MULTIPLE CLASSIFICATION SYSTEMS FOR ECONOMIC DATA: CAN A THOUSAND FLOWERS BLOOM? AND SHOULD THEY?

by

Robert H. McGuckin*

CES 91-8 December 1991

Abstract

The principle that the statistical system should provide flexibility--possibilities for generating multiple groupings of data to satisfy multiple objectives--if it is to satisfy users is universally accepted. Yet in practice, this goal has not been achieved. This paper discusses the feasibility of providing flexibility in the statistical system to accommodate multiple uses of the industrial data now primarily examined within the Standard Industrial Classification (SIC) system. In one sense, the question of feasibility is almost trivial. With today's computer technology, vast amounts of data can be manipulated and stored at very low cost. Reconfigurations of the basic data are very inexpensive compared to the cost of collecting the data.

Flexibility in the statistical system implies more than the technical ability to regroup data. It requires that the basic data are sufficiently detailed to support user needs *and* are processed and maintained in a fashion that makes the use of a variety of aggregation rules possible. For this to happen, statistical agencies must recognize the need for high quality microdata and build this into their planning processes. Agencies need to view their missions from a multiple use perspective and move away from use of *a* primary reporting and collection vehicle.

Although the categories used to report data must be flexible, practical considerations dictate that data collection proceed within a fixed classification system. It is simply too expensive for both respondents and statistical agencies to process survey responses in the absence of standardized forms, data entry programs, etc. I argue for a basic classification centered on commodities--products, services, raw materials and labor inputs--as the focus of data collection. The idea is to make the principle variables of interest--the commodities--the vehicle for the collection and processing of the data. For completeness, the basic classification should include labor usage through some form of occupational classification.

In most economic surveys at the Census Bureau, the reporting unit and the classified unit have been the establishment. But there is no need for this to be so. The basic principle to be followed in data collection is that the data should be collected in the most efficient way--efficiency being defined jointly in terms of statistical agency collection costs and respondent burdens.

Keywords: SIC, Longitudinal Microdata, Economic Classification

[*] The opinions and conclusions expressed in this paper are those of the author and do not necessarily represent those of the U.S. Bureau of the Census. This paper was presented at the 1991 International Conference on the Classification of Economic Activity in Williamsburg, Virginia,

November 6-8, 1991. Special thanks are due to Mark Doms, Timothy Dunne, John Haltiwanger, Sang Nguyen, and Jack Triplett for their comments and insights. Rebecca Turner provided excellent typing skills.

I. INTRODUCTION

More than 35 years ago Daniel Suits remarked that "[T]he proper evaluation of classifications can only be made in terms of the objective to be achieved by the use of the resulting classes, and different objectives generally require different classifications."[1] A more recent commentator raises these issues again in discussing the current SIC. Jack Triplett (1990) argues that "many--though not all--of the criticisms of the SIC reflect real problems that arise from its lack of a positive conceptual structure." He goes on to say that there are multiple legitimate concepts for classifying economic data and argues that for particular uses "one must choose one SIC concept."[2]

This paper discusses the feasibility of providing flexibility in the statistical system to accommodate multiple uses of the industrial data now primarily examined within the SIC system. "Multiple uses" refers to groupings of the data based on different aggregation concepts or rules. For example, a simple aggregation concept might group products in order of establishment energy usage or classify firms by age. More complicated concepts include groupings of establishments based on the cross-elasticities of supply for their products or groupings

[1] This quotation opened Daniel Suits' comment on the paper "Census Principles of Industry and Product Classification, Manufacturing Industries" presented over 35 years ago at an NBER Conference on Business Concentration and Price Policy (1955).

[2] Emphasis in original.

of products involving cross-elasticities of both demand and supply. The important point is that the rule(s) for grouping the data are determined by the use for which the data are being developed.

In one sense, the question of feasibility is almost trivial. With today's computer technology, vast amounts of data can be manipulated and stored at very low cost. Once the basic microdata are collected and processed, it is technically straightforward to provide numerous reconfigurations of the data.[3] Moreover, reconfigurations of the basic data are very inexpensive compared to the cost of collecting the data.

Flexibility in the statistical system requires that the basic data are sufficiently detailed to support user needs <u>and</u> are processed and maintained in a fashion that makes the use of a variety of aggregation rules possible. Viewed in this way, the provision of flexibility to satisfy user needs is not a trivial exercise.

[3] Perhaps the main constraint on the extent to which the system can provide flexibility arises from the need to maintain confidentiality for individual reporting units. There are two interrelated factors involved, the level of detail that can be released and complementary disclosures. If there is only one classification system, then it is relatively straightforward to develop disclosure standards. Such rules have been employed successfully by the Census Bureau for many years.

Once multiple systems are introduced, complementary disclosure must be considered. This refers to the fact that once aggregation is released, not only does a new aggregate have to pass the disclosure rules viewed as if only one aggregate is released, but it also must be evaluated to ensure that the earlier release, coupled with the new release, do not together permit disclosure of individual data. Discussion of this issue is beyond the scope of this paper.

Although the categories used to report data must be flexible, practical considerations dictate that data collection proceed within a fixed classification system. It is simply too expensive for both respondents and statistical agencies to process survey responses in the absence of standardized forms, data entry programs, etc. I argue for a basic classification centered on commodities--products, services, raw materials and labor inputs--as the focus of data collection. The idea is to organize the collection process around the principle variables of interest--inputs and outputs.

Lists of services and commodities produced and consumed by economic entities provide a straightforward vehicle for the collection and processing of the data. For completeness, the basic classification should include labor usage with some form of occupational classification. Detailed information on worker type is not currently collected in establishment surveys. For example, production and non-production workers are the only breakout identified in the censuses and surveys in manufacturing. This level of detail needs to be greatly expanded.

In most economic surveys at the Census Bureau, the reporting unit and the unit of analysis have been the establishment. But there is no need for this to be so. The basic principle to be followed in data collection is that the data should be collected in the most efficient way--efficiency being defined jointly in terms of statistical agency collection costs and respondent burdens.

For analysis purposes, the outputs and inputs should be linked to the most detailed production unit at which inputs are transformed into outputs. For many inputs and outputs, the analysis unit will remain the geographically distinct establishment. In areas such as pipelines and banking characterized by network structures, the establishment concept may have to be modified. Determining the appropriate reporting and analysis units is important and difficult. Yet, it is logically independent of the basic input/output classification system.

In Section II, I argue that the statistical systems used for economic surveys are not organized to provide sufficient flexibility to satisfy user needs. The product/materials data that most closely corresponds to the input/output classification system proposed here are inadequate. Simply put, too little emphasis has been placed on the basic microdata. I attribute this to a widespread failure to recognize the importance of microdata analysis. Because of this, an appendix extending my brief comments in the text has been added. Section III discusses the input/output classification system proposed for data collection and processing. Section IV provides some comments on the appropriate reporting unit and briefly reviews the main problems in implementing aggregation rules to categorize production units (e.g., establishments). Section V concludes by observing that while the focus of data collection efforts and the way economic data are processed would change dramatically with the proposed system, it is likely that the statistical system would continue to primarily report data with two principle classifications--industry and commodity based systems. Despite the similarity of reporting systems, the quality of the reported statistics as well as a variety of new statistics will be different under the new procedures.

II. FLEXIBILITY IN THE STATISTICAL SYSTEM FOR ECONOMIC SURVEYS

The statistical system for economic surveys exists to satisfy user's needs. To best serve users, a wide range of products and the ability to serve special needs and requests are necessary. Statistical agencies now offer a wide range of products designed for many different uses (and users). For example, the Census Bureau publishes economic data on both an industry basis (SIC) and a product basis. The Census Bureau also provides categorizations of the data by size and geographical location as well as on the basis of establishments and enterprises. Special groupings of the data on a cost reimbursable basis are also provided.

Despite the existence of the basic attributes of a flexible system--a wide range of products and the ability to serve special needs and requests--users are not satisfied with economic statistics.[4] While specific complaints vary greatly, it appears that the system is not flexible enough. Flexibility requires a very detailed classification scheme for the basic "commodities"--the things to be aggregated. It also requires high quality microdata. Economic statistical programs have neither attribute in sufficient quantities to satisfy user's needs.[5]

Product/Material Detail in the SIC

The current SIC system groups establishments into industries based on four levels of hierarchical detail. As an example, meatpacking has the four-digit SIC 2011, which is part of the two-digit SIC 20, food. SICs beginning with 2 and 3 refer to manufacturing, while other digits in the first place signify sectors such as agriculture, transportation, and services. While there have been numerous revisions, this basic structure has

[4] See Triplett (1991) and the various articles cited.

[5] While commentators have cited many factors as sources of inadequacies in economic statistical programs, I stress three interrelated factors in explaining the existence o f inadequate microdata. First, as discussed by Triplett (1991), statistical agencies have not emphasized studies that evaluate the usefulness of the statistics developed and reported in the SIC system. Evaluation studies require active programs involving the microdata and regular contacts with analytical users of the data.

Second, and reinforcing the first factor, most economic surveys and programs focus on cross-section statistics. The cross-section perspective means that once the data for a particular year are aggregated and published, there are few incentives to work with the microdata. Relatedly, an explicit longitudinal perspective in data collection would improve data quality--through improved imputation and editing possibilities--and expand the range of products that agencies could offer.

Finally, efforts to satisfy customer demands are, for the most part, made within the context of the existing SIC. This has led to lower quality microdata for use in implementing aggregation concepts than would be expected if basic commodity lists were the focus of data collection and processing.

remained since the early 1950s, and similar systems are in use throughout the world.

The SIC system is an industrial classification system structured primarily around a supply side perspective that classifies establishments using similar production processes. Nonetheless, there seems little doubt that the current SIC system represents somewhat of a hodgepodge of principles reflecting many compromises and adjustments. Recent articles by Abbott and Andrews (1988, 1990), and Triplett (1990) amply demonstrate this point.

The Census Bureau has developed a complementary system of product and kind of service classes for use in data collection and some reports. The product codes, at least for manufacturing, are tied hierarchically to the SIC system. The product system is based on seven-digit product codes aggregated into five-digit product classes that are grouped within the four-digit SIC industries. There also exists a materials or input classification system characterized by six-digit codes. The product and material codes represent an extension of the SIC in the sense that they group the basic lists of seven-digit products according to the industry primarily responsible for their production.[6]

[6] Product information is collected in the Census of Manufactures (CM) and the Current Industrial Reports (CIR) program. The product detail collected in the CM supplements the variety of special CIR surveys, many of which were originally developed and funded by specific private agencies.

A key point is that the product and materials classifications are only loosely linked together and are not maintained in forms conducive to multiple use. While both of these commodity classifications are tied hierarchically to the SIC system, the files are processed separately, and a standardized concordance between the systems is not available. Linkages between the two can be made at a three- or four-digit level of industrial detail, but to do so requires a laborious hand process. This limits the possibilities for analysis. Moreover, the product-materials classifications are not maintained in a way that facilitates comparisons through time. One would envision a much different situation if the basic commodity information were the primary focus of data collection and the SIC was viewed simply as one of many possible reporting frameworks.

Lack of Priority for Microdata

The importance of microdata and of developing flexibility in the statistical system was noted by Solomon Fabricant (1955) in connection with comments on the SIC: "...[T]he Bureau [of the Census] can produce a valuable body of source material merely by making new arrangements of data now in its files. Generally speaking, this means providing breakdowns and cross classifications of various sorts. Aggregates are only the beginning of information." Yet 15 years later, F.M. Scherer

(1980) complained, "The data, collected at an expense of tens of millions of dollars, lie unanalyzed in Census Bureau files. Though less apt to draw headlines than Congressional junkets and the overpayment of welfare recipients, this state of affairs is equally wasteful."

Part of this problem can be traced to confidentiality issues which prevent widespread access to the underlying data for analytical users at statistical agencies. The research and data development program at the Center for Economic Studies (CES) now supports a good deal of the research and data work that Scherer and Solomon were concerned with, including the creation of longitudinal microdata files. Moreover, there are similar longitudinal microdata files being developed all over the world.[7] Thus, arguing that the statistical system pays too little attention to the microdata may be unfair, particularly in light of legitimate confidentiality issues. However, despite the creation of CES and recent improvements in survey design, the emphasis given to analysis and the underlying microdata in statistical processing--particularly longitudinal linkages and the basic product and material classifications--is still too low. As Triplett (1992, forthcoming) has forcefully argued, uses of the data beyond primary sponsor specified and supported aggregations are not emphasized at statistical agencies.

[7] Canada, the Netherlands, Israel and France are some examples of countries involved in these efforts.

A major difficulty in using the microdata for both analysis and publication in alternative forms is that the data collection and processing systems are geared to the production of cross-section statistics within the SIC system. If the major focus of data collection is centered on a particular reporting system, in this case the SIC, and little emphasis is placed on alternative classifications or longitudinal data products, then there is little reason to focus on the underlying microdata. In turn, alternative classifications of the underlying microdata are either impossible or expensive to produce.

For example, consider the response of an agency sponsor at a Census Bureau's Annual Research Conference (ARC) when the quality of the microdata used to develop national research and development estimates was questioned. It was argued that criticism was unfair because the goal of the survey was a national estimate, not use of the microdata to analyze, in this case, the relationship of productivity to R&D spending. This view is not untypical of economic surveys, which are often designed to produce a set of aggregate cross-section statistics for macroeconomic analysis.

As another example, until fairly recently there was little attention to maintaining information on the edits and imputations undertaken in the processing of economic survey data. Corrections were made to totals at the time of publication and these adjustments were never carried back to the microdata.

Similarly, longitudinal linkages tying reporting units together over time have not received much attention in processing. Thus, even when analytical users had been granted access, their ability to carry out microdata based research was severely hampered. While the situation is better today, flexibility with its concomitant emphasis on the microdata still needs improvement.

The importance of the microdata collected in the various censuses and surveys cannot be overstated. Not only are they the source of the aggregate statistics produced by the system, they provide the basis for evaluating the usefulness of the statistics in the analysis of particular issues. Microdata also serve as the raw material for research and new statistics not envisioned when the original data are collected. (Because I think the usefulness of microdata is often not appreciated, I have added a brief Appendix illustrating these points.)

It is also important to recognize that not all the uses to which the data will be put, and consequently the precise aggregations necessary, can be decided in advance. And they don't have to be. Once the data are collected, it is technically straightforward to develop new groupings of the information if accessibility to the microdata is maintained.

The flexibility problem is not simply a Census Bureau problem or unique to it. The data collection process involves all the statistical agencies, and Census Bureau collection strategies are driven by the demands of other agencies for data to support the national income accounts and productivity measures. But the data, once collected, can be used for a wide variety of research and policy purposes. In addition, the microdata support studies that evaluate the aggregate statistics generated at the Census Bureau. If the detailed microdata are not a part of a statistical agency's missions, then the quality of the microdata is likely to continue to be a low priority item.

III. DATA COLLECTION IN A MULTIPLE USE ENVIRONMENT

The detail and quality of the microdata collected in economic surveys are not now sufficient to satisfy all the demands of users. A change in perspective--from making the SIC system the main focus of data collection to more reliance on basic classifications of commodities--would tend to produce the higher quality microdata necessary to support both a wider range and a higher quality of statistical products. The principal variables of interest--the commodities--would represent the finest level of detail in the statistical system. Adoption of a commodity-based system would not preclude the use of the current SIC system (with some modifications to account for the greater level of commodity detail) for particular uses.

A commodity classification based on the inputs and outputs used in the production of goods and services offers a useful way to organize the collection of data. In fact, inputs to any process actually represent outputs of other processes. For example, natural resources used in manufacturing are produced by the mining and extraction industries.

In principle, these commodities are the most basic economic units of economic theory. They represent a detailed specification of a transaction. In practice, the commodities would likely consist of a list of goods and services and, as suggested in the introduction, a classification of labor types. Inclusion of labor types among the basic inputs to the system would mean that occupational information could be integrated with materials, energy inputs, and purchased services. This basic list would necessarily incorporate prior aggregations across time, space, and characteristics. For example, blue and black ball point pens, sold for delivery today and next week, at Suitland and in the District of Columbia would likely be grouped as one commodity.

Construction of these basic lists is not trivial. Nonetheless, records of product sales and purchases of specific materials would likely be kept by business entities in great detail. Such lists are currently used in collecting foreign trade data. Many private organizations keep very detailed product listings. For example, CorpTech has a list of over 3,500 high-tech product codes. According to the CorpTech Handbook (1986), the list was developed because of "the inadequacies of the SIC codes in relation to high technology products." Also, the Bureau of Labor Statistics (BLS) maintains commodity lists to use in reporting producer price indexes.

An advantage of developing consistent product lists is that as new products are introduced, they would more easily be identified then they are today. This would require continuous monitoring of trade association and other lists. Such information would also need to be augmented with survey information that could be obtained in the context of record-keeping surveys. Finally, the basic commodity lists would need to be constructed and maintained in a longitudinally consistent way.

The controversies associated with the SIC system would likely be reduced with the adoption of a basic commodity-oriented collection system. Reporting classifications would still generate some controversy, particularly if user's desires had to be prioritized because of confidentiality concerns and information availability. But with flexibility and some expansion of existing opportunities for access to the microdata for evaluation studies, I anticipate that most user's needs could be satisfied.

I see little in the way of disadvantages. The proposed design could be implemented in manufacturing by simply focusing more processing resources on the input/output basics. Developing economy-wide lists of commodities would require some difficult analysis of output measures in the service, transportation, banking, and other industries. Measuring output by deflated revenues is not likely to be satisfactory. In addition, a real problem in service sectors such as banking is deciding on the importance of geographical detail about the distribution of services that is needed. These problems would have to be faced irrespective of the processing system.

I realize that in focusing my comments on development of a basic commodity list, I have bypassed hard choices with regard to which aggregation concept(s) apply. I think that a demand side aggregation rule that groups commodities according to their substitutability makes most sense for developing the basic commodity classifications. However, the basic groups must be quite detailed. The problems associated with the classic example of beet and cane sugar will not go away just because these two products are perfect substitutes from the demand side. Separate categories for each product are necessary if their production functions are very different and users are interested in this difference. Similarly, steel and aluminum, cans and bottles, wood and metal desks and a host of other substitutable products would need to be treated as separate commodities. To give some idea of the kind of detail that is necessary, consider that in examination of "high-tech" trade, the Census Bureau develops information on over 25 different semi-conductor chips: Some are advanced technology products and some are not.

As these comments suggest, the implementation of conceptual frameworks to group the basic commodities would not be easy. Purely demand-side systems could be developed from clustering algorithms based on information on prices such as that collected to develop producer price indexes. See Jaditz (1990) for an example of this type of procedure. While it appears relatively straightforward to aggregate products on the basis of demand substitutability, practical problems occur.

A quick review of the ongoing and voluminous debate on how to define markets for antitrust purposes will be enough to give anyone pause who seeks to come up with a simple rule for grouping products according to a demand-based aggregation concept.[8] The problem with clustering or correlation analysis of prices is that without good time-series data, spurious correlation is a major problem. In many cases, a hedonic approach could help. These problems do not mean that grouping the commodities on the basis of substitution possibilities should be abandoned. They do mean that the classification should be detailed enough to minimize the possibility that close substitutes are not placed in the same classification if they are produced with very different input structures. I think the sugar example suggests that some analyst judgement will be necessary.

[8] See, for examples, "Symposium on Mergers and Antitrust" (1987), also Werden (1983) and White (1987).

To summarize, the advantages of the commodity list approach are the following:

1. Data collection is focused around the basic unit of economic analysis - the commodity.
2. Data collection and processing are based on a relatively stable classification system.
3. Multiple aggregation rules can be applied to group commodities, establishments, firms, industries, geographic areas, energy usage, etc.
4. Historical comparability can be maintained even if desired aggregate groupings change, since the entire historical series can be retabulated. Thus, desired changes in classifications can be accommodated without destroying historical time series.
5. Comparisons to international and other classifications of data can be made, since any aggregation concept can be applied to the product-based system.
6. A wide range of new data products can be accommodated because of the longitudinal structure of the basic data.
7. Policy makers and other users will determine the primary way of how the data will be reported, but the collection will be in the hands of statistical authorities.

8. Evaluation studies of statistical products will involve substantive studies of their usefulness by analytical users and statistical authorities.

9. Problems associated with emerging industries--such as the failure of the system to provide a separate classification for computers--could easily be avoided.

With detailed lists of commodities, the focus of collection activities, statistical agencies still need to determine the classification and reporting units and what level of detail is obtainable. The choice in both instances involves considerations of how much detail is maintained in the records of individual respondents and how much of a reporting burden would be required. We turn to these issues next.

IV. THE UNIT OF ANALYSIS

In discussions of classification issues, one must distinguish between the classification unit and the reporting unit. The reporting unit refers to the point of contact for collection and need not be the same as the classification unit, the unit of analysis. If, for example, a multi-establishment firm maintains records for each establishment at central headquarters, this may mean that reporting forms and contacts between the statistical agency and the firm are best handled at the firm level. If the records are kept at the establishment, then this may be the best place to obtain the data. It may make good sense to collect at both levels. The basic principle to be followed is that the data are collected in the easiest way--easiest being defined jointly in terms of agency collection costs and firm reporting burdens.

Traditionally, the reporting unit and the unit of analysis for most economic surveys have been the establishment, although the firm has been in some instances. The establishment is the smallest reporting unit for economic statistics, representing a distinct physical location and set of activities "of concern in management policy decisions."[9] This definition, it should be noted, does not imply that management authority necessarily resides at the establishment, only that the establishment is an economic unit in the sense of a production function for a good or service.[10]

Since I see the problem with the current system as a lack of sufficient detail, I do not think a move away from the establishment, or its equivalent, as the basic unit of analysis in most industries should be made unless absolutely necessary.[11]

[9] Conklin and Goldstein (1955), pp. 21.

[10] As noted earlier, in some instances the Census Bureau splits an establishment if two or more distinct activities are carried on at the same location.

[11] There are many reasons to seriously consider reorganizing processing procedures along the Canadian lines so that data collection and processing is more closely tied to the various activities of a company and its subsidiaries. Discussion of this issue is well beyond the scope of this paper.

In addition to its advantage in terms of detail, the establishment represents a fixed location, independent of ownership status. This has a number of advantages in maintaining linkages across time. If firms or divisions are made the unit of analysis, the quality of the linkages will depend heavily on the agency's ability to track changes in ownership and management structures.

For industries such as pipelines or those with complicated networks such as banking, adjustments to the physical location criterion may be necessary. However, without some detailed consideration, including empirical work, of how to measure outputs in these industries, any change seems premature. Moreover, for those industrial sectors in which the establishment concept remains valid, it seems counter-productive to move away from it. The operable principle is that the link between the inputs used to make the output(s) of a production process, whether it be a bank transaction or the manufacture of a widget, should be measured as close to the level appropriate for production function analysis as possible.

Aggregating Production Units

While commodities should be the basic unit of analysis, a very important category of aggregations involves grouping and classifying production units. The desire for a classification that links together demand and supply arises from the most basic issues of interest to economists: What determines the behavior of an industry, including the products that it sells, the size and efficiency distribution of its productive units, the prices charged, and the factors determining its rate of technological advance? As Triplett (1990) states, "It is inevitable that the information so collected ... links inputs and outputs in an explicit or implicit production-oriented way."

A number of the difficulties in applying aggregations associated with productive units can be traced to the fact that establishments produce multiple products. If each productive unit produced products that are close substitutes, then the problem of defining commodity classes and industry classes would be the same: Find an appropriate commodity class and assign each producer uniquely to that class. Unfortunately, products and industries do not line up this way. Many establishments produce multiple products, not all of which would normally be grouped into a class of competing products.[12]

[12] In practice, SIC industries have been defined in ways that ensure that the primary product specialization ratio is relatively high so that establishments in the same industry produce closely related products. However, even when establishments are grouped according to primary products, recent research by Streitwieser (1991) indicates that the secondary products produced in these establishments are not similar. On the other hand, 72 percent of manufacturing output and 85 percent of establishments have primary industry specialization ratios above 90 percent.

The production of multiple products at establishments would not be a problem if one could allocate the factors of production to the various outputs of the establishment. Under this circumstance, it would be possible to split the establishment into components and to aggregate inputs by each of its outputs. This procedure is, in fact, done for some manufacturing establishments. But this is not possible without imputation or estimation rules because, in many cases, businesses do not keep sufficiently detailed records on input usage. Thus, practical considerations suggest it is difficult, if not impossible, to develop a conceptually clean aggregation linking products and inputs.

Problems also arise in productive unit aggregations based on input usage. In principle, it would be possible to start with the assumption that technologies can be differentiated by the set of inputs they use. For this approach to be theoretically valid, the production technology must conform to assumptions similar to those employed in the input-output model. The input-output model specifies one of the simplest production functions: A fixed factor input-output relationship between a homogeneous output and a set of inputs. Neither the homogeneous output nor the lack of substitution possibilities among inputs is likely to be satisfied in practice.

This approach is used in Abbott and Andrews (1990). They employ clustering algorithms--statistical techniques that group data based on the "distance" between input vectors--to group establishments by input usage with some success. Even though they work at a four-digit SIC industry level, rather than at the more detailed commodity level, and ignore labor, they derive reasonable groupings. A variant of this methodology has also been employed by Gollop and Monahan (1989 and 1991) to develop indexes of diversification based on the "closeness" of an establishment's products.

Despite the impossibility of cleanly implementing economic classification concepts, such concepts can and should be used in practice. The application of conceptual frameworks to real world data involves analyst judgment. If the basic commodity data are available, various analytic procedures can be used to guide analysts in deriving groupings of the data. In fact, the impossibility of adopting a completely algorithmic approach makes it critical that a conceptual framework be in place to guide analytic judgments. The important need is for detailed microdata--a sufficiently detailed commodity classification system and a narrowly defined economic unit--maintained in a longitudinally consistent way. Not only will multiple aggregations be supported, but the data for evaluation studies of the usefulness of the aggregations will be available.

V. CONCLUDING COMMENTS

More than 35 years ago two Census Bureau employees, Maxwell Conklin and Harold Goldstein (1955), described in great detail the dilemmas associated with devising a classification system on the basis of the two main groupings of aggregation concepts outlined by Triplett (1990): demand-side (product-based) and supply-side (input-based) systems. Conklin and Goldstein described the compromises inherent in the new SIC in terms of a third aggregation concept, one which plays a prominent role in industrial organization analysis, including competition policy. It involves the definition of an economic market that takes into account considerations of both supply- and demand-side concepts. This aggregation rule is described by Abbott and Andrews (1990) as a multiple indicator's approach because it involves the application of more than one aggregation rule.

I find it interesting that many of the examples used today to illustrate inconsistencies in the current SIC system were used in the Conklin and Goldstein article and accompanying commentary. For example, Conklin and Goldstein use beet and cane sugar and tin cans and glass containers to illustrate demand side substitutes included in different industries because the classification system groups on the basis of homogeneity of production. They illustrate the existence of non-competing products in the same industry with drill and lathe presses. This suggests that the basic problems associated with using one system for all purposes have not changed over time.

Agencies need to view their missions from a multiple use perspective and build flexibility into the statistical system. Flexibility requires high quality microdata. It also means that the almost exclusive reliance on the current SIC as both the primary reporting and collection vehicle must change.[13]
A practical approach to achieving flexibility is to use a well-defined commodity-based classification as the basis for data collection.[14] Such a system consistently maintained over time would permit the implementation of many aggregation rules, since it is clear that reconfigurations of the underlying data are inexpensive.

Moreover, it is clear that many data users will continue to use a basic industrial system similar in form, if not in content, to the current SIC. A supplementary commodity reporting system will also likely be maintained. Both systems are part of current

[13] More emphasis on flexibility should tend to reduce the incentives to develop collection strategies focused on narrowly defined uses and users in primary economic surveys such as the Annual Survey of Manufactures (ASM). ASM procedures are designed to produce a specific set of cross-section industry statistics, with most emphasis placed on output. While industry output is arguably the most important statistic produced in the survey, resources are focused on the largest establishments since the smaller establishments contribute little to industry output. The ASM design reduces the quality of the information to examine such things as the growth and survival of small firms. In fact, the decision to reduce over 20,000 establishments from the sample in 1979 was undertaken because they were not needed for an accurate estimate of industry output. No other objective appears to have been seriously considered at that time. See Waite and Cole (1989).

[14] In fact, though inadequate, a basic product and material system is now used at the Census Bureau. Moreover, the original SIC was developed from commodity lists created in the mid 1940s.

reporting at the Census Bureau.[15] Thus, the focus of changes envisioned in this paper are directed to the collection and processing systems. Moving to a commodity based system will require more attention to the principal variables collected, the microdata, and concomitantly provide flexibility. This will greatly expand the range of data products available and will improve the quality of important national statistics in all areas.

Finally, I think it is important to recognize that in drawing up sampling designs and collection procedures agencies tend to focus on their most important users--sponsoring Government agencies. Given limited resources, collection strategies are adopted to minimize sample requirements for the particular objectives of the sponsoring agency. If broader goals are desired, then the tradeoffs--in terms of variance increases for a primary statistic such as output--that are required to satisfy the needs of users seeking to understand small business growth need to be built explicitly into the decision framework.

[15] Many other statistical agencies, including Statistics Canada, report on both an industrial and commodity basis.

Appendix - A Note on the Importance of Microdata

Aggregations of the basic commodity information help to reduce the myriad of individual detail to manageable proportions. For example, a time series of 30 industry observations, rather than 30 times 100, or 3000 individual establishment observations reduces the complexity of the analysis dramatically. Unfortunately, in this aggregation process information is lost. Microdata are of crucial importance to the evaluation of the usefulness of particular aggregations of the data.[16]

For some problems this loss of detail may not matter in the sense that the phenomena under study may be sufficiently understood without reference to the underlying microdata. The difficulty is that without analysis of the microdata it is virtually impossible to evaluate the extent of the aggregation error.[17] Moreover, since the economy changes which are perfectly acceptable aggregate measures at one point in time may be misleading in another point in time. This means that aggregation

[16] The size distribution of business units is highly skewed with a small number of large units accounting for large proportions of output. This makes public use microdata files virtually impossible to create. See McGuckin and Nguyen (1991).

[17] If the 'industry' is the true unit of analysis, then this statement is too strong. Industry output is viewed as a draw from a hypothetical distribution of outputs which are subject to some random error generating process. In this case, the variance of the output serves as a measure of heterogeneity associated with the representative industry. Unfortunately, economic model building does not usually use the industry (the aggregate) as the basic decision unit. Rather, the firm of establishment is used and thus the criticism is valid.

must be approached with substantial caution and continually be reevaluated.

Evaluation of Statistics: Bias Issues

It is not easy to provide a general demonstration of the size of aggregation bias because the extent to which aggregation bias is present is model specific.[18] That is, the exact error depends on application or use of the data. In earlier work, McGuckin (1990), I argued that the homogeneity of establishment behavior assumed in empirical studies based on aggregate data is not evident in the detailed data. A legitimate response is that this showing is not determinative since even though the behavior of the individual units to be aggregated is idiosyncratic, the covariances between the omitted establishment-level variables and the aggregate variables in the model are so small that a bias in the estimated relationship is negligible. Unfortunately, without good quality microdata, it is impossible to verify this type of conjecture. Moreover, correcting for such measurement error is difficult at best without access to the microdata.

[18] My own interest in aggregation stems from a lucid article by Theil (1957) which first made me realize the potential importance of aggregation error in the statistical analysis of economic models. Work with the microdata over the last 5 years at CES has rekindled my interest in specification issues and convinced me that aggregation is a much bigger problem than is generally realized. "How big?" is still an open question, a question that continues to drive a substantial portion of the research agenda at CES.

While we do not yet have definitive answers, recent work at CES goes beyond the simple demonstration of idiosyncratic behavior. We now know that in certain circumstances the aggregation biases can be large. For example, recent work by Olley and Pakes (1991) finds substantial differences between estimated productivity relationships in the recently deregulated telecommunication industry depending on whether the micro or macrodata are used. As another example, consider the behavior of inventories, an important indicator variable in business cycle analysis. As documented by various authors, the behavior of measured inventories appears to defy economic logic. While explanations have been put forward for unpredictable inventory movements, recent work by Schuh (1991) with microdata indicates that firms have two very different types of behaviors--some "smooth" production and some "bunch" it.

Compositional Effects

The problem with exclusive use of aggregate statistics is not simply one of aggregation bias in the sense of inferior estimates of economic relationships such as the elasticities of a production function, inventory adjustment coefficients or wage equation parameters. With aggregate data alone it is impossible to examine the differential effects of policies on the entities classified within the aggregate. Examining individual changes is necessary if particular components of an aggregate movement are important.

In a related vein, McGuckin and Peck (1991) find that the average manufacturing industry had over a third of its measured output changes associated with establishments switching industry classification in the 1981-1982 period. Preliminary evidence suggests that the reclassifications tend to occur in census years and when Annual Survey of Manufactures (ASM) panels change, suggesting that the sampling procedures are introducing spikes in the aggregate output series. Moreover, while some of this change is associated with processing lags for new entrants, the vast majority of the effect is linked to relatively large establishments. Of particular interest in terms of whether industry output change reflects a representative firm, there is no correlation between the growth of the reclassified establishments and those continuing in the industry.

The Importance of Longitudinal Microdata

The importance of examining individual changes with longitudinal data is illustrated in recent work on job turnover by Haltiwanger and Davis (1990, 1991). They find that establishments characterized by job creations and destructions have very different behavioral patterns and that these patterns are important in both a time-series (business cycle) and cross-section (across establishments and industries) sense. The importance of such gross change measures is also supported by the work of Dunne, Roberts, and Samuelson (1988, 1989) dealing with the entry and exits of firms and plants. They also find very different behavior across types of entrants, suggesting important compositional effects.

While the mechanisms at work are not yet completely understood, there are several reasons for expecting the gross changes to be important measures of economic impacts. First, change typically requires resources and therefore measures of gross change provide a basis for measuring and understanding such costs. Second, since change affects performance, measures of change that differ among economic units provide important information on competitiveness. Finally, as mentioned earlier, it is important for policy analysis to know how broadly based is the behavior reflected in an aggregate statistic. Change measures--which represent new aggregations of the underlying microdata--capture the heterogeneity of establishment behavior within individual SICs and thereby provide useful policy information.

Non-"Industry" Aggregation Rules

The above examples take as given the SIC and focus on the issue of aggregation bias and what additional information can be extracted from the underlying microdata. But, consider a different type of aggregation, one that for example simply ranks establishments by their energy intensiveness. This kind of tabulation might group all producers into high, medium, and low energy consumers on the basis of their BTUs per unit of output. While this type of classification might be satisfactory for some questions--for example, how concentrated is the distribution of energy usage?--for other questions the information is next to useless. Suppose that the real issue of concern is the size and structure of producers with high sulfur emissions. Then a focus only on large energy users is inappropriate since sulfur emissions depend on the source of the energy--coal is high, nuclear is low--and the type of pollution abatement equipment in use at the establishment.

One can construct innumerable examples of this sort in which the obvious way to examine the data is through a reclassification or aggregation of the underlying microdata. This example which plays on today's environmental concerns seems obvious. But, would it have been obvious 10 or 20 years ago?

REFERENCES

ABBOTT, THOMAS A. III, and STEPHEN H. ANDREWS, "The Classification of Manufacturing Industries: An Input-Based Clustering of Activity," Center for Economic Studies Discussion Paper #90-7, August, 1990.

ANDREWS, STEPHEN H. and THOMAS A. ABBOTT III, "An Examination of the Standard Industrial Classification System of Manufacturing Activity Using the Longitudinal Research Database," Proceedings, Bureau of the Census Fourth Annual Research Conference, March, 1988.

BUGENHAGEN, ROGER H., "Implementing the Revised Standard Industrial Classification for the 1987 Economic Censuses," presented at the Annual Meeting of the American Statistical Association, Chicago, Illinois, August, 1986.

BUGGE, PAUL, "Healing the SIC: The 1987 Standard Industrial Classification Revision," presented at the Annual Meeting of the American Statistical Association, Chicago, Illinois, August, 1986.

CARLSTROM, CARLAXEL, "Revision of the Swedish SIC of All Economic Activities," Proceedings, Bureau of the Census Fourth Annual Research Conference, March, 1988.

Census Advisory Committees of the American Statistical Association, on Population Statistics, of the American Marketing Association, and of the American Economic Association, "Minutes and Report of Committee Recommendations," October, 1988.

CONKLIN, MAXWELL R., and HAROLD T. GOLDSTEIN, "Census Principles of Industry and Product Classification, Manufacturing Industries," in Business Concentration and Price Policy, National Bureau of Economic Research, Princeton University Press, 1955.

DAVIS, STEVEN J. and JOHN HALTIWANGER, "Gross Job Creation and Destruction: Microeconomic Evidence and Macroeconomic Implications," NBER Macroeconomics Annual, 1990.

DHRYMES, PHOEBUS J., "The Structure of Production Technology: Productivity and Aggregation Effects," CES Discussion Paper #91-5.

DUNNE, TIMOTHY, MARK J. ROBERTS, and LARRY SAMUELSON, "The Patterns of Firm Entry and Exit in the U.S. Manufacturing Sector, 1963-1982," The Rand Journal of Economics, Winter, 1988, pp. 495-515.

FABRICANT, SOLOMON, Comment on Conklin and Goldstein's "Census Principles of Industry and Product Classification, Manufacturing Industries," in Business Concentration and Price Policy, National Bureau of Economic Research, 1955.

FAULHABER, GERALD, "Discussion," Proceedings, Bureau of the Census Second Annual Research Conference, March, 1986.

FELDMAN, STANLEY J., JENNIFER SCHEURING, and PAUL FELDMAN, "Moving Towards an Improved Micro-Based Classification of Economic Activity: A Report on the Structure and Implementation of a Revised SIC and Related Economic Clusters," mimeo, December, 1990.

GOLLOP, FRANK M. and JAMES L. MONAHAN, "From Homogeneity to Heterogeneity: An Index of Diversification," Technical Paper 60, U.S. Department of Commerce, Bureau of the Census, February, 1989.

GOLLOP, FRANK M. and JAMES L. MONAHAN, "A Generalized Index of Diversification: Trends in U.S. Manufacturing," Review of Economics and Statistics, 1991, pp. 318-330.

GOVONI, JOHN and JAMES L. MONAHAN, "The Longitudinal Establishment Data File: Access and Limitations," Presented at the Census Advisory Committees of the American Economics Association, April, 1986.

JADITZ, TED, "Economic Markets and the Standard Industrial Classification," Bureau of Labor Statistics Working Paper #205, October, 1990.

KIMMEL, SHELDON, "Price Correlation and Market Definition," Economic Analysis Group Discussion Paper #87-8, U.S. Department of Justice, 1987.

KOTTKE, FRANK J., Comment on Conklin and Goldstein's "Census Principles of Industry and Product Classification, Manufacturing Industries," in Business Concentration and Price Policy, National Bureau of Economic Research, 1955.

MAKUC, DIANE M., BENGT HAGLUND, DEBORAH D. INGRAM, JOEL C. KLEINMAN, and JACOB J. FELDMAN, "Use of Cluster Analysis to Identify Health Care Service Areas," American Statistical Association, 1990 Proceedings of the Social Statistics Section.

MCGUCKIN, ROBERT H. and SANG V. NGUYEN, "Public Use Microdata: Disclosure and Usefulness," Journal of Economic and Social Measurement, Volume 16, Number 1, 1990, pp. 19-39.

MCGUCKIN, ROBERT H. and GEORGE PASCOE, "The Longitudinal Research Database: Status and Research Possibilities," Survey of Current Business, November, 1988, pp. 30-37.

MCGUCKIN, ROBERT H., "Longitudinal Economic Data at the Census Bureau: A New Data Base Yields Some Fresh Insights on Some Old Issues," in Analysis of Data in Time, Proceedings of the 1989 International Symposium, Statistics Canada, October, 1989.

MONAHAN, JAMES L., "Alternative Ways of Classifying Economic Activities," Presented to the Census Advisory Committees of the American Marketing Association and the American Economic Association, 1986.

NATRELLA, VITO and JOEL POPKIN, "International Comparability of Industry Classification Systems," presented at the Annual Meeting of the American Statistical Association, Chicago, Illinois, August, 1986.

POPKIN, JOEL, "Recommendation and Description of the Principles Upon Which a Revised Industrial Classification System Should be Built," mimeo, 1991.

POPKIN, JOEL, "Monitoring Economic Performance in the 21st Century: Measurement Needs and Issues," mimeo, 1991.

POWELL-HILL, PAM and PATRICIA BUCKLEY, "Service Sector in the Next 10-50 Years: Jobs, Lifestyles, and Data Needs," Proceedings, Bureau of the Census Second Annual Research Conference, March, 1986.

SALOP, S., "Symposium on Mergers and Antitrust," Journal of Economic Perspectives, Volume 1, Number 2, Fall, 1987, pp. 3-12.

SCHERER, F.M., "The Structure of U.S. Industry," Second Edition Industrial Market Structure and Economic Performance, 1979, pp. 77.

SCHERER, F.M., Industrial Market Structure and Economic Performance, Second Edition, 1979, p. 77.

SCHUH, SCOTT D., "Inventory Models and Aggregation: An Investigation with Company-Level M3 Data," presented to the American Economic Association Advisory Committee, April, 1991.

STIGLER, G.J. and R.A. SHERWIN, "The Extent of the Market," Journal of Law and Economics, Vol. XXVIII, 1985, pp. 555-585.

SUITS, DANIEL B., Comment on Conklin and Goldstein's "Census Principles of Industry and Product Classification, Manufacturing Industries," in Business Concentration and Price Policy, National Bureau of Economic Research, 1955.

THEIL, H., "Specification Errors and the Estimation of Economic Relationships," Review of the International Statistical Institute, Volume 25, 1957, pp. 41-51.

TRIPLETT, JACK E., "The Federal Statistical System's Response to Emerging Data Needs," Journal of Economic and Social Measurement, forthcoming, 1992.

TRIPLETT, JACK E., "The Theory of Industrial and Occupational Classification and Related Phenomena," Proceedings, Bureau of the Census Annual Research Conference, March, 1990.

U.S. Department of Commerce, Bureau of the Census, "1989 Recordkeeping Practices Survey," December, 1990.

U.S. Office of Management and Budget, Standard Industrial Classification Manual, (1972 and 1987), Government Printing Office, Washington, D.C.

WAITE, PRESTON J. and STACEY J. COLE, "Selection of a New Sample Panel for the Annual Survey of Manufactures," presented at the Annual Meeting of the American Statistical Association, Houston, Texas (1980).

WERDEN, G., "Market Delineation and the Justice Department's Merger Guidelines," Duke Law Journal, Number 521, 1983.

WERDEN, GREGORY J., "Four Suggestions on Market Delineation," Economic Analysis Group Discussion Paper #90-5, U.S. Department of Justice, 1990.

WERDEN, GREGORY J. and LUKE M. FROEB, "Correlation, Causality, and All That Jazz: The Inherent Shortcomings of Price Tests for Antitrust Market Delineation," Economic Analysis Group Discussion Paper #91-6, 1991.

WHITE, L., "Antitrust and Merger Policy: Review and Critique," Journal of Economic Perspectives, Volume 1, Number 2, Fall, 1987, pp. 13-22.

WHITE, LAWRENCE, Comment on Triplett's "The Theory of Industrial and Occupational Classifications and Related Phenomena," and Abbott and Andrews' "The Classification of Manufacturing Industries: An Input-Based Clustering of Activity," Proceedings, Bureau of the Census Annual Research Conference, March, 1990.

WORDEN, G., "Company Reporting for Segment of Business: Improving Industrial Statistics," Presented to the Census Advisory Committees of the American Marketing Association and the American Economic Association at the Joint Advisory Committee Meeting, 1986.