An Analysis of the Geographic Concentration of Industry in Spain[†]

Olga Alonso-Villar[*]
José-María Chamorro-Rivas
Xulia González-Cerdeira

Universidade de Vigo

May 2001

Abstract:
This paper seeks to analyse the extent of geographic concentration in Spanish industry. To that end the concentration index derived from a model of industrial localisation proposed in Maurel & Sédillot (1999) is used, and a comparison is made with other indices used in literature. Starting from the model proposed, an in-depth sectorial and geographic study is made of the spillovers generated by proximity between businesses. The data used are taken from the *Encuesta Industrial de Empresas* (Industrial Survey of Businesses) and cover the period from 1993 to 1999. The results confirm that there is major geographic concentration in a number of industries with widely varying characteristics, including high-tech businesses and those linked to the provision of natural resources as well as traditional industries. It is also observed that for most sectors spillovers between companies go beyond the provincial level, and that in some cases those spillovers affect not just businesses in the same sector but also those in related sectors.

Keywords: geographic concentration, spillovers, industry

[†] We are grateful for financial aid received from the Spanish Ministry of Science and Technology ("Ministerio -0613-C02-01 and SEC1999-1236-C02-01, and from the Regional Government of Galicia ("Xunta de Galicia") via PGIDT00PXI30001PN.

[*] Address for correspondence: Departamento de Economía Aplicada, Universidade de Vigo, Campus Lagoas-Marcosende s/n, 36200 Vigo, Spain. Tel. 34 986 81 25 07; Fax. 34 986 81 24 01; E-mail: ovillar@uvigo.es.

1. Introduction

In the early 1990's a new line of research was opened up which used certain ideas from earlier economists (Marshall (1890)) as the basis for studying the agglomeration of economic activity with a more rigorous and formal approach. The resulting literature forms part of what is known as "New Economic Geography", and takes up the approach of the circular, cumulative causation theory put forward in Myrdal (1957). This theory puts great emphasis on feedback in processes of economic growth: the author sustains that if a critical threshold of development is passed at a particular point, whatever the cause, an even greater process of concentration of economic activity may ensue. Insofar as many companies may already have set up in a particular location, this may attract new investment, which will generate higher levels of income and, in turn, consumption, and thus enable more businesses to be created. In the final analysis this leads to serious inequalities between regions.

These ideas are taken up in the seminal work of Krugman (1991), where a theoretical framework is proposed which can explain why some regions concentrate the greater part of industrial activity in detriment to others, even though there are no *a priori* differences between them in resources or technology (the arguments traditionally put forward to explain inter-regional differences in comparative advantage theory). This work has resulted in a considerable literature concerned with studying processes of agglomeration of economic activity. Such agglomeration can be intensified at some stages by processes of economic integration (Ottaviano & Puga (1998)) or by new communications technologies (Warf (1995)). [1]

In our paper we seek to analyse the geographic concentration of industry in Spain between 1993 and 1999, and to study the spillovers across businesses which may be behind that concentration. First of all we look at whether localisation patterns vary widely from one industry to another. Secondly, we attempt to compare patterns in Spain with those observed

[1] According to Ottaviano & Puga (1998), one conclusion which can be drawn from this literature is that processes of integration favour increased concentration of economic activity in their early stages, but subsequently lead to greater dispersal.

elsewhere in order to draw general conclusions, i.e. to determine what industries tend to concentrate to a greater degree in different countries, and what industries are generally more widely dispersed. Thirdly, we attempt to discover what role spillovers may play in these processes of concentration. Finally, we study the evolution of the manufacturing firms throughout the period.

The method we use is that proposed in Maurel & Sédillot (1999), which enables us not only to determine the degree of concentration of each sector of industry but also to analyse the localisation spillovers involved.[2] Two points are analysed in regard to spillovers: first of all a study is made of concentration in each sector at regional and provincial level in an attempt to determine whether or not the spillovers between plants in the sector in question go beyond provincial level. Secondly, we run a sectorial analysis to determine whether the spillovers are highly restricted in nature and are found only between businesses in the same sector, or whether they also affect businesses in related sectors.[3]

The paper by Maurel & Sédillot (1999) discusses the similarities between the index proposed by its authors (which is referred to here as M-S) and that put forward earlier in Ellison & Glaeser (1997) (referred to here as E-G). However, it does not analyse the differences between the two indices, so it is hard to perform an empirical analysis to determine why they do not always coincide: it may happen that M-S concludes that a sector is highly concentrated while E-G concludes that it is not, and *vice versa*. We therefore analyse the differences between these indices here and indicate the aspects of concentration on which each one places most emphasis. We also compare these indices with the Gini index, which is also widely used in literature.[4]

[2] A distinction is drawn in the relevant literature between localisation spillovers (across businesses in the same sector) and urbanisation spillovers (across businesses in different sectors). See Henderson *et al*. (1995) and Glaeser *et al*. (1992) among others.
[3] This method is not suitable for a more exhaustive analysis to determine whether urbanisation spillovers exist.
[4] Callejón (1997) and Callejón & Costa (1995) also study the concentration of industry in Spain, though their methods differ from ours and the period covered is earlier (1981-1992). Another noteworthy study is Paluzie *et al*. (2000), which seeks to identify the determining factors in the localisation of industrial concerns.

The paper is structured as follows. Section 2 defines the M-S index and presents its main characteristics. Section 3 discusses the differences between M-S and E-G and compares both indices with that of Gini. Section 4 presents the data used and studies the concentration of industry in Spain in 1999. This section also analyses spillovers across companies, drawing a distinction between the geographical and sectorial scope of these spillovers. Section 5 shows the evolution of the M-S index along the period 1993-1999. Section 6 gives the main conclusions reached.

2. The Concentration Index in Maurel & Sédillot (1999)

General Approach

This section defines the concentration index, $\hat{\gamma}$, which we use throughout this paper to attempt to determine the degree of spatial concentration of Spanish industry. Concentration is analysed sector by sector, i.e. a concentration index must be defined for each sector considered. In what follows we therefore assume that businesses belong to the same sector.

Taking Ellison & Glaeser (1997) as a reference, Maurel & Sédillot (1999) proposes a model of industrial location according to which plants in a particular sector decide to locate in particular geographical regions either because of the natural conditions of those regions or because of spillovers which may arise from the proximity of other plants in the same sector which are already located there.[5]

An outline of the most significant elements of the model proposed in M-S are presented below, with as little recourse as possible to technical details. The random variable U_{ij} is defined, which takes a value of 1 if plant j is located at location i, and 0 otherwise. It is assumed that all pairs of plants j and k in the sector have the same joint distribution for their binary responses (U_{ij}, U_{ik}), such that:[6]

[5] This model does not discriminate between these two possible causes of the localisation decision.
[6] This 2-dimensional random variable is made up of two non-independent Bernouilli variables.

$$E(U_{ij}) = E(U_{ik}) = x_i,$$
$$prob(U_{ij} = U_{ik} = 1) = x_i^{\;2} + x_i(1 - x_i)g,$$
$$prob(U_{ij} = U_{ik} = 0) = (1 - x_i)^2 + x_i(1 - x_i)g,$$
$$prob(U_{ij} = 1, U_{ik} = 0) = prob(U_{ij} = 0, U_{ik} = 1) = x_i(1 - x_i)(1 - g).$$

This means that all the plants in the sector have the same probability, denoted by x_i, of locating at a particular location *i*. [7] Moreover, from the foregoing expressions it can be deduced that $g = corr(U_{ij}, U_{ik})$ for $j \neq k$, i.e. the correlation between the locations of plants j and k is precisely g, a parameter which shows both the interdependence of plant location decisions due to their interests in natural advantages and the existence of spillovers between them, $g \in [-1,1]$.

As can be deduced from the above probability distribution, the probability that any two plants in the sector will choose the same location, *p*, [8] can be written as a linear function of the parameter g so that by proposing an estimator for the said probability for *p* an estimator can be obtained for g, which is what ultimately interests us, as will be shown below. M-S's paper proposes an estimator for *p* which leads to an estimator for g as follows:[9]

$$\hat{g} = \frac{\dfrac{\sum_i s_i^{\;2} - \sum_i x_i^{\;2}}{1 - \sum_i x_i^{\;2}} - H}{1 - H},$$

where i denotes the location,[10] s_i is the proportion of employment in the sector accounted for by location *i*, x_i is the proportion of industrial employment at location *i*, and H is the

[7] This probability depends on the size of the location, measured in terms of aggregate industrial employment there, so that if one location has twice as much employment than another, the probability of a plant in the sector analysed choosing to locate there is twice as high as at the other location. In other words, x_i is the proportion of industrial employment at *i*.

[8] This probability, *p*, is precisely $\sum_i prob(U_{ij} = 1, U_{ik} = 1) = \sum_i x_i^{\;2} + g\left(1 - \sum_i x_i^{\;2}\right)$.

[9] $\hat{p} = \sum_i \dfrac{\sum_{j,k \in i} z_j z_k}{\sum_{j,k} z_j z_k}$, where $j, k \in i$ denotes the plants in the sector that choose to locate at location *i*.

[10] At empirical level the location may be a natural district, department, province, region, state, etc.

Herfindahl index for the sector, which is given by $H = \sum_j z_j^{\,2}$, where z_j is the proportion of employment in the sector accounted for by plant j. H thus shows concentration of output, i.e. whether or not the sector's output is concentrated in just a few plants. If all the employment in the sector is concentrated in one plant, H takes a value of 1, and if there are many plants of similar sizes it is close to 0.

This index is similar to that proposed previously by E-G, which is expressed as

$$\hat{g}_{EG} = \frac{\dfrac{\sum_i (s_i - x_i)^2}{1 - \sum_i x_i^{\,2}} - H}{1 - H}.$$

Both indices are non biased estimators of parameter g, though M-S has the advantage that it comes from a simpler probabilistic location model. At empirical level there are also differences between the two indices, as they do not necessarily emphasise the same points in assessing concentration. This is discussed in a later section.

Now let us look at why both indicators can be used as concentration indices. First of all, the first terms of the numerators of both $\hat{g}$ and $\hat{g}_{EG}$ can be interpreted as primary indices (according to their terminology) of geographic concentration, insofar as they measure the differences between spatial distribution in the sector (given by s_i) and the industrial aggregate (given by x_i). As M-S show, the expectations of both primary indices can be written as $H + g(1 - H)$.[11] Thus, g is actually showing the excess of primary concentration, i.e. that part of geographic concentration which is above the concentration of production (given by H). Moreover, using either $\hat{g}$ or $\hat{g}_{EG}$, if a sector is randomly distributed through the different geographic units, or if there are no spillovers across the various plants in a sector, these indices average zero, regardless of how concentrated production is in a small number of plants. However this is not true if we directly use the primary spatial concentration index, as deduced from the mean value given above. The fact that indices $\hat{g}$ and $\hat{g}_{EG}$ have this property makes them especially suitable for measuring spatial concentration.

This method can be adapted, as proposed in M-S, not just to study the concentration of a sector but also to check for spillovers across businesses. What follows looks at both the geographical and the sectorial scope of these spillovers, i.e. first an analysis is run to see if spillovers extend beyond the sub-regional scope, and then it is studied whether spillovers occur across businesses in the same sector or in related sectors.

Geographical Scope of Spillovers

This section looks at the geographic aspect of spillovers, i.e. the approach developed above is used as a basis for determining whether spillovers, represented by g, extend beyond the sub-regional scope. M-S propose comparing the probability that two plants in a sector will locate together when the geographic unit of reference is smaller than a region (which in Spain means taking the province as a unit), as considered in the previous section, i.e. taking into account that plants decide their location in a single stage, and then obtaining that probability through the two-stage process described below. In stage one the plants must decide in what region they wish to locate, and in stage two in what area of that region (i.e. in what province). Modelling this location process calls for the calculation of conditioned probabilities that could be obtained from the joint probability distribution presented above.

Given that the probability obtained via one-stage location must coincide with that obtained in two-stage location, it can easily be shown that

$$g_{p1} = \overbrace{l_{r1} g_{r1}}^{reg} + \overbrace{\sum_r l_r g_{p(r)2}}^{subreg} + \overbrace{\sum_r m_r g_{r1} g_{p(r)2}}^{cross},$$

where $g_{p1}, g_{r1}, g_{p(r)2}$ are, respectively, the correlation in plant location decisions in the sector at sub-regional (provincial) level in a single stage, at regional level, and at sub-regional level

[11] Note that the fraction of employment in the sector at a location can be written in terms of random variables, and hence the primary indices can also be considered as random variables. $s_i = \sum_j z_j U_{ij}$.

in two stages (i.e. after a region has been selected). On the other hand, $l_{r1} + \sum_r l_r + \sum_r m_r = 1$, with

$$l_{r1} = \frac{\sum_r x_r(1-x_r)\Omega_r}{1-\sum_r x_r^2\Omega_r},\ l_r = \frac{x_r^2(1-\Omega_r)}{1-\sum_r x_r^2\Omega_r},\ m_r = \frac{x_r(1-x_r)(1-\Omega_r)}{1-\sum_r x_r^2\Omega_r},$$

where Ω_r denotes the sum of the squares of the weights of the sub-regions of region *r* measured in terms of the proportion of aggregate employment in the region.

Spillovers at sub-regional level can therefore be written as a weighted mean of spillovers with a regional scope, spillovers with a solely sub-regional scope and the crossed product of both. Using the above expression it can thus be calculated what proportion of spillovers is due to each factor.

As shown in M-S's paper, when the scope of geographical spillovers extends beyond the limit of the smallest geographical unit we should observe a greater concentration in the higher area level ($g_{p1} < g_{r1}$).

Sectorial Scope of Spillovers

This section studies the sectorial scope of spillovers. To that end their effect is permitted to extend not only to plants in the same sector but also to production plants in related sectors. We thus assume that these spillovers will affect the location decisions of both the plants in a certain subsector (e.g. a 3-digit level classification in the CNAE (Spanish National Classification of Economic Activities)) and those in other subsectors of the same sector (2-digit level in the CNAE). These spillovers which affect plants belonging to different subsectors can be measured from the value of the correlation g_0.[12] Given that g_0 is not expressed in the paper by M-S, we have derived the expression in the Appendix.

[12] In this case $g_0 = corr(U_{ij}, U_{ik})$, with *j* and *k* being two plants belonging to different subsectors of the same sector.

$$\hat{g}_o = \frac{\dfrac{\sum_i s_i^2 - \sum_i x_i^2}{1-\sum_i x_i} - H - \sum_l \hat{g}_l w_l^2 (1-H_l)}{1-\sum_l w_l^2},$$

with $\hat{g}_l$ being the concentration index for subsector *l*, w_l the proportion of employment accounted for by subsector *l* within the sector and H_l the Herfindahl index for subsector *l*.

On the other hand, as M-S affirm, it can be shown that

$$g = \frac{\overbrace{\sum_l g_l w_l^2 (1-H_l)}^{\text{intra}} + \overbrace{g_0 \left(1-\sum_l w_l^2\right)}^{\text{inter}}}{1-\sum_l w_l^2 H_l},$$

i.e. the correlation between plants in the sector, g, can be written as a weighted mean of the correlation between plants in the same subsector, denoted by g_l for each subsector *l*, and the correlation between plants in different subsectors, g_0.

Using this expression, we can calculate what part of the concentration within the sector, $\hat{g}$, is due to intra-sectorial spillovers (within the same subsector), and what part to inter-sectorial spillovers (between plants in different subsectors of the same sector).

3. Concentration Indices

Although we concentrate in this paper on the use of the M-S index, we compare the results with those obtained with the E-G and Gini indices. This section presents the similarities and differences between these indices. All three aim to measure the geographic concentration of a sector, taking industrial activity as a whole as their point of reference. Thus, the first two indices differ basically in the way in which their primary indices are obtained. M-S calls for the calculation of the differences between $\sum_i s_i^2$ and $\sum_i x_i^2$, which are taken as reflecting the divergences between the territorial location of the sector in terms of employment and that of the industrial aggregate. $\sum_i (s_i - x_i)^2$ appears in the calculation of the primary index of E-G, which also takes into account the differences between what happens at sectorial level and in the industry as a whole, though in this case these differences are calculated location by location. The Gini concentration index also measures the extent to which the spatial distribution of a sector differs from that of the industry as a whole.[13]

For any of these indices the concentration will therefore show the divergences between what happens at sectorial level and at aggregate level, so that if the geographic distribution of a sector coincides with that of the industry as a whole, that sector is said not to be concentrated. As shown in our presentation of the method proposed by M-S, both the M-S and E-G indices measure geographical concentration beyond the concentration of output in just a few plants (measured by the Herfindahl index), which means they have advantages over the Gini index.

It must be borne in mind that the first two indices differ in the degree of importance which they allocate to divergences between the sector analysed and industrial activity as a whole. A location in which the percentage of the sector is greater than that of total industrial activity is a positive factor in the M-S index, while one in which the contrary is true is a negative factor (note that the first term of the numerator of the M-S index can be written as $\sum_i (s_i - x_i)(s_i + x_i)$). Moreover, if the location has a high level of aggregate industrial employment and an even higher level in terms of the sector, its contribution to the index is very great, while if it has little industry, even though the weight of the particular sector in question is greater, its contribution is positive but small (though higher than its contribution to the E-G index, since in the former case it would contribute $(s_i - x_i)(s_i + x_i)$ and in the

[13] The Gini index is calculated by ordering the various units of territory in accordance with the Hoover-Balassa index, which measures the ratio s_i / x_i. The x-axis represents the cumulative proportions of industrial employment as a whole, and the y-axis the cumulative proportions for the sector under study. The Gini index measures the quotient between the area between the corresponding Lorenz curve and the 45-degree line and the area below this line. Specifically, the Gini index would take the form $\sum_{i=1}^{n-1}(p_i - q_i) / \sum_{i=1}^{n-1} p_i$, where p_i denotes the cumulative proportion of employment in the sector and q_i the cumulative proportion in industry for the first i units of territory in the ranking obtained via the Hoover-Balassa index (see Brülhart (2000)). A sector which is distributed in a similar fashion to the industry as a whole gives a value for Gini's index of

latter it would contribute $(s_i - x_i)(s_i - x_i)$). This is why the M-S index takes on high values when the sector is localised at those locations where there is most industry, as shown in some sector discussed below. But if the sector is localised at locations with little industrial weight the index shows little concentration. However the E-G index takes into account the divergences between the sector percentage and the industrial aggregate in each location, regardless of the direction of the difference, and the contribution to the index value is the same in both cases. Moreover, if a location has more employment in one sector than for industry as a whole its contribution to this index is less than its contribution to M-S.[14]

4. Concentration of Industry between 1993 and 1999

The Data

The data used in the analysis are taken from the *Encuesta Industrial de Empresas* (EI) drawn up by the INE (Spanish National Institute of Statistics). The analysis covers the years from 1993 to 1999, but results are presented only for 1999, because the performance of the sectors was observed to be similar throughout the period (see Table 1).[15] A more detailed analysis of this evolution will be presented in the next section. The EI provides data on employment with a maximum territorial breakdown to provincial level and a maximum sectorial breakdown to the 3-digit level as per the CNAE. Table 2 shows the sectors available.

This analysis uses on the one hand information at regional level (regional autonomous communities, denoted by CCAA) with a sectorial breakdown to two and three digits, and on the other hand information at provincial level with a breakdown to two digits (in data broken down to three digits there is a major lack of information due to high levels of

zero. We have also calculated the Gini index, taking population distribution as a reference, and the results are very similar. Correlation between the two indices is high.

[14] The correlation of the array obtained from the M-S index with that obtained from the E-G index in the Spanish case is around 56%. With the Gini index it is 68%. Using French data, the paper by M-S finds higher correlations on the order of 90% for the two first indices.

secrecy in statistics).[16] The analysis performed using 2-digit sectorial classification takes in a total of 30 sectors, 5 of which were eliminated on grounds of lack of information in practically all locations with positive values.[17] The degree of secrecy in statistics varies in the remaining sectors, averaging around 5% for employment at provincial level and 2% at CCAA level.

Given the lack of information in some of the remaining sectors we have assigned a value to those locations for which statistics are secret by using the information available at a more aggregate level. Thus, for each sector at two-digit level the first step is to calculate employment associated with the whole of the locations with secret statistics via the difference between total employment in the sector in Spain as a whole and employment in those localisations for which information is available. This difference is allocated among these locations according with their weights in industrial employment. When the sample is corrected with a breakdown to the 3-digit level (respectively, breakdown to provincial level) the information available at 2-digit level (respectively, CCAA level) is taken into consideration, so that attributed employment is as close to actual employment as possible.[18] Using this sample forestalls the possibility that the geographic units considered may differ when a comparison is drawn between regional and provincial results, producing discrepancies as a result of the non availability of information at provincial level for a sector for which information is available at CCAA level.

[15] In 1993 the survey was modified in two important points: the survey unit changed from establishments to businesses, and the CNAE-93 sectorial classification was adopted. The period analysed begins in 1993 so that homogenous data are available for the full period.

[16] The INE provides no information on sectors in a localisation (province or CCAA) when there are less than 4 plants.

[17] Sectors 11, 12, 13, 16 and 23 have been eliminated. These sectors cover part of the *mining and extraction* industry, *tobacco* and *coke plants/ oil refineries/ nuclear waste treatment*. In sectors 11 and 12 the INE provides no aggregate information for the sector. In the rest the number of CCAA in which statistics are secret is 9, 11 and 13 out of 17, respectively.

[18] When the sample is corrected at provincial level, CCAA information (if any) is considered and the weight of the province within the CCAA is taken into account. In correcting the sample to a sectorial breakdown level of three digits, employment information from the 2-digit sector to which it belongs is taken into account.

Concentration at Provincial Level

We now go on to discuss the analyses performed at provincial level with a sectorial breakdown of two digits as per the CNAE. In grouping sectors according to their degree of concentration we have followed the M-S concentration index, but the results are compared with those obtained via the E-G and Gini indices. To enable us to compare the concentration figures we obtain with those presented in M-S for France and those obtained by E-G for the USA, we consider the same criteria proposed in both these cases: index values (both M-S and E-G) lower than 0.02 are taken as low concentration, values from 0.02 to 0.05 represent intermediate concentration and values higher than 0.05 are taken as high concentration.

Table 3 shows the M-S, E-G and Gini concentration indices with the corresponding array of sectors obtained with each of them, plus the Herfindahl index. The most highly concentrated sectors according to M-S are the following: *Preparation, tanning & finishing of leather* (19),[19] *Office machinery & computer equipment* (30), *Textiles* (17), *Electronic materials, radio, TV and communications* (32), *Mining & extraction of anthracite, coal, lignite and peat* (10), *Publishing & graphic arts* (22), *Medical, precision & optical instruments and watch-making* (33), and the *Chemical industry* (24).[20] These sectors are characterised by the concentration of most of their activity in just a few provinces, generally Barcelona and Madrid.[21]

The results seem fairly robust, in view of the degree to which the three indices used coincide. In fact, the E-G and Gini indices also place these sectors among the most highly concentrated, though there are exceptions: sector 24 is considered as having intermediate level concentration under E-G and low under Gini. Employment in sector 24 is

[19] In this case the Herfindahl index informs us that employment is distributed across many plants, but in spite of this a high spatial concentration is observed.
[20] The concentration of the *Recycling* sector (37) must be viewed with precaution in view of the high degree of secrecy in statistics in this sector.
[21] The *Preparation, tanning & finishing of leather* sector and *the Office machinery & computer equipment* sector have experienced a very high concentration between 1993 and 1999. The *Measuring & precision instruments* and *Publishing & graphic arts* sectors are also characterised by an increase in concentration during the period.

concentrated precisely in provinces such as Barcelona and Madrid, which are already highly industrialised, which explains the divergence between the two indices: M-S is highly sensitive to patterns of this type, while E-G is not.

The sectors which show up as having low concentration under all three indices used are: *Foodstuff and beverage industry* (15), *Manufacture of metal products other than machinery and equipment* (28), *Manufacture of furniture, other manufacturing industries (toys, jewellery, musical instruments & sports articles)* (36) and *Production and distribution of electricity, gas, steam and hot water* (40).[22]

However, there are some sectors in which the indices analysed show certain contradictions in classification. For instance *Mining & extraction of non metallic minerals (stone, sand, minerals for fertiliser and salts)* (14) is the least concentrated sector according to M-S, but has an intermediate concentration under E-G.[23] Similar divergences are found in sectors 20, 26, 27 and 35.

A more in-depth look at the causes of these discrepancies shows that they are due to different nuances in the definitions of the two indices, along the lines of those mentioned above. For example a look at the distribution of employment in sector 14 leads us to deduce that it is not heavily concentrated in the most highly industrialised provinces, and thus is rated lower under M-S than under E-G. Barcelona and Madrid account for less than 12% of the employment in this sector, but between them they have 33% of employment in industry. Nor is there any other province in which the quota of employment is outstandingly high. Similar patterns can be found in the other four sectors, i.e. the divergences between the M-S and E-G indices for these sectors are due to the relatively high number of localisations in less industrialised provinces.

[22] It should be noted that the Gini index does not put sector 40 among the lowest. This is probably due to the high degree of concentration of output at a small number of plants, as can be deduced from the Herfindahl index.
[23] The Gini index in these cases does not show very high figures.

Finally we have a group of sectors which are classed as being of intermediate concentration under M-S and low under E-G. The sectors involved are 18, 21, 25 and 31. A more exhaustive analysis reveals that the concentration of these four sectors does not appear to be low, given that a major part of the employment in them is located in provinces with a considerable industrial weight and in provinces known for their specialisation in these sectors. This is why M-S gives higher values than E-G.

Geographic Scope of Spillovers

Thus far we have analysed concentration using the province as our geographic area of reference. The model proposed in E-G indicates that the concentration index should not in theory be sensitive to the definition of the geographical area (CCAA or province in our case). The authors assume that spillovers are local in scope, and that their area of influence is therefore limited to the smallest unit of territory. However, they also indicate that this assumption could be somewhat restrictive in practice, given that spillovers may have larger scopes of action.[24]

Indeed, if spillovers may have greater scopes it is easy to realise that concentration at regional level should be lower than at provincial level. We have drawn the following figures to show this.[25]

[insert Figures 1 & 2]

Each CCAA, marked by a thick line, has the same number of provinces, the limits of which are shown by the thin lines. The points represent the plants in each province. A look at these illustrations shows that if the plants in a sector are widely distributed across CCAA's but not within each CCAA (Figure 1), we would expect to find a greater concentration at provincial than at regional level, since the neighbouring provinces in the same CCAA do not contain plants working in the industry in question. However, if spillovers transcend the

[24] Remember that parameter g could be interpreted as the degree of spillovers, as it measures the extent to which the location of one plant may be conditioned by the location of another, $g = corr(U_{ij}, U_{ik})$, with *j* and *k* being two plants in the same sector.

local (provincial) scope neighbouring provinces should benefit from the location of plants in the sector in question, and concentration at provincial level may be expected to be lower than at regional level (Figure 2).

If these results are confirmed, we should observe higher levels of concentration at CCAA level than at provincial level in Spain. A comparison of the calculated results of the concentration index with the two levels of geographic aggregation suggests that in general the index value is slightly higher at CCAA level. This is the case in 17 of the 25 sectors analysed, so it seems that the scope of the spillovers in these sectors is broad.[26] However the arrays resulting at regional and provincial level do not differ substantially, except in *Manufacturing of motor vehicles, trailers and semi-trailers* (34), which ranks higher at provincial level and has a higher index value. Observation of data enables us to state that employment in this sector is distributed across several CCAA's, but within them is localised in one or two provinces.

By using the two stage model presented above we can obtain the proportion of spillovers due to provincial and regional scope.The ranking of industries derived from the one stage model at the provincial level (see table 3) is roughly the same as the one obtained with the two stage model (table 4) but now we are able to separate the contribution of broader and closer spillovers. As we can see in table 4, 12 of the 20 sectors considered show a higher proportion of regional spillovers, especially in *textiles, paper, chemical, machinery, other manufacturing industries*, and *electricity, gas & water*. Thus, it seems clear that in these sectors spillovers go indeed beyond the provincial scope.

Sectorial Scope of Spillovers

So far spillovers across plants in the same sector at 2-digit level have been analysed. We now go on to discuss the results when a distinction is drawn within each sector between spillovers affecting plants in the same subsector and those affecting plants in different

---

[25] This is formally demonstrated in M-S.

[26] Sectors 30, 40, 22, 32, 33, 34, 10 and 18 have higher index values at provincial level, while for the rest the position is reversed. This situation is found in 5 sectors under the E-G index.

subsectors of the same sector. To that end the results obtained are compared with the breakdown of industries at the 2- and 3-digit levels. Given the high degree of secrecy in statistics involving information at provincial level and a breakdown to three digits, the analysis is restricted solely to CCAA level.

By analysing intra- and inter-sectorial spillovers the following results can be inferred (see Table 5). Of the sectors which are most highly concentrated at provincial level, sectors *19 (tanning and leather*), 32 (*electronics material and radio & TV sets*, etc.), *33 (measuring & precision instruments*, etc.), 22 (*publishing & graphic arts*, etc.) and 37 (*recycling*) stand out as having sectorial spillovers which are greater than inter-sectorial spillovers. From this it can be deduced that the plants in these sectors benefit more from proximity to other plants in the same sector than from proximity to plants from related sectors. This may be because subsectors maintain a certain independence one from another.

It should be noted that, even though sectors 35 and 36 show up a low concentration index, their corresponding intra-sectorial spillovers are very high. This is mainly due to the high concentration level that their subsectors have, which implies that spillovers are very intensive within subsectors.

However, agglomeration in sectors 17 (*textile* industry) and 24 (*chemical* industry) is due more to spillovers across plants in different subsectors of their sector than to those within each subsector. This may be because of the input-output relationship between the different subsectors in the former case (*preparation and spinning of textile fibres, finishing of textile products*, etc.), and to the use of skilled labour and research facilities common to different subsectors (*basic chemicals, pesticides, paint, pharmaceuticals, soaps*, etc.) in the latter. In other words, the inter-relationship between subsectors of sectors 17 and 24 could be greater than in other sectors, and spillovers between sectors therefore show up as having a greater weight than those within sectors.

5. Evolution of industrial concentration in the period 1993-1999

To analyse the geographic concentration of industry in Spain between 1993 and 1999, we calculate the M-S index at provincial level with a sectorial breakdown of 2-digits, as in previous sections. We are particularly interested in analysing how the industrial agglomeration has developed along the period, and whether there exists a tendency to a greater/lower geographic concentration or not. From the results, we can confirm that there is not a general long-run tendency to concentration. However, data show that some sectors have experienced remarkable changes in their levels of concentration (see Table 1 and Figures 3-5).

[insert Figure 3]

Figure 3 shows the evolution of the M-S index between 1993 and 1999 for each manufacturing sector. The x-axis represents the index value in 1993, with respect to the median, and the y-axis the mean rate of change. Sectors 30 and 19 have strongly increased their concentration in the period. An exhaustive analysis of sector 30 allows us to observe that Madrid has gained employment, whereas Valencia has lost an amount of employment similar to that gained by Madrid. It seems therefore that there has been a relocation of the sector and, since Madrid had a high share of manufacturing in 1999 (11%), it is reasonable to expect this change in the industrial location to be associated with a higher value of the index. With respect to sector 19, we observe that Alicante, which already had a high share of firms in 1993, has experienced a remarkable employment increase.

This figure also allows us to classify sectors in four groups. The first group includes those sectors in the 4th quadrant, which had a concentration level below (or equal to) the median in 1993 and experienced an increase in concentration throughout the period. These sectors are 15, 20, 26, 35, 36, 41, 28, 30, 21 and 18. The groups in the 3rd quadrant are those with an M-S index below the median in 1993 and an decreasing concentration level thoughout the period. These are sectors 14, 40, 27, and 29. Sectors in the right-hand quadrants are those with an index over the median in 1993, among them we can distinguish between sectors 21, 25, 22, 33 and 19, which show an increase in the industrial concentration, and sectors 34, 31, 37, 24, 17, 32 and 10 which show a reduction. This last sector suffered the greatest lost of concentration, since several establishments in the highest specialised provinces (such as Leon and Oviedo) were closed. Also, sector 32 lost weight in Madrid and sector 17 in Barcelona, this leads to lower concentration indices since those locations which have lost weight were highly industrialised.

[insert Figures 4 & 5]

Figures 4 and 5 show the evolution of the manufacturing sectors with more detail, distinguishing between those sectors that in 1993 were below the median (Fig. 4) and those above (Fig. 5). Between the former we should notice that sector 30 shows a great increase in concentration until 1996, but afterwards it tends to decrease. The rest of the sectors in the group show a roughly stable concentration pattern. Between those sectors above the median, we can observe that sector 19 shows a different evolution path: its geographic concentration has increased all over the period, even though in the last year it seems to have decreased. Also, we should highlight that concentration in sector 10 has decreased all over the period, even though the highest change took place in 1998.

6. Conclusions

This paper analyses the extent of geographic concentration in Spanish industry, based on data from the *Encuesta Industrial de Empresas*, between 1993 and 1999. The results confirm the interdependence which exists among businesses as regards location decisions in a large number of sectors. This is reflected in a major geographic concentration of the output of those sectors.

The sectors which show up as most highly concentrated include especially those for which geographic location is strongly determined by access to raw materials (*mining & extraction*); traditional sectors (*textiles* and *leather*), those based on high technology (*IT* and *electronics*), for which technological spillovers seem to be important, and those which require skilled labour (e.g. the *chemical* industry or *publishing & graphic arts*). In particular, *IT* and *leather* have experienced a remarkable increase in their concentration levels in the period. It should be noted that the *textiles* and *leather* sectors are also highly concentrated in other countries, such as France and the USA, as evidenced by papers such as Maurel & Sédillot (1999) and Ellison & Glaeser (1997). From this it can be inferred that these sectors tend to concentrate to a greater extent than others. A comparison between Spain and France shows similarities also in *mining & extraction* and in *optical instruments & watch-making*, which are also highly concentrated.

In Spain, as in France and the USA, the least concentrated sectors include *the manufacturing of furniture* and *metal products*. Other sectors with low concentration levels in Spain include *foodstuffs & beverages* and *production & distribution of energy*, which are less dispersed in other countries, or for which no information is available in the aforementioned papers.

The geographic reference unit for part of our analysis of concentration is the province. However spillovers may extend beyond the administrative limits of provinces. If so, as discussed herein, we should obtain higher levels of concentration at higher geographic levels, i.e. at regional (CCAA) level. This is indeed what we find in 17 of the 25 sectors analysed.

Another important result of our paper comes from our analysis of the scope of intra- and inter-sectorial spillovers. Interest here is centred on analysing whether businesses benefit more from proximity to other businesses in their same sector or those in other closely related sectors. Our results suggest that in sectors such as the *textile* industry (17) and the *chemical* industry (24) concentration is due more to spillovers across companies belonging to different subsectors (but all within the same sector at a 2-digit breakdown level) than to those within each subsector. In sector 17 this may be due to the input-output relationship between different subsectors (*preparation and spinning of textile fibres*, *manufacture of fabrics, finishing of textile products*, etc.), while in sector 24 it is due more to the use of skilled labour or research facilities common to various subsectors (*basic chemicals, pesticides, paint, pharmaceuticals, soaps*, etc.). In other words, the degree of inter-relation between subsectors in sectors 17 and 24 could be greater than that of other sectors, thus resulting in spillovers between sectors having more weight than those within a sector. Similar results for *textiles* and part of the *chemical* sector are observed in France. However in other sectors, such as *tanning & leather*, *precision instruments & watch-making*, spillovers across companies in the same subsector are observed in both countries.

Table 1: M-S index in the period 1993-1999

Year	Mean	Median	Deviation
1993	0,053	0,026	0,085
1994	0,058	0,028	0,091
1995	0,060	0,027	0,087
1996	0,064	0,029	0,096
1997	0,061	0,027	0,093
1998	0,059	0,024	0,088
1999	0,058	0,031	0,086

Table 2: Sectors at 2-digit level

Sector	CNAE 2-digit	Number of subsectors
Mining, extraction & agglomeration of anthracite, coal, lignite and peat	10	3
Extraction of crude oil and natural gas; activities in services related to oil and gas fields other than prospecting.	11	2
Uranium and thorium ore extraction	12	1
Metal mineral ore extraction	13	2
Extraction of non metallic and energy-destined mineral ores	14	5
Foodstuff products and beverage industry	15	9
Tobacco industry	16	1
Textile industry	17	7
Garment-making and fur industry	18	3
Preparation, tanning & finishing of leather; manufacture of leather and travel goods; accessories and footwear.	19	3
Wood and cork industry other than furniture, basket-making and mat making	20	5
Paper industry	21	2
Publishing, graphic arts & reproduction of recorded media	22	3
Coke, oil refining and nuclear fuel treatment plants	23	3
Chemical industry	24	7
Manufacture of rubber goods and plastics	25	2
Manufacture of other non metallic mineral products	26	8
Metallurgy	27	5
Manufacture of metal products other than machinery & equipment	28	7
Machinery and mechanical equipment construction industry	29	7
Manufacture of office machinery and computer equipment	30	1
Manufacture of electrical material and machinery	31	6
Manufacture of electronic material, manufacture of radio, TV and communication equipment and sets	32	3
Manufacture of medical & surgical, precision and optical equipment and instruments and watch-making	33	5
Manufacture of motor vehicles, trailers and semi-trailers	34	3
Manufacture of other transport material	35	5
Manufacture of furniture, other manufacturing industries	36	6
Recycling	37	2
Production & distribution of electricity, gas, steam and hot water	40	3
Catchment, treatment and distribution of water	41	1

Table 3: Concentration indices in 1999

Sector	M-S		E-G		Gini		Herfindahl	
14	-0,044	(1)	0,036	(15)	0,378	(10)	0,0022	(12)
40	-0,033	(2)	-0,003	(1)	0,475	(14)	0,0348	(23)
15	-0,032	(3)	0,012	(8)	0,192	(1)	0,0005	(3)
20	-0,028	(4)	0,022	(13)	0,273	(3)	0,0005	(4)
26	-0,024	(5)	0,032	(14)	0,310	(6)	0,0008	(7)
27	-0,007	(6)	0,046	(18)	0,528	(16)	0,0116	(17)
35	-0,004	(7)	0,041	(17)	0,577	(18)	0,0146	(18)
36	0,001	(8)	0,015	(11)	0,316	(7)	0,0004	(2)
28	0,009	(9)	0,004	(4)	0,202	(2)	0,0002	(1)
41	0,010	(10)	0,013	(9)	0,441	(12)	0,0250	(21)
34	0,019	(11)	-0,003	(2)	0,481	(15)	0,0255	(22)
29	0,020	(12)	0,011	(6)	0,309	(5)	0,0010	(10)
18	0,031	(13)	0,014	(10)	0,387	(11)	0,0008	(8)
25	0,032	(14)	0,007	(5)	0,321	(8)	0,0034	(14)
21	0,034	(15)	0,004	(3)	0,280	(4)	0,0029	(13)
31	0,040	(16)	0,011	(7)	0,348	(9)	0,0041	(15)
37	0,058	(17)	0,017	(12)	0,582	(19)	0,0173	(19)
24	0,111	(18)	0,037	(16)	0,451	(13)	0,0018	(11)
33	0,123	(19)	0,062	(20)	0,599	(20)	0,0076	(16)
22	0,128	(20)	0,056	(19)	0,531	(17)	0,0007	(6)
10	0,178	(21)	0,320	(25)	0,944	(25)	0,1714	(25)
32	0,179	(22)	0,086	(22)	0,747	(22)	0,0218	(20)
17	0,182	(23)	0,087	(23)	0,676	(21)	0,0009	(9)
30	0,221	(24)	0,074	(21)	0,906	(24)	0,1342	(24)
19	0,235	(25)	0,292	(24)	0,790	(23)	0,0005	(5)

Table 4: Shares of regional and sub-regional spillovers

Sector[27]	Concentration index, γ	Fraction of regional spillovers (%)	Fraction of sub-regional spillovers (%)	Fraction of the cross product (%)
14	-0.032	41.9	62.8	-4.7
15	-0.025	49.3	54.1	-3.4
17	0.160	87.2	5.0	7.8
18	0.036	44.5	48.8	6.6
19	0.211	66.2	12.8	21.0
20	-0.025	42.7	59.9	-2.6
21	0.030	94.3	-1.2	6.9
22	0.082	64.5	25.5	10.0
24	0.084	89.6	5.9	4.5
25	0.032	66.9	26.5	6.5
26	-0.012	-16.9	115.4	1.4
27	0.022	30.1	66.0	3.9
28	0.012	57.4	40.0	2.6
29	0.018	95.7	1.9	2.4
31	0.028	77.3	16.8	5.9
33	0.077	71.0	19.1	9.9
34	0.021	5.0	93.8	1.2
35	0.018	51.3	75.4	3.4
36	0.003	110.9	-17.3	6.4
40	-0.021	171.6	-128.9	57.3

[27] Sectors 10, 30, 32, 37 and 41 have not been considered in the analysis since some regions have not employement in these sectors, which makes the calculation of the concentration indices for these regions ($\gamma_{p(r)2}$) impossible.

Table 5: Intra- and inter-sectorial spillovers[28]

Sector	γ	Intra	Inter
14	-0,025	-0,194	1,194
15	-0,023	0,207	0,793
17	0,262	0,102	0,898
18	0,030	2,916	-1,916
19	0,262	0,938	0,062
20	-0,020	-0,257	1,257
21	0,053	0,625	0,375
22	0,099	1,782	-0,782
24	0,141	0,312	0,688
25	0,041	0,842	0,158
26	0,004	2,198	-1,194
27	0,013	0,848	0,152
28	0,013	1,103	-0,103
29	0,032	0,691	0,309
31	0,041	0,223	0,777
32	0,152	0,907	0,093
33	0,103	1,364	-0,364
34	0,002	2,604	-1,604
35	0,007	11,71	-10,71
36	0,007	18,36	-17,36
37	0,070	0,627	0,373
40	-0,067	-0,501	1,501

[28] These spillovers have been written as a percentage, i.e. the ratio between intra (or inter) spillovers and γ.

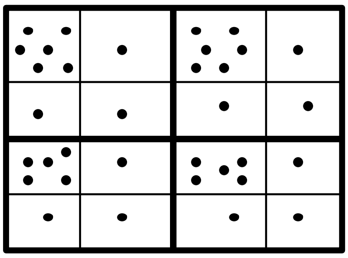

Figure 1

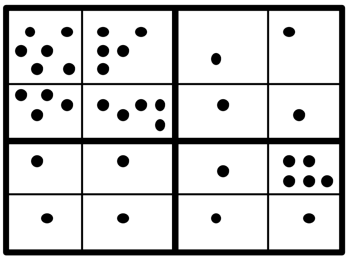

Figure 2

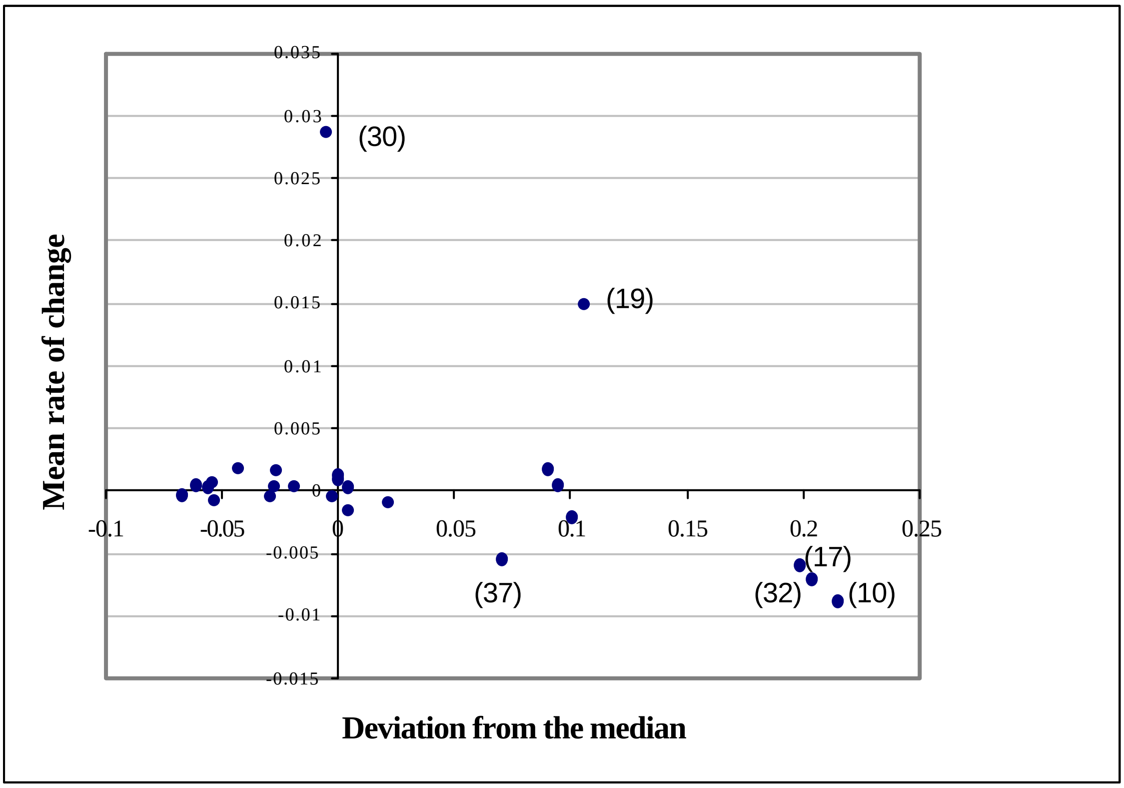
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
(30)
(19)
(37)
(17)
(32)
(10)
Mean rate of change
Deviation from the median

Figure 3

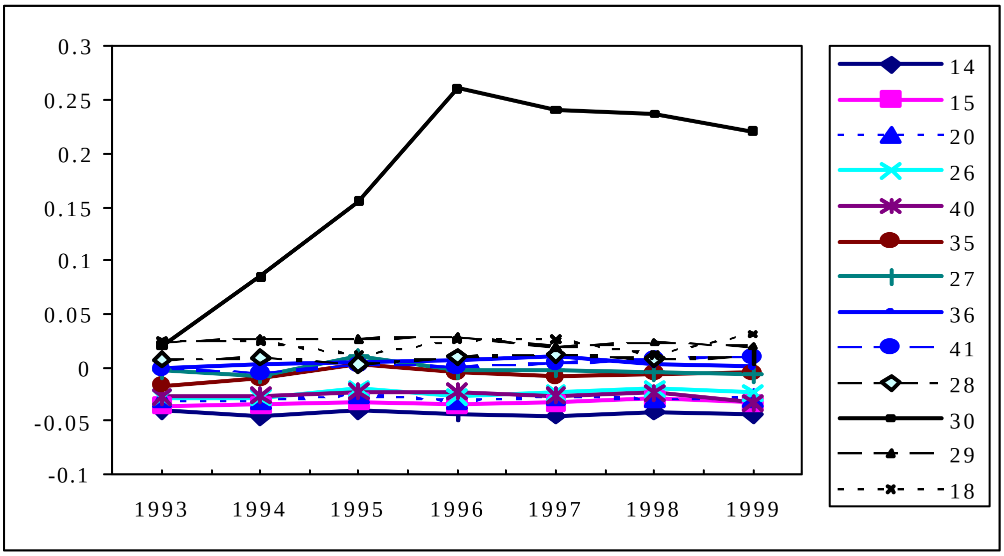

Figure 4

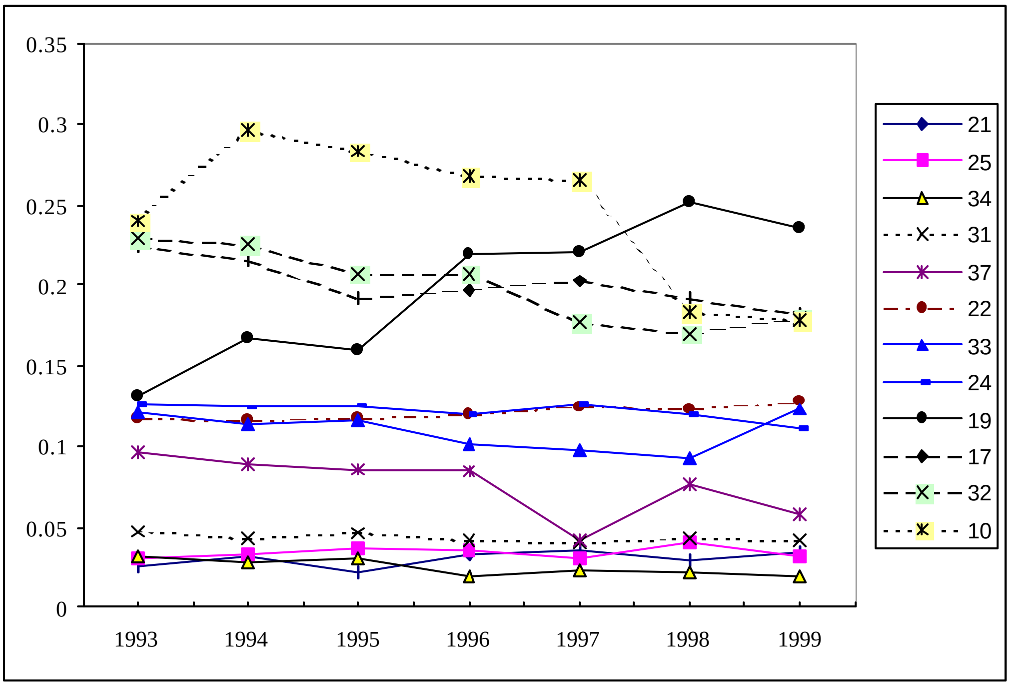

Figure 5

Appendix

Lemma 1. Let p be the probability that two plants in an industry locate in the same area, and $g = corr\left[U_{ij}, U_{ik}\right]$, where j and k represent two plants of sector r, $j \neq k$. It can be shown that

$$p = g\left(1 - \sum_i x_i^{\,2}\right) + \sum_i x_i^{\,2} \;.$$

Besides, $\hat{p} = \sum_i \frac{\sum_{\substack{j,k\in i \\ j,k\in r}} z_j z_k}{\sum_{j,k\in r} z_j z_k}$ is an estimator of p.[29] From which it follows that

$$\hat{g} = \frac{\hat{p} - \sum_i x_i^{\,2}}{1 - \sum_i x_i^{\,2}}$$

is an estimator of g.[30]

Proof. See Maurel & Sédillot (1999).

Lemma 2. Let us assume that sector r has two subsectors, l and l'. Then,

$$\sum_{\substack{j\in l \\ k\in l'}} z_j z_k = 1 - \left(\sum_{j\in l} z_j\right)^2 - \left(\sum_{k\in l'} z_k\right)^2 ,$$

where z_j is the employment share of plant j in sector r.

Proof. Taking into account that $\sum_{j\in r} z_j = 1$, we can write that

$$1 = \left(\sum_{j\in r} z_j\right)^2 = \sum_{j\in l} z_j^{\,2} + \sum_{j\in l'} z_j^{\,2} + \sum_{\substack{j\in l \\ k\in l'}} z_j z_k + \sum_{j,k\in l} z_j z_k + \sum_{j,k\in l'} z_j z_k \;. \qquad \text{(A1)}$$

A straightforward calculation shows that

---

[29] By $j,k \in i$ we mean that j and k locate in the same geographic area i.

[30] Analogous expressions can be found when $j \in l, k \in l'$, l and l' being two subsectors in sector r, $l \neq l'$. In this case, we denote by $g_0 = corr\left(U_{ij}, U_{ik}\right)$. Othewise, that is, if $j,k \in l$, we denote by $g_l = corr\left(U_{ij}, U_{ik}\right)$.

$$\sum_{\substack{j\in l\\k\in l'}} z_j z_k = 1 - \left[\sum_{j\in l} z_j^{\,2} + \sum_{j,k\in l} z_j z_k\right] - \left[\sum_{j\in l'} z_j^{\,2} + \sum_{j,k\in l'} z_j z_k\right],$$

which leads to the expression we wanted to obtain.

Proposition 1. We propose an estimator of the probability of two plants in industry r choosing the same location as given by the following expression

$$\hat{p}_0 = \frac{\sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}(1-H_l)\hat{\gamma}_l + \sum_i x_i^{\,2} \sum_{l\in r} w_l^{\,2}(1-H_l)(\hat{\gamma}_l - 1)}{1 - \sum_{l\in r} w_l^{\,2}},$$

where s_i is the employment share of sector r in location i, H is the Herfindahl index in sector r, w_l is the share of subsector l in sector r employment, H_l is the Herfindahl index in subsector l, $\hat{\gamma}_l$ is the geographic concentration index in sector l, and x_i is the proportion of the whole manufacturing employment in location i.

Proof. The estimator of p_0 we use is analogous to the one proposed by Maurel & Sédillot (1999) for the case in which plants belong to the same sector[31]

$$\hat{p}_0 = \sum_i \frac{\sum_{\substack{j,k\in i\\ j\in l,\, k\in l'}} z_j z_k}{\sum_{j\in l,\, k\in l'} z_j z_k},$$

where z_j denotes the share of plant j in employment sector r.

Step 1. We first prove that $\sum_{j\in l, k\in l'} z_j z_k = 1 - \sum_{l\in r} w_l^{\,2}$.

Using Lemma 2 when more than two subsectors exist, we can write

$$\sum_{j\in l, k\in l'} z_j z_k = 1 - \sum_{l\in r}\left(\sum_{j\in l} z_j\right)^2. \qquad \text{(A2)}$$

[31]In order to simplify notation, by j and k we mean two plants belonging to the same sector r without making it explicit in the equation. By $j \in l, k \in l'$ we mean that j belongs to subsector l, while k does to subsector l', l and l' being two subsectors of sector r.

Denoting by z_{jl} the proportion of plant j's employment with respect to subsector l, it follows that $z_j = w_l z_{jl}$. Introducing this expression in equation (A2) we obtain that

$$\sum_{j\in l,k\in l'} z_j z_k = 1-\sum_{l\in r}\left(\sum_{j\in l} w_l z_{jl}\right)^2 = 1-\sum_{l\in r}\left(w_l\sum_{j\in l} z_{jl}\right)^2 = 1-\sum_{l\in r} w_l^2\left(\sum_{j\in l} z_{jl}\right)^2 = 1-\sum_{l\in r} w_l^2 .$$

Step 2. Now we are going to prove that the numerator in $\hat{p}_0$ can be written as

$$\sum_i \sum_{\substack{j,k\in i\\ j\in l,\, k\in l'}} z_j z_k = \sum_i s_i^2 - H - \sum_{l\in r} w_l^2 (1-H_l)\left[\hat{g}_l\left(1-\sum_i x_i^2\right)+\sum_i x_i^2\right].$$

Analogously to the steps followed to obtain (A1) we have that

$$\sum_{\substack{j,k\in i\\ j\in l,k\in l'}} z_j z_k = \left(\sum_{j\in i} z_j\right)^2 - \sum_{j\in i} z_j^2 - \sum_l \sum_{\substack{j,k\in i\\ j,k\in l}} z_j z_k .$$

Using the above expression and taking into account that $\sum_{j\in i} z_j = s_i$, and $\sum_i \sum_{j\in i} z_j^2 = H$, we have that

$$\sum_i \sum_{\substack{j,k\in i\\ j\in l,k\in l'}} z_j z_k = \sum_i s_i^2 - H - \sum_i \sum_{l\in r} \sum_{\substack{j,k\in i\\ j,k\in l}} z_j z_k .$$

Since $z_j = w_l z_{jl}$, we can write

$$\begin{aligned}\sum_i \sum_{\substack{j,k\in i\\ j\in l,k\in l'}} z_j z_k &= \sum_i s_i^2 - H - \sum_i \sum_{l\in r} \sum_{\substack{j,k\in i\\ j,k\in l}} w_l^2 z_{jl} z_{kl}\\ &= \sum_i s_i^2 - H - \sum_i \sum_{l\in r} w_l^2 \sum_{\substack{j,k\in i\\ j,k\in l}} z_{jl} z_{kl}\\ &= \sum_i s_i^2 - H - \sum_{l\in r} w_l^2 \sum_i \sum_{\substack{j,k\in i\\ j,k\in l}} z_{jl} z_{kl}.\end{aligned}$$

The estimator of the probability, p_l, of two plants in subsector l choosing the same location is

$$\hat{p}_l = \frac{\sum_i \sum_{\substack{j,k\in i\\ j,k\in l}} z_{jl} z_{kl}}{\sum_{j,k\in l} z_{jl} z_{kl}} .$$

Using this estimator the above expression can be written as

$$\sum_{i}\sum_{\substack{j,k\in i\\ j\in l,\, k\in l'}} z_j z_k = \sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\left(\hat{p}_l \sum_{j,k\in l} z_{jl} z_{kl}\right).$$

Besides,

$$1 = \left(\sum_{j\in l} z_{jl}\right)^2 = \sum_{j\in l} z_{jl}^{\,2} + \sum_{j,k\in l} z_{jl} z_{kl}\,.$$

Hence

$$\sum_{i}\sum_{\substack{j,k\in i\\ j\in l,\, k\in l'}} z_j z_k = \sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\hat{p}_l\left(1 - \sum_{j\in l} z_{jl}^{\,2}\right).$$

Since $H_l = \sum_j z_{jl}^{\,2}$, it follows that

$$\sum_{i}\sum_{\substack{j,k\in i\\ j\in l,\, k\in l}} z_j z_k = \sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\hat{p}_l\left(1 - H_l\right).$$

Using Lemma 1 in subsector l we have that

$$\hat{p}_l = \hat{g}_l\left(1 - \sum_i x_i^{\,2}\right) + \sum_i x_i^{\,2}\,,$$

from which we get to Step 2

$$\sum_{i}\sum_{\substack{j,k\in i\\ j\in l,\, k\in l}} z_j z_k = \sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\left(1 - H_l\right)\left[\hat{g}_l\left(1 - \sum_i x_i^{\,2}\right) + \sum_i x_i^{\,2}\right].$$

Step 3. Finally, we use Steps 1 and 2 in $\hat{p}_0$ and after

$$\hat{p}_0 = \frac{\sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\left(1-H_l\right)\left[\hat{g}_l\left(1 - \sum_i x_i^{\,2}\right) + \sum_i x_i^{\,2}\right]}{1 - \sum_{l\in r} w_l^{\,2}}$$

$$= \frac{\sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\left(1-H_l\right)\hat{g}_l + \sum_{l\in r} w_l^{\,2}\left(1-H_l\right)\hat{g}_l \sum_i x_i^{\,2} - \sum_{l\in r} w_l^{\,2}\left(1-H_l\right)\sum_i x_i^{\,2}}{1 - \sum_{l\in r} w_l^{\,2}}$$

$$= \frac{\sum_i s_i^{\,2} - H - \sum_{l\in r} w_l^{\,2}\left(1-H_l\right)\hat{g}_l + \sum_i x_i^{\,2}\sum_{l\in r} w_l^{\,2}\left(1-H_l\right)\left(\hat{g}_l - 1\right)}{1 - \sum_{l\in r} w_l^{\,2}}.$$

Theorem 1. The estimator of the correlation between the location decision of two plants belonging to different subsectors of the same sector, $g_0 = corr\left(U_{ij}, U_{ik}\right)$, can be written as

$$\hat{g}_0 = \frac{G - H - \sum_{l \in r} w_l^2 (1 - H_l)\hat{g}_l}{1 - \sum_{l \in r} w_l^2},$$

where $G = \frac{\sum_i s_i^2 - \sum_i x_i^2}{1 - \sum_i x_i^2}$.

Proof. Using Lemma 1 when plants belong to different subsectors, we have that $\hat{g}_0 = \frac{\hat{p}_0 - \sum_i x_i^2}{1 - \sum_i x_i^2}$, where $\hat{p}_0$ is the estimator of the probability of two plants in different subsectors choosing the same location. Using expression $\hat{p}_0$ in Proposition 1, we can rewrite $\hat{g}_0$ as

$$\hat{g}_0 = \frac{1}{1 - \sum_i x_i^2}\left[\frac{\sum_i s_i^2 - H - \sum_l w_l^2 (1 - H_l)\hat{g}_l + \sum_i x_i^2 \sum_{l \in r} w_l^2 (1 - H_l)(\hat{g}_l - 1)}{1 - \sum_{l \in r} w_l^2} - \sum_i x_i^2\right]$$

$$= \frac{1}{1 - \sum_i x_i^2}\left[\frac{\sum_i s_i^2 - H - \sum_{l \in r} w_l^2 (1 - H_l)\hat{g}_l + \sum_i x_i^2 \sum_{l \in r} w_l^2 (1 - H_l)(\hat{g}_l - 1) - \sum_i x_i^2 \left(1 - \sum_{l \in r} w_l^2\right)}{1 - \sum_{l \in r} w_l^2}\right]$$

$$= \frac{1}{1 - \sum_i x_i^2}\left[\frac{\sum_i s_i^2 - \sum_i x_i^2 - H - \sum_{l \in r} w_l^2 (1 - H_l)\hat{g}_l + \sum_i x_i^2 \sum_l w_l^2 (1 - H_l)(\hat{g}_l - 1) + \sum_i x_i^2 \sum_l w_l^2}{1 - \sum_l w_l^2}\right]$$

$$= \frac{\dfrac{\sum_i s_i^2 - \sum_i x_i^2}{1-\sum_i x_i^2} + \dfrac{1}{1-\sum_i x_i^2}\left[-H - \sum_{l\in r} w_l^2 (1-H_l)\hat{g}_l + \sum_i x_i^2 \sum_{l\in r} w_l^2 (1-H_l)(\hat{g}_l - 1) + \sum_i x_i^2 \sum_{l\in r} w_l^2\right]}{1-\sum_{l\in r} w_l^2}$$

$$= \frac{1}{1-\sum_{l\in r} w_l^2}\left\{G + \frac{1}{1-\sum_i x_i^2}\left[\left(-1+\sum_i x_i^2\right)\sum_{l\in r} w_l^2 (1-H_l)\hat{g}_l\right]\right.$$

$$\left. + \frac{1}{1-\sum_i x_i^2}\left[-H - \sum_i x_i^2 \sum_{l\in r} w_l^2 (1-H_l) + \sum_i x_i^2 \sum_{l\in r} w_l^2\right]\right\}$$

$$= \frac{1}{1-\sum_{l\in r} w_l^2}\left\{\left[G - \sum_{l\in r} w_l^2 (1-H_l)\hat{g}_l\right] + \frac{1}{1-\sum_i x_i^2}\left[-H + \sum_i x_i^2 \sum_{l\in r} w_l^2 H_l\right]\right\}$$

$$= \frac{G - H - \sum_{l\in r} w_l^2 (1-H_l)\hat{g}_l}{1-\sum_{l\in r} w_l^2}.$$

References

Callejón, M. (1997): "Concentración geográfica de la industria y economías de aglomeración", Economía Industrial, 317, pp. 61-68.

Callejón, M. and Costa, M.T. (1995): "Economías externas y localización de las actividades industriales", Economía Industrial, 305, pp. 75-86.

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

Glaeser, E.; Kallal, H.; Scheinkman, J. and Shleifer, A. (1992): "Growth in cities", Journal of Political Economy, 100, pp. 1126-1152.

Henderson, V., Kunkoro, A. and Turner, M. (1995): "Industrial development in cities", Journal of Political Economy, 103, pp. 1067-1090.

Krugman, P. (1991): "Increasing returns and economic geography", Journal of Political Economy, 99, pp. 483-499.

Marshall, A. (1890): Principles of Economics, 8th edn 1986, Macmillan, Basingstoke.

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

Myrdal, G. (1957): Economic Theory and Under-developed Regions. London: Duckworth.

Ottaviano, G. and Puga, D. (1998): "Agglomeration in the global economy: a survey of the 'new economic geography' ", The World Economy, 21, pp. 707-731.

Paluzie, E., Pons, J. and Tirado, D. (2000): "Regional integration and specialization patterns in Spain", Documents de Treball de la Divisió de Ciències Jurídiques, Econòmiques i Socials, E00/62, Universitat de Barcelona.

Warf, B. (1995): "Telecommunications and changing geographies of knowledge transmission in the late 20th century", Urban Studies, 32, pp. 361-378.