La versión final está publicada en:
https://www.tandfonline.com/doi/abs/10.1080/09535310600844375
Cuyo doi es: https://doi.org/10.1080/09535310600844375

CESR184370

Queries

Bárbara Díaz, Laura Moniche & Antonio Morillas

Q1 Please clarify sense. 'A summary of the weight in the Spanish economy of the sectors, when classified the preferred cluster, that sheds some more light on the classification...'

Q2 Please clarify sense'... a membership equal to one for the cluster where it belongs to ...' Does this mean the following? '... a membership equal to one for the cluster to which it belongs...'

Q3 Oosterhaven (2002) is not in the References. Is this Oosterhaven & Stelder (2002)?

Q4 Korte & Oberhofer (1970) and Lantner (1972) are not cited in the text.

Q5 Please provide initial for Campbell (1975)

Economic Systems Research
Vol. 18, No. 3, 299–318, September 2006

Routledge
Taylor & Francis Group

A Fuzzy Clustering Approach to the Key Sectors of the Spanish Economy

BÁRBARA DÍAZ*, LAURA MONICHE** & ANTONIO MORILLAS***

**BISC, University of California at Berkeley, USA, **Departamento de Estructura Económica, Universidad de Málaga, Spain, ***Departamento de Estadística y Econometría, Universidad de Málaga, Spain*

(Received October 2005; revised May 2006)

ABSTRACT *The search for key sectors in an economy has been and still is one of the more recurrent themes in input–output analysis. When using clustering techniques, sectors can only belong to a group, having a particular performance. But, actually, the same sector could be important from different perspectives at the same time, to a different degree. So, a fuzzy clustering approach is needed. In this work we propose a multidimensional approach to classify the productive sectors of the Spanish input–output table for 1995, based on three groups of variables: those related to their productive integration, others measuring their specific weight in the economy and finally some showing their economic dynamic. We also incorporate into the analysis the technological level, which being a categorical variable presents special methodological problems. All these questions are tackled applying a robust and fuzzy clustering analysis, which gives as a result a classification of sectors illustrating the role that each one plays in the Spanish economy.*

KEY WORDS: Key sectors, fuzzy clustering, input–output analysis, Spanish economy

1. Introduction

The search for the important sectors in an economy started, within the input–output analysis, in a very elementary way. The buy and sell direct relationships among sectors or the use of the multiplier based on the inverse of Leontief's model would not nowadays be considered sufficient criteria to identify key sectors. Nevertheless, Hirschman (1958) considered the seminal work by Chenery and Watanabe (1958) and Rasmussen (1956) as a way to measure what he called linkages in his development theory. Probably the discussion over causality in the economy (Simon, 1952; Wold, 1954) contributed to the idea that

Correspondence Address: Antonio Morillas, Departamento de Estadística y Econometría, Facultad de Ciencias Económicas y Empresariales, Universidad de Málaga, C/ El Ejido s/n, 29071 Málaga, Spain. Email: morillas@uma.es.

0953-5314 Print/1469-5758 Online/06/030299–20
DOI: 10.1080/09535310600844375

linkages effects, using causality or hierarchy as a criterion, were a determinant of the economic relevance of a sector.

One of the first controversial debates was on which matrix was more suitable to be used, the input matrix from Leontief's model or the output matrix proposed by Ghosh (1958). The discussion on this issue includes the papers by Augustinovics (1970), Cella (1984), Oosterhaven (1988) and Dietzenbacher (1997). Other analysed aspects were those related to the aggregation problem (Hewings, 1974) or to the necessity of normalizing and weighting the multipliers (relative importance of the gross value added, employment or final demand of the sectors). In Oosterhaven and Stelder (2002) and Oosterhaven (2004) these last ones are called net multipliers.[1]

Another approach to the study of the importance of sectors and/or coefficients is the extraction method, introduced by Strassert (1968), with the elimination of a sector (i.e. a row and a column) from an input–output table. Cella (1984) proposed a less drastic solution, with the possibility of distinguishing between forward and backward effects. Others have used partitioned matrices, see Dietzenbacher (1992), Dietzenbacher *et al.* (1993), Sonis *et al.* (1995), Dietzenbacher and van der Linden (1997), Miller and Lahr (2000) and Cai and Leung (2004), or relate extraction to graph theory (Aroche-Reyes, 2002).

All these methods underlie in some way Hirschman's old idea on linkages and the equilibrium–disequilibrium dialectic in growth. However, without doubt, they are also related to the no less old concepts of polarization, growth poles and driving force industries (Perroux, 1949) and with those of domination and asymmetries in economic relationships, studied by several French economists such as Perroux (1973) and Lantner (1974).

This interest to study the hierarchy among sectors was also fuelled from other perspectives like triangularization of the interindustry matrix (e.g. Chenery and Watanabe, 1958; Simpson and Tsukui, 1965; Fukui, 1986), analyzing block-causal structures or applying decomposition methods (filters, reduced graphs, skeleton graphs, articulation points). Typically, the concepts are close to graph theory (Lantner, 1974; Campbell, 1975; Morillas, 1983a; Hauknes, 1999) and the resulting blocks can be considered as clusters or industrial groups.

In most of these works one can already glimpse the importance of the industries networks development, where a growth impulse does not so much follow from the hierarchy of autonomous driving sectors but rather from the recursive interdependence of clusters or groups of industries. The studies on industrial clusters are almost as numerous, and as diverse in their considerations on what should be taken into account. The seminal works by Czamanski (1974) and, subsequently, Porter (1990) caused a growing interest in industrial clustering from different areas of knowledge. Industrial clustering[2] is focused basically on analysing the relationships among sectors, although occasionally it incorporates other variables (such as employment, investment, exports) or it includes aspects such as technological innovation (Schmookler, 1966; Scherer, 1982, 2003; Papaconstantinou *et al.*, 1996; Sakurai *et al.*, 1997; DeBresson and Hu, 1999; Verspagen, 1999; Dietzenbacher and Los, 2002) or the sector's strategic value in growth models (Los, 2001).

Apart from the referred graph theory and the maxim method (Roelandt *et al.*, 1999; Oosterhaven *et al.*, 2000; Broersma, 2001; Peeters *et al.*, 2001; Hoen, 2002), the applications in the input–output context have made use of very different analysis methods, including some multivariate statistical techniques as well. One of the first applications of multivariate analysis to study industrial groups through an input–output table was

Czamanski (1974), in which a principal component analysis is carried out on the US input–output table for 1963, with the aim of grouping sectors on the basis of their buy and sell profiles. Roepke *et al.* (1974) made a similar application to the Ontario 1965 tables. Latham (1976) and Harrigan (1982) criticised these methods considering that they only grouped vertically linked sectors. Ó hUallacháin (1984) proposes a posterior analysis of the components found in the inter-industrial flux matrix, to verify the complementarities of the grouped sectors, supporting the suitability of these statistical methods. Feser and Bergman (2000) differentiate in a group what they consider key and secondary sectors, attending to their factorial load,[3] in an application that aims to identify the industrial clusters in the manufacturing sector of the US, using the 1987 table and Czamanski's coefficients. In later works Feser *et al.* (2001a, b) carry out a study of the spatial concentration of employment for each group.[4]

Rey and Mattheis (2000) elaborate a consensus methodology to combine the results of some principal components and clustering techniques and incorporate, for the first time, a non-hierarchical clustering to determine industrial clusters.[5] Lainesse and Poussart (2005) introduce a hybrid proposal, based on the former one, which adds a supplementary analysis of indicators with the aim of including or excluding sectors according to their complementarities.

Our approach is difficult to locate in a specific line of those previously mentioned. From a conceptual point of view it is nevertheless unquestionable that it is fed from all of them. We do not sustain our study on the links among sectors as a key sector analysis or even an industrial clusters study would do. But we do not neglect them either, because the incorporation of a synthetic quantitative measure of them (including the multiplier and its dispersion along with other two qualitative measures – i.e. the cohesion grade and the interrelationship index) takes into account aspects related to sector interdependencies. We do not consider either the spatial aspects or the existence of common customers or providers, which is typical for this kind of analysis. Instead, we propose a strategic vision for development in the identification of the important sectors. In doing so, we use criteria such as the sectors' innovative capacity, their exporting potential and dynamic, or the generation of high salary rents. These criteria have, in general, not been used jointly, and to our knowledge they have never been considered from a dynamic perspective, as introduced in this paper. Finally, we want to emphasize that an additional contribution of this paper lies (as will become clear later) in the methodological procedures used, which are based on a fuzzy multidimensional cluster analysis that has never been used before in this context.

The aim of the statistical cluster analysis is to classify objects in a certain number of sets so that the elements are as similar as possible within each set and dissimilar to elements belonging to other sets. Each one of the objects can only belong to a determinate set. In fuzzy cluster analysis it is quite different because there are no sharp boundaries between clusters and an object can simultaneously belong to several clusters. Membership degrees between zero and one are used for each object instead of assigning each element to one of the clusters.

This fact allows each sector to belong to any cluster with a different degree, adding flexibility to the classification task and making it possible for each one of them to have simultaneously different performances. This diversity has been considered of interest (Feser and Bergman, 2000; Dridi and Hewings, 2003; Lainesse and Poussart, 2005) and a fuzzy clustering is probably the most appropriate way to get it (see Beynon *et al.*, 2005, for a recent application to input–output analysis by using fuzzy output multipliers for the ranking of sectors).

Before introducing some fuzzy logic concepts and fuzzy clustering methods, it is necessary to remember that in most of the classical clustering methods one could only consider numerical variables.[6] That is, the variables should be measured at least on an interval scale. However, the so-called *relational* clustering methods allow the use of ordinal and nominal variables, since in these methods the input information represents a dissimilarity measure among elements or objects. Because both quantitative and ordinal variables are used in this paper for the sector classification in the Spanish economy, the alternative approach to traditional clustering will be a *relational fuzzy clustering*, where a crucial problem will be how to obtain the distance matrix between objects.

2. Methodology and Statistical Sources

Zadeh (1965) established the basis of fuzzy logic making a fundamental contribution to the representation of human knowledge, giving the basis for its proper treatment and formalization, taking into account the imprecision that governs it and its approximate nature. Within fuzzy logic, two of the principles of classic logic are overcome. The first one is the contradiction principle (i.e. 'at the same time *B* and not-*B*'), implying that opposed concepts cannot overlap. In terms of sets, the intersection of the two sets *B* and not-*B* is the empty set according to the contradiction principle. The second one is the bivalence law. Any proposition has to be true or false, because there are only two 'truth values': true or false. This implies the middle excluded law: *A* has to be either *B* or not-*B* (in other words, the union of the two sets *B* and not-*B* is the universal set). These two laws had kept classic (bi-valued) logic away from some aspects of the real world.

A 'crisp' or classic set is defined so that the elements in a universe are divided in two groups: members (those that certainly belong to the set) and non-members (those that do not belong to it). There is a precise and clear distinction between members and non-members of the class or category represented by the set. Yet, we know that in real world cases, many of the categories that we use do not fit this representation. In natural language we talk about expensive cars, low rents, wealthy regions, solvent customers, or important (key) sectors. In these cases, the transition between members and non-members appears gradually. To cope with this, a fuzzy set can be mathematically defined to assign a value to each element of the universe that represents its membership to that set. This value is the grade to which an element is compatible with the concept represented by the fuzzy set. In other words: for any element j, its membership to a set i is given a certain value u_{ij}. For a classic or 'crisp' set, u_{ij} has only two values, 0 and 1, meaning respectively non-membership and membership. For fuzzy sets, u_{ij} can have any value between 0 and 1. As one can see, in fuzzy logic, precise reasoning is only a limit case of approximate reasoning, so that fuzzy logic is a natural extension of classic logic.

Then, in a fuzzy set, different degrees of membership are allowed. Its characteristic function is called the membership function, and can take any value in the real interval [0,1]. The fuzzy set i is defined as follows:

$$i = \left\{ (j, u_{ij}) : j \in U, u_{ij} \in [0,1] \right\}$$

Fuzzy sets have been used in very different domains, and many classic modelling, classifying and decision-making techniques have incorporated them. We can find in the literature fuzzy regression, fuzzy inference systems, fuzzy clustering and fuzzy decision-making among others.

In fuzzy clustering, each cluster is a fuzzy set. It has its origins in Ruspini (1969) who pointed out several advantages of using fuzzy clustering. That is, a membership value equal or close to one would identify 'core' points in a cluster, while lower membership values in a cluster would identify boundary points. 'Bridge' points, in Nagy's (1968) terminology, may be classified within this framework as 'undetermined points with a degree of indeterminacy proportional to their similarity to core points'. Bridges or strays between sets originated problems of misclassification in classic cluster analysis.

Zadeh (1977) provides an outline of a conceptual framework for cluster analysis and pattern classification based on the theory of fuzzy sets, since at that time there was still no unified theory. Bezdek (1981) subsequently generalized Dunn's (1973) work, introducing a fuzzy version of the well known c means clustering algorithm. In Kaufman and Rousseeuw (1990) a different objective function is used. Their dissimilarity or distance measure makes it more robust than fuzzy c means (Rousseeuw, 1995). Hathaway *et al.* (2000) present a generalization of fuzzy c means to allow the use of different distances (norms).

Hathaway *et al.* (1989) shows that there is a dual relational problem of object fuzzy c means for the special case in which the relational data coincide with the Euclidean distances among their elements. Their method is called RFCM (relational fuzzy c means). Later, Hathaway and Bezdek (1994) introduce a non-Euclidean relational fuzzy clustering (NERFCM). Their objective is to modify RFCM to be effective with arbitrary relational data that show dissimilarities among objects. In this direction, the fuzzy analysis clustering in Kaufman and Rousseeuw (1990) can be used with both kinds of input data, since the distances in the objective function are distances among objects and not to the centres of the clusters. Dave and Sen (2002) propose another robust version of fuzzy analysis clustering called robust non-Euclidean fuzzy relational clustering (robust-NE-FRC). In our case, we have used the NERFCM algorithm (Hathaway and Bezdek, 1994).[7] The objective function to minimize is:

$$\min \sum_{i=1}^{k} \frac{\sum_{j=1}^{N} \sum_{l=1}^{N} u_{ij}^{m} u_{il}^{m} r_{jl}}{2 \sum_{t=1}^{N} u_{it}^{m}}$$

where N is the number of cases (sectors), $m > 1$ is a parameter controlling the amount of fuzziness in the partition, k is the number of clusters, u_{ij} is the membership of case j to cluster i, with $u_{ij} \in [0, 1]$ and $\Sigma_{i=1}^{k} u_{ij} = 1$. Memberships cannot be negative and each sector has a constant total membership, distributed over the different clusters. By convention, this total membership is normalized to one, and r_{jl} measures the relationship between cases j and l.

The algorithm for the NERFCM proceeds in several steps. In the first stage, the membership matrix is initialized for a given number (k) of clusters. With these memberships, the *mean vectors* are then determined for each one of the k clusters.[8] In the third step, distances for each element to the *mean vector* of each cluster are calculated. It is in this step where a modification from RFCM takes place, ensuring that the distances are not negative.[9] In the fourth step, the memberships are updated. This process is carried out until a convergence rate in the membership matrix is reached. The results of this algorithm are given in a matrix containing the final membership values of each sector to each cluster.

Before applying the NERFCM algorithm, several preliminary steps are necessary, however. The first involves pre-processing of the variables, while the second provides the relational matrix on which the NERFCM will operate. Step three is to comply with the need for an initialization of the membership values for the NERFCM objective function. This is done by using a robust crisp clustering method. These three steps are exposed in more detail in the next paragraphs.

With respect to the first step, recall that in cluster analysis the presence of multicollineality among variables causes the affected ones to be more represented in the similarity measure. For this reason, and given the multidimensional nature of our approach, we will use as variables the ones resulting from a previous principal component analysis applied to our database. This allows avoiding the multicollineality problem when obtaining the similarity matrix, and also provides a better conceptual identification of the resulting clusters, reducing the number of variables for the classification task.

Secondly, since NERFC operates on relational data, for the calculus of the distance matrix among sectors involving the principal components in Step 1 (which are quantitative variables) and the technology variable (which is an ordinal variable) we use the *daisy* method proposed by Kaufman and Rousseeuw (1990). This method computes all pairwise dissimilarities among objects as follows:

$$r_{jl} = \frac{1}{nok} \sum_{h} w_{jl}^{h} d_{jl}^{h}$$

where *nok* is the number of non-zero weights, and

$$w_{jl}^{h} = \begin{cases} 0 & \text{if } h \text{ is missing for } j \text{ or } l, \text{ or } h \text{ is as ymmetric binary and both values are } 0 \\ 1 & \text{otherwise} \end{cases}$$

Depending on the type of variable h:

$$d_{jl}^{h} = \begin{cases} 0 & \text{if } x_{jh} = x_{lh} \\ 1 & \text{if } x_{jh} \neq x_{lh} \end{cases} \text{ if } h \text{ is a nominal or binary variable,}$$

$$d_{jl}^{h} = \frac{|R_{jh} - R_{lh}|}{\max R_h - \min R_h} \text{ if } h \text{ is an ordinal variable (} R \text{ indicates the rank)}$$

$$d_{jl}^{h} = \frac{|x_{jh} - x_{lh}|}{\max x_h - \min x_h} \text{ if } h \text{ is an interval scaled variable.}$$

Finally, the NERFC algorithm needs an initialization matrix for the memberships that it will subsequently optimize. For this, we have used the results provided by the crisp partitioning around medoids (PAM) clustering method (Kaufman and Rousseeuw, 1990), which was used in Rey and Mattheis (2000).[10] This is a robust version of the well-known k means method (MacQueen, 1967) and is less sensitive to the presence of outliers. The PAM method selects k representative or typical objects called medoids among the observations. A medoid is an object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. Each observation is assigned to the closest medoid. The k medoids selected should minimize the objective function (which is the sum of the dissimilarities of all objects i to their nearest medoids m_i).[11] That is,

$$O = \min \sum_{i=1}^{k} \sum_{j \in i} r_{jm_i}$$

The PAM method sequentially selects k centrally located objects, used as initial medoids. Then, it computes the swapping cost of interchanging a selected object with an unselected one as $S = O' - O$. If $S < 0$, the objective function is reduced with the new medoid, and this is done until there is no benefit in changing the latest medoids.

To select the best number of clusters k for the PAM method we use the silhouette width method, proposed in Rousseew (1987), which not only makes the grouping more robust but also helps to select the suitable number of groups, a question that is always important in cluster analysis. The silhouette width for a given object (sector) j is calculated as $s_j = (b_j - a_j)/[\max(a_j, b_j)]$, where a_j is the average distance between j and all the other observations in its cluster i (i.e. the 'within' dissimilarity), b_j is the minimum average distance of j to objects in other clusters (i.e. the 'between' dissimilarity). The silhouette width has values between -1 and 1, and we have $-1 \leq s_j \leq 1$. If $s_j = 1$, the 'within' dissimilarity is much smaller than the 'between' dissimilarity, so object j has been assigned to an appropriate cluster. The case $s_j = 0$ means $a_j \approx b_j$ and it is not clear whether j should be assigned to A or B. This can be considered an intermediate case. If $s_j = -1$, object j is misclassified.

Next, the average silhouette width for each cluster and an overall average silhouette width are calculated. The overall average silhouette width is the average of s_j for all objects in the whole dataset. The largest overall average silhouette indicates the best clustering. Therefore, the number of clusters with maximum overall average silhouette width is considered the optimal number of clusters.

The variables we use, their statistical sources and respective definitions, are shown in Table 1. We have distinguished between three concepts or blocks. The economic integration block contains four variables. It captures both the intensity and the quality of the sectors' interdependence. The variables are: the multiplier derived from the relative influence graph,[12] its dispersion as measured by the variation coefficient, the cohesion grade[13] and a topologic integration index,[14] obtained for each sector. The last two variables express qualitative aspects related with the relative position of a sector in the exchange structure. The economic weights block (with four variables) summarizes the relative participation in the national output, in exports and in wage rents and employment generation. The economic potential block (with five variables) includes a dynamic vision of the former variables through the observed trends plus the technological innovation capacity along with the information given by Eurostat (1998a, b) and the high technology indicators of the Spanish Statistics National Institute (INE) given in Appendix 1.

The basis for our study is the Spanish 1995 input–output table. The initial 70 branches (sectors) have been reduced to 66 in the paper for two reasons. First, the non-existence of technology data for the first three branches, which belong to the primary sector. Second, branch 70 (Homes with domestic personnel) has no relationships with the other sectors.

Table 1. List of variables

Variable	Definition	Comments
Economic integration block		
MULTIPLIC	Multiplier	Addition by columns of Ghosh's inverse matrix, Augustinovics' (1970) multiplier.
CVINVERS	Inverse of the multipliers' variation coefficient	Each sector is characterized by the inverse of the dispersion, represented by its column variation coefficient.
COHESIO	Cohesion Grade	Number of times that a sector is a path (from graph theory.)
Ri	Total productive integration Index	A topological offer-demand integration index for each sector (see Note 18).
Economic weights block		
PRODUC	Output	Output at basic price by product from the Spanish symmetric table.
EXPORT	Exports	The source is the Spanish symmetric table for 1995, INE (2000).
EMPLEO	Full-time equivalent employment	Since these data are only available by industry (use table), we have applied the correction coefficient that presents the compensation of employees by products.
SALARIOS	Mean salaries	It is estimated with corrected paid employment (also with use table) and data of compensation of employees (from the symmetric table.)
Economic potential block		
TENPROD	Output trend	It is the real output variation rate for each sector for the period 1995–2000 (source: INE a).
TENEMPL	Employment trend	It is the full-time equivalent employment variation rate for each sector respect to the initial period, from 1995 to 2000. We used the implicit sectoral deflator of the Gross Valued Added at basic prices (VABpb in the National Accounting of the Spanish National Institute of Statistics, INE a).
TENSALAR	Mean salary trend	It is the real mean salary variation rate for each sector for the period 1995–2000. We used the consumption prices index (IPC) variation during that period (14,2) to deflate salaries for 2000.
TENEXPO	Exportation trend	It is the real exportation variation rate for each sector for the period 1995–2000.The procedure used to deflate is the one used for output.
TECNO	Technological innovation capacity	Sectors classification attending to their technological level (high, medium or low technology.)

3. The Sectoral Grouping Keys for the Spanish Economy

We have extracted the subjacent principal components to avoid the possible effects of multicollineality on the similarity matrix, as we mentioned earlier. We have finally found in the rotated solution that three components with an eigenvalue slightly greater than one explain about 88% of the variance (41%, 24% and 23% respectively). Considering only the more important factor loadings, the first of these components is mainly integrated by the variables multiplier, its dispersion index and the sector output. We term this component *economic relevancy*. The second component (to be termed *integration*) mainly considers the two qualitative measures based on graph theory, i.e. cohesion and productive integration indexes. The third component, with exports and their trend as the more important variables, will be identified as *export base*. The other variables seem not to be of sufficient importance (or they are redundant) to explain the data cloud variability.

With these three numerical variables and the ordinal variable (termed *technology*) we have carried out the fuzzy analysis clustering mentioned earlier. The best option given by the silhouette plot in the PAM method (as shown in Figure 1) has three clusters with 32, 23 and 11 sectors respectively. This option will be subsequently used as an input into the NERFCM algorithm. Note that the average silhouette width indicates an ill-delimited classification. There is a lot of fuzziness, which prompts fuzzy clustering.

The solution obtained using the fuzzy clustering[15] allows going deeper into the real meaning of the assignment of a sector to a given group and featuring it with greater knowledge. Table 2 shows only those sectors with a membership value equal to or greater than 0.5 in a cluster. It would be indicative of the priority (although not exclusively) of classifying such a sector to a certain cluster. In this situation are 24 sectors of cluster 1, 15 of cluster 2, and 9 of cluster 3. Even accepting this threshold, it means that the other 18 sectors (i.e. more than 27% of the sectors) can be considered as sectors that may be classified simultaneously from different perspectives. This reasonable result would not have been obtained if we had used any other crisp (as opposed to fuzzy) clustering technique.[16]

In cluster 1 there are eight sectors that are difficult to classify. In particular, this holds for sectors 42 (Railway transportation services), 4 (Other mining and quarrying) and 37 (Construction). The first two of them do not have a clear profile in any of the components. The third one, being clearly a member of cluster 1 since it is the sector with the largest economic relevancy, presents serious difficulties because it ranks at number 40 on ‘integration’.

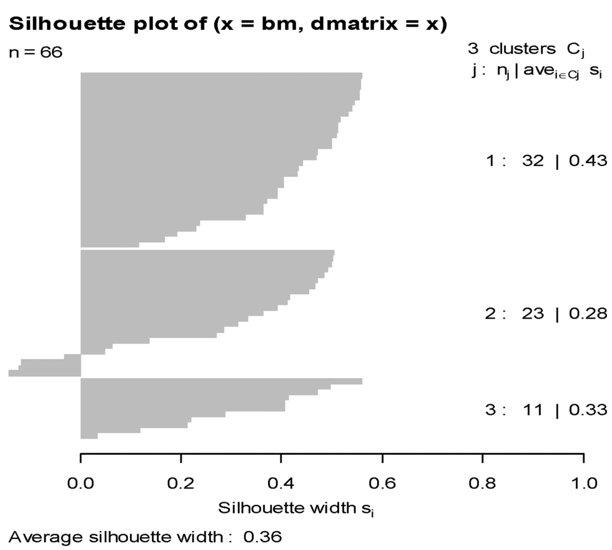

Figure 1. Silhouettes for the best classification

Table 2. Clusters and membership values

CLUSTER	Membership ≥ 0.5			Membership < 0.5			
	Sectors			Sector	Membership	Alternative cluster	Membership
Cluster 1	5	18	43	58	0.4993	Cluster 2	0.3769
Local base:	6	19	46	48	0.4744	Cluster 2	0.4175
- Economic Relevancy	7	21	51	23	0.4619	Cluster 2	0.3897
- Integrated	8	27	52	42	0.4475	Cluster 2	0.4366
- Low Exporting Base	14	35	55	4	0.4462	Cluster 2	0.4242
	15	38	56	44	0.4445	Cluster 2	0.3021
	16	39	57	41	0.4357	Cluster 2	0.3361
	17	40	59	37	0.3786	Cluster 2	0.3739
Cluster 2	2	50		61	0.4965	Cluster 1	0.3106
Services, Final Consumption:	3	60		1	0.4781	Cluster 1	0.4037
- Low Integration	9	62		22	0.4743	Cluster 1	0.4090
- Low Economic Relevancy	10	63		11	0.4727	Cluster 1	0.3276
- Low Exporting Base	12	64		24	0.4614	Cluster 1	0.3723
	13	65		25	0.4522	Cluster 1	0.4322
	36	66		26	0.4093	Cluster 1	0.3623
	49			45	0.3993	Cluster 1	0.3708
Cluster 3	28	34		20	0.4600	Cluster 1	0.2942
Exporting Base:	29	47				Cluster 2	0.2458
- Technological, Exporting	30	53		33	0.3900	Cluster 2	0.3233
- Low Economic Relevancy	31	54				Cluster 1	0.2867
- Medium-High Integration	32						

Taking into account the sectors with higher membership value in cluster 1, it would have the highest mean for ‘economic relevancy’ and ‘integration’, but the lowest for ‘export base’. All the sectors with higher membership value in this cluster have a low technological level. These sectors could be considered to form the local base of the Spanish economy. Among them, we find those with more economic relevancy and the ones showing a greater interdependency.

Appendix 2 shows that sectors belonging to cluster 1 are in manufacturing (e.g. Wood and cork, Wearing apparel and furs; Printed matter and recorded media), some leisure services and those related to sports, mainly because of their high level of integration. Curiously, among the sectors with the highest level of membership in cluster 1, the two sectors with highest possibility of belonging to this group are Construction (37) and Hotel and restaurant services (41). Nevertheless, they rank first and second respectively on economic relevancy. The reason for this can only be that they are clear outliers attending to the first component or the variables forming this component.

It can be seen that four sectors of cluster 2, which are wrongly classified in the silhouette plot, have memberships very close to the ones they have in cluster 1. Those sectors are 1 (Hard coal, lignite and peat), 22 (Cement, lime and plaster), 25 (Other non-metallic mineral products), and 45 (Air transport services). They rank on places 42 to 45 on ‘integration’. Especially fuzzy is the classification of sectors 45 and 25, with membership values very similar in the other clusters. Sector 45 presents a good behaviour on ‘export base’ and sector 25 on ‘economic relevancy’, especially because of its recent growth trend and its high multiplier. Cluster 2 is basically distinguished for being less integrated, with ‘economic relevancy’ and ‘export base’ below the mean (below zero in the respective components). Its technological level is also low. The key to interpret this cluster is its lack of integration and slight relevancy as a driving force for the economy, both for the poor economic boost that they are able to generate and for their low exporting capacity.

The best-classified sectors in cluster 2 are those oriented to final demand, especially to public and private consumption (basically the nutritional products industry). Among those with a low membership value are sector 11 (Prepared animal feeds and Other food products) and 26 (Basic metals), which were erroneously classified in the silhouette plot. They rank fourth and sixth respectively on ‘economic relevancy’ and are among those with better exporting behaviour. They are, as in the previous case, outliers on ‘export base’ and ‘economic relevancy’.

In cluster 3 there are two sectors that could also belong to clusters 1 and 2. They are Manufacture of chemicals and chemical products (sector 20) and Motor vehicles (33). Both of them have a medium technology level and a large export component while at the same time they rank seventh and third respectively on ‘economic relevancy’. Sector 20 is the best one on ‘integration’ while sector 33 scores low on ‘integration’. The principal features of cluster 3 are that it has the largest exports and the highest technological levels. Its score on ‘integration’ is above average, but it is the least relevant in terms of ‘economic relevancy’. In our opinion, the key to interpreting them is that they are exporting sectors with medium and high technology. It could be concluded that they are the true key sectors from the perspective of a growth based on foreign demand and innovation. That is, from the perspective of a very desirable scenario of development for the Spanish economy (see the results in Appendix 2 for the 11 sectors in cluster 3, i.e. sectors 29, 53, 47, 54, 31, 32, 34, 30, 28, 20, and 33).

Table 3. Participation of clusters in the Spanish economy

	N	output	exports	employment	salaries	mean salary*
Cluster 1	32	64.2%	35.3%	55.9%	45.9%	95%
Cluster 2	23	22.8%	17.2%	37.6%	35.7%	102%
Cluster 3	11	13.0%	47.5%	6.5%	18.3%	110%
Total	66	100%	100%	100%	100%	100%

(*) Mean salary of the cluster as percentage of the total mean salary.

A summary of the weight in the Spanish economy of the sectors, when classified the preferred cluster, that sheds some more light on the classification is given in Table 3. **Q1** Cluster 1, with more than 64% of the output, is the one that provides more employment and salaries, and is also important it terms of export volumes. Nevertheless, there is a less than proportional relation to its production value and its salaries are 10 percentage-points below the employment. They are the sectors with lowest salaries, as can be seen in the last column of the table. Cluster 2 instead shows employment and salaries larger than its output weight, but its exports are small in comparison to the output. Its (apparent) mean productivity is lower and its mean salary is higher than for cluster 1.

Finally, cluster 3 represents only 13% of national output, but it brings more than 47% of the exports. This is 12 percentage-points more than cluster 1, whose number of sectors is three times as large. Besides, its productivity and mean salary are the largest, with a large difference with the other clusters. Its mean salary is 10% larger than the mean salary of the Spanish economy. In summary, we can assure that cluster 3 by far performs the best, both from the perspective of exports and technology and from the point of view of productivity and high salaries. They would constitute the more important sectors of the Spanish economy for a growth strategy based in innovation and competitiveness.

4. Conclusions

In spite of the time since the first studies, the procedures to identify the key sectors (or clusters of sectors) for an economy are still focusing on single aspects obtained from the interindustry matrix. Nevertheless, the consideration of a sector as an important one cannot be based solely on these aspects. Certainly, in the literature on industrial clusters there has been a progressive introduction of other variables related to indicators that are relevant to evaluate their economic potential, such as technology, employment, salaries, etc. However, in our opinion, there has not been much emphasis on this multivariate approximation. In this paper we have focused on this multivariate character, using variables that are related to the intensity and structure of the relationships among sectors, their economic strength, technological level, and recent dynamics.

As Rey and Matheis (2000) point out, no other clustering methods than hierarchical ones have been applied. In their work, a non-hierarchical clustering method is applied for the first time. Our contribution is the use, for the first time in key sector analysis to our knowledge, of an algorithm that (next to being robust) classifies sectors in a fuzzy way. It allows for flexibility when interpreting clusters and provides interesting nuances, as we have shown.

In our application to the Spanish economy, we have found three clusters. The first one can be considered as a group of local base sectors (with large economic weights and a large integration). The second cluster, with less integrated sectors, is basically oriented to internal demand. The third cluster is identified by its orientation to foreign demand and its higher technological level. Some additional considerations made us conclude that this is the group of sectors that are the most important when developing a positive growth strategy for the Spanish economy.

Besides improving and widening the database to other relevant aspects, there are some considerations that would be desirable to explore in the near future. In the literature it is mentioned that this kind of analysis usually gives a few mega clusters. This is what we have found in our application, indeed. However, we have seen that most of the sectors that were difficult to classify can be considered outliers for the univariate distributions of the indicators we used.

Anyone who has worked with an input–output table knows by their own experience that almost any measure defined over rows or columns of the interindustry matrix will produce this kind of result. Something similar may happen with the sectoral information provided from outside the table that could be used, since the behaviour of different sectors is very asymmetric. Surprisingly, the literature has never related the results for the mega clusters to the existence of outliers. Moreover, outliers not only affect the cluster analysis but, with no doubt, they also affect the results of any multivariate application that has been proposed in the literature before (including Czamanski's work based on principal components extraction and our own approach in this paper).

The presence of outliers can be a really important problem, both for the extraction of the components (because of its incidence on the correlations matrix) and for the definition of clusters. This serious problem has not been explored yet, to our knowledge, in an input–output analysis context. Its treatment is not an easy task, but it is undoubtedly a necessary and promising field of research.

Acknowledgements

This work has been partially supported by the Spanish State Secretary of Education and Universities and The Social European Fund under grant number EX2003-0563.

Notes

[1]See also the comments by de Mesnard (2002) and Dietzenbacher (2005). A good and up-to-date synthesis of key sectors is given by Robles Teigeiro and Sanjuán (2005).

[2]See Jacobs and de Man (1996) and Rosenfeld (1996, 1997) for definitions and additional information.

[3]It could be considered the first antecedent of our proposal to use fuzzy clustering to get non-mutually exclusive groups.

[4]A good summary of these works is given by Lainesse and Poussart (2005).

[5]The applications previously made to input–output analysis of these statistical techniques (Blin and Cohen, 1977; Cabrer *et al.*, 1991; Abott and Andrews, 1990) mostly focus on aggregation problems.

[6]Two-step clustering (Zhang *et al.*, 1996; Chiu *et al.*, 2001), which was recently incorporated in the statistical software SPSS, admits the possibility of using nominal variables. It uses as theoretical hypothesis normality for the numerical variables and a multinomial distribution for the nominal variables. This makes it, in our opinion, less suitable for handling the kind of variables present in input–output studies (variables associated to sectors in an input–output table are strongly asymmetric and far from similar to a normal distribution), and in general less robust than the methods that do not rest on these theoretical hypotheses.

[7]The *fanny* algorithm of Kaufman and Rousseeuw (1990), used in Dridi and Hewings (2003), has serious problems dealing with certain types of data.

[8]The elements of the mean vectors are basically calculated as the proportion of the membership of each element *j* (elevated to *m*) to cluster *i* in relation to the sum of memberships (elevated to *m*) of all elements to cluster *i*.

[9]This modification is active whenever some negative value for the distances is encountered, allowing the use of non-Euclidean matrices. This is the difference between RFCM and NERFCM. For details on this modification, called β-spread transformation, see Hathaway and Bezdek (1994, p. 431). For a detailed description of NERFCM see Hathaway and Bezdek (1994, pp. 433–434).

[10]The results from the PAM method are used to initialize the membership matrix assigning for each element *j* a membership equal to one for the cluster where it belongs to and a membership of zero for the rest of the Q2 clusters. It is thus a binary matrix with zeros and ones.

[11]Since our matrix is already a dissimilarity matrix, these are the distances used within the algorithm. In other cases the metric to compute these distances needs to be specified.

[12]This multiplier comes from the relative influence graph associated with the distribution coefficients matrix and can be interpreted in terms of elasticities (Lantner, 1974, 2001; Mougeot *et al.*, 1977; Morillas, 1983a; and de Mesnard, 2001). Its expression is: $\hat{x}^{-1}\Delta x = (I - D)^{-1}\hat{e}\hat{y}^{-1}\Delta y$. Since it is weighted by the export rate from outside the graph structure ($\hat{e}$ = final demand coefficient), it could be said that it is very close to the net multiplier mentioned by Oosterhaven (2002, 2004). The interindustry matrix has been deprived of self- Q3 consumption, to reinforce the interdependencies analysis.

[13]The cohesion concept in an economic model was first introduced by Rossier (1980) and applied for the first time to input–output analysis by Morillas (1983a).

[14]See Morillas (1983b). Extending the global indicator proposed in that work, the integration index for a given sector is defined as

$$R_i = \frac{1}{2n-1} \sum_{\substack{j=1 \\ i \neq j}}^{n} \left(\frac{1}{e_{ij}} + \frac{1}{e_{ji}} \right)$$

where e_{ij} is an element of the distances matrix.

[15]See Appendix 2 for a full table with all the memberships of each sector to each cluster.

[16]Only sectoral grading on the basis of the intensity of correlations as suggested by Czamanski's method or on the basis of the factorial loads as applied in Feser and Bergman (2000) can approach our result to some extent.

References

Abbott, T.A. and Andrews, S.H. (1990) The classification of manufacturing industries: an input based clustering of activity, *Discussion Paper CES 90-7*, U.S. Census Bureau, Washington, DC. Available at: http://www.ces.census.gov/paper.php?paper=100140&PHPSESSID=838fec06e9652892337cde662d45ed5a

Aroche-Reyes, F. (2002) Structural transformations and important coefficients in the North American economies, *Economic Systems Research*, 14, pp. 257–273.

Augustinovics, M. (1970) Methods of international and intertemporal comparison of structure, in: A.P. Carter and A. Bródy (Eds) *Contributions to Input–Output Analysis* (Amsterdam: North Holland).

Beynon, M.J., Munday, M. and Roberts, A. (2005) Ranking sectors using fuzzy output multipliers, *Economic Systems Research*, 17, pp. 237–253.

Bezdek, J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms* (New York: Plenum).

Blin, J. and Cohen, C. (1977) Technological similarity and aggregation in input–output systems, *Review of Economics and Statistics*, 59, pp. 82–91.

Broersma, L. (2001) The role of services in innovative clusters. *Discussion Paper* within the framework of the Research Programme Structural Information; Provision on Innovation in Services (SIID) for the Ministry of Economic Affairs, Directorate for General Technology Policy. Available at: http://www.ggdc.net/pub/SIID_papers/innovative_clusters.pdf.

Cabrer, B., Contreras, D. and Miravete, E.J. (1991) Aggregation in input–output tables: how to select the best cluster linkage, *Economic Systems Research*, 3, pp. 99–109.

Cai, J. and Leung, P. (2004) Linkage measures: a revisit and a suggested alternative, *Economic Systems Research*, 16, pp. 65–85.

Campbell (1975) Application of graph theoretic analysis to interindustry relationships, *Regional Science and Urban Economics*, 5, pp. 91–106. Q5

Cella, G. (1984) The input–output measurement of interindustry linkages, *Oxford Bulletin of Economics and Statistics*, 46, pp. 73–83.

Chenery, H.B. and Watanabe, T. (1958) International comparison of the structure of production, *Econometrica*, 26, pp. 487–521.

Chiu, T., Fang, D., Chen, J., Wang, Y. and Jeris, C. (2001) A robust and scalable clustering algorithm for mixed type attributes in large database environment, *Proceedings of the Seventh ACM SIGMOD International Conference on Knowledge Discovery and Data Mining*, p. 263.

Czamanski, S. (1974) *Study of Clustering of Industries* (Halifax, Canada: Institute of Public Affairs, Dalhousie University).

Dave, R.J. and Sen, S. (2002) Robust fuzzy clustering of relational data, *IEEE Transactions on Fuzzy Systems*, 10, pp. 713–727.

DeBresson, C. and Hu, X. (1999) Identifying clusters of innovative activity: a new approach and a toolbox, in: *OECD, Boosting Innovation: The Cluster Approach* (Paris: OECD).

Dietzenbacher, E. (1992) The measurement of interindustry linkages: key sectors in the Netherlands, *Economic Modelling*, 9, pp. 419–437.

Dietzenbacher, E. (1997) In vindication of the Ghosh model: a reinterpretation as a price model, *Journal of Regional Science*, 37, pp. 629–651.

Dietzenbacher, E. (2005) More on multipliers, *Journal of Regional Science*, 45, pp. 421–426.

Dietzenbacher, E. and van der Linden, J. (1997) Sectoral and spatial linkages in the EC production structure, *Journal Regional Science*, 37, pp. 235–257.

Dietzenbacher, E. and Los, B. (2002) Externalities of R&D expenditures, *Economic Systems Research*, 14, pp. 407–425.

Dietzenbacher, E., van der Linden, J. and Steenge, A.E. (1993) The regional extraction method: EC input–output comparisons, *Economic Systems Research*, 5, pp. 185–206.

Dridi, C. and Hewings, G.J.D. (2003) Sectors associations and similarities in input–output systems: an application of dual scaling and fuzzy logic to Canada and the United States, *Annals of Regional Science*, 37, pp. 629–656.

Dunn, J.C. (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, *Journal of Cybernetics*, 3, pp. 32–57.

Eurostat (1998a) *Ressources humaines en haute technologie. La mesure de l'emploi dans les secteurs manufacturiers de haute technologie; une perspective européenne*. Serie Statistiques en bref. Recherche et développement (Luxembourg: Eurostat).

Eurostat (1998b) Employment in high technology manufacturing sectors at the regional level, *Document Eurostat/A4/REDIS/103* (Luxembourg: Eurostat).

Feser, E.J. and Bergman, E.M. (2000) National industry cluster templates: a framework for applied regional cluster analysis, *Regional Studies*, 34, pp. 1–19.

Feser, E.J., Koo, J., Renski, H. and Sweeney, S.H. (2001a) Incorporating spatial analysis in applied industry cluster studies, *Working paper*, University of North Carolina, Chapel Hill, NC. Available at: http://www.planning.unc.edu/pdf/FeserPUBS/EDQ%20Revised.pdf.

Feser, E.J., Sweeney, S.H. and Renski, H.C. (2001b) A descriptive analysis of discrete U.S. industrial complexes, Prepared for the *Journal of Regional Science*. Available at: http://www.planning.unc.edu/pdf/FeserPUBS/Complexes.pdf.

Fukui, Y. (1986) A more powerful method for triangularizing input–output matrices and the similarity of production structures, *Econometrica*, 54, pp. 1425–1433.

Ghosh, A. (1958) Input–output approach in an allocation system, *Economica*, 25, pp. 58–64.

Harrigan, F. (1982) The relationship between industrial and geographical linkages, *Journal of Regional Science*, 22, pp. 19–31.

Hathaway, R.J., Bezdek, J.C. and Hu, Y. (2000) Generalized fuzzy *c*-means clustering strategies using LP norm distances, *IEEE Transactions on Fuzzy Systems*, 8, pp. 576–582.

Hathaway, R.J. and Bezdek, J.C. (1994) NERF c-means: non-Euclidean relational fuzzy clustering, *Pattern Recognition*, 27, pp. 429–437.

Hathaway, R.J., Davenport, J.W. and Bezdek, J.C. (1989) Relational duals of the *c*-means algorithms, *Pattern Recognition*, 22, pp. 205–212.

Hauknes, J. (1999) Norwegian input–output clusters and innovation patterns, in: T. Roelandt and P. den Hertog (Eds) *Boosting Innovation: The Cluster Approach* (Paris: OECD).

Hewings, G.J.D. (1974) The effect of aggregation on the empirical identification of key sectors in a regional economy: a partial evaluation of alternative techniques, *Environment and Planning*, 6, pp. 439–453.

Hirschman, A.O. (1958) *The Strategy of Economic Development* (New Haven, CT: Yale University Press).

Hoen, A. (2002) Identifying linkages with a cluster-based methodology, *Economic Systems Research*, 14, pp. 131–146.

INE (2000) *Tablas input–output de la Economía Española de 1995* (Madrid: Spanish Statistics National Institute) www.ine.es.

INE (a) Contabilidad Nacional. Base 1995. www.ine.es.

INE (b) INEbase en la metodología expuesta en Investigación y desarrollo tecnológico. Indicadores de alta tecnología. Available at: http://www.ine.es/inebase/cgi/um?M = %2Ft14%2Fp197&O = inebase&N = &L = .

Jacobs, D. and de Man, A.-P. (1996) Clusters, industrial policy and firm strategy: a menu approach, *Technology Analysis and Strategic Management*, 8, pp. 425–437.

Kaufman, L. and Rousseeuw, P.J. (1990) *Finding Groups in Data* (New York: Wiley).

Korte, B. and Oberhofer, W. (1970) Triangularizing input–output matrices and the structure of production, *European Economic Review*, 1, pp. 482–511. **Q4**

Lainesse, L. and Poussart, B. (2005) *Methode de reperage des filieres industrielles sur le territoire quebecois dasee sur les tableux d'entrees-sorties* (Montréal, Canada: Institut de la Statistique du Québec). Available at: http://www.stat.gouv.qc.ca/publications/savoir/methode_reperage_pdf.htm.

Lantner, R. (1972) L'analyse de la dominance economique, *Revue d'Economie Politique*, 2, pp. 216–283. **Q4**

Lantner, R. (1974) *Théorie de la Dominance Economique* (Paris: Dunod).

Lantner, R. (2001) Influence graph theory applied to structural analysis, in: M.L. Lahr and E. Dietzenbacher (Eds) *Input–Output Analysis: Frontiers and Extensions* (Basingstoke, UK: Palgrave Macmillan).

Latham, W.R. (1976) Needless complexity in the identification of industrial complexes, *Journal of Regional Science*, 16, pp. 45–55.

Los, B. (2001) Identification of strategic industries: a dynamic perspective, *Paper* presented at the 41st European Regional Science Association Congress, Zagreb. Available at: http://www.ub.rug.nl/eldoc/ccso/200202/200202.pdf.

MacQueen, J.B. (1967) Some methods for classification and analysis of multivariate observations, in: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297 (Berkeley, CA: University of California Press).

de Mesnard, L. (2001) On Boolean topological methods of structural analysis, in: M.L. Lahr and E. Dietzenbacher (Eds) *Input–Output Analysis: Frontiers and Extensions* (Basingstoke, UK: Palgrave Macmillan).

de Mesnard, L. (2002) Note about the concept of 'net multipliers', *Journal of Regional Science*, 42, pp. 545–548.

Miller, R.E. and Lahr, M.L. (2000) A taxonomy of extractions, in: M.L. Lahr and R.E. Miller (Eds) *Regional Science Perspectives in Economic Analysis; A Festschrift in Memory of Benjamin H. Stevens* (Amsterdam: Elsevier Publishers).

Morillas, A. (1983a) *La teoría de grafos en el análisis input–output* (Malaga: Secretariado de Publicaciones, Universidad de Málaga).

Morillas, A. (1983b) Indicadores topológicos de las características estructurales de una tabla input–output. Aplicación a la economía andaluza, *Investigaciones Económicas*, 20, pp. 103–118.

Mougeot, M., Duru, G. and Auray, J.P. (1977) *La Structure Productive Francaise* (Paris: Económica).

Nagy, G. (1968) State of the art in pattern recognition, *Proceedings IEEE*, 56, pp. 836–882.

Ó hUallacháin, B. (1984) The identification of industrial complexes, *Annals of the Association of American Geographers*, 74, pp. 420–436.

Oosterhaven, J. and Stelder, D. (2002) Net multipliers avoid exaggerating impacts: with a bi-regional illustration for the Dutch transportation sector, *Journal of Regional Science*, 42, pp. 533–543.

Oosterhaven, J. (1988) On the plausibility of the supply-driven input–output model, *Journal of Regional Science*, 28, pp. 203–217.

Oosterhaven, J. (2004) On the definition of key sectors and the stability of net versus gross multipliers, *SOM Research Report 04C01*, Research Institute SOM, University of Groningen, The Netherlands. Available at: http://som.rug.nl.

Oosterhaven, J., Eding, G.J. and Stelder, T.M. (2000) Clusters, linkages and regional spillovers: methodology and policy implications for the two Dutch mainports and the rural North, *SOM Research Report 99C69*, Research Institute SOM, University of Groningen, The Netherlands. Available at: http://som.rug.nl.

Papaconstantinou, G., Sakurai, N. and Wyckoff, A. (1996) Embodied technology diffusion: an empirical analysis for 10 OECD countries, *STI Working Papers* 1996/1 (Paris: OECD).

Peeters, L., Tiri, M. and Berwert, A. (2001) Techno-economic clusters in Flanders and Switzerland: an input–output analysis, Working paper CEST 2001/9, Center for Science and Technology Studies, Bern.

Perroux, F. (1949) Economic spaces: theory and application, *Quarterly Journal of Economics*, 64, pp. 89–104.

Perroux, F. (1973) *Pouvoir et Economie* (Paris: Bordas).

Porter, M.E. (1990) The competitive advantage of nations, *Harvard Business Review*, 68, pp. 77–93.

Rasmussen, P.N. (1956) *Studies in Intersectoral Relations* (Amsterdam: North Holland).

Rey, S. and Mattheis, D. (2000) Identifying Regional Industrial Clusters in California. Volumes I-III, *Reports* prepared for the California Employment Development Department, San Diego State University. Available at: http://irsr.sdsu.edu/%7Eserge/node7.html.

Robles Teigeiro, L. and Sanjuán, J. (2005) Sectores claves. Coeficientes grandes y coeficientes importantes, *Paper* presented at I Jornadas de Análisis Input–output, Oviedo.

Roelandt, T.J.A., den Hertog, P., van Sinderen, J. and van den Hove, N. (1999) Cluster analysis and cluster policy in the Netherlands, in: OECD, *Boosting Innovation: The Cluster Approach* (Paris: OECD).

Roepke, H., Adams, D. and Wiseman, R. (1974) A new approach to the identification of industrial complexes using input–output data, *Journal of Regional Science*, 14, pp. 15–29.

Rosenfeld, S.A. (1996) *Overachievers, Business Clusters that Work: Prospects for Regional Development* (Chapel Hill, NC: Regional Technology Strategies).

Rosenfeld, S.A. (1997) Bringing business clusters into the mainstream of economic development, *European Planning Studies*, 5, pp. 3–23.

Rossier, E. (1980) *Economie Structural* (Paris: Económica).

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational Applied Mathematics*, 20, pp. 53–65.

Rousseeuw, P.J. (1995) Discussion: fuzzy clustering at the intersection, *Technometrics*, 37, pp. 283–285.

Ruspini, E. (1969) A new approach to clustering, *Information and Control*, 15, pp. 22–32.

Sakurai, N., Papaconstantinou, G. and Ioannidis, E. (1997) Impact of R&D and technology diffusion on productivity growth: empirical evidence for 10 OECD countries, *Economic Systems Research*, 9, pp. 81–110.

Scherer, F.M. (1982) Interindustry technology flows and productivity growth, *Review of Economics and Statistics*, 64, pp. 627–634.

Scherer, F.M. (2003) Technology flows matrix estimation revisited, *Economic Systems Research*, 15, pp. 327–358.

Schmookler, J. (1966) *Invention and Economic Growth* (Cambridge, MA: Harvard University Press).

Simon, H.A. (1952) On the definition of the causal relation, *Journal of Philosophy*, 49, pp. 517–528.

Simpson, D. and Tsukui, J. (1965) The fundamental structure of input–output tables: an international approach, *Review of Economics and Statistics*, 47, pp. 434–446.

Sonis, M., Guilhoto, J.J.M., Hewings, G.J.D. and Martins, E.B. (1995) Linkages, key sectors, and structural change: some new perspective, *Developing Economies*, 33, pp. 233–270.

Strassert, G. (1968) Zur bestimmung strategischer sektoren mit Hilfe von Input–Output Modellen, *Jahrbücher für Nationalökonomie und Statistik*, 182, pp. 211–215.

Verspagen, B. (1999) The role of large multinationals in the Dutch technology infrastructure: a patent citation analysis, *Mimeo*, University of Maastricht/MERIT, The Netherlands.

Wold, H. (1954) Causality and Econometrics, *Econometrica*, 22, pp. 162–177.

Zadeh, L.A. (1965) Fuzzy sets, *Information and Control*, 8, pp. 338–353.

Zadeh, L.A. (1977) Fuzzy sets and their application to pattern classification and clustering analysis, in: J. van Ryzin (Ed.) *Classification and Clustering* (New York: Academic Press).

Zhang, T., Ramakrishnon, R. and Livny, M. (1996) BIRCH: an efficient data clustering method for very large databases, in: *Proceedings of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, pp. 103–114.

Table A1. High and medium-high technology sectors listed by the Spanish Institute of Statistics (INE)

CNAE-93 (NACE Rev. 1)	Sectors
	High Technology Manufacturing Sectors
244	Manufacture of pharmaceuticals, medicinal chemicals and botanical products
30	Manufacture of office machinery and computers
321	Electronic components
32-32.1	Manufacture of radio, television and communication equipment and apparatus excluding electronic components
33	Manufacture of medical, precision and optical instruments, watches and clocks
35.3	Manufacture of aircraft and spacecraft
	Medium-High Technology Manufacturing Sectors
24-24.4	Manufacture of chemicals and chemical products excluding manufacture of pharmaceuticals, medicinal chemicals and botanical products
29	Manufacture of machinery and equipment n.e.c.
31	Manufacture of electrical machinery and apparatus n.e.c.
34	Manufacture of motor vehicles, trailers and semi-trailers
35-35.3	Manufacture of other transport equipment excluding manufacture of aircraft and spacecraft
	High Technology Service Sectors
64	Post and telecommunications
72	Computer and related activities
73	Research and development

Source: INE (b, p. 5).

Table A2. Membership values to each cluster

Sector	Name	Cluster 1	Cluster 2	Cluster 3
1	Hard coal, lignite and peat	0.4037	0.4781	0.1182
2	Crude petroleum and natural gas	0.2655	0.6019	0.1325
3	Metal ores	0.2965	0.5274	0.1760
4	Other mining and quarrying	0.4462	0.4242	0.1296
5	Coke, refined petroleum products and nuclear fuel	0.5292	0.3300	0.1409
6	Electricity supply	0.5287	0.3053	0.1661
7	Gas, steam and hot water supply	0.5590	0.3263	0.1147
8	Collection, purification and distribution of water	0.5898	0.2872	0.1230
9	Meat industry	0.2985	0.5829	0.1186
10	Dairy products and ice cream	0.2654	0.6260	0.1085
11	Prepared animal feeds and Other food products	0.3276	0.4727	0.1997
12	Beverages	0.2937	0.5825	0.1238
13	Tobacco products	0.2874	0.5419	0.1707
14	Textiles	0.5861	0.2690	0.1448
15	Wearing apparel; Furs	0.6469	0.2398	0.1133

(*Table continued*)

Table A2. Continued

Sector	Name	Cluster 1	Cluster 2	Cluster 3
16	Leather and leather products	0.5244	0.3170	0.1586
17	Wood and cork	0.6588	0.2419	0.0993
18	Paper products	0.6313	0.2450	0.1236
19	Printed matter and recorded media	0.6462	0.2463	0.1075
20	Manufacture of chemicals and chemical products	0.2942	0.2458	0.4600
21	Rubber and plastic products	0.5718	0.2671	0.1611
22	Cement, lime and plaster	0.4090	0.4743	0.1168
23	Glass	0.4619	0.3897	0.1485
24	Ceramic articles	0.3723	0.4614	0.1663
25	Other non metallic mineral products	0.4322	0.4522	0.1157
26	Basic metals	0.3623	0.4093	0.2283
27	Fabricated metal products	0.5724	0.2816	0.1459
28	Machinery and equipment n.e.c.	0.2524	0.2146	0.5330
29	Office machinery and computers	0.1648	0.1568	0.6784
30	Electrical machinery and apparatus n.e.c.	0.2247	0.2039	0.5714
31	Radio, television and communication equipment and apparatus	0.1822	0.1963	0.6215
32	Medical, precision and optical instruments; watches and clocks	0.2075	0.1800	0.6125
33	Motor vehicles	0.2867	0.3233	0.3900
34	Other transport equipment	0.2133	0.2001	0.5866
35	Furniture; other manufactured goods n.e.c.	0.6374	0.2423	0.1203
36	Recycling	0.2630	0.6102	0.1268
37	Construction	0.3786	0.3739	0.2475
38	Sale, maintenance and repair of motor vehicles	0.5871	0.3132	0.0997
39	Wholesale trade and commission trade	0.5003	0.3150	0.1846
40	Retail trade	0.5019	0.3152	0.1829
41	Hotel and restaurant services	0.4357	0.3361	0.2282
42	Railway transportation services	0.4475	0.4366	0.1158
43	Other land transportation services	0.6225	0.2648	0.1126
44	Water transport services	0.4445	0.3021	0.2534
45	Air transport services	0.3708	0.3993	0.2299
46	Supporting and auxiliary transport services	0.6356	0.2465	0.1179
47	Telecommunication services	0.1865	0.1769	0.6366
48	Financial intermediation	0.4744	0.4175	0.1081
49	Insurance and pension funding	0.3378	0.5608	0.1015
50	Activities auxiliary to financial intermediation	0.2992	0.5879	0.1129
51	Real estate services	0.5215	0.3301	0.1484
52	Renting services	0.5460	0.3407	0.1133
53	Computer and related services	0.1669	0.1555	0.6776
54	Research and development services	0.1867	0.1736	0.6398
55	Other business services	0.5158	0.3260	0.1583
56	Market education services	0.5760	0.2837	0.1403
57	Market health and social work services	0.5077	0.3112	0.1811
58	Market sewage and refuse disposal services, sanitation and similar services	0.4993	0.3769	0.1237

(Table continued)

Table A2. Continued

Sector	Name	Cluster 1	Cluster 2	Cluster 3
59	Market recreational, cultural and sporting services	0.6299	0.2622	0.1079
60	Other services	0.2846	0.6068	0.1086
61	Public administration and defence; compulsory social security	0.3106	0.4965	0.1928
62	Non-market education services	0.2913	0.5441	0.1645
63	Non-market health and social work services	0.3011	0.5220	0.1768
64	Non-market sewage and refuse disposal services, sanitation and similar services	0.2907	0.5392	0.1700
65	Membership organization services n.e.c.	0.2868	0.5437	0.1695
66	Non-market recreational, cultural and sporting services	0.2841	0.5536	0.1623