ESTUDIOS DE ECONOMÍA APLICADA VOL. 23 - 2, 2005. PÁGS. 451-463

Creating Industry Classifications by Statistical Cluster Analysis

PENEDER, MICHAEL

Austrian Institute of Economic Research (WIFO), Vienna. E-mail: michael.peneder@wifo.ac.at

ABSTRACT

Much emphasis on the sectoral perspective in economics originates in the observation of the diverse *and* contingent nature of competitive behaviour, where a firm's performance depends on the capability to match its organisation and strategy to the technological, social and economic restrictions imposed by the business environment. For this reason, analytically based industry classifications are frequently applied in empirical studies on competitive performance, technological development, international trade, and industrial economics. This paper shows how the use of statistical cluster analysis and the generation of empirically based sectoral classifications can provide new and valuable tools for research into the sectoral contingency of industrial development.

Keywords: sectoral classifications, taxonomy, statistical cluster analysis, industrial development.

Generación de clasificaciones industriales mediante técnicas estadísticas de análisis cluster

RESUMEN

El estudio de la diversidad del comportamiento competitivo en las distintas ramas de actividad se ha convertido en uno de los principales objetivos del análisis sectorial de forma tal que el mayor o menor nivel de éxito de una determinada empresa esta condicionado por su capacidad para adaptar su organización y estrategias competitivas a las restricciones especificas de su entorno, tanto tecnológico, como socio-económico. Por este motivo, las clasificaciones sectoriales obtenidas de forma analítica son frecuentemente utilizadas en los estudios empíricos de competitividad, desarrollo technologic, comercio internacional o economía industrial.

En este artículo se muestra cómo el uso del análisis cluster para generar calsificaciones sectoriales puede convertirse en una valiosa herramienta para la investigación aplicada al desarrollo sectorial.

Palabras Clave: Clasificación sectorial, análisis cluster, desarrollo industrial.

Clasificación JEL: B41, C81, L10, O12.

Artículo recibido en marzo de 2005 y aprobado en julio de 2005.

Artículo disponible en versión ⓔlectrónica en la página www.revista-eea.net, ref.: ⓔ-23212.

ISSN 1697-5731 (online) – ISSN 1133-3197 (print)

1. INTRODUCTION

More than any other economic discipline, industrial economics stresses the diverse *and* contingent nature of competitive behaviour. Within the confines of its paradigms, competitive performance depends on the capability to match a firm's organisation and strategy to the technological, social and economic restrictions imposed by its business environment. The intention of this paper is to demonstrate how the generation of empirically based sectoral classifications can provide new and valuable tools for research into the sectoral contingency of industrial development.

To begin with, we can distinguish two major reasons for the creation and use of analytically based industry classifications: First, sectoral taxonomies facilitate investigations into the impact of specific characteristics of the market environment on economic activity. Substituting structural knowledge for exhaustive information about single attributes, the intractable diversity of real-life phenomena is condensed into a smaller number of salient types. Classifications thus direct our attention towards a few characteristic dimensions, according to which relative similarities or differences can be identified. They allow us to take account of heterogeneity, but simultaneously force us to be selective.

Second, from a purely practical perspective, the taxonomic approach is particulary useful when referring to data that are not easily available in a comparable format across countries or firms. The reason is that it builds upon data from those entities, which offer the best coverage of specific attributes and then produces typical profiles of the relevant variables. The resulting classification can then be applied to other data of economic activity, which are available on a broader comparable basis (for example, value added, employment, or foreign trade data).

The process of classification is generally defined as the ordering of cases in terms of their similarity. According to Bailey (1994), classifications themselves can be distinguished by (among others) the following characteristics: They can be labelled either as typologies or taxonomies; monothetic or polythetic; synchronic or diachronic. The term *typology* refers specifically to a conceptual classification, the cells of which represent type concepts rather than empirical cases. Conversely, the term *taxonomy* refers to a classification of empirical entities based upon quantitative analysis. In this sense, one can also distinguish *monothetic* classes, in which all the cases included in a certain category are identical with respect to every relevant dimension. No exceptions or further differentiations are allowed. Such a neat and (idealised) categorisation is typical of qualitative categorisations, whereas empirical classifications generally come up with *polythetic* classes. Here, the cases are not identical with respect to all variables, but rather are grouped according to the generally strongest similarity. In other words, the existence of large individual variations within the given categories of a classification is taken for granted. Finally, classifications are called *synchronic* (or phenetic), if they refer to the characteristics of an observation at a certain point in time. Conversely, classifications are called *diachronic* (or phyletic), if they are based upon characteristic patterns of change or evolution. Moreover, we generally expect that our classifications are *exhaustive* and *mutually exclusive*, thereby demanding the existence of one (but only one) appropriate class for each observation.

Two general approaches to the quantitative identification of individual observations into classes can be distinguished. A 'cut-off' procedure by which a certain discriminatory edge is defined exogenously by the researcher is the more frequently applied method. The sole advantage of this approach lies in its simplicity. In choosing not to use more powerful statistical tools, the underlying structure within the data is more or less presumed, rather than explored. Although this approach can be defended as long as the classifications are built upon one or two variables only, it is generally inept for the categorisation of a data profile of larger dimensions. Statistical cluster analysis is the obvious alternative. It is specifically designed for classifying observations on behalf of their relative similarities with respect to a multidimensional array of variables. It is a powerful tool for the creation of sectoral taxonomies and thus deserves a more detailed discussion in the next section.

Analytically based industry classifications are frequently applied in empirical studies on competitive performance, technological development, international trade, and industrial economics. Peneder (2003a) provides a critical survey of major classifications applied within these various fields, while Peneder (2005) additionally highlights the intellectual roots within the Neo-Schumpeterian research tradition. In contrast, this paper primarily intends to stimulate the methodological discussion. The next section on statistical cluster analysis starts with general concepts and definitions but then focuses on a number of critical choices that have to be made during that process. In the final section, I will present an illustrative example which additionally tries to give some idea about how the results from statistical cluster analyses might be validated in terms of their economic meaning.

2. STATISTICAL CLUSTER ANALYSIS

Definitions and aim

Statistical cluster analysis is defined as "the art of finding groups in data" (Kaufmann and Rousseuw, 1990) such that the degree of "natural association" (Anderberg, 1973) is (i) high among members within the same class (*internal cohesion*) and (ii) low between members of different categories (*external isolation*). In practice, internal cohesion and external separation are not definite requirements, but rather general objectives. Their fulfillment is a matter of degree and depends on the nature of the data as well as the clustering techniques applied.

Cluster analysis offers a sophisticated statistical tool for the exploration and classification of multivariate data, but it is important to acknowledge that it remains a heuristic method, which requires the researcher to make a number of choices that critically affect the final outcomes.

Once the variables are chosen, the clustering procedure starts with a given data matrix of i = 1, ..., *n* observations for which characteristic attributes *x* are reported for j = 1, ..., *p* variables. The initial data set of the dimension *n* x *p* is then transformed into a symmetric (dis)similarity matrix of dimensions *n* x *n* observations with d_{ih} being the coefficients of (dis)similarity for observations x_i and x_h.

$$D_{n,n} = \begin{bmatrix} 0 & \dots & & & & 0 \\ d_{21} & 0 & \dots & & & \\ d_{31} & d_{32} & 0 & \dots & & \vdots \\ \vdots & & \vdots & & & \\ & \dots & d_{ih} & \dots & & \\ & & \vdots & & & \\ d_{n1} & d_{n2} & \dots & & d_{n(n-1)} & 0 \end{bmatrix} \qquad \textbf{(1)}$$

For any observations x_i , x_h and x_g with i, h, and g = 1, ..., n, located within measurement space **E**, the desired formal properties of the (dis)similarity matrix $\mathbf{D}_{nn}$ are defined as follows (Anderberg, 1973, p. 99):

1. $d_{ih} = 0$ if and only if $x_i = x_h$, i.e. for all observations the distance from itself is zero and any two observations with zero distance are identical;
2. $d_{ih} >= 0$, i.e. all distances are non-negative;
3. $d_{ih} = d_{hi}$, i.e. all distances are symmetric; and finally
4. $d_{ih} <= d_{ig} + d_{hg}$, known as the triangle inequality, which states that going directly from x_i to x_h is shorter than making a detour over object x_g.

The combination of the first and second properties assures that $\mathbf{D}_{nn}$ is fully specified by its values in the lower triangle. The fourth property establishes that E is an Euclidean space and that we can correctly interpret distances by applying elementary geometry. Any dissimilarity function that fulfills the above four conditions is said to be a *metric*.

In this spirit, the *Euclidean distance* e_{ih} appears to be the most natural measure of (dis)similarity, due to its direct application of the Pythagorean theorem:

$$euc_{ih} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{hj})^2} \qquad 0 \le euc_{ih} < \infty \qquad \textbf{(2)}$$

Operating with the squared differences, the Euclidean measure will, for example, rank two observations with a difference of 1 unit in the first variable and 3 units in the second variable as farther apart than two observations with a difference of 2 units in both variables. In other words, it is sensitive to outliers. Alternatively, the closely related Manhattan or *city block distance* prescribes equal importance to any unit of dissimilarity, because it simply calculates the sum of the absolute lengths of the other two sides of the triangle:

$$cityb_{ih} = \sum_{j=1}^{p} \left| x_{ij} - x_{hj} \right| \qquad 0 \le cityb_{ih} < \infty \qquad \textbf{(3)}$$

Kaufmann and Rousseeuw (1990, p. 12) use the image of a city in which the streets run vertically and horizontally to explain the peculiar name. The Euclidean measure corresponds to the shortest geometric distance ‘a bird could fly’ straight from point x_i to point x_h, whereas the use of the Manhattan measure is consistent with the distance that ‘people have to walk’ around the city blocks. Both measures in (2) and (3) fulfill the requirements of a metric.[1]

When we are interested in the ‘shape‘ of objects rather than in the absolute size of differences, alternative measures can be more helpful. The following two measures of similarity, called *angular separation* in (4) and the *correlation coefficient* in (5), are most frequently used:

$$ang_{ih} = \frac{\sum_{j=1}^{p} x_{ij} x_{hj}}{\sqrt{\sum_{j=1}^{p} x_{ij}^2 \sum_{j=1}^{p} x_{hj}^2}} \qquad -1{,}0 \le ang_{ih} \le 1{,}0 \qquad \textbf{(4)}$$

$$corr_{ih} = \frac{\sum_{j=1}^{p} x_{ij} x_{hj} - (1/p)(\sum_{j=1}^{p} x_{ij} \sum_{j=1}^{p} x_{hj})}{\sqrt{\left\{\left[\sum_{j=1}^{p} x_{ij}{}^2 - (1/p)(\sum_{j=1}^{p} x_{ij})^2\right]\left[\sum_{j=1}^{p} x_{hj}{}^2 - (1/p)(\sum_{j=1}^{p} x_{hj})^2\right]\right\}}} \qquad -1{,}0 \le corr_{ih} \le 1{,}0 \ \textbf{(5)}$$

Both angular separation and the correlation coefficient measure the cosine of the angle between two vectors. The essential difference between the two is that the former is based on deviations from the origin, whereas the latter operates with deviations

[1] They are special cases of a general dissimilarity function called the *Minkowski metric*.

from the mean of the variables of an observation. As a consequence, the correlation coefficient is unaffected by mere size displacements (i.e. the uniform addition of a constant to each element). The correlation coefficient is therefore less discriminating than the angular separation measure.[2]

In addition to the above examples, the literature provides a variety of other (dis)similarity functions that are applied in statistical cluster analysis. For extensive surveys see, for example, Romesberg (1984) and Gordon (1999). The following Section presents a simple numerical example plus geometric visualisation that demonstrates, how the choice of various measures affects the values of the final (dis)similarity matrix $\mathbf{D}_{nn}$. The example is taken from Peneder (2004).

A simple numerical example can demonstrate the differences between the four (dis)similarity functions. Table 1 provides the values for five hypothetical objects *I* through *V* for the three variables *A*, *B* and *C*. Table 2 reports the calculated (dis)similarity for four different measures. Figure 1 offers an additional geometric visualization of the two-dimensional case, in which we only consider the variables *A* and *B*. Objects are characterized in brackets according to their respective co-ordinates. The straight line between two cases corresponds to the Euclidean distance, whereas the city block distance equals the length of the connecting horizontal and vertical lines. The two rays that go from the origin to the respective cases determine the angular separation measure.

Table 1: A numerical example

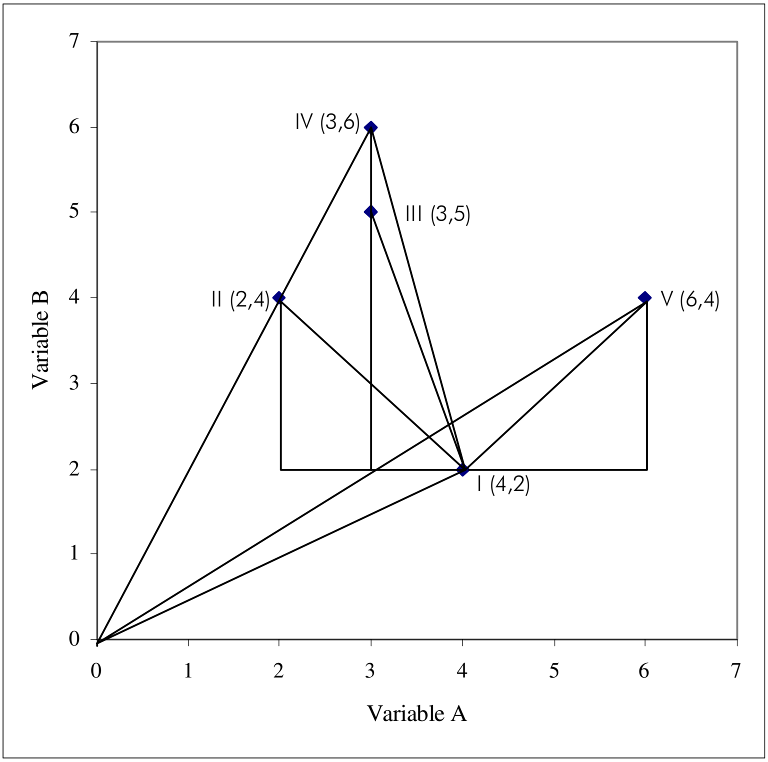
Objects	*Numerical values of variable*		
	A	*B*	*C*
I	*4.0*	*2.0*	*1.0*
II	2.0	4.0	3.0
III	3.0	5.0	3.0
IV	3.0	6.0	4.5
V	6.0	4.0	3.0

Table 2: Comparing measures of (dis)similarity of the numerical example

Measure	*Comparison of (dis)similarity between object I and ..*			
	II	*III*	*IV*	*V*
Euclidean	3.46	3.74	5.41	3.46
City block	6.00	6.00	8.50	6.00
Angular s	0.77	0.83	0.77	0.98
Correlation	-0.65	-0.19	-0.65	1.00

[2] Since correlation-type measures can take negative values, they do not strictly fulfill the above requirements of a metric. Anderberg (1973, p 113f) discusses the "limited metric character" of the correlation coefficient. However, these measures can be transformed to take values between 0 and 1 by defining angih* = (1+angih)/2 and corrih* = (1+corrih)/2) (see Gordon, 1999, p. 21).

Figure 1: A geometric illustration of differences in (dis)similarity measures



Note: The straight lines between two objects determine the Euclidean distance, the connected horizontal and vertical lines the city block distance and the two rays from the origin the angular separation measure. Source*: Peneder (2004).*

The first interesting observation is that the city block measure treats objects *II* and *III* as equally distant from *I*, whereas the Euclidean measure regards the latter as more distant. The simple reason is that we move from a quadratic to a rectangular shape. In contrast, both the angular separation and the correlation coefficient say that relative to *I*, *III* is more similar than *II*. Secondly, for both the Euclidean and the city block distance, case *IV* is more dissimilar to *I* than is case *III* or case *II*. However, when we apply angular separation or the correlation coefficient, *IV* is just as similar to *I* as is *II*, since both locate on the same ray from the origin. Finally, case *V* is an extreme example of the differences between size- and shape-oriented measures. Whereas *I* and *V* are clearly distant in the sense of Euclidean or city block measures (mirroring the distance between *I* an *II*), the two cases are highly similar in the measure of angular separation and even identical, if we apply the correlation coefficient. The reason is that for case *V,* we only add a constant of two units to each of the variables.

Since the correlation coefficient is insensitive to mere size displacements, both cases are treated as identical.

The next crucial step concerns the choice of how to group objects into separate categories, i.e. we must choose what clustering algorithm to use. Again a variety of approaches is possible.[3] Among the clustering algorithms that are most widely used, we must distinguish between two general approaches. The first is the *partitioning* method, which breaks objects into a distinct number of non-overlapping groups. The most common of them, which is also applied here, is the so called *k-means* technique. The second approach is the *hierarchical cluster analysis*, which is either divisive or agglomerative, i.e. dividing or combining hierarchically related objects into clusters.

For the *k-means* method, the set of observations is divided by a pre-defined number of clusters *k*. For example, *k* nearly equal-sized segments can be formed as an initial partition. Cluster centers are computed for each group, which are the vectors of the means of the corresponding values for each variable. The objects are then assigned to the group with the nearest cluster center. After this, the mean of the observations are recomputed and the process is repeated until convergence is reached. This is the case, when no observation moves between groups and all have remained in the same cluster of the previous iteration.

In contrast to the *k*-means method, *hierarchical* cluster analysis enables us to determine the boundaries between clusters at different levels of (dis)similarity. Preserving a higher degree of complexity in the output produced, hierarchical techniques require a heuristic interpretation of the surfacing patterns. Dendrograms (or ‘cluster trees‘) support this by means of graphical representation. As with *k*-means cluster analysis, any of the above measures of distance can be applied. When groups with more than one object merge, various methods differ in the way they determine what the (dis)similarity between groups precisely is. The most popular and intuitively appealing choice is the *average linkage* method, whereby the average (dis)similarity between all the observations is compared for any pair of groups. Alternatively, the *complete linkage* method compares the (dis)similarity between the observations which are farthest apart, whereas the *single linkage* method takes the (dis)similarity of the nearest neighbors in any pair of groups into account.

The choice between the different linkage methods directly relates to the objectives of internal cohesion and the external isolation of clusters (mentioned at the beginning of this section). Single linkage aim only for *external isolation*, implying that any observation is more similar to some other object within the same cluster than to any other objects outside. Due to this property, single linkage methods frequently fail to

[3] Anderberg (1973, p. 23) remarked, that "one of the most striking things about the many methods in the literature is the high degree of redundancy when applied to a set of data. The ideal would be to have a small stable of algorithms minimally duplicative among themselves but collectively representative of all the general types of classifications that might be produced by all other algorithms put together."

reveal much structure within the data. The reason is that observations tend to join one common and expanding cluster, which leads to undesirable ‘chaining‘ effects. Conversely, the complete linkage method aims at *internal cohesion*. This leads to compact classes, which, however, need not be externally isolated. The average linkage method avoids both extremes and seeks a compromise between the aims of internal cohesion and external isolation.

To conclude, this methodological section has demonstrated the multitude of potentially very influential choices researchers have to make during the clustering process. In order to be credible, any classification should therefore be backed by a comprehensive documentation of the critical choices and a detailed explanation of how the graphical representations were interpreted. The next section gives a specific example of how that might be done in practice.

3. AN ILLUSTRATIVE EXAMPLE: HUMAN RESOURCES IN THE ‘NEW ECONOMY’[4]

The rapid advance of new information technologies (IT) is a major cause of qualitative transformations in modern production systems. IT personnel is the fundamental category of human capital formation in the process of dissemination and adoption of computers and related equipment. It drives the progress in computer related technologies of the IT producing sectors and enables the actual realisation of productivity gains among IT user industries. The much quoted ‘new economy’ or ‘digital revolution’ also leaves some pronounced imprints on the overall formation of human capital, which we may trace in at least two dimensions. First, the structural change towards the ‘new economy’ favours the growth of specific computer related *occupations*. Second, it tends to raise the demand for higher levels of workforce *education*. Together, occupational and educational attributes characterise the IT labour intensity of a firm, an industry, or the aggregate economy.

In the present analysis we are interested in occupational and educational characteristics of workforce composition, i.e. the share and educational level of IT labour. Data sources are the UK Labour Force Survey and the US Current Population Survey, with annual data on workforce composition available for both employment and wages. The data cover 39 sectors in the USA and the United Kingdom from 1979 until 2000. The annual data are pooled by calculating three (four) year averages from 1979 onwards. The workforce composition is represented by (i) employment and wage shares for IT labour in the total workforce and (ii) the share of personnel with higher education (university degrees) among IT labour.

[4] This section briefly summarises the work documented in Peneder (2003b).

Data for the different time periods enter as independent observations in the first part of the analysis, so that the initial data matrix comprises four variables and 546 observations (i.e. two countries times seven periods times 39 sectors). In order to give equal weights to all variables and eliminate the impact of specific time and country effects on the clustering process, the initial data matrix is standardised with respect to the total variation across industries for each country and year.

The current investigation proceeds through an elaborate three-stage clustering process, which combine *k*-means in the first and agglomerative hierarchical methods in the second and third steps of the analysis. The *k*-means method produces a first partition, which reduces the large initial data set for better use in the second step of hierarchical clustering. The second stage results in an interim classification. The third stage relies again on hierarchical clustering but uses the specific *time profile* of cluster identification in the interim classification as new variables. The sectoral taxonomy presented here is therefore a rare instance of a 'diachronic' classification.

The purpose of the first step is to condense information and segregate outlying observations into separate clusters without imposing a strong structure on the overall outcome yet. The cluster centres of the first partition are then entered as individual observations in the second step of hierarchical analysis, which is based on the average linkage method and the City block measure of distance. The other algorithms discussed in the previous methodological section were used to assess its robustness.[5] Overall, the patterns were reasonably robust and produced an interim classification of six separate categories, which represent a descending order of IT-labour intensity.

In the third and final stage of the cluster analysis I transformed the data into a matrix of 39 industries as observations and the cluster identification for the respective time periods and countries as variables. Focusing only on the City block measure and assuming equal distances between classes, both average and complete linkage again produced almost identical results, whereas the single linkage method failed due to 'chaining'. Inspection of the data and the graphical representation (not displayed here, see Peneder 2003b) showed that the two outliers of 'computer related services' and 'computers and office machinery' represent distinct categories within a genuinely longtailed distribution. I therefore split the general category of IT producers into manufacturing (ITP/manuf.) and services (ITP/serv.). Conversely, the numerous IT user industries were seperated into the category of 'dynamic IT user with a high and growing IT-labour intensity' (ITU/high) and 'other IT user' industries (ITU/other). Among the dynamic IT user sectors we find, for example, the chemicals industry, motor vehicles and air transport, as well as telecommunication, financial intermediation, and the business services.

[5] Essentially identical cluster trees appear when Euclidean distances replace the City block measure, or the complete linkage method is applied instead of average linkages. Despite some differences, both angular separation and the correlation coefficient preserved a similar order of associations. The single linkage method suffered from chaining effects.

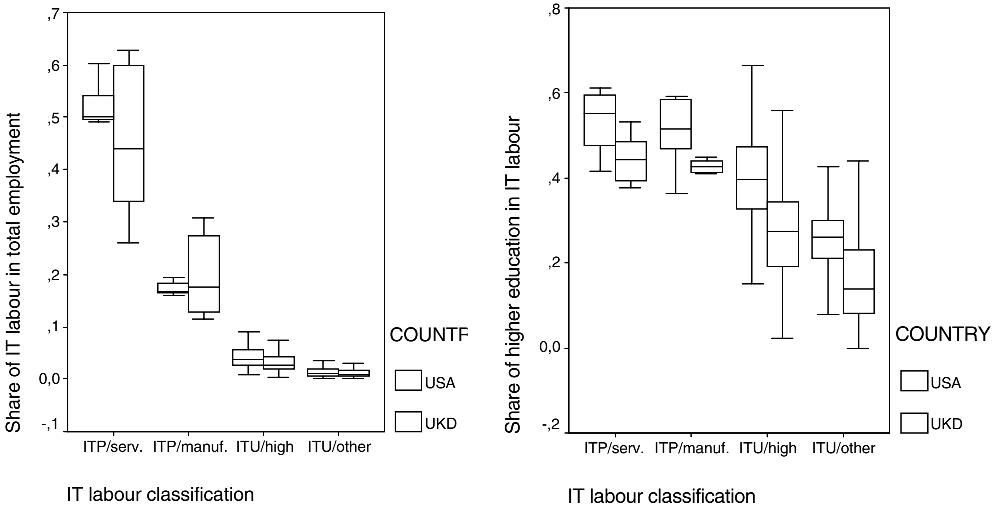
A sensible industry classification must also present interpretable structures. The boxplot charts in Figure 2 and 3 are particularly useful for that purpose, since they simultaneously display information about the shape and dispersion of the chosen attributes. The box itself comprises the middle 50 percent of observations. The line within the box is the median. The lower end of the box signifies the first quartile, while the upper end of the box corresponds to the third quartile. In addition, the lowest and the highest lines outside the box indicate the minimum and maximum values. The observations have been split into the four different classes and are additionally separated by country.

The boxplots allow several important observations, which help to validate and interpret the cluster outcome: First, with respect to the proper identification of the seperate categories, the extremely skewed distribution demonstrates why it is absolutely necessary to distinguish between IT producer and IT user industries (Van Ark, 2001). The first two groups consist only of the two outlying cases of computers and computer services. Since both are IT producing sectors, they naturally exhibit a very high IT-labour intensity. The third and the fourth category represent IT-user industries. Both differ with respect to the share of persons with higher education among its IT workforce, which is much larger in the third than in the fourth group.

Second, concerning the robustness of the cluster solution with respect to differences between countries, the boxplots indicate that these hardly affect the distribution and relative order of industry groups in the chosen attribute. Similarly, with respect to the time dimension, all the four industry classes experienced a rather uniform increase in the share of higher education among its IT labour and three of the four classes exhibit a similar and progressive pattern for the overall share of IT personnel in total employment over time.

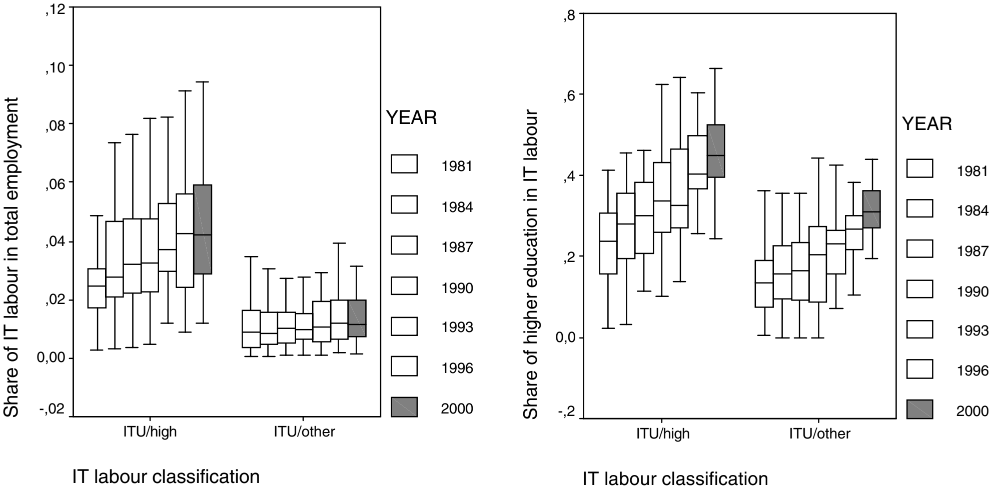
Third, with respect to the particular time profile, the class of ‘other IT user industries’ comprises a remarkable group of sectors, for which the occupational composition of the workforce has been almost unaffected by the rapid advance of information technologies during the 1980s and 1990s (Figure 3). The explicit mapping of the time dimension in the third stage of the cluster analysis appears to have produced a remarkable observation, which to my knowledge has not yet received any notable emphasis in the literature on the ‘new economy’ phenomenon. In sharp contrasts to popular believes about a more or less uniform dissemination of new information technologies in all sectors, these industries not only show no signs of catching-up from low initial levels but fall further behind in terms of IT personnel. Consequently, the passage of time even appears to further reinforce the separation between the industry types instead of blurring or reversing their order.

Figure 2: Boxplots of workforce composition by country



Source: Peneder (2003b)

Figure 3: Boxplots of workforce composition by 3-year averages (up to indicated years)*



**) 4-year average for the final period.*
Source: Peneder (2003b)

4. CONCLUSIONS

Competitive performance depends on the capability to match a firm's organisation and strategy to the technological, social and economic restrictions imposed by the business environment. Without denying the heterogeneity among individual actors and firms, sectoral taxonomies stress specific characteristics of the competitive environment and their impact on economic activity. Substituting structural knowledge for exhaustive information about single attributes, the intractable diversity of real-life phenomena is thus condensed into a smaller number of salient types. This paper discussed the method of classification, putting especial emphasis on a number of critical choices that have to be made for that purpose. Finally, the paper presented an illustrative example that demonstrates the use and validation of statistical cluster analysis in practice.

REFERENCES

ANDERBERG, M.R., (1973), *Cluster Analysis for Applications*, Academic Press, New York.

BAILEY, K.D. (1994), *Typologies and Taxonomies. An Introduction to Classification Techniques*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-102, Thousand Oaks, CA: Sage.

GORDON, A.D., (1999), *Classification*, 2nd ed., Chapman & Hall, Boca Raton.

KAUFMANN, L., ROUSSEEUW, P.J., (1990), *Finding Groups in Data. An Introduction to Cluster Analysis*, Wiley, New York.

PENEDER, M. (2005), Sectoral Taxonomies: Identifying Competitive Regimes by Statistical Cluster Analysis, forthcoming in Hanusch H. and Pyka A., *The Elgar Companion to Neo-Schumpeterian Economics*, Edward Elgar, Cheltenham UK.

PENEDER, M. (2004), A Sectoral Taxonomy of Educational Intensity. Statistical Cluster Analysis and Validation, Working Paper: EPKE-WP-26.

PENEDER, M. (2003a), Industry Classifications. Aim, Scope and Techniques, *Journal of Industry, Competition and Trade*, 3 (1-2), pp. 109-129.

PENEDER, M., (2003b), The Employment of IT Personnel, *National Institute Economic Review*, No. 184, April 2003, pp. 74-85.

ROMESBURG, H.C. (1984), *Cluster Analysis for Researchers*, Waldsworth Inc., Belmont.

VAN ARK, B. (2001), The Renewal of the Old Economy: An International Comparative Perspective, mimeo.