

THE EMPLOYMENT OF IT PERSONNEL

Michael Peneder*

The rapid advance and diffusion of new information technologies left pronounced imprints on the formation of human capital. Firstly, it favoured the growth of specific computer related occupations, secondly it raises the demand for higher levels of workforce education. Although IT equipment is the showcase of a general purpose technology, pervading the production process in any kind of economic activity, there is substantial heterogeneity among industries in terms of IT-labour intensity. The article demonstrates that these differences are sufficiently systematic to establishing a new sectoral classification, which captures substantial portions of the total variation in the above dimensions. More specifically, the results contrast popular beliefs about a uniform dissemination of new information technologies in all sectors.

Introduction

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.

Recent empirical research pays increasing attention to the complementarity between human capital and computer technologies. To cite just a few examples, Falk and Seim (2001) report a positive relation between the IT investment to output ratio and the employment of high-skilled workers for a panel of German companies; Chun (2003) demonstrates a positive interaction of demand for educated workers with IT adoption and use for a US industry panel from 1960 to 1996. Such evidence for 'skill-biased technological change' is usually interpreted as a direct consequence of the sharp decrease in quality-adjusted real computer prices, which cause the demand for all complementary input factors to go up as well. But based on empirical evidence from US firm-level data, Bresnahan, Brynjolfsson and Hitt (2002) also offer a more elaborate explanation, arguing that the demand for labour skills is embedded in a three-way system of complementarities among IT capital, new products and services, and new organisational practices. It should be noted, however, that all these studies are concerned with the educational composition of the workforce at large.

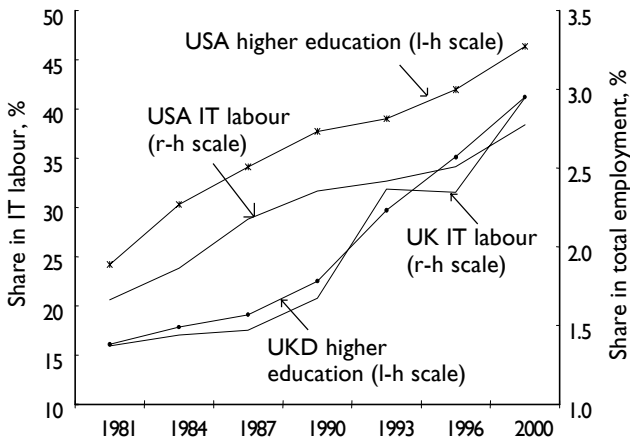
In contrast, this article focuses more narrowly on IT labour. The reason is that the digital revolution leaves some pronounced imprints on the overall formation of human capital in at least two dimensions. First, it favours the growth of specific computer related *occupations* (National Research Council, 2001). Second, it raises the demand for higher levels of workforce *education* (Autor, Katz and Krueger, 1998). Together, occupational and educational attributes characterise the IT labour intensity of a firm, an industry, or the aggregate economy.

The aggregate development of IT-labour intensity in the USA and the United Kingdom is consistent with this emphasis. Displaying 3-year averages¹ (until the indicated years), chart 1 shows that both the share of IT occupations in the total workforce and the share of persons with higher education among the IT personnel sharply increased between 1979 and 2000. From the beginning of the 1980s up to the end of the past century, the share of IT labour in total employment grew from 1.7 per cent to 2.8 per cent in the USA and more than doubled from 1.4 per cent to 3.0 per cent in the United Kingdom. Similarly, the share of persons with higher education among the IT workforce rose from 24.2 per cent to 46.8 per cent in the USA, while the United Kingdom experienced an increase from 16.1 per cent to 41.9 per cent.

Although IT equipment is a showcase example of a general purpose technology pervading the production process in any kind of economic activity, there is still

*Österreichisches Institut für Wirtschaftsforschung. I am grateful to Mary O'Mahony for providing the data and reviewing the paper; Eva Sokoll for her data-related assistance at WIFO; as well as Martin Falk and Serguei Kaniovski for their careful reading and helpful comments on an earlier draft.

Chart I. Aggregate development of IT labour, 3-year average^(a) (until indicated year)



Note: (a) Except for the last period ranging from 1997–2000.

much heterogeneity among industries in terms of IT-labour intensity. As this article demonstrates, these differences are sufficiently systematic to establish a new industry classification that captures substantial portions of the total variation in the above dimensions.

The basic reason for creating industry classifications is to facilitate investigations into the impact of specific characteristics of the market environment on economic activity. Therefore, such taxonomies are most frequently applied in the disciplines of industrial economics, technological development, or international trade, where they offer additional structure to a wide range of potential empirical applications. The general purpose is to select essential characteristics of technology and markets, condensing the vast heterogeneity of competitive environments into a smaller number of salient types. Directing our attention towards a few characteristic distinctions, industry classifications enable us to take account of heterogeneity, but, simultaneously, force us to be selective. Peneder (2003a) surveys the various aims, scope, and techniques applied in contemporary industry classifications. Peneder (2001, 2002, 2003b) or Kaniowski and Peneder (2002) are recent examples of analytic applications that range from micro-level studies on entry, exit and the age distribution of firms up to macro-panel estimations of the impact of industrial structure on aggregate growth.

From a purely practical perspective, the taxonomic approach is most useful precisely when it refers to data

that are not easily available in a comparable format across countries or firms. The reason is that it builds upon data offering the best coverage of specific attributes from those countries, 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 internationally comparable basis (for example, value added, employment, or foreign trade data). Unfortunately, for the same reason many classifications are based only on data for a single country and time period, lacking control of the heterogeneity between different economies and over time. In contrast, the new classification presented here is not only based on data for the USA and the United Kingdom, but also comprises annual data from 1979 until 2000. This enables us to consider their dynamic properties and take account of heterogeneity and interaction effects in the final cluster validation.

Throughout this article, a variety of statistical cluster techniques will be applied. Statistical cluster analysis is defined as ‘the art of finding groups in data’ (Kaufman and Rousseeuw, 1990), such that the degree of ‘natural association’ (Anderberg, 1973) is high among members within the same class (*internal cohesion*) and low between members of different categories (*external isolation*). Cluster analysis is a sophisticated tool for exploration and classification of multivariate data. Nevertheless it is a heuristic method, which requires the researcher to make a number of choices that can critically affect the outcome.

In order to be credible, any classification should therefore be backed by a full documentation of the critical choices and a detailed explanation of how the graphical representations were interpreted. This is my intention for the remainder of this article, which is organised as follows. The next section explains the data and its sources, and the third section the dissimilarity measures and clustering algorithms applied throughout the analysis. The fourth section briefly guides us through the three separate stages of the current clustering analysis. The fifth section validates the outcome and presents the final taxonomy and the sixth section summarises and concludes.

Data and the selection of variables

The most sensitive part of cluster analysis is the initial decision on the dimensions against which individual cases should be assessed. In the present analysis we are

Table 1. Definition of IT occupations

<i>United Kingdom, Standard Occupational Classification 1990</i>	
126	Computer systems manager
214	Software engineer
320	Computer analyst, programmer (incl. robot programmer)
490	Computer operator (incl. data processor, VDU operator, data entry clerk, database assistant)
526	Computer engineer, installation and maintenance (incl. computer repairer)
<i>USA, Occupational Classification from the 1980 Census</i>	
64	Computer systems analyst and scientist
65	Operations and systems researcher and analyst
229	Computer programmer
233	Tool programmer, numerical control
304	Supervisor, computer equipment operator
308	Computer operator
309	Peripheral equipment operator
385	Data entry keyer
525	Data processing equipment repairer

Source: Mason et al. (2003).

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 (LFS) and the US Current Population Survey (CPS), with annual data on workforce composition available for both employment and wages. The UK LFS is based on a total sample of 60,000 households per quarter, while the US CPS covers 50,000 households on a monthly basis. The data were harmonised into a common industry breakdown of 39 sectors and aggregated according to the chosen definition of IT occupations. Adhering to a narrow definition, only occupations with a direct responsibility for computer hardware and software are treated as IT personnel (table 1). Occupations such as communications equipment operators, electrical engineers, or electronic repairers, which occasionally appear in broader definitions of IT workers (see, for example, National Research Council, 2001) are not included here.

The need to maintain reasonable sample sizes limits the choice of variables and the degree of sectoral disaggregation. While the total number of IT labour is well represented in the chosen disaggregation of 39 sectors, the further dissection of IT labour into various educational categories produces a lot of random variation in the data, probably due to small sample sizes. For that reason, the share of persons with university degrees in total IT labour is the only variable on educational attainment. In addition, the annual data was pooled by calculating three-year averages from

1979 onwards plus a four-year average for the latest period from 1997 to 2000.

The workforce composition is represented by employment and wage shares for IT labour in the total workforce and the persons with higher education (university degrees) among IT labour. Since employment and wages are highly correlated, the four variables span only two independent dimensions of attributes. This redundancy in the variables serves to further mitigate the impact of noisy variations due to small sample sizes. Because of its symmetry, the lack of independence among the respective pairs of employment and wage shares is not a concern.

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. Finally, the particular time profile will be added as an additional dimension at the end of the analysis.

Measures of dissimilarity and clustering algorithms

Before starting a cluster analysis, one must decide on the particular measures of association and the clustering algorithm to use for the analysis. To begin with the measures of association, the *Euclidean distance* (L2) between any pair of observations is given by the root of the sum of squared differences over all the attributes. The Euclidean distance is very sensitive to outliers. In contrast, the closely related *City block* measure (L1) is the sum of the absolute distances for each attribute and hence accrues equal importance to any unit of dissimilarity. When we are more interested in the 'shape' of objects rather than in the absolute size of their differences, *angular separation* and the *correlation coefficient* are more appropriate. Both 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 from the mean of the variables under observation.

In addition to the above examples, the literature provides a variety of other dissimilarity functions that

can be applied in statistical cluster analysis. For extensive surveys see, for example, Romesberg (1984) and Gordon (1999). Unfortunately, there is no general guideline how to establish the superiority of one measure over another. In some instances, *a priori* conceptual considerations about the nature of the variables and the desired properties of the classification might provide sufficient guidelines for the decision as to which measure should be implemented. In general, it is desirable to experiment with more than one function in order to gain an idea of the robustness to variations in the concepts of measurement. For the purpose of the current classification, the four measures mentioned above are sufficient.

There is also a variety of different clustering algorithms, which can be used to group objects into separate categories. The *k-means* method is a popular example of the partitioning approach, where the set of observations is divided into a pre-defined number of clusters k . Cluster centres 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 centre. In the next step, the mean of the observations is recomputed and the process is repeated until convergence is reached. This is the case when no observation moves between groups and all remain in the same cluster of the previous iteration.

As a second approach, we also apply *agglomerative* hierarchical clustering algorithms. In contrast to the partitioning methods, the outcome is a hierarchical structure where observations are united at different levels until all belong to a single trunk. Again, there are various methods, which differ in the way in which they precisely determine the extent of dissimilarity between groups. The most popular and intuitively appealing choice is the *average linkage* method, whereby the average dissimilarity between all observations is compared for any pair of groups. Alternatively, the *complete linkage* method compares the dissimilarity between the most distant observations of two groups, whereas the *single linkage* method focuses on the dissimilarity of the nearest neighbours in any pair of groups. Although results are presented for the average linkage method, the other algorithms are used to assess their robustness such that these shall not be an artefact of a single method.

A three-stage clustering process

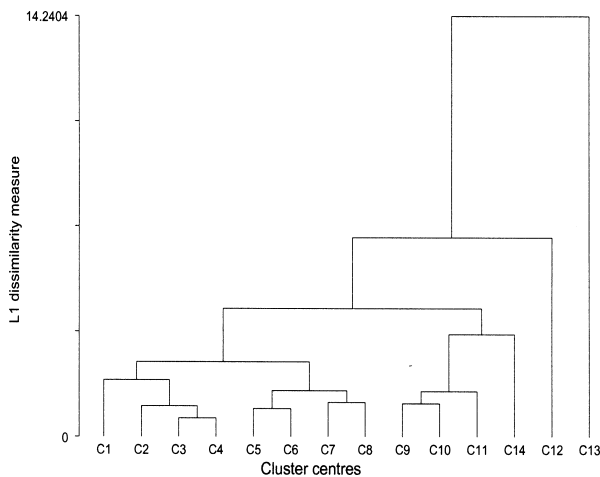
Lacking any general prescription for optimal solutions,

statistical cluster analysis is highly choice-dependent. Each exogenous decision by the researcher eliminates information and constrains the range of possible outcomes. A crude once-and-for-all cut into the data might obscure valuable patterns. In a sense, cluster analysis therefore requires an archeologist's method of carefully lifting various layers of soil and dust before detecting and salvaging any findings. For similar reasons, 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 purpose of the first step is to condense information and segregate outlying observations into separate clusters without yet imposing a strong structure on the overall outcome. At this stage, the *k-means* method offers the advantage that the initial case assignments remain reversible during the course of iterations. But as mentioned before, the researcher has to determine the number of clusters *a priori*. For reasons of internal consistency, we have applied the following self-binding rule-of-thumb since Peneder (1995): "*Choose the lowest number k that maximises the quantity of individual clusters l which include more than 5 per cent of the observed cases*". With 546 observations, the 5 per cent benchmark criterion amounts to 27 observations. Running the *k-means* algorithm on a dissimilarity matrix made up of Euclidean distances between any pair of observations for all values of k ranging from 2 until 35, the lowest number that fulfils the above rule turns out to be $k = 14$ with $l = 11$. The initial partition thus segregates 11 clusters of more than 27 objects, plus three smaller groups of outlying cases.

Inspection of the data reveals that the outliers cannot be dismissed as gross errors in the data, but represent a genuinely longtailed distribution.² The outliers are therefore relevant to our classification. Two clusters are of particular interest, because each of them comprise exactly all the fourteen observations for one specific sector. The first is the manufacturing industry of 'computer and office machinery' and is characterised by very high values for both dimensions of IT-related workforce composition. The second sharply

Chart 2. Dendrogram of second stage clustering, LI (city block) distance and average linkage



distinguished cluster of ‘computer and related activities’ refers to IT services. It exhibits a similar share of persons with university degrees among its IT workforce, but its share of IT personnel in total employment exceeds those of all other sectors by far.

As a further refinement, the fourteen cluster centres of the first partition are entered as individual observations in the second step of hierarchical analysis. In contrast to the k -means method, hierarchical analysis enables the determination of the boundaries between clusters at different levels of dissimilarity. 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. The branches at the bottom of the chart each represent one entity of the fourteen cluster centres (C1 to C14), while the final trunk at the top represents the entire set of objects. As we move upwards on the chart, the degree of association between objects is higher, the sooner they are connected by a common branch. Conversely, objects or groups are more dissimilar, the longer they remain disconnected. The result is a hierarchical structure, where the degree of relative association is the weaker, the more the respective branches appear separated.

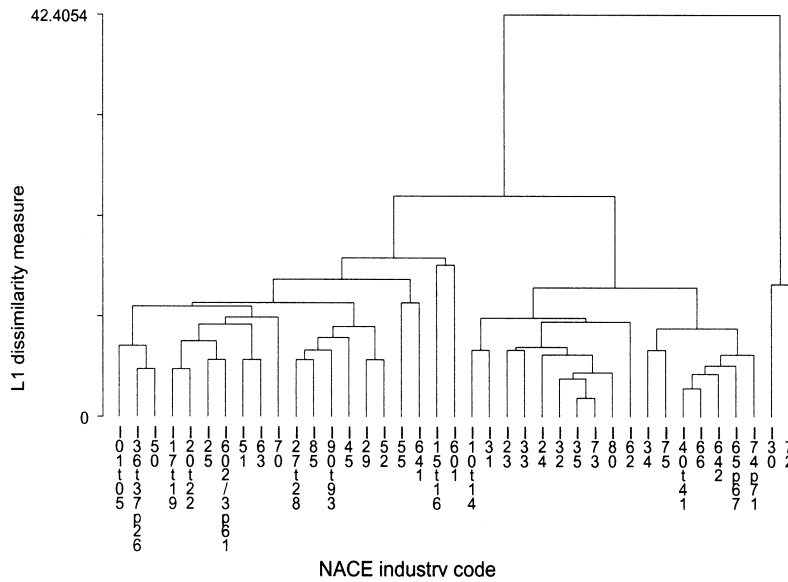
The dendrogram in chart 2 is based on the average linkage method and the city block measure of distance. Essentially identical cluster trees appear when Euclidean distances replace the city block measure, or the complete linkage method is applied instead of average

linkages. Despite some differences, both angular separation and the correlation coefficient preserved a similar order of associations. The single linkage method produced similar outcomes but suffers from chaining effects, especially if Euclidean distance or the city block measure were used. Overall, the patterns presented in chart 2 are reasonably robust. The labels at the bottom of the trees refer to the fourteen cluster centers of the initial partition in step one of the analysis.

This second stage of the clustering process has identified an interim classification of six separate categories. To begin with, the above mentioned two outlying clusters (C12 and C13) were isolated and appear in the immediate vicinity in nine out of the twelve possible outcomes. At the same time, the distance between the two clusters is considerable and exceeds that between the other groups. Consequently, we classify them into two small but separate categories. In all the twelve possible outcomes, objects C9 to C11 plus C14 form a robust and clearly identified separable cluster. While C7 and C8 appear together ten times, the association between C5 and C6 is much less stable and appears only four times. Lacking any stronger alternative association, we put the cluster of observations C5 to C8 into a common category. Objects C1 to C4 are not always that close in the other dendrograms, but C1 joins C2 in ten out of the twelve possible outcomes. Hence, I keep them separate and classify one group, which comprises C3 together with C4, and another encompassing C1 and C2. Overall, the six categories represent a descending order of IT labour intensity based both on the occupational and the educational characteristic.

To recapitulate, this interim classification of six separate clusters is based upon a data matrix of four variables of IT labour intensity and 546 cases, which are treated as independent observations. This implies that each of the 39 industries can belong to different classes depending on the specific time period or whether it refers to the United Kingdom or the USA. Standardisation of the data has eliminated aggregate time trends for each country, so that the overall patterns are remarkably robust. This is especially true for the two outlying classes of IT producers in both services and manufacturing, where no variation appears at all. Still, this falls short of establishing a general taxonomy, which in principle applies irrespective of the specific period or country under investigation. For this purpose, one can exploit the fact that the new classification condenses information from a multivariate attribute

Chart 3. Dendrogram of third stage clustering, L1 (city block) distance and average linkage



space into one single ordinal scale of six separate classes ranging from 1 (highest) to 6 (lowest IT labour intensity).

In the third and final stage of the cluster analysis the data is transformed 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,³ both average and complete linkage again produced almost identical results, whereas the single linkage method failed due to the typical problem of chaining objects without revealing any distinguishing structure among them. Chart 3 presents the dendrogram for the average linkage method. The labels at the bottom of the chart give the respective NACE industry codes.

After close inspection of the graphical representation in chart 3 and the according attribute values, the following separation into four final classes appeared to offer the most robust and consistently interpretable aggregation of the 39 sectors. First, the two outlying sectors NACE ‘72’ and ‘30’ establish two related but nevertheless separate classes. Then the group of objects from NACE ‘10t(ill)14’ to ‘74p(lus)71’ join another category, which exhibits a high IT labour profile if compared to the other non-IT producing industries. Avoiding the creation of too many classes with relatively minor differences between them, all other industries are put together into one residual category.

Cluster validation

The statistical cluster analysis has grouped industries according to their dissimilarity across a multivariate range of attributes. A sensible industry classification must also present interpretable structures. Before defining the final taxonomy, we therefore have to validate the results. The boxplots in charts 4 and 5 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 per cent 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 four different classes and are additionally separated by country.

The extremely skewed distribution supports the distinction 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. As the boxplots in chart 4 illustrate, both display approximately equal shares of persons with higher education among their IT workforce, but the share of IT labour in total employment is much larger for computer services than

Chart 4. Boxplots of workforce composition by country

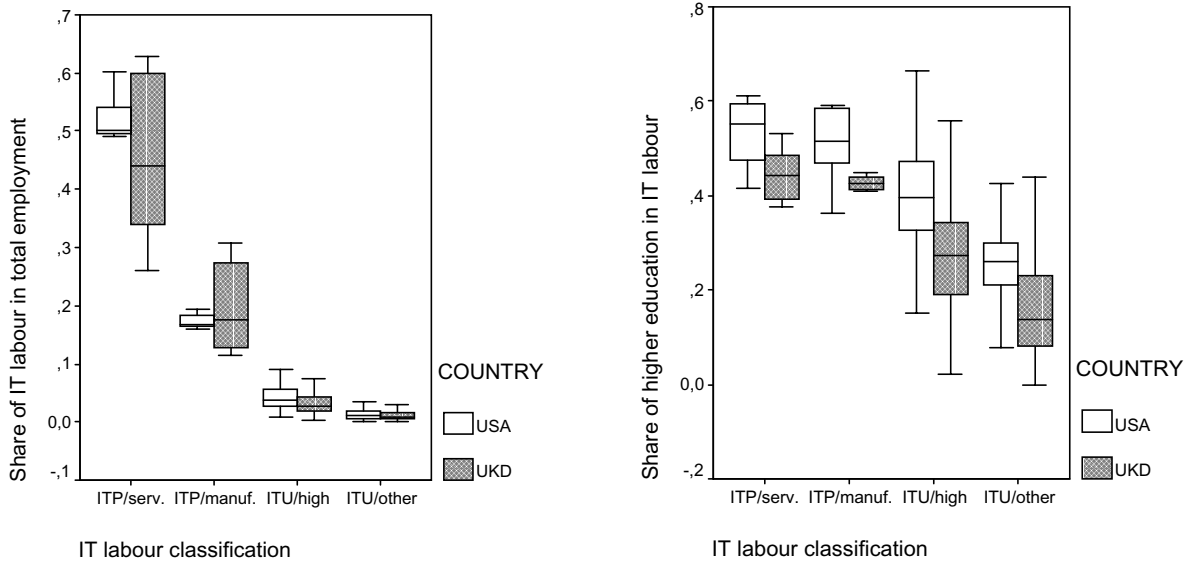
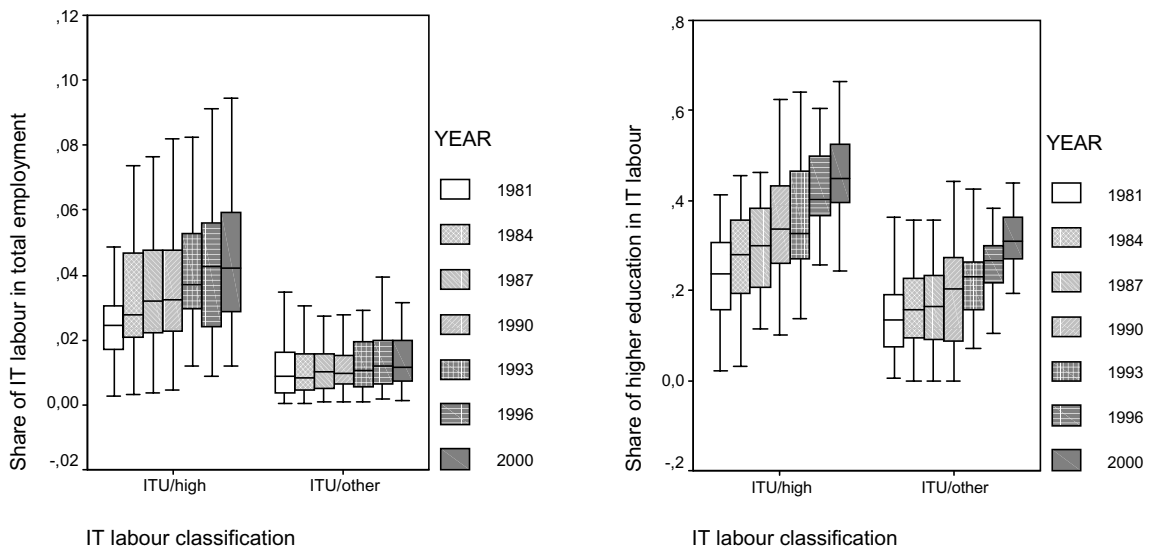


Chart 5. Boxplots of workforce composition by 3-year^(a) averages (up to indicated years)



^(a) 4-year average for the final period.

for computer manufacturing. Taking account of that difference, we put them into the two separate classes of *IT producer/services* (ITP/serv.) and *IT producer/manufacturing* (ITP/manuf.). 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. While differences

regarding the share of IT labour in total employment look minor on the grand scale of chart 4, the two groups are clearly distinct if we take the narrower focus in chart 5. Not only does the third group exhibit higher shares of IT labour in total employment, it also shows a marked growth over time. But the same does not apply to industries in the fourth category. This is a remarkable result. Because of the explicit mapping of the time

The IT labour industry classification (NACE industry codes in brackets)

1. *IT producer – services* (ITP/serv.): Computer and related activities (72).
2. *IT producer – manufacturing* (ITP/manuf.): Computers and office machinery (30).
3. *Dynamic IT user with a high and growing IT-labour intensity* (ITU/high): Mining and quarrying (10–14); Mineral oil refining, coke and nuclear fuel (23); Chemicals (24); Electrical machinery and apparatus (31); Radio, television and communication (32); Instrument engineering (33); Motor vehicles (34), Other transport equipment (35), Electricity, gas and water supply (40–41), Air transport (62); Telecommunications (642); Financial intermediation (65, 67), Insurance and pension funding (66), Research and development (73); Other business services (71, 74), Public administration and defence, incl. compulsory social security (75); Education (80).
4. *Other IT-user industries* (ITU/other): Agriculture, forestry and fishing (01–05), Food, drink and tobacco (15–16), Textiles, leather, footwear and clothing (17–19), Wood, products of wood and cork; Pulp, paper and paper products, printing and publishing (20–22), Rubber and plastics (25), Non-metallic mineral products, furniture, miscellaneous manufacturing (26, 36–37), Basic metals and fabricated metal products (27–28), Mechanical engineering (29), Construction (45), Sale, maintenance and repair of motor vehicles and motor cycles (50), Wholesale trade (51), Retail trade (52), Hotels and catering (55), Railways (601), Other inland transport, Water transport (602–603, 61), Supporting and auxiliary transport activities, activities of travel agencies (63), Post and courier activities (641), Real estate (70), Health and social work (85), Other community, social and personal services (90–93).

dimension in the third stage of the cluster analysis, the taxonomy successfully separated industries according to their development over time.

While educational levels show a similar rise in all four classes, two important stylized facts appear for the share of IT personnel. First, the diffusion of IT as measured by the shares of IT personnel in the total workforce is not evenly distributed among IT-user industries. Second, those industries which already exhibited a higher share already at the beginning of the 1980s have subsequently expanded their IT personnel faster. Because of the higher levels in both dimensions of IT labour intensity and the faster growth in the share of IT labour in the total workforce, one can characterise the third category as dynamic IT-user industries with a *high and growing IT labour intensity* (ITU/high). This group comprises seventeen industries, while the fourth category of *other IT-user* industries (ITU/other) comprises the twenty remaining sectors. The final taxonomy thus classifies 39 sectors into four separate industry types (see Box above).

In addition to a meaningful economic interpretation, industry classifications should be reasonably robust with respect to time and spatial boundaries, especially if they are meant for use in international comparative analyses. Although the share of persons with university degrees is generally lower in the United Kingdom than in the USA, the boxplots indicate that differences between the two countries hardly affect the distribution and relative order of industry groups in that attribute. Similarly, with respect to the time dimension, all four

industry classes experienced a rather uniform increase in the share of higher education among 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. The remarkable exception is the class of other IT-user industries, for which the occupational composition of the workforce is almost unaffected by the rapid advance of information technologies during the 1980s and 1990s. Consequently, the passage of time appears to further reinforce the separation between the industry types instead of blurring or reversing their order.

Going beyond mere visual validation, the analysis of variance (ANOVA) and simple OLS regressions of educational intensity on the four industry types test the discriminatory power of the new taxonomy more strictly. Obviously, the F-ratios, which confirm that for all variables the taxonomy indeed discriminates significantly between observations, are not the issue in table 2. This result is almost trivial, since the taxonomy was created explicitly for that purpose. Otherwise, we would have had to worry about violations of the necessary homoscedasticity assumption. But taking a closer look at the coefficients and the respective t-values (in brackets) of simple OLS regressions, we see that in each of the eight variables all the remaining industry types differ significantly from the comparison group of ITU/other industries. In addition to this positive validation of the discriminatory power of the new industry classification, table 1 also presents a summary of average workforce composition. The coefficient for the constant term corresponds to the mean shares in the

Table 2. OLS regression on IT labour intensity, 1979 to 2000

Industry type	Share of IT labour in total workforce				Share of higher education in IT labour			
	USA		United Kingdom		USA		United Kingdom	
	Employment	Wages	Employment	Wages	Employment	Wages	Employment	Wages
Constant	0.0129 (9.77)	0.0146 (10.81)	0.0113 (4.46)	0.0150 (5.03)	0.2610 (33.79)	0.3492 (39.33)	0.1626 (16.27)	0.2304 (19.55)
ITP/serv.	0.5114 (84.44)	0.5085 (82.39)	0.4467 (38.63)	0.4997 (36.67)	0.2708 (7.65)	0.2676 (6.58)	0.2812 (6.14)	0.3000 (5.56)
ITP/manuf.	0.1605 (26.50)	0.1742 (28.23)	0.1889 (16.34)	0.1919 (14.08)	0.2501 (7.07)	0.2324 (5.71)	0.2586 (5.65)	0.2745 (5.08)
ITU/high	0.0290 (14.85)	0.0291 (14.59)	0.0220 (5.92)	0.0211 (4.81)	0.1432 (12.53)	0.1508 (11.49)	0.1211 (8.21)	0.1432 (8.24)
Observations	272	272	273	273	272	272	273	273
F-ratio	2545.8	2460.3	586.8	501.0	72.6	57.3	37.1	33.8
Prob>F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
R-squ.	0.9661	0.9650	0.8638	0.8482	.4482	0.3906	0.2925	0.2739
Adj,R-squ.	0.9657	0.9646	0.8623	0.8465	0.4420	.03838	0.2846	0.2658
Root MSE	0.0156	0.0159	0.0297	0.0352	0.0914	0.1051	0.1182	0.1395

Note: ITP/serv. = IT producer / services; ITP/manuf. = IT producer / manufacturing; ITU/high = IT user with dynamic and high IT-labour intensity; ITU/other is the comparison group.

comparison group of ITU/other. The mean shares for the other industry types can easily be calculated by adding the respective coefficients to the constant.

Finally, we can infer from the R^2 and the F-ratio which variables the taxonomy discriminates better than others. For example, we see that the share of explained between-group variation is particularly large for the occupational attributes (i.e. the share of IT labour in the total workforce), whereas we still find much within group variation in the workforce composition for the educational variables (i.e. the share of higher education in IT labour). One likely reason is that the latter dimension is exposed to more random variations, because of the necessarily smaller sample sizes (since it relies only on IT labour with university degrees instead of total IT personnel). Additionally, systematic time and country effects might be more prevalent due to national differences in the educational system and its expansion. Conversely, we would generally expect that occupational attributes are more closely related to technology and markets.

Table 2 also shows that despite their high overall correlation, the variables which rely on wage shares exhibit somewhat more unexplained within-group variation compared to those with employment shares as measure of workforce composition. Finally, in each instance the F-ratio is considerably higher for the USA than for the United Kingdom. Since both countries had a symmetric impact on the clustering process, the higher degree of unexplained variation might indicate a more

pronounced change in the attributes over time, which would be consistent with the developments displayed in chart 1. Alternatively, it might also hint at the poorer quality of the UK data.

Going one final step further in the cluster validation, we are additionally interested in the relative importance of the industry types versus the impact of variation between the two countries and the passage of time. For this purpose, table 3 presents the results of a 3-way ANOVA OLS regression with interaction effects on each of the four attributes used in the clustering process as dependent variable.

First, as regards the occupational dimension measured by the share of IT labour in total employment, time variation certainly matters. The mean share is significantly lower in the years up to 1990 compared to the dropped period of 1997 to 2000. From 1990 onwards the time dummies are not significant. However, the abrupt change in the coefficients after 1990 probably indicates a break in the data around that time. The country dummy displays a significant and negative coefficient for the USA, which reflects the higher share of IT labour in the United Kingdom for the final period. However, for most prior periods this is offset by the interaction term between year and countries, where before 1990 the coefficients are significantly positive for the USA. Thus, the interaction of the time and spatial dimension is consistent with the aggregate picture of chart 1.

Table 3. Three-way ANOVA, OLS regression on time, country and industry effects, 1979 to 2000

	Share of IT labour in total workforce				Share of higher education in IT labour			
	Employment		Wages		Employment		Wages	
	Coefficient	t-value	Coefficient	t-value	Coefficient	t-value	Coefficient	t-value
Constant	0.0180	4.40	0.0248	5.63	0.2952	21.40	0.3639	21.79
Year ^(a)								
1981	-0.0131	-2.31	-0.0160	-2.63	-0.2188	-11.59	-0.2305	-10.09
1984	-0.0094	-1.66	-0.0141	-2.32	-0.1981	-10.49	-0.2058	-9.00
1987	-0.0118	-2.08	-0.0180	-2.96	-0.1972	-10.45	-0.1940	-8.49
1990	-0.0104	-1.83	-0.0168	-2.77	-0.1708	-9.05	-0.1518	-6.64
1993	-0.0018	-0.32	-0.0028	-0.46	-0.1254	-6.64	-0.1110	-4.86
1996	-0.0009	-0.16	-0.0009	-0.15	-0.0486	-2.58	-0.0448	-1.96
Country ^(b)	-0.0088	-1.77	-0.0139	-2.61	0.0478	2.53	0.0828	3.62
Industry type ^(c)								
ITP/serv.	0.5185	32.50	0.5815	33.98	0.2760	12.09	0.2838	10.27
ITP/manuf.	0.2313	14.50	0.2493	14.57	0.2544	11.14	0.2535	9.17
ITU/high	0.0300	5.84	0.0328	5.96	0.1317	17.90	0.1466	16.45
Year ^(a) * Country ^(b)								
1981	0.0199	3.00	0.0210	2.96	0.0543	2.03	0.0302	0.93
1984	0.0132	2.00	0.0195	2.75	0.0708	2.65	0.0473	1.46
1987	0.0197	2.98	0.0271	3.82	0.1049	3.93	0.0753	2.33
1990	0.0166	2.52	0.0239	3.37	0.1036	3.88	0.0650	2.01
1993	0.0018	0.28	0.0024	0.33	0.0722	2.70	0.0413	1.28
1996	0.0020	0.30	0.0008	0.12	0.0104	0.39	-0.0003	-0.01
Year ^(a) * Industry type ^(c)								
1981-ITP/serv.	-0.1304	-6.18	-0.0902	-3.98				
1981-ITP/manuf.	-0.0685	-3.24	-0.0995	-4.39				
1981-ITU/high	-0.0172	-2.52	-0.0241	-3.30				
1984-ITP/serv.	-0.0642	-3.04	-0.1293	-5.71				
1984-ITP/manuf.	-0.0819	-3.88	-0.1089	-4.81				
1984-ITU/high	-0.0118	-1.74	-0.0178	-2.44				
1987-ITP/serv.	-0.1602	-7.59	-0.1996	-8.81				
1987-ITP/manuf.	-0.0826	-3.91	-0.1050	-4.64				
1987-ITU/high	-0.0103	-1.51	-0.0156	-2.14				
1990-ITP/serv.	-0.1168	-5.53	-0.1411	-6.23				
1990-ITP/manuf.	-0.0570	-2.70	-0.0783	-3.46				
1990-itu/high	-0.0082	-1.20	-0.0120	-1.64				
1993-itp/serv.	-0.0105	-0.50	0.0050	0.22				
1993-itp/manuf.	0.0079	0.37	-0.0038	-0.17				
1993-ITU/high	-0.0041	-0.61	-0.0069	-0.95				
1996-ITP/serv.	-0.0206	-0.97	-0.0176	-0.78				
1996-ITP/manuf.	-0.0146	-0.69	-0.0064	-0.28				
1996-ITU/high	-0.0041	-0.60	-0.0058	-0.80				
Country ^(b) * Industry type ^(c)								
ITP/serv.	0.0647	5.73	0.0088	0.73				
ITP/manuf.	-0.0284	-2.52	-0.0177	-1.46				
ITU/high	0.0069	1.91	0.0079	2.02				
Observations	545		545		545		545	
F-value	221.8		215.4		65.4		53.0	
Prob>F	0.0000		0.0000		0.0000		0.0000	
R ²	0.9418		0.9402		0.6647		0.6164	
Adj. R ²	0.9376		0.9358		0.6545		0.6048	
Root MSE	0.0206		0.0221		0.0834		0.1009	

Notes: (a) 3-year averages up to indicated year, '2000' is the comparison group. (b) USA versus United Kingdom as comparison group. (c) ITU/other is the comparison group.

Most important for our purpose, the industry types show the expected positive coefficients and ranking order, exhibiting by far the strongest level of significance with very similar outcomes for both employment and wage shares as measures of workforce composition. There is some significant interaction with country effects for employment shares but few for wages. More interestingly, the year industry type interaction is particularly strong. Compared to the group of ITU/other industries and given the positive direct industry effects combined with the negative direct effects of the year dummies, all the three remaining industry types display a significant negative interaction for the years prior to 1990. This tells us that the industry effects are generally smaller for the earlier years, which is consistent with the stylized fact already illustrated in chart 5. There the group of ITU/other stood out by its surprisingly flat time profile, while all other industry types showed a strong and steady growth of their IT-labour shares. In short, this implies that the group with the lowest share of IT labour fell further behind. In principle any significant interaction terms for the industry types signal troubles with the robustness of a general classification. In this case, however, over time the interaction between industry types and years tends to further strengthen the initial partition of our cluster analysis.

Turning to the additional educational dimension of higher education in the IT workforce, the overall share of explained variation is much smaller than for the purely occupational dimension before. Again this suggests a larger portion of noisy variation in the educational attributes because of small sample sizes. However, compared with table 2, the share of explained variation has risen considerably after the inclusion of year and time effects, which seems to reflect a greater importance of national differences in the educational system and a stronger time trend. The year effects are negative with a regular and significant rise of shares over time. The country effect is significant and positive, consistent with the higher share of persons with university degrees in the USA. The industry effects display the highest coefficients for the two IT producer industries, whereas the coefficient for IT user with a dynamic and high share of IT labour are also significant and positive. The interaction between years and country are significant and positive, capturing the uneven pattern in the aggregate development of that variable as shown in chart 1.

The interaction terms between industry types and years

as well as industry type and country are not significant and were therefore not included in the final regression. This reassures us that the strong variations between countries and over time are contained by standardisation of the variables, which were used for the clustering process. The taxonomy is less powerful in discriminating with respect to the educational attribute than it was before (regarding the occupational variable). Our crude separation of industries into four categories nevertheless appears to be particularly robust with respect to both the time and country variations of that educational variable.

Summary and discussion

This article develops a new taxonomy based on sectoral characteristics of IT labour intensity, which is defined, first, by the occupational attribute of an industry's share of IT personnel in the total workforce and, second, by the share of higher education among its IT labour as an educational attribute. Both employment and wage shares represent workforce composition. The data cover 39 sectors in the USA and the United Kingdom from 1979 until 2000. The taxonomy is established by a three-staged statistical cluster analysis, using the dynamic profile of prior cluster identifications for the final partition. Cluster validation by means of boxplot charts helped to interpret the outcome. In a series of 3-way ANOVA regressions with additional country, time, and interaction effects, the industry types explain substantial portions of the overall variation. Further inspection of the interaction terms suggests the overall robustness of the taxonomy with respect to differences between the two countries and over time.

The new taxonomy consists of four classes. First, computer services and, second, computer manufacturing are both IT-producing industries with the highest IT labour intensity. Being extreme outliers, each establishes a category of its own. The third group comprises seventeen IT-user industries, which are characterised by a comparatively high and dynamic profile of IT labour intensity. For them the advance and diffusion of new information technologies has left a strong imprint on the workforce, both in terms of occupations and the rising share of persons with higher education. A certain characteristic of the fourth category of other IT-user industries sharply contrasts popular beliefs about a more or less uniform dissemination of new information technologies in all sectors. The twenty industries belonging to this class, display not only lower levels of educational attributes,

but also exhibit a surprising lack of dynamics in the occupational attribute. While the share of IT labour in the total workforce grew steadily in all other sectors, these industries show no signs of catching-up from low initial levels but fall further behind. Two possible explanations seem plausible. First, these industries might be intrinsically less inclined towards information processing activities and therefore need less IT personnel. Second, they might systematically outsource more of their IT-related tasks to the specialised IT-producing services (see, for example, Altinkemer, Chaturvedi, and Gulati, 1994). It is, however, likely that both causes interact. That is to say that, in the group of other IT-user industries, outsourcing might be more important, because the demand for information processing activities remains below a critical scale, where internal production would pay off.

The general purpose of the new industry classification is to pick essential characteristics of technology and markets, condensing the vast heterogeneity of competitive environments into a smaller number of salient types. In particular, when the underlying attributes of IT labour intensity are not directly observable, the taxonomy offers a potentially useful analytic structure to a wide range of empirical applications.

NOTES

- 1 Except for the last period ranging from 1997 to 2000.
- 2 For a discussion see Hampel (2002).
- 3 We actually treat the interim classification as a new scale, against which we want to measure changes in the time dimension. We implicitly assume that, if an industry moves between any two neighbouring classes, it always covers the same distance on that scale. If it moves two classes, it covers twice that distance, and so on.

REFERENCES

Altinkemer, K., Chaturvedi, A., Gulati, R. (1994), 'Information systems outsourcing: issues and evidence', *Journal of Information*

- Management*, 14, 4, pp. 252–68.
- Anderberg, M.R. (1973), *Cluster Analysis for Applications*, New York, Academic Press.
- Autor, D.H., Katz, L. and Krueger, A.B. (1998), 'Computing inequality: have computers changed the labor market?', *The Quarterly Journal of Economics*, 113, 4, pp. 1169–213.
- Bresnahan, T. F., Brynjolfsson, E.B. and Hitt, L.M. (2002), 'Information technology, workplace organization, and the demand for skilled labor: firm-level evidence', *The Quarterly Journal of Economics*, 117, 2, pp. 339–76.
- Chun, H. (2003), 'Information technology and the demand for educated workers: disentangling the impacts of adoption versus use', *The Review of Economics and Statistics*, 85, 1, pp. 1–8.
- Falk, M. and Seim, K. (2001), 'The impact of information technology on high-skilled labor in services: evidence from firm-level panel data', *Economics of Innovation and New Technology*, 10, 4, pp. 289–324.
- Gordon, A.D. (1999), *Classification*, 2nd edn, Chapman & Hall, Boca Raton.
- Hampel, F. (2002), 'Some thoughts about classification', in Jajuga, K., Sokolowski, A. and Bock, H.-H., *Classification, Clustering, and Data Analysis. Recent Advances and Applications*, Berlin, Springer, pp. 5–26.
- Kaniowski, S. and Peneder, M. (2002), 'On the structural dimension of competitive strategy', *Industrial and Corporate Change*, 11, 3, pp. 257–79.
- Kaufmann, L. and Rousseeuw, P.J. (1990), *Finding Groups in Data. An Introduction to Cluster Analysis*, New York, Wiley.
- Mason, G., Robinson, K., Forth, J. and O'Mahony, M. (2003), 'Industry-level estimates of ICT and non-ICT employment, qualifications and wages in the UK and USA, 1979–2000', mimeo, National Institute of Economic and Social Research.
- National Research Council (2001), *Building a Workforce for the Information Economy*, Washington DC, National Academy Press.
- Peneder, M. (1995), 'Cluster techniques as a method to analyse industrial competitiveness', *International Advances in Economic Research*, 1, 3, pp. 295–303.
- (2001), *Entrepreneurial Competition and Industrial Location*, Cheltenham, Edward Elgar.
- (2002), 'Entry, age and sectoral specialisation of Viennese firms', *Austrian Economic Quarterly*, 2002, 3, pp. 109–20.
- (2003a), 'Industry classifications. Aim, scope and techniques', London, EPKE Working Paper, 2002–07 (also in *Journal of Industry, Competition and Trade*, forthcoming).
- (2003b), 'Industrial structure and aggregate growth', *Structural Change and Economic Dynamics* (forthcoming).
- Romesburg, H.C. (1984), *Cluster Analysis for Researchers*, Belmont, Waldsworth Inc.
- van Ark, B. (2001), 'The renewal of the old economy: an international comparative perspective', mimeo.