THE TECHNOLOGICAL STRUCTURE OF U.S. MANUFACTURING

Thomas A. Abbott III, Rutgers University, and
Stephen H. Andrews, Bureau of the Census[1]
Thomas A. Abbott III, 92 New Street, Newark, NJ 07102

Abstract

Understanding the relationship of production technologies across manufacturing industries is vital for analyzing dynamic economic activity because firms and establishments often change industries in response to economic conditions. Most researchers use the Standard Industrial Classification (SIC) system, with its hierarchical 2, 3 and 4-digits, to identify these changes. However, as discussed in Andrews-Abbott (1988), the SIC is replete with problems, and may not provide a good basis for detecting dynamic changes in economic activity. For example, if two industries are technologically very similar, changes from one industry to the other may reflect changes in coding rather than real changes in economic activity at the establishment.

This paper examines the technological relationships across industries. It begins by defining each industry's production technology, and develops a continuous measure of the distance between industries. We find empirically that the SIC does not do a particularly good of grouping industries with similar production technologies. Next we use these measures of technological distance to cluster industries and form a new, technologically based, classification system. While our technological approach yields results which are similar to the SIC in many regards, there are important differences between the two classifications in terms of the industrial categories which emerge and the amount of information lost in the process of aggregation. Thus, we conclude that much can be learned about the dynamic interactions between firms and establishments by looking at industry and establishment classification in a less rigid fashion.

1. Introduction

Over the past several years, we have written several papers examining the way in which establishments are grouped to form industries and how these industries are grouped to form an industrial classification system, see Andrews-Abbott (1988), Abbott-Andrews (1990,1993), Abbott (1992). Throughout these papers, we have argued that classification issues are very important for economic analysis because the classification system colors the way researchers look at the data. Furthermore, because every classification results in the loss of information, an "optimal" classification can only be defined with respect to a particular use of the data. Therefore, different uses of the data require different methods of classification. In this paper, we focus on production technologies, and examine the relationship between the production technologies of different 4-digit industries. Our analysis is conducted in three steps.

First, we introduce a continuous measure of the technological distance between pairs of industries which is consistent with the economic theories of production, and commonly used parametric representations of the production and cost functions.

Second, we use this measure of distance to evaluate how well the SIC groups industries which are close together to form Industrial Groups (3-digit SIC) and Major Industrial Groups (2-digit SIC). Our analysis shows that although the SIC does a better job than randomly assigning industries into the 2-digit categories, the average distance between pairs of industries within the same 2-digit Major Group is only slightly less than the average distance of pairs of industries in different 2-digit groups. Thus, there remains considerable room for improvement when using the SIC to measure technological similarities across industries.

Third, we use our distance measure and a clustering algorithm to form our own technologically based,

---

[1] *Assistant Professor at the Graduate School of Management, Rutgers University and Research Associate at the Center for Economic Studies, Bureau of the Census; and Branch Chief, Industry Division, Bureau of the Census, respectively. The opinions expressed are those of the authors and do not represent the views of Rutgers University or the Bureau of the Census.* *The authors would like to thank Robert McGuckin, Tim Dunne, Sang Nguyen, and Mark Doms for their helpful comments on an earlier draft. As usual, responsibility for remaining errors rests solely with the authors.*

industrial classification system -- minimizing the amount of information lost due to the aggregation. Based on our analysis, we conclude that:

1) many current 4-digit industries are quite close technologically, and could be grouped with very little information loss.

2) some industries have radically different technologies from all of the other industries and should never be grouped.

3) it does not make sense to try to group all manufacturing industries into only 20 categories, because one looses too much information and the resulting classes are heterogeneous.

We discuss each of these points in turn.

2. Measuring Technological Distance

Our study of the relationship of production technologies across U.S. Manufacturing begins by making a fundamental assumption that the production technology of an industry can be characterized by its vector of input shares. As many authors have noted[1], using the input share vector to define the production technology can be justified by an assumption of a Cobb-Douglas production function, since in competitive equilibrium, the input shares exactly equal the coefficients of the production function. However, the input shares also provide valuable information on the production technology in other contexts as well. For example, under the Leontief technology, the quantity of each input is fixed. If we assume a competitive input market, differences in input shares correspond directly to differences in the production technologies. Even in more general "flexible functional forms" such as the Translog production or cost functions, the input shares are linearly related to the parameters of the model. Thus, examining input shares is consistent with many of the models currently used to describe production technologies.

Implementing this approach requires a comprehensive list of the inputs used by each industry. Using the information collected in the 1987 Census of Manufacturing, we constructed input vectors for 456 of the 459 4-digit industries based on a sample of over 66,000 establishments.[2] These vectors consist of data on the different types of labor (divided into production workers and non-production workers), energy (split into electricity and other fuels), capital (defined as the residual), and detailed information on 360 types of material inputs. A more detailed discussion of the construction of these input vectors can be found in Abbott-Andrews (1993).

Table 1 presents statistics on selected factor shares across our sample of 456 industries. As shown in the table, there is a great deal of variability in the shares of each of the major inputs across the different industries. For example, fuels ranged from a share of .02 percent to 21 percent depending upon the industry; and production workers ranged from a low of 1 percent to a high of 37.4 percent. Unfortunately, the table also reveals a problem with defining the share to capital as a residual. In the Metal Heat Treating Industry (3398), the share of capital was negative; that is, the cost of all of the other inputs exceeded the total value of production. Overall, we find that materials had the highest average share (nearly 43 percent), followed by Capital (35 percent) and production workers (12 percent.) The remaining inputs had relatively small average shares -- although they were nevertheless important for some industries. Given this widespread variation in factor shares across industries, one would anticipate that there might be some industries which had very similar production technologies and others which were very different.

Turning to the question of measuring the technological distance between industries, because of the use of input shares, each production technology can also be represented as a point on the unit hyperplane. Although several measures of the distance between points on a plane can be constructed, we have chosen to use the Euclidean distance measure in this paper, in part because it is consistent with the clustering algorithms used later. Our first step was to construct distance measures for all possible combinations of the 456 4-digit industries. This yielded a total of 103,740 pair-wise comparisons. Table 2 provides summary statistics on these distance measures. In particular, the average distance between any two industries was .36; and the range went from .02 to 1.11 -- although the numbers in and of themselves are not very interesting. It is, however, interesting, to look at which industries are closest together and whether the current SIC does a good job of grouping those industries. Specifically, if the SIC groups industries with similar technologies, one would expect that the average distance between pairs of industries within the same 2-digit Major Group would be much lower than the average distance between pairs of industries cutting across 2-digit Major Groups. Table 2 presents statistics for these two subsamples as well. Although one can easily reject the hypothesis that

the two groups have the same mean; it is surprising that they are qualitatively not very different (.30 versus .37). One would have expected a much larger difference between the two groups. Moreover, if one looks at the range; one sees an enormous overlap.

After examining the 100 pairs of industries which are closest together according to our distance metric (the extreme tail of the distribution with less than .1 percent of the comparisons), we found that 21 of these pairs cut across 2-digit boundaries. Table 3 presents these pairs of industries, their ranking, and the distance between them. Two industries in particular appear frequently on this list. Industry 2542, Partitions and Fixtures nec, is very close to six of the industries in major group 34 - Fabricated Metal Products; and Industry 3821, Laboratory Apparatus and Furniture, is very close to five of the industries in major group 35 - Industrial Machinery and Equipment. Thus, changing the major group assignments of these two industries would eliminate over half of this list.

Based on this analysis, we conclude that although the current SIC does a better job than randomly grouping industries, it fails to successfully group some of the industries which are quite close together. Distance measures, like those developed here, can be used to find areas where the SIC is particularly weak and can serve as a basis for making minor revisions to the current system. In addition, these same distance measures can be combined with clustering algorithms to develop new classifications based on the similarities in production technologies, as discussed below.

3. A Technology Based Industrial Classification

In this section we examine the use of hierarchical clustering methods for grouping the 4-digit industries into higher levels of aggregation. Conceptually, the process begins with each industry in its own cluster. At each step of the process, the two clusters which are closest together are combined to form a single cluster, reducing the total number of clusters by one. This process continues until all industries are in a single cluster. This process of aggregation results in information being lost because, as industries are grouped together, their individual technology vectors are replaced by the average vector for the entire group.

Competing methods for clustering differ primarily in how the distance between groups of industries is measured.[3] In our analysis, we chose to use Ward's method, see Ward (1963)), because, at each step of the process, it minimizes the amount of information lost. Ward's method measures the difference between two groups as the weighted difference between the two mean vector. Using this method, one can measure the information loss as the ratio of the sum of squared distances from each cluster mean (i.e. the within variance in an ANOVA decomposition) to the total sum of squares. From this, one can construct an R-square measure for the information retained as one minus the information loss.

Table 4 provides the R-square statistic for several levels of aggregation using this approach. As shown in the table, one could reduce the number of clusters substantially without loosing much information. Specifically, one could cut the number of industries from 456 to only 329 and loose only 1 percent of the information on the individual production technologies. Such a reduction in the number of industries might produce significant savings for the collection and processing of the data, and might eliminate many of the establishments "switching" industries. At the other end of the table, one can't help but notice that aggregating all of manufacturing into only 20 categories results in a tremendous loss of information on the production technologies (well over half of the information is lost.) The drop in the R-square is particularly dramatic in going from 55 clusters to 20; thus we recommend against aggregating beyond about 55 clusters. After examining these 55 clusters more closely,[4] we discovered that 16 of these clusters consist of only a single 4-digit industry, despite the fact that Ward's method tends to result in evenly distributed clusters. Thus, these 16 production technologies are clearly distinct and it would be misleading to force them into clusters with the other industries, as currently done by the SIC.

Table 5 provides a list of these 16 industries. One thing that these industries have in common is the fact that they are closely tied to a single primary material input. For example, the primary input for Cane Sugar Refining is raw sugar cane (72% input share) which no other industry uses. Likewise, Creamery Butter uses 80% milk and cream, Soybean Oil Mills use 75% raw soybeans, and Primary Copper uses 73% raw copper ore. Thus, these industries tend to be isolated along a single dimension in our input space and are very far away from the other industries. Forcing these industries into clusters with other industries, as done by the SIC, results in great distortions to their input vectors.

Table 4 also provides a comparison of the information

retained by the current SIC. As shown in the table, our optimal classification retains significantly more information at comparable levels of aggregation -- affirming our basic proposition that if one wants to study production technologies and changes in economic activity, one should use a classification system designed to preserve that information.

4. Conclusions and Future Research

The analysis presented in this paper shows that the Standard Industrial Classification system does not do a good job of grouping together industries which are technologically close together, and forces together industries which are quite distinct. Whether this should be interpreted as a weakness of the SIC or a weakness of the present methodology is a subject for additional research. For, it is clear that changing the list of inputs used to define the production technologies or using alternative distance measures would change the specific results presented. However, as a basis for raising questions and pointing towards specific sections of the SIC which may need additional examination, we believe that the methodology is sound.

Furthermore, although it is clear that one would not want to mechanically following the clustering procedures outlined here to construct a new classification, insights into the relationships of production technologies across industries can clearly be obtained from the analysis of the data in this fashion. Through the clustering analysis, we were able to identify sets of industries which could be grouped together without loosing much information, as well as sets of industries which had very distinct technologies. The former results suggest that the current 4-digit industry definitions are too narrow; while the latter result suggests that the 2-digit level is too aggregated to be useful. Unfortunately, there are no natural breaks in the R-square during the aggregation, and thus a definitive conclusion about the number of "industries" is not possible.

1. See for example Gollop (1986), Chambers (1988), and Gollop-Monahan (1989).

2. In choosing the establishments used to construct the aggregate input vector for the industry, we restricted our attention to only those establishments which: a) reported detailed (6-digit) materials consumed, and b) had a specialization ratio of at least 95% (i.e. 95% or more of the value of shipments from the establishment were in products which were primary to the industry.) The latter restriction was used to insure that all materials inputs consumed by the establishment were used to produce output for that specific industry, and avoid potential contamination from diversification in production. Such diversification in production may indicate a failure in the current definition of the industry, but such an examination would be beyond the scope of the current study.

3. See Anderberg (1973), Fisher (1969) and Hartigan (1975) for discussions of alternative clustering algorithems.

4. Abbott-Andrews (1993) presents the complete results of our clustering efforts. It includes the complete hierarchically structure broken into 20, 55, 139, and 200 clusters; as well as a cross-reference between the technological classification and the Standard Industrial Classification.

Table 1: Average Factor Shares

Variable	N	Mean	Std Dev	Min	Max
FUELS	456	0.00932	0.01832	0.0002	0.2082
ELECTRICITY	456	0.01376	0.01877	0.0011	0.2561
OTHER WORKERS	456	0.07477	0.04318	0.0058	0.2335
PROD WORKERS	456	0.12342	0.05691	0.0108	0.3740
CAPITAL	456	0.35297	0.10762	-0.0472	0.7492
MATERIALS	456	0.42576	0.13599	0.0877	0.8946

Table 2: Average Distance Between Industry Pairs

	N	Mean	Std Dev	Min	Max
Total	103740	0.3619	0.1547	0.0213	1.1102
Across	97170	0.3660	0.1520	0.0441	1.1102
Within	6570	0.3015	0.1799	0.0213	1.1049

Table 3: Closest Pairs of Industries Crossing Major Groups

RANK	IND1	IND2	DISTANCE
9	3499	3593	0.044088
20	**2542**	3493	0.051971
25	**2542**	3496	0.055211
27	3535	**3821**	0.055462
29	3452	3593	0.056240
34	**2542**	3495	0.058259
39	3645	3999	0.058831
40	3499	3644	0.059894
43	3569	**3821**	0.060364
45	3589	**3821**	0.060597
52	3532	**3821**	0.062740
62	**2542**	3431	0.064407
65	3324	3675	0.064790
70	3423	3568	0.065640
74	3554	**3821**	0.066162
83	3699	3829	0.067726
91	3676	3822	0.068977
94	3593	3644	0.069589
95	3317	3412	0.069803
96	**2542**	3452	0.069901
97	**2542**	3444	0.069927

Table 4: Information Retained in Aggregation

Number of Clusters	R-Square Optimal	SIC
456	1.000	1.00
369	.995	
329	.99	
275	.98	
200	.955	
139	.92	.58
55	.75	
20	.48	.27

Table 5: Industries with Distinct Production Technologies

2011 Fresh and frozen meat from animals slaughtered
2015 Poultry and egg processing
2021 Creamery butter
2041 Flour and other grain mill products
2044 Rice milling
2062 Cane sugar refining
2074 Cottonseed oil mills
2075 Soybean oil mills
2095 Roasted coffee

2141 Tobacco stemming and redrying

2411 Logging

2833 Medicinals and botanicals

2911 Petroleum refining

3295 Minerals and earths, ground or otherwise treated

3331 Primary copper

3398 Metal heat treating

REFERENCES

Abbott, T.A. (1992), "Alternative Classification Systems for the Year 2000: A Discussion," Proceedings of the 1991 International Conference On the Classification of Economic Activity. U.S. Department of Commerce, Economics and Statistics Administration, Bureau of the Census, p. 457-460.

Abbott, T.A. and Andrews, S.H. (1990), "The Classification of Manufacturing Industries: An Input-based Clustering of Activity," Sixth Annual Research Conference Proceedings, U.S. Bureau of the Census.

----------. (1993), "The Structure of U.S. Manufacturing: A Technological Perspective," Journal of Social and Economic Measurement. forthcoming.

Anderberg, M.R. (1973), Cluster Analysis for Applications. New York: Academic Press.

Andrews, S.H. and Abbott, T.A. (1988), "An Examination of the Standard Industrial Classification of Manufacturing Activity Using the Longitudinal Research Data Base," Fourth Annual Research Conference Proceedings, U.S. Bureau of the Census.

Chambers, R.G. (1988), Applied Production Analysis: A Dual Approach. New York: Cambridge University Press.

Fisher, W.D. (1969), Clustering and Aggregation in Economics. Baltimore: Johns Hopkins Press.

Gollop, F.M. (1986), "Evaluating SIC Boundaries and Industry Change over Time: An Index of Establishment Heterogeneity," Second Annual Research Conference Proceedings, U.S. Bureau of the Census.

Gollop, F.M. and Monahan, J.L. (1989), "From Homogeneity to Heterogeneity: An Index of Diversification," Technical Paper 60, U.S. Bureau of the Census.

Hartigan, J. (1975), Clustering Algorithms. New York: John Wiley.

Leontief, W.W. (1967), "An Alternative to Aggregation in Input-Output Analysis in National Income Accounts," Review of Economics and Statistics, August.

----------. (1986), Input-Output Economics. Second Edition, New York: Oxford University Press.

Ward, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, 58, 236-244.

DISCUSSION

David Wroe, Central Statistical Office
Great George Street, London SW1P 3AQ

As a student, when I first came across the concepts of industrial and products classifications, I naively accepted the idea that the industrial classification related to the nature of the production process going on in the business while the product classification related to the nature of the output. Alas, the distinction no longer seems quite as clear as it did. Indeed hearing different people talking about industrial classifications I am reminded of an incident in the autobiography of the English mathematician and philosopher Bertrand Russell whose life spanned much of the twentieth century. On volunteering to join the British army in the first world war, he was asked by the recruiting sergeant to give his religion. "Agnostic," Russell replied. "I'll put you down, sir, as Church of England," said the recruiting sergeant. "After all we all believe in the same god."

The three papers that have been presented bring out well the fundamental issues which beset the concept of an industrial classification. The papers are all based on US experience, and in that sense I fear that as an English person I may have been rash to accept the invitation to intrude into a dispute on this side of the Atlantic. But on the other hand the authors have been bold enough to expose their concerns about the US industrial classification. Moreover the issues they raise are important internationally.

The first of the papers, by Harvey Monk and Cynthia Farrar, provides a very useful historical perspective. The experience they describe illustrates very clearly problems which exist both in the present classification and in the way that the classification is used. That experience contains also many salutary warnings about the issues and pitfalls which those trying to produce better results need to address.

Work in the 1930's started on the basis that there were broad industrial categories such as agriculture, forestry and fishing, manufacturing, retail trade, etc and then different committees worked out detailed classifications within these broad categories. In particular the initial approach was essentially a top-down approach - with preconceived ideas about a basic structure, which might or might not turn out subsequently to fit well with the detailed groupings.

Each revision has accepted that the purpose of the US system has remained the same:

- promote comparability of data;
- facilitate collection and presentation;
- cover the entire field

There were also classification principles adopted by the Technical Committee charged with the revision. Maintaining the continuity of the major Federal statistical series was, of course, also another important consideration. The paper illustrates, among other points, the fact that often the data available fall far short of those required. For some sectors there was detailed information on the inputs and productive processes, in other cases there was at best information about outputs. Even so, it is worth noting that even in the manufacturing sector the categorisation of industry seems to have been related to "the products which define the industry".

The paper also describes how industry codes are assigned to establishments. It would be interesting to hear a little more about the basis on which the computer assigned codings are made, and in particular the extent to which the process is driven by the pattern of output, and, for example, what other data from the quinquennial censuses are particularly important. Another major problem referred to is the fact that different organisations are involved in implementing the classification - leading for example to lack of comparability in output and employment estimates for particular industries. One is tempted to conclude that the issue here is one of machinery of government, rather than a manifestation of weaknesses in the classification (though I have to admit that the necessary adjustments are not ones we have found easy to achieve in the United Kingdom).

The 1987 review sought to avoid radical changes to the classification. Such radical change, or at least a "fresh slate examination," is the subject of the paper by Jack Triplett. The committee established by the Office of Management and Budget is required

to give particular emphasis to the "conceptual foundations" of classification systems - in part with a view to improve data on services and to improve international comparability of industrial statistics -at least in relation to Canada and Mexico, though it is to be hoped that the committee will also look more broadly.

Very properly the committee, ECPC, are asking for what uses industrial statistics are required. In a sentence which is a model of restrained condemnation Jack Triplett tells us that asking this question marks perhaps the ECPC's "greatest departure from past work on classification". He contrasts two different approaches to industrial classification. The first focuses on the production process, so that establishmentss would be grouped together according to the production process, or economic activity, in which they are involved. The alternative approach focuses on the output of the establishment. Basically the two approaches correspond to the difference between, on the one hand, inputs and the production process and, on the other, outputs and their uses: in Jack Triplett's terms between "production oriented" and "demand based" concepts.

The paper demonstrates clearly that neither of these approaches has been followed consistently in the existing classification. As shown by the Canadian study referred to in the paper only a minority of industries are such that both approaches lead to the same result. This and the US results to follow are valuable, important contributions in this field.

How then should the ECPC proceed? Should they seek a conceptually pure approach? Are they worrying unnecessarily? Is it at all realistic to maintain just one or other of the two approaches? Could the data be collected? These are issues others will have views about, but as discussant let me first offer one or two thoughts.

As people who want to be taken seriously in government, in academic circles, and in the wider community, it is imperative that we try to answer the questions "What are these statistics for? What is this classification attempting to achieve?"

For some purposes users will be interested in the production oriented approach. For example, homogeneity in the production process is a major assumption underlying the use of input-output tables and related approaches in much of our work on short term estimates of gdp etc. The production orientated approach seems necessary there. On the other hand, in demand studies the process by which goods are produced is usually less relevant than the nature of the products. The simple-minded view that there should be one classification by activity, or production process, which we call an industrial classification, and a second classification of products, or of commodities, seems to me to be indisputable.

But before I try to develop this point a little further I would like to comment on some of the findings in the third of these interesting papers, that by Thomas Abbott and Stephen Andrews. They, implicitly at least, accept that the industrial classification should focus on the production process - the first of Jack Triplett's two alternative approaches. They then explore in an illuminating way whether in practice the US SIC succeeds in grouping together 4 digit industries according to whether they are similar technologically. The concept of distance they use derives from the vector of input shares - the proportion each input makes up among total inputs. This seems a very appropriate technique, with, as explained in the presentation as many as 365 different categories of input being distinguished. Presumably the more detail that is used, the more discriminatory the process. But while the technique would seem appropriate for rejecting the grouping of particular industries, it is not clear that we could accept that industries were similar (ie should be grouped) simply because the distance between the vectors of input shares was the same. Possibly the authors could say more about whether they found that the process occasionally grouped together industries which we would intuitively regard as rather different industries. However I should add that the authors claim only that the procedure is a device for raising questions, and not a device for determining which industries should be grouped. It is presumably an approach which could also be used at the establishment level as well as at the industry level. It would be helpful as well to know whether the authors consider that the same approach could be extended to the services sectors, possibly adapting

the input vector to give weight to the input of human capital to help in characterising different service industries.

However, even if the approach can be extended to services, we are still left with the question "Is the nature of the productive process really all that we are concerned about?" In other words, if two industries have the same inputs when described in a certain way, should we regard them as the same industry even if the outputs are different? I would question seriously whether in practice users would want to regard two such industries as identical. This leads then to the question of whether we can classify the production process without some regard to the output.

Perhaps this point, though, offers us a way of addressing some of the other issues the papers raise? The industrial classification should I suggest distinguish different productive processes. But on the basis of the results in the paper perhaps one should conclude that this approach can be applied only at a relatively detailed level - perhaps at the four digit level. Wherever possible, statistical results should be made available at that detailed level for those with the need or appetite for detailed results. For those many who require a more summarised approach, for example with groupings into broad categories such as manufacturing and retailing, possibly we should accept that broader groupings will have to be based on the nature of the product, having regard certainly to demand based considerations, though again I question whether these latter would capture entirely all the distinctions which are being sought eg between agiculture and manufacturing, between manufacturing and services etc.

It would be illuminating to see how far a conceptually consistent approach starting from the production oriented approach takes us, and I very much hope that we shall have an opportunity to find out. The three papers, taken together, perhaps reinforce the view that conceptually pure approaches may provide only a good start, but will not alone get the ECPC to an outcome which meets the bulk of users' requirements. This is perhaps what the Business Research Advisory Committee to the BLS was saying to the ECPC.

No doubt the authors will tell me if I am misusing their results or am underestimating the force of some of their arguments. Before they do so I would however like to thank them for three very clear, useful and, above all, stimulating papers.