the so-called triangle inequality in order to eliminate redundant or counter-intuitive links. If there are two links (paths) in a network connecting a couple of nodes, the preserved link (path) has a greater weight in terms of the Minkowski metric. One could assume that as the weight of the link (path) increases, it becomes easier to capture the interrelationship between the two nodes, and that the alternative, lighter link (path) becomes redundant or even counter-intuitive and should be pruned from the network (IVC, 2005).

Two parameters (*r* and *q*) influence the topology of a pathfinder network. The *r*-parameter influences the weight of a path (which is based on the Minkowski metric), while the *q*-parameter defines the number of links in alternative paths (i.e. the length of a path) up to which the triangle inequality must be maintained.

A network of *N* nodes can have a maximum path length of $q = N-1$. With $q = N-1$ the triangle inequality is maintained throughout the entire network.[21]

Four disperse industries (all from services) show co-location relationships but only with a few industries. In order to better illustrate such inter-industry relationships, we present some self-organizing maps. These maps show that there is a dual situation regarding co-location relationships: first, most industries do not have such relationships and, second, there is a smaller number of industries that tend to co-locate frequently.

5. Conclusions

With this paper we have contributed to extant literature on cluster identification by designing a procedure to identify groups of industries that tend to cluster together and to analyse whether this behaviour can be explained in terms of vertical integration or by common location determinants shared by those industries. This distinction allows detailed analysis of firm location determinants and our results show that diversified clusters are not casual and are strongly determined by industry characteristics. In particular, it means that firms need "specific" neighbours in order to maximise their performance.

The methodology proposed in this paper allows the main reasons driving cluster formation to be better explained, but much more work needs to be done in this area, particularly to identify cluster size and thus better capture cluster borders. This methodology involves dividing spaces into homogeneous cells of equal size. This procedure must be handled with care because cell size influences the number and characteristics of the identified clusters. Specifically, bigger cells are more lilkely to contain a cluster, whereas smaller cells are more likely to have fewer inter-industrial clusters because the number of firms in each cell will be smaller. Given that in this paper we have assumed equal sizes for all the clusters, it would appear that using flexible sizes fits better with the real

[21] See Schvaneveldt (1990) for additional technical details.

distribution of economic activity and is therefore a promising line for future research.

This is just a first attempt to better identify the forces driving cluster formation. Consequently, we have studied several types of clusters in order to provide a general overview of this phenomenon. However, this is just a starting point and further work needs to be done, in particular to cover industry specific characteristics that influence the location decisions of firms. We therefore plan to extend our analysis of specific types of clusters (both specialised and diversified) to cover several types of urban / rural environments that are hypothesised to influence such agglomerative behaviour. Finally, as we mentioned beforehand, industry aggregation is also important and, despite the computational constraints that make it unfeasible to work with such disaggregate industry-levels, we need to carry out further research to accurately determine whether our results are robust to different industry aggregation levels.

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.

Barlet, M., Briant, A. and Crusson, L. (2009): "Location patterns of services in France: A distance-based approach", Paris School of Economics Working Paper.

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. (2006): "Identification of Local Industrial Clusters in Germany", *Regional Studies* **40 (9)**: 991-1004.

Brenner, T. (2004): *Local industrial clusters: existence, emergence and evolution*, Routledge: London.

Chen, C. (1999): "Visualizing semantic spaces and author co-citation networks in digital libraries", *Information Processing and Management* **35(3)**: 401-420.

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.

IVC (2005)

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.

Schvaneveldt, R.W. (ed.) (1990): *Pathfinder Associative Networks: Studies in Knowledge Organization*, Norwood, NJ: Ablex.

Schvaneveldt, R.W., Dearholt, D.W. and Durso, F.T. (1988): “Graph Theoretic Foundations of Pathfinder Networks”, *Computer Mathematics Applications* **15 (4)**: 337-345.

Silverman, B. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.

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 equipment, 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

The 10 industries with the lowest co-location index values									
Code industry *y*	**Code industry *i***	***X***	**Mean**	**STD**	**Index**	**X-2S**	**X+2S**	**Concentrated**	**Dispersed**
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

The 10 industries with the highest co-location index values									
Code industry *y*	**Code industry *i***	***X***	**Mean**	**STD**	**Index**	**X-2S**	**X+2S**	**Concentrated**	**Dispersed**
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: Inter-industry linkages according to the co-location index

The 10 industries with the lowest co-location index values							
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
The 10 industries with the highest co-location index values							
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
2	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 of euros.
Source: Spanish Input – Output Table (INE) and own calculations.

Table 4: Co-location relationships by industries (P-A Index)

Code	Industry	(1) Below median (0,787)	(2) Between median (07,87) and mean (0,840)	(3) Between mean (0840) and 1	(4) Over 1
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 equipment, 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 services	17	4	6	0
28	Other services	15	4	7	1

Source: own elaboration and NOMBRE SOFTWARE.

Table 5: Types of co-location among industries (P-A Index)

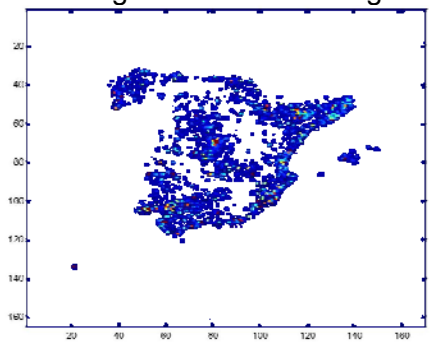
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1		**4**	**4**	3	**4**	1	3	2	**4**	3	**4**	3	1	1	3	**4**	**4**	**4**	**4**	**4**	**4**	1	3	3	2	1	3	3
2			**4**	3	**4**	3	3	3	**4**	3	**4**	3	1	1	3	3	**4**	**4**	**4**	**4**	**4**	2	**4**	3	3	2	3	**4**
3				3	**4**	2	3	3	**4**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

Source: Own elaboration

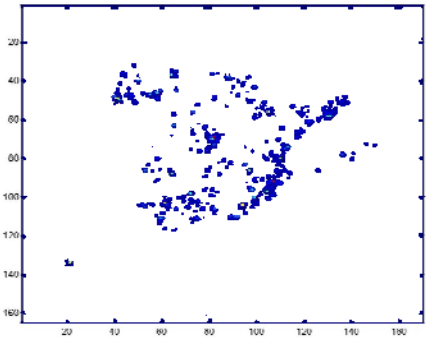
Maps

Map 1: Spatial distribution of firms included in the data set

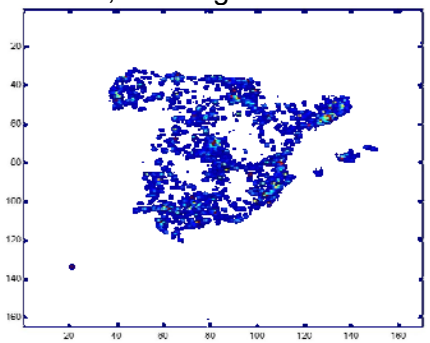
1.- Agriculture and fishing



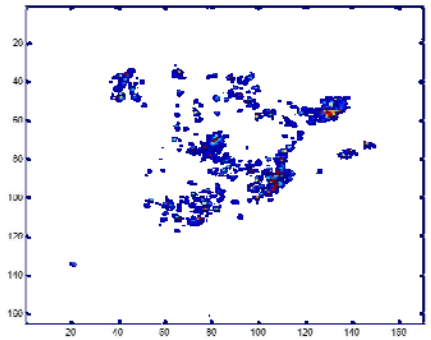
2.- Extractive activities



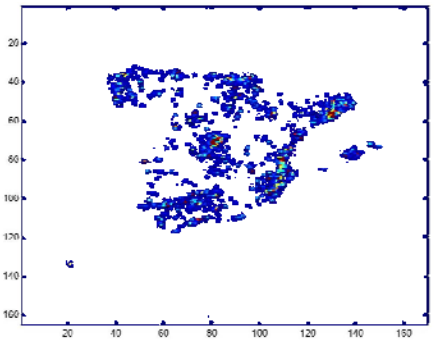
3.- Food, beverages and tobacco



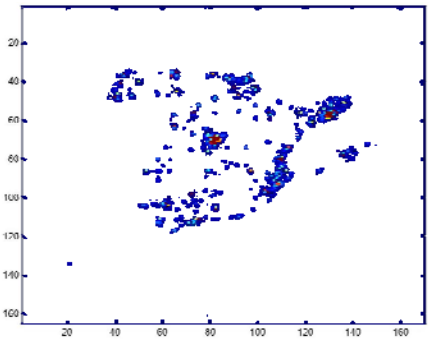
4.- Textiles, leather clothes and shoes



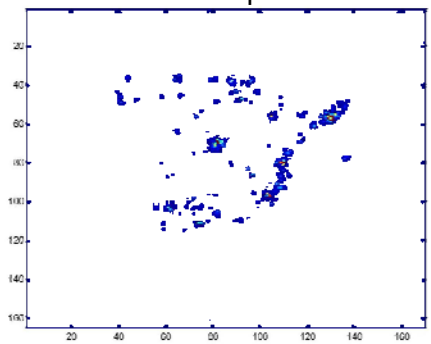
5.- Wood, furniture and other manufactures



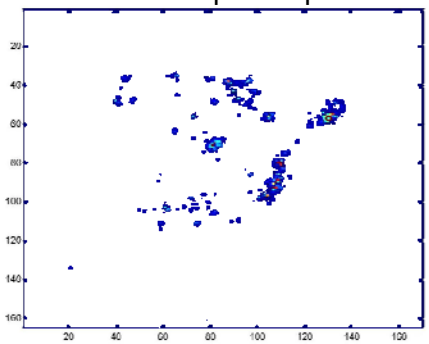
6.- Paper and publishing



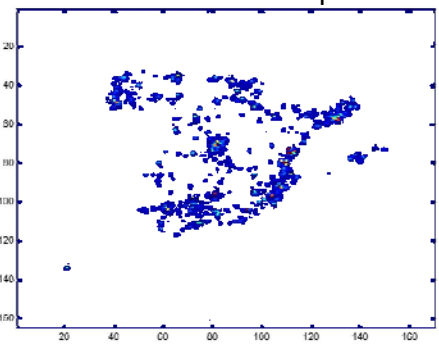
7.- Chemical products



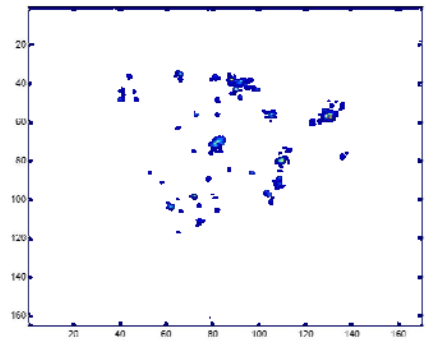
8.- Rubber and plastic products



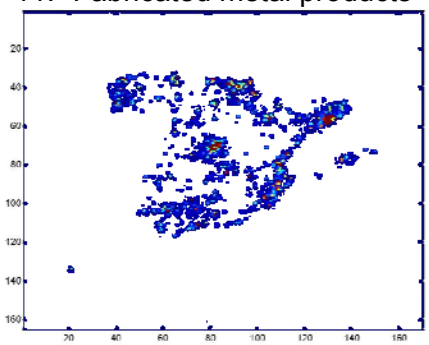
9.- Non-metallic mineral products



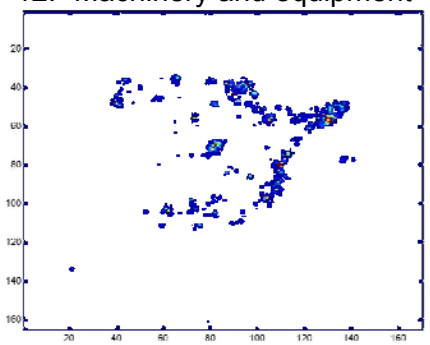
10.- Basic metals



11.- Fabricated metal products



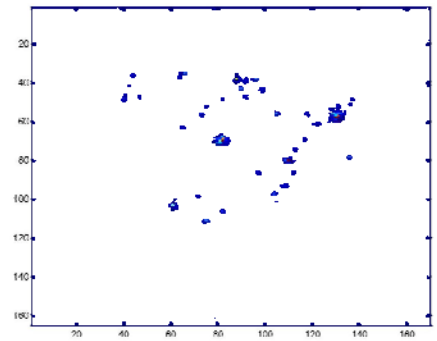
12.- Machinery and equipment



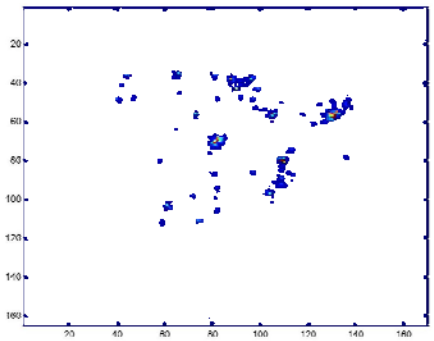
13.- Office machinery, computers and medical equipment, precision and optical instruments

14.- Electrical machinery and apparatus

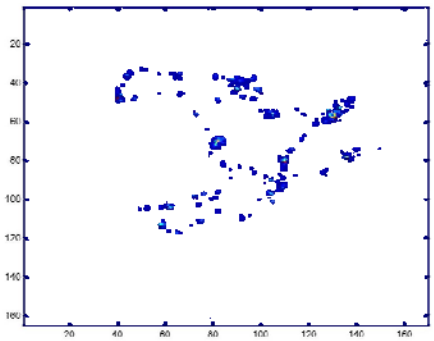
15.- Transport equipment



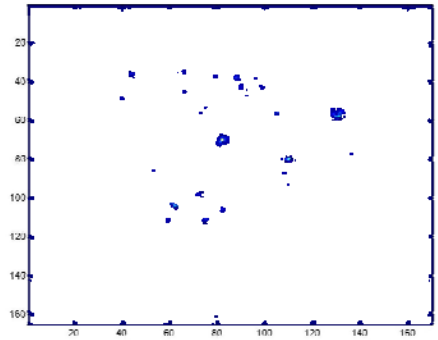
16.- Recycling



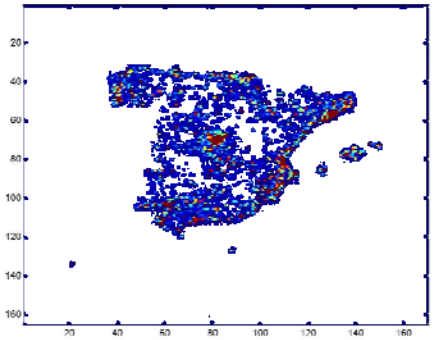
17.- Construction



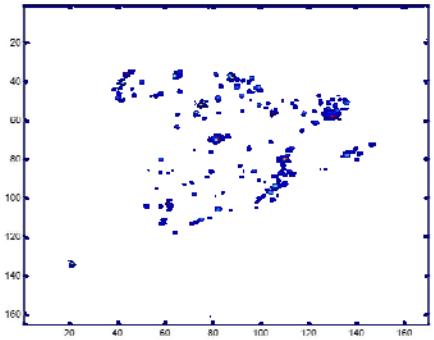
18.- Electricity and water distribution



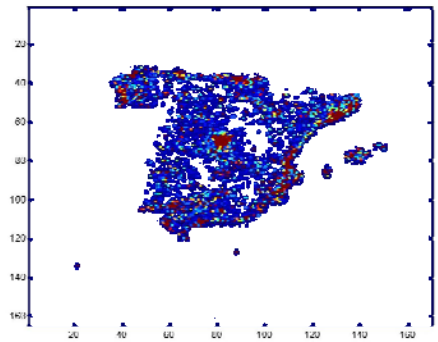
19.- Trade and repair



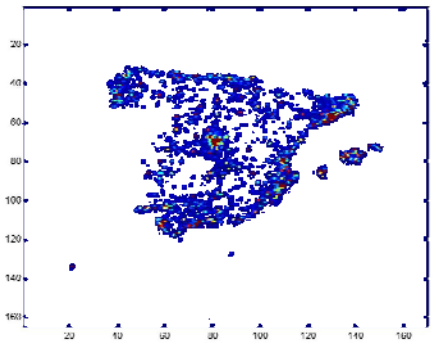
20.- Hotels and restaurants



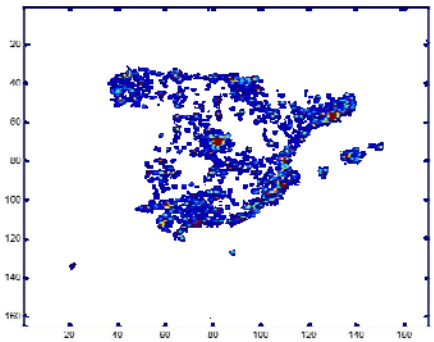
21.- Transport and communications



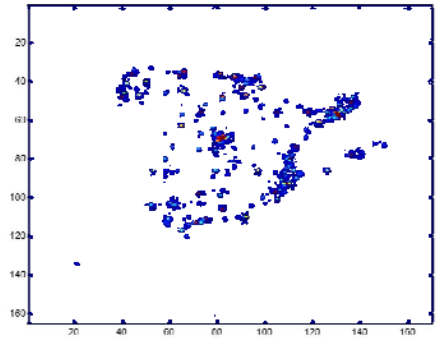
22.- Financial intermediation



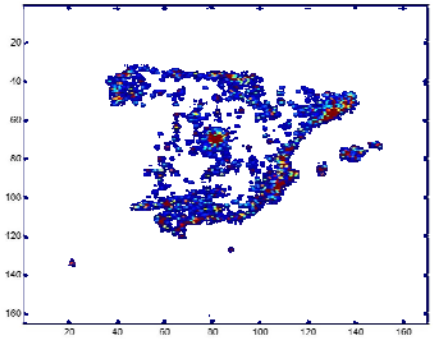
23.- Real estate activities



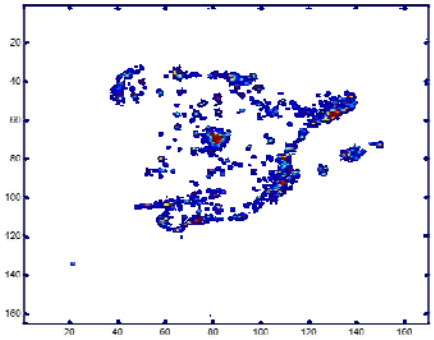
24.- Business services



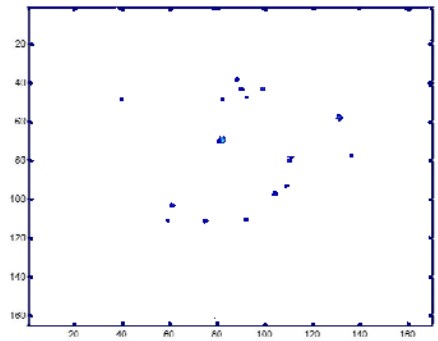
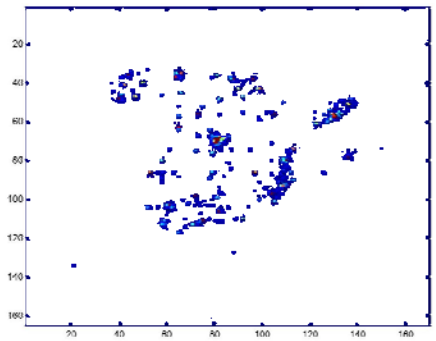
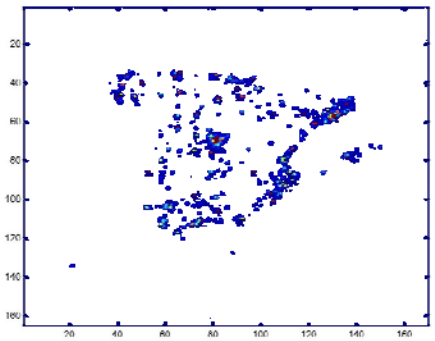
25.- Public administration



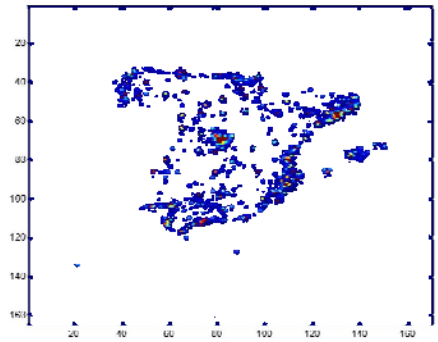
26.- Education



27.- Health and veterinary activities, social services

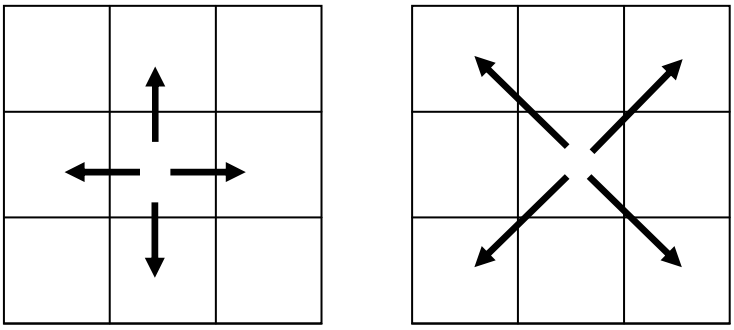
28.- Other services



Source: Own elaboration.

Figures

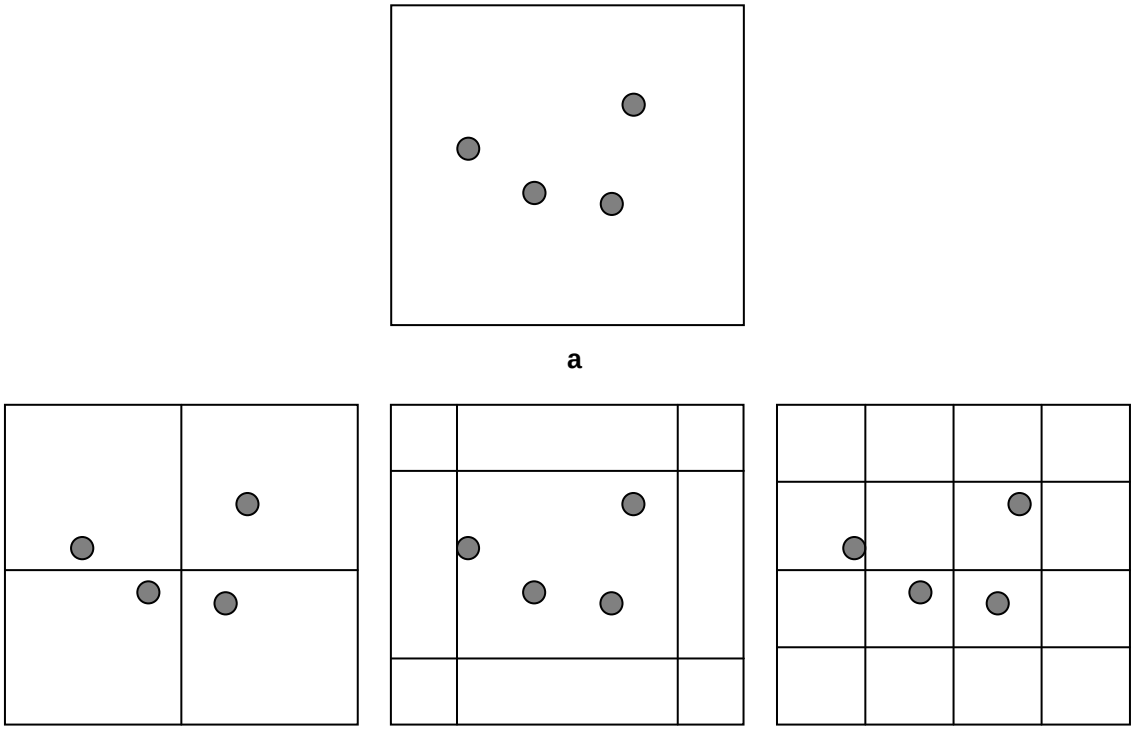
Figure 1: Contiguity in a square map



Rook contiguity | Bishop contiguity

Source: Own elaboration.

Figure 2: Modifiable Area Unit Problem (MAUP)

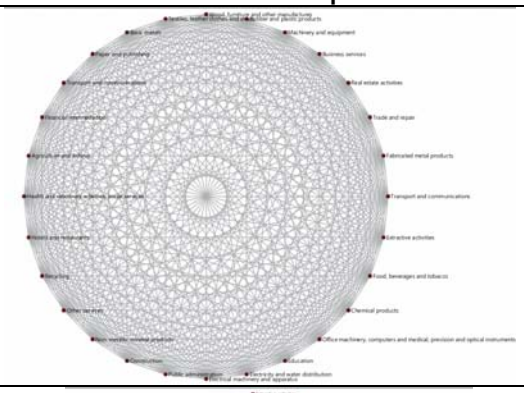
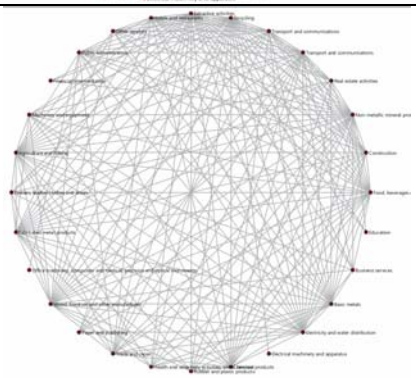
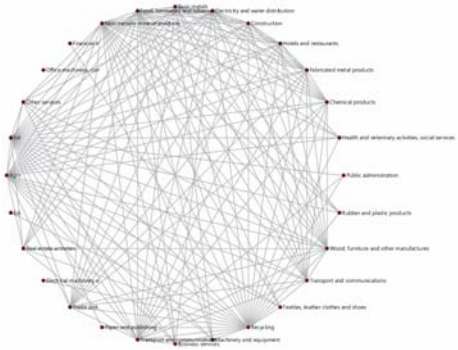
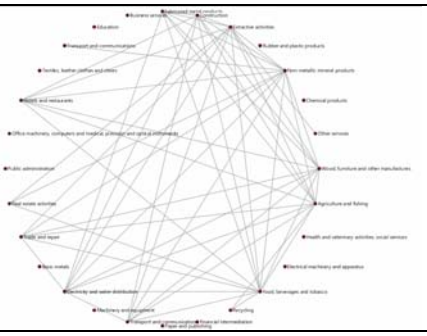


a

b Minimum concentration | **c** Maximum concentration | **d** Intermediate concentration

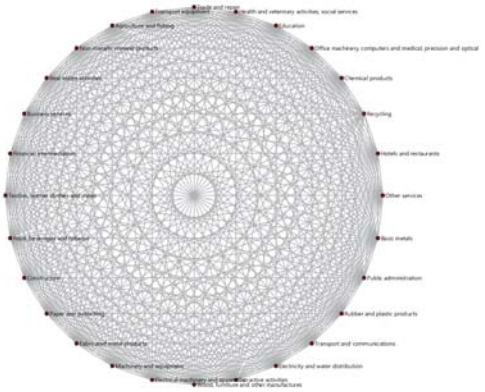
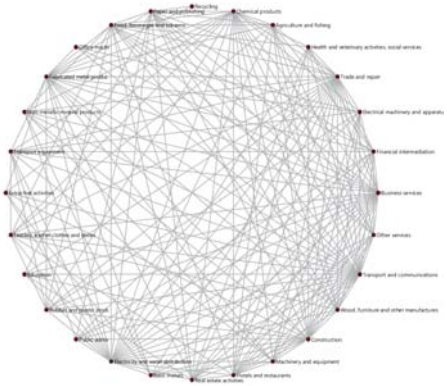
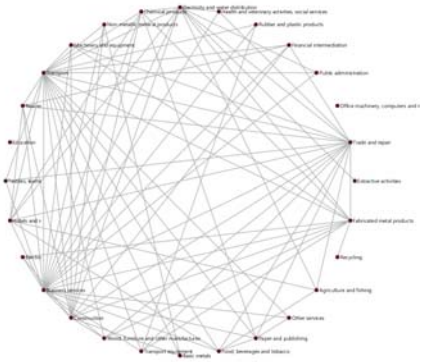
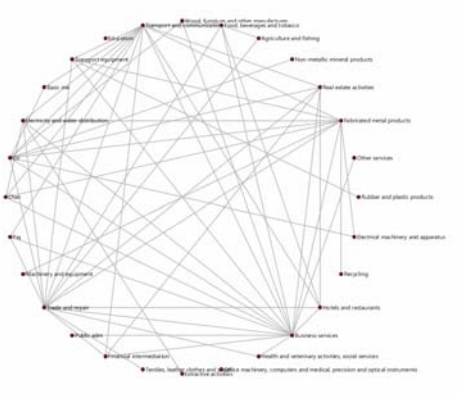
Source: Arbia (2001)

Figure 3a: Sectorial connectivity. 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: 3.8571428571428577 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

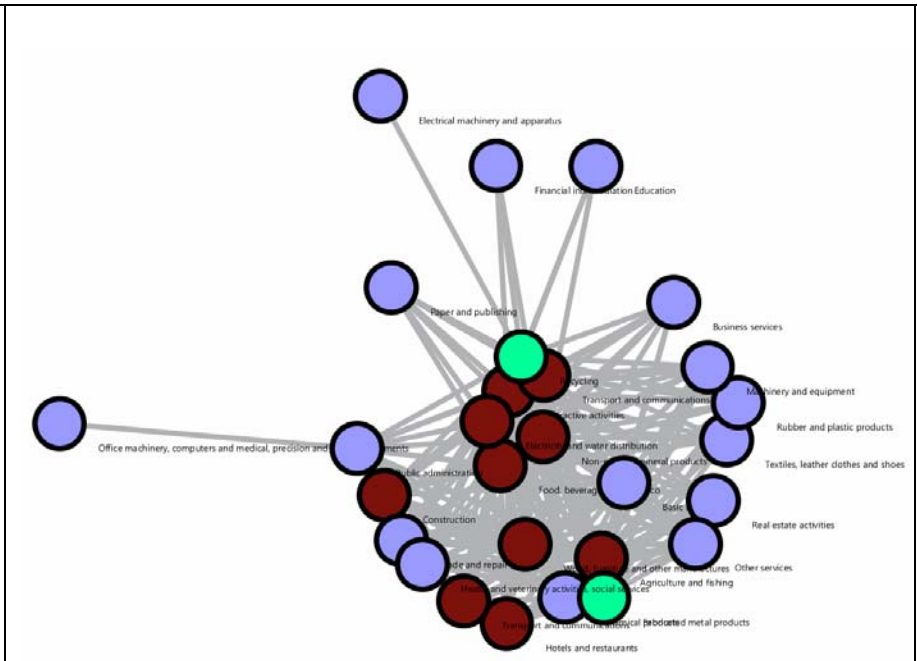
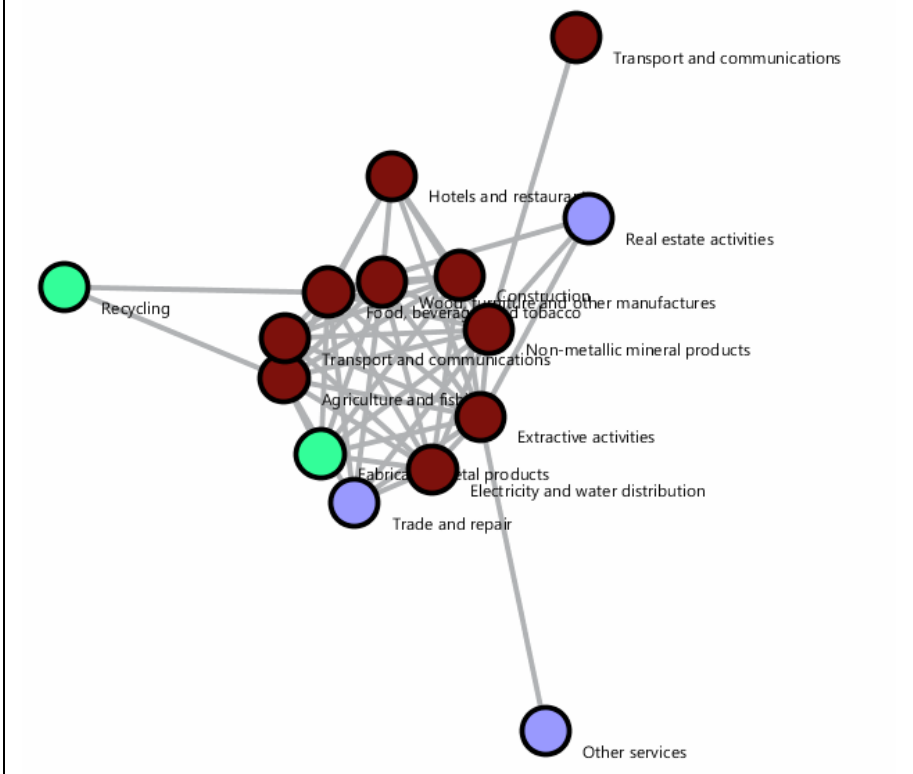
Source: Own elaboration using NWB.

Figure 3b: Sectorial connectivity (TIO). Radial Tree Graphs

	Radial Tree/Graph	Nodes: 28
All		Edges: 373 Isolated nodes: 0 Average degree: 26.64285714285715 This graph is weakly connected. There are 1 weakly connected components. (0 isolates) The largest connected component consists of 28 nodes. Density (disregarding weights): 0,98677 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 957,6619 densities (weighted against observed max) weight: 0,04199
>300		Edges: 196 Isolated nodes: 0 Average degree: 14.000000000000004 This graph is weakly connected. There are 1 weakly connected components. (0 isolates) The largest connected component consists of 28 nodes. Density (disregarding weights): 0,51852 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 910,24418 densities (weighted against observed max) weight: 0,03991
>Mean 970,5		Edges: 88 Isolated nodes: 2 Average degree: 6.285714285714285 This graph is not weakly connected. There are 3 weakly connected components. (2 isolates) The largest connected component consists of 26 nodes. Density (disregarding weights): 0,2328 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 746,39497 densities (weighted against observed max) weight: 0,03273
>1500		Edges: 59 Isolated nodes: 2 Average degree: 4.2142857142857135 This graph is not weakly connected. There are 3 weakly connected components. (2 isolates) The largest connected component consists of 26 nodes. Density (disregarding weights): 0,15608 Additional Densities by Numeric Attribute densities (weighted against standard max) weight: 653,85132 densities (weighted against observed max) weight: 0,02867

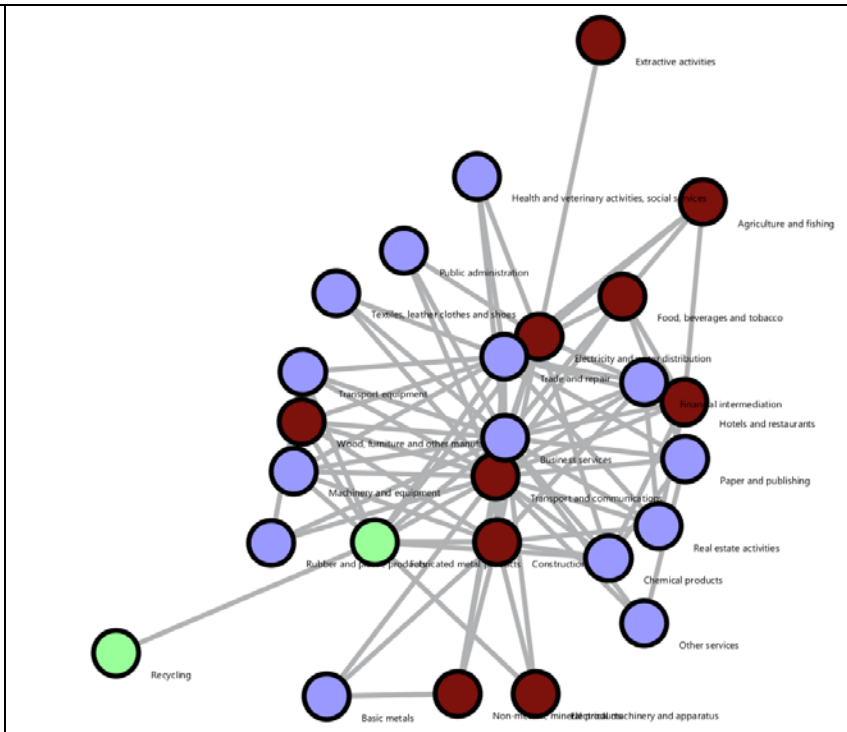
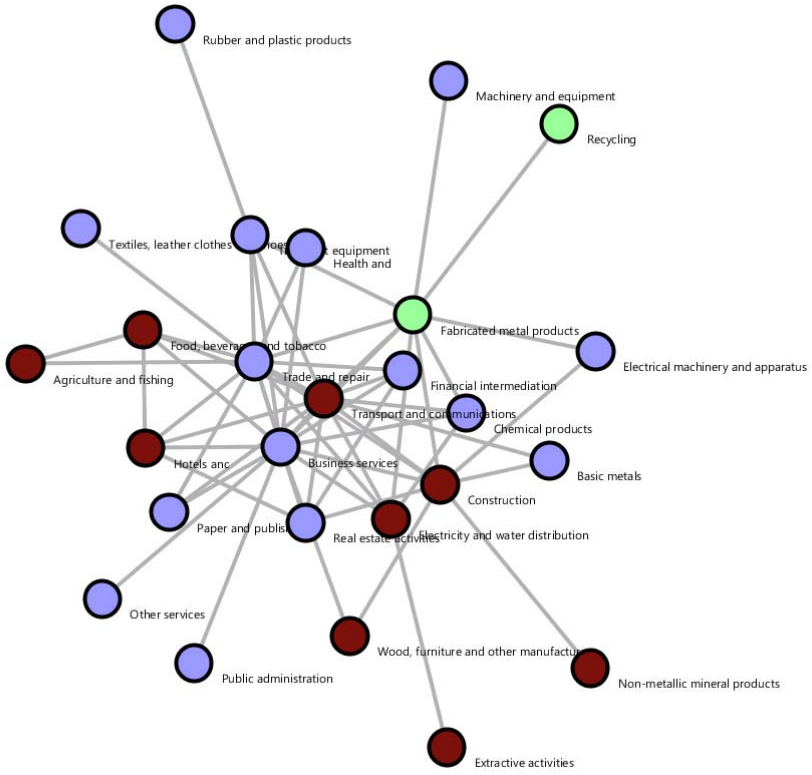
Source: Own elaboration using NWB.

Figure 4a: Sectorial proximity. Spring Graphs

>0,78714		Edges: 206 Isolated nodes: 0
>1		Edges: 54 Isolated nodes: 13

Note: Colours refer to dispersed (●), concentrated (●) and random (●) industries.
Source: Own elaboration using NWB.

Figure 4b: Sectorial proximity (TIO). Spring Graphs

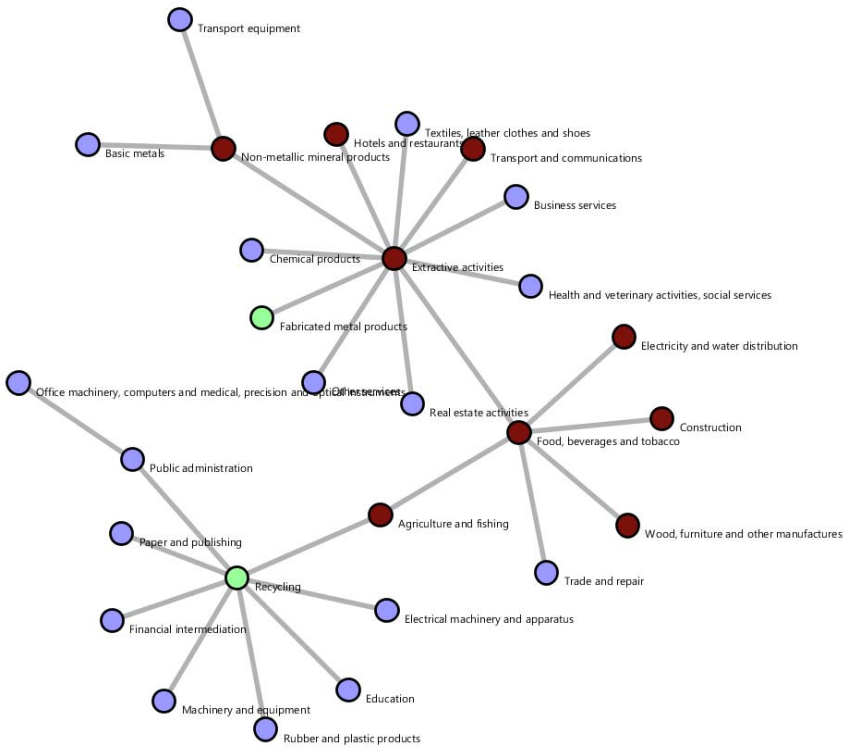
>Mean 970,5		Edges: 88 Isolated nodes: 2
>1500		Edges: 59 Isolated nodes:

Note: Colours refer to dispersed (●), concentrated (●) and random (●) industries.
Source: Own elaboration using NWB.

Figure 4c: Sectorial proximity. Spring Graphs

MST-Pathfinder Network Scaling was selected.
Implementer(s): Chintan Tank
Integrator(s): Chintan Tank
Documentation: **https://nwb.slis.indiana.edu/community/?n=AnalyzeData.MSTPathfinderNetworkScaling**

Input Parameters: Pablo&Arauzo v0.3 (378to27)
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight



Network Analysis Toolkit (NAT) was selected.
Implementer(s): Timothy Kelley
Integrator(s): Timothy Kelley
Reference: Robert Sedgewick. Algorithms in Java, Third Edition, Part 5 - Graph Algorithms. Addison-Wesley, 2002. ISBN 0-201-31663-3. Section 19.8, pp.205
Documentation: https://nwb.slis.indiana.edu/community/?n=AnalyzeData.NetworkAnalysisToolkit
This graph claims to be undirected.

Nodes: 28
Isolated nodes: 0
Node attributes present: label
Edges: 27
No self loops were discovered.
No parallel edges were discovered.
Edge attributes:
Nonnumeric attributes:
Example value edge_ty...

Numeric attributes:

min	max	mean weight
0,78714	1,48598	1,07185

This network seems to be valued.

Average degree: 1.9285714285714297
This graph is weakly connected.
There are 1 weakly connected components. (0 isolates)
The largest connected component consists of 28 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0,07143
Additional Densities by Numeric Attribute
densities (weighted against standard max) weight: 0,07656
densities (weighted against observed max) weight: 0,05152

MST-Pathfinder Network Scaling was selected.
Implementer(s): Chintan Tank ; Integrator(s): Chintan Tank
Documentation: https://nwb.slis.indiana.edu/community/?n=AnalyzeData.MSTPathfinderNetworkScaling

Input Parameters:
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight (DATA P&A v03)
Fast Pathfinder Network Scaling was selected.
Implementer(s): Chintan Tank
Integrator(s): Chintan Tank
Documentation: **https://nwb.slis.indiana.edu/community/?n=AnalyzeData.FastPathfinderNetworkScaling**

Input Parameters:
Weight Attribute measures: SIMILARITY
Edge Weight Attribute: weight
Value of R parameter: 1.0
Value of R parameter = Infinity ?: true
For undirected networks, use MST Pathfinder network scaling algorithm for faster results.

Note: Colours are referred to dispersed (●), concentrated (●) and random (●) industries.
Source: Own elaboration using NWB.

Annexes

Annex 1: List of industries

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

Source: SABI.