THE ROLE OF MICRODATA IN THE NATIONAL ECONOMIC AND SOCIAL ACCOUNTS*

BY RICHARD RUGGLES AND NANCY D. RUGGLES

Yale University and National Bureau of Economic Research

Microdata sets—samples of data relating to individual reporting units—can provide a valuable extension of the national economic accounts as they presently exist, making it possible to meet many of the criticisms being leveled at the accounts over their failure to include much nontransactions information that is essential to the evaluation of economic and social performance. To serve this purpose, however, the microdata sets must be integrated with the aggregate accounts, and with one another. A microdata set relating to any given sector should add up (with appropriate weighting) to the economic constructs for that sector in the national accounts, and the microdata set for one sector should be articulated as appropriate with those of other sectors. This paper discusses techniques for constructing such microdata sets, including necessary adjustments of the macro accounts, techniques of alignment of microdata with the macro accounts, and the creation and development of synthetic microdata sets. Synthetic matching and other techniques of merging data sets are discussed. The paper concludes with a consideration of the methodological implications of the integration of microdata and national accounts.

National income accounting has come of age. Most nations produce national income accounts on an official basis, and it has been seven years since the United Nations introduced its revised Systems of National Accounts.[1] But discontent with national accounts has increased even as they have become widely accepted and have continued to be refined. The neglect of nonmarket activity and environmental factors, for example, has raised serious questions about the appropriateness of national accounting aggregates for measuring economic and social performance, while the omission of most social and demographic information from the accounts has led to the social indicators movement in partial answer to this obvious inadequacy.

The deficiencies of the national accounts are particularly irksome for those concerned with the design and evaluation of policy in such areas as taxation, government expenditures, and other programs or regulations aimed at improving the distribution of income, reducing discrimination, or improving the quality of life. Dealing as they do with market transaction flows at an aggregative level, the national accounts as presently construed do not provide the needed social and demographic information, or information on the distribution of benefits among different social or regional groups.

The response of the traditional national accountant to the increasing demand for information has been to increase the amount of detail by disaggregating the major transactions flows for the different sectors, and to provide supplementary information than can in some degree be related to the accounts. In particular, more detailed data are being provided for the government sector, linking the revenue and outlays as shown in the government budget with the national income accounts so that the aggregative impact of various government programs on the

*Research supported by National Science Foundation Grant GS-33956.

[1] United Nations Statistical Office, *A System of National Accounts*, Studies in Methods, Series F, No. 2, Rev. 3 (New York, 1968. Sales no. E. 69. xvii. 3).

economy can be traced. More detailed industrial breakdowns of output, wages, and profits are being increasingly provided for the enterprise sector; this is of interest especially to those who are following the development of specific industries, and who wish to be able to predict changes in the level and composition of industrial activity. The basic industrial statistics are often supplemented by such things as surveys of intended investment expenditures. There has been less development of the national economic accounts in the direction of disaggregation of the household sector by social and demographic group. Although work is progressing on regional breakdowns of government, enterprise, and household data, most countries do not have very much regional data available on a regular basis.

For some purposes, detailed cross tabulations involving social and demographic data are quite useful. As Richard Stone has demonstrated,[2] Markov transformation matrices can be used to project some of the social and demographic changes that may be expected to occur. The procedure involves the development of rather elaborate multi-dimensional cross tabulations to include the information germane to a specific type of analysis (e.g., the analysis of the educational process). The System of Social and Demographic Statistics of the United Nations has gone in this direction.[3]

Neither the disaggregation of the macroeconomic accounts in the SNA or the more detailed and elaborate social and demographic statistics along the lines of the SSDS, however, provide the kind of detailed information required for the design and evaluation of specific policies and programs. For instance, the analysis of alternative tax measures requires computation of the effects of specific tax regulations on different enterprises and households, in terms both of the distribution of the tax burden and of the total amount of revenue that would be generated. In an ongoing tax system, much of the basic data required for such analysis can be obtained from samples of the tax returns of enterprises and households. Data of this type is quite different in nature from data obtained by disaggregation. In samples of microunit data all of the information relevant to a specific microunit is available as a separate and distinguishable set, but in disaggregated data individual microunits cannot be observed as separate entities.

The specific type of microunit for which information is required will differ from problem to problem. In analyzing revenue sharing between central and local government, for example, budgetary data for local governments and information on the social and economic characteristics of the population living in each community will be needed; in this instance the microunit would be the local government, with associated data about the individual community, much of which already exists. Data sets in which the microunit is the household will have wide applicability in many different kinds of policy assessments. Programs which are related to income maintenance, housing, health, education, labor force participation, and discrimination can use household microunit data to analyze the nature of

[2] Richard Stone, "A System of Social Matrices", *Review of Income and Wealth*, Series 19, No. 2, June 1973.

[3] Richard Stone, "An Integrated System of Demographic, Manpower, and Social Statistics and Its Links with the System of National Economic Accounts", United Nations, New York (E/CN. 3/394, 28 May 1970).

the population with which they are dealing and to examine the distributive impact and costs of alternative proposals. In the United States, samples of the basic population census records contain the core of the required information about households. In addition, the Current Population Survey yields a continuous flow of information about unemployment and other household socioeconomic information, as do samples of certain administrative files such as the social security records. In general, the administrative records of government agencies contain large amounts of information for many different kinds of microunits.

The Need for Integration of Microdata with National Economic Accounts

Although microdata sets derived from administrative records are a valuable tool for the analysis and evaluation of government policy, they are by their very nature partial, in the sense that they refer to the activities and characteristics of microunits in one part of the economy only; they do not reflect the interrelation of different parts of the economy. Presently existing microdata sets have usually resulted from administrative requirements for information to carry out operating functions of specific government agencies, rather than from the implementation of a carefully designed statistical information system. Thus they do not fit into an overall established framework, and they are difficult to relate to one another since they use different concepts and classification procedures. Nevertheless, their usefulness is so great that they will continue to increase in importance as a statistical source, and unless some technique can be developed to integrate them with the national economic accounts, statistical offices will be faced with the situation in which the most used forms of statistical data lie outside the statistical system. At the same time, because administrative data are not adequately linked either with data about the total economy or with other administrative data, more general types of analyses will be impeded, and conflicts will arise among the different sources of data.

On the other hand, if methods of integration can be developed, it will become possible to employ more sophisticated models which take the behavior of microunits into account, while analyzing the interdependence among them. By integrating the different microdata sets into a general system, furthermore, conflicts among them can be resolved much in the same way as different sources of aggregated economic data are now reconciled within the national economic accounting framework.

To integrate microdata with the national economic accounts, it is first necessary to adjust the sectors and economic constructs in the national economic accounts to form an appropriate framework into which the microdata sets can be fitted. Second, techniques of aligning microdata sets with the macro accounts are needed to insure consistency between the aggregates obtained from the microdata and the economic constructs in the national accounts. Third, techniques are needed to reconcile and integrate microdata from different sources.

Sectoring and Economic Constructs for Integration

The sectors of the national income accounts have never been rigorously defined to reflect microunit behavior. Thus for example, the personal income account traditionally has included not only the activities of individual households, but also the accounts of non-profit institutions. Personal income, therefore, is not the total income of persons; it also includes the income of certain designated non-profit organizations such as private universities and mutual insurance companies. If there is to be integration between microdata and the national accounts, the sectoring of the national income accounts must correspond to the major types of microunits. This suggests that, in addition to a purified personal income account, an explicit enterprise account is needed, in which enterprises are partitioned by legal form of organization (e.g., corporate, noncorporate, government enterprises, non-profit institutions, etc.). The government sector should be classified so as to identify governmental units which have revenue and outlay budgets and which make budgetary decisions. Thus different levels of government should be recognized explicitly, and independent governmental bodies separated out. Only if these principles of sectoring are adopted can we expect to be able to integrate the information on households, enterprises, and different levels of government with the more aggregative information in the national accounts. The economic constructs in the national accounts, in turn, should be defined so that they represent aggregations of the data observable at the microunit level. Where economic constructs include transaction flows arising in different sectors of the economy, as does gross capital formation, they should be decomposed into subtotals which reflect the composition by sector. Thus total capital formation would be decomposed into enterprise, government, and household sector capital formation.

The revised United Nations System of National Accounts has made substantial progress toward making the accounts more compatible with microunit information. Yet there has been no explicit recognition of this need in the design of the accounts, and further modification is required in both the sectoring and the economic constructs if the accounts are to accommodate microdata. At the most aggregate level, little recognition is given in the UN system to the institutional decision-making units in the economy; rather the major economic constructs are summary totals which are defined in purely functional terms (e.g., production, consumption, and capital accumulation). The reorientation of the accounts with a view to integrating microdata would emphasize even at the most summary levels the activities of important decision-making units, such as governments, enterprises, and households. This does not mean that attention would be diverted from functional categories; but rather that even at the most summary level the consumption of households would be differentiated from that of government, and the capital formation of households, government, and enterprises would be differentiated. The differences in the content of the functional categories for the different sectors are so pronounced that combining them into summary overall economic constructs is of doubtful utility.

Alignment of Microdata with the Macro Accounts

National economic accounting data are derived from a variety of sources, and the stock in trade of the national accountant is to draw upon these sources to produce a set of consistent estimates that are in accord with the best available information about the economy. If different sources give conflicting estimates, it is the function of the national accountant to resolve these conflicts in the statistically most valid way. The same kinds of problems arise in integrating microdata into the national accounting system. In most cases, microdata sets are samples of observations which must be blown up to yield estimates for the universe which they represent. Even where the sectoring of the national accounts corresponds precisely to the universe represented by a microdata set, it will often be found that the resulting blowup of a microdata sample differs from the data in the national economic accounts. It then becomes necessary to find the best possible way to reconcile, or align, them.

There are of course many reasons why microdata sets are inconsistent with national accounts data. The concepts in the microdata set may differ from the concepts used in the national accounts. Thus for example, a microdata set which purports to contain information on the wage and salary receipts of households may not take into account the fringe benefits which are included in the national accounts; to agree with the national accounts the wage data in the microdata set must be adjusted to include these benefits. In other cases, the coverage of the microdata set may not be complete, and it will be necessary either to reweight the sample or to include additional observations. Deficiencies in a microdata set can often be identified by comparing tabulations of it with the detailed breakdowns of the national accounts. Thus for example the geographical and industrial distribution of wages which results from the tabulation of the microdata should match the regional and industrial breakdown of wages in the national accounts. Quite sophisticated adjustments to the raw data may have to be made to bring them into complete alignment with the national accounts. The alignment works both ways, however: national accountants will also find that microdata sets can contribute to improving the quality and consistency of the aggregate estimates.

The Creation and Development of Synthetic Microdata Sets

If the integration of microdata sets with the macro accounts were dependent upon the prior existence of fully satisfactory microdata, the outlook would not be promising. Few if any countries have the kind of basic microdata which would be required. The problem, however, can be viewed in quite a different light. On a purely theoretical basis, one can postulate a microdata set for each sector in the national accounts. It is the task of the national accountant to deduce the characteristics of the appropriate microdata sets so that they are consistent with all known information.

For example, let us assume that some aggregative and cross-tabulated data of a fragmentary sort exist for a given country, and the objective is to generate a microdata set for all the households in that country which would be consistent with all known information about its economy and population. Presumably, the population size will be known, so that it would be possible to create an initial microdata set of this size, or in the case of large populations, a representative sample. At this stage the microdata set would contain only a given number of persons, without age, sex, race, location, or any other characteristics. The next step is to attach additional types of information to it. Thus for example, if census information is available on the distribution of the population among regions and/or cities, each of the individual observations can be assigned a geographic location in such a way that the resulting regional distribution will correspond to the known census information. Each observation can also be assigned a sex, since the approximate division of the population between males and females is known. Even rough approximations of the sex ratio will usually be satisfactory, although where the sex ratio is known to be atypical for specific geographic areas, e.g., those including military bases, this could be taken into account. In similar manner, each case can be assigned an age and race based upon what is known about the age and race distribution. Then, taking into account the age, sex, and race characteristics of the observations, individuals can be combined into households on the basis of what is generally known about family size and composition. The work status of individuals, including occupation and industry, can be assigned to be consistent with what is known about industries and economic activities in the various regions, and wages can be assigned in terms of the known industry, occupation and geographical status. Where there is genuinely no information on a given topic, assignment on a random basis will not do violence to anything that is known.

Once a synthetic microdata set has been created for households, synthetic microdata sets can also be generated for other sectors. Thus for example, since the household microdata set contains information on the location of individuals and the industry and occupation in which they are employed, synthetic establishments which employ these individuals can be created. Since the occupation and industry of individuals as employees will have been assigned on the basis of known industrial and geographic information, the household microdata set will already have been aligned with what is known, and assembling individuals into establishments merely requires making assumptions about the size distribution of establishments. Past censuses or industry samples may give a basis on which to develop such distributions. Because such establishments employ people in a given location, they will, of course, have locational characteristics. Rough estimates of their total sales can be developed by applying coefficients for cost of materials to the establishment employment data and using percentage gross margins to mark up the wage bill and cost of materials to get value of shipments. The necessary coefficients relating cost of materials to employment and the percentage gross margins can generally be obtained from fragmentary sources of information about different industries. In some instances, it might even be possible to go further and break the cost of materials down by type of material for the various industries, and the sales by type of product, so as to provide the basis for a rough input-output analysis.

With respect to local governments, communities containing both households and establishments have already been created, and what remains is merely to educate the children, provide fire and police protection, clean the streets, and collect the trash. In return, of course, these households and establishments pay taxes to their local government. In most cases, at the local government level even the citizenry has some idea of the level of taxes and types of services provided, and such general information, together with the information generated for individuals and enterprises in various communities, can be used to develop rough estimates of the activities of local governments even in the absence of any explicit local government budget data. In a large country the number of local government units is very substantial, and postulating their existence and the general nature of their activities even on the basis of the most casual empiricism is not as ridiculous as it would appear. For the central government, the government budget will provide some information in almost every country that can be used in conjunction with the synthetic microdata sets.

The synthetic microdata sets created in this way will, of course, not refer to any real microunits, but they will be consistent with all known economic, social, and demographic information. In other words, the synthetic microdata sets are used as vehicles onto which any available information can be mapped; they serve as repositories. When new information is found to be inconsistent with the synthetic microdata sets, revisions will have to be made in the microdata sets to embody it. As more and more becomes known about households, enterprises, and governments, the synthetic microdata sets will become more and more stable. They will thus correspond more and more closely, statistically, to the information content of real microdata sets, even though at the microunit level they still contain no real microunits.

The Reconciliation and Integration of Microdata Sets

In actual practice, the statistician seldom faces a complete absence of midrodata, even in statistically primitive countries. The more usual situation is one in which a large variety of different kinds of information is available from many different sources, and the task is that of deciding how to reconcile and integrate these different types of information at the microunit level.

If two microdata sets contain exactly the same items of information, differing only in alignment either with each other or with known macrodata, the problem resolves itself into determining which data set is statistically more valid, or whether adjustments can be introduced to reconcile them. The differences may be due to sampling biases, different methods of obtaining information, or minor differences in concepts. Two samples of data will rarely refer to precisely the same time frame, and the information may have been obtained from different categories of respondents. Problems such as these are quite similar to those the national accountant faces with macrodata, where different sources of data often yield quite different results. In such instances a good understanding of the differences in sources and methods employed in collecting the information is needed for a satisfactory evaluation of their relative reliability and validity.

Two data sets relating to similar microunits seldom do contain exactly the same items of information, however. Thus for example, the 1970 U.S. Census Public Use Sample contains over 125 items of social and economic information on individuals (including housing information) and the 1969 Federal Income Tax sample contains over 75 items of information relating to income and taxes. The census data lack the detailed income tax information which appears in the tax records, and the tax records lack the rich social and demographic information contained in the census file. In addition, the census sample contains a large number of individuals whose income is below the tax filing limits and who are thus excluded from the tax records. On the other hand, because the tax sample is much denser for the high income groups, very much more accurate information on these groups is available in the tax file than in the census records. The census records, in fact, do not differentiate among individuals having incomes in excess of $25,000. If satisfactory methods can be found to integrate these two bodies of data, the resulting set of information would be much richer than either alone. If the same individuals were represented in both files and could be identified, an exact matching of individuals in the two files would provide the desired integration. However, quite aside from the problems of confidentiality and disclosure, many of the most useful microdata sets are samples, and it is not likely that the same individual will be found in more than one. Thus exact matching is not a general solution.

One method of integrating two microdata sets would be to make a synthetic match between specific observations in one data set and specific observations in a second data set. In the above example, each household in the census sample could be synthetically matched with an appropriate tax return selected from the sample of federal individual tax returns. The census household records contain information on wages, other income, and home ownership, which can be related to the same sort of information in the tax file to provide a basis for synthetic matching. Studies at both the Brookings Institution and the U.S. Department of Commerce have demonstrated the feasibility and usefulness of such synthetic matching.

It can be argued that instead of relating information in two files by synthetically matching the characteristics of specific cases, imputation by use of a multivariate regression analysis would be preferable. Where a single variable is to be imputed, this method does have its advantages, especially when dealing with small samples. The particular multivariate analysis used to perform the estimation can be developed and tested with reference to a number of different sets of information.

But the technique of imputation by regression is considerably less satisfactory in transferring complex sets of information. Thus for example, if consumer expenditure patterns are to be imputed to a sample containing other social and demographic information, a problem arises in that the outlays are all highly interrelated. A separate estimate for each type of outlay would produce an inconsistent expenditure pattern for any specific individual. One of the major objectives of collecting budget information, furthermore, is to study the interrelationships among expenditure items—interrelationships which would be lost if each outlay were imputed independently. Although it might be possible to design a model which would take into account for each item of outlay the elements which had already been imputed, thus attempting to preserve the information on interrelationships in the original sample, such a model would be extremely complex, especially if the actual relationships were not well approximated by a linear or log linear additive model. A much simpler and usually more satisfactory way of proceeding would be to transfer complete sets of budget information from observations in one sample to observations in the other sample by a matching process, thus retaining the integrity of the sets of information in both samples.

The use of a matching process has important methodological implications. First, imputation by regression would normally result in assigning mean values, whereas the matching technique reproduces the distribution of values in the original data set. For a single imputation the mean value may be desirable, but for repeated imputations the use of mean values destroys the observed variance. Second, matching does not require the advance determination of a specific functional relationship. Non-linear relationships will automatically be handled as efficiently as linear relationships, without explicit recognition that the relationships are non-linear. This is in marked contrast with the regression technique, which requires determination of the precise functional form in advance. The success of the matching techniques does, however depend on the data being quite dense, so that similar cases can be found in both data sets. In those instances where the functional form is well known and the data are scattered so that matching is difficult, regression analysis may provide more valid imputations, but with large bodies of data where similar cases do exist, imputation by matching has the virtues of retaining the distributional characteristics of the original sample and reflecting the basic relationship more accurately.

The Technique of Synthetic Matching

The process of synthetic matching involves comparing values of the matching variables in one data set to the values of the same matching variables in another data set in order to bring together similar observations from the two data sets. The central question in this process resolves itself into the choice of criteria to determine a synthetic match. Where the values of the matching variables in sample *A* precisely correspond to the values of the matching variables in sample *B* there is no problem. In such an instance the observations in files *A* and *B* having identical values for the matching variables can be synthetically matched on a stochastic basis. Without introducing additional matching information it is not possible to do better than this. The real problem arises when the values of the matching variables in the two data sets differ somewhat, and it becomes necessary to decide which combination of matching values is most satisfactory.

Conceptually, a distance function could be constructed to express the difference between the values of all of the matching variables for each pair of observations in data sets *A* and *B*. The object of such a procedure would be to find for each observation in data set *A* that observation in data set *B* which has the smallest distance measure. To construct such a distance function, a measure of what is meant by the difference between the values of the matching variables is required.

In principle, the matching variables are intermediate, in the sense that their function is to bring the non-matching variables together synthetically. Although it is true that the two data sets contain no information about the joint distribution of the non-matching variables conditional on the matching variables, information is available on the joint distributions of the matching variables and the non-matching variables in each data set, and this information is relevant to the creation of a satisfactory distance function. If outside information on the joint distribution of the non-matching variables is available it could, of course, also be introduced as part of the matching criteria; but this possibility is not being considered here. If the synthetic matching is undertaken for a specific analytic purpose, certain non-matching variables may be very much more important than others, so that different weights might be attached to the different variables. Thus for example if the purpose of matching the two data sets is to analyze the interrelationships among demographic and economic variables, these variables may be emphasized. But if the purpose is to create data sets designed to serve a wide variety of uses, much as the national economic accounts provide data for many types of aggregative analysis, a more general approach is needed.

One such approach to developing distance functions makes use of multivariate regression analysis, with the dependent variables the non-matching variables and the independent variables the matching variables, to determine the weights to be attached to each of the matching variables to get the best explanation of the non-matching variables. From such information a distance function can be constructed. Horst Adler's paper describing Statistics Canada's matching of two surveys illustrates the use of such a procedure.[4]

Okner's merge of the Survey of Economic Opportunity files with the tax model files[5] in effect also created a distance function, by assigning consistency scores for various criteria and then requiring that matching be carried out in accordance with these consistency scores. The initial step in this process was to group the units in each file into "equivalence classes", broad categories which were considered to be very important for the synthetic matching process. Within these equivalence classes narrow income class bands were defined, and within these bands consistency scores were used to define acceptable matches, which were then made on the basis of sampling probabilities.

The work by Edward Budd and Daniel Radner in the U.S. Department of Commerce on merging the Current Population Survey files and the tax model files[6] differs somewhat from Okner's approach. The Budd-Radner approach depends on the rank order of observations in the two files within broad equivalence classes. In effect the process ranks both files within fairly broad wage classes, and within these, by self employment income and property income. The actual synthetic match is achieved by splitting the records in each file so that the weight for two records with the same rank in a particular subclass is the same. It should be noted that this technique of synthetic matching using rank order in the two files takes care of the alignment problem, on the assumption that the general ordering of information in the two files is correct and that the alignment problem is one of level.

More recently, a research project at the National Bureau of Economic Research has been developing a technique of synthetic matching based on a strategy of sorting and merging microdata files.[7] The basic principle behind this

[4] Horst Adler, "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey, 1970", *Annals of Economic and Social Measurement*, Vol. 3, No. 2, April 1974, pp. 373–394.

[5] Benjamin Okner, "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File", *Annals of Economic and Social Measurement*, Vol. 1, No. 3, July 1972, pp. 325–42.

[6] Edward C. Budd, "The Creation of a Microdata File for Estimating the Size Distribution of Income", *Review of Income and Wealth*, Series 17, No. 4, December 1971, pp. 317–34.

[7] "A Strategy for Merging and Matching Microdata Sets", Progress Report on NSF Grant Project GS-33956, Nancy and Richard Ruggles, *Annals of Economic and Social Measurement*, Vol. 3, No. 2, April 1974, pp. 353–72.

technique is the determination of the proper intervals of the matching variables to be used to define a sorting tag, which in turn determines the cell boundaries within which observations are considered to be statistically equivalent. Ideally, one would like to have assurance that within a specified interval of a matching variable the distributions of the non-matching variables are invariant. In other words, intervals are determined in such a way that if a given interval were to be broken into smaller intervals, the distribution of non-matching variables for the samples of observations in each of these smaller intervals would not be significantly different, in a statistical sense, from each other. Thus, each matching variable is divided into intervals, each of which represents a homogeneous grouping at some designated level of significance. It is possible to examine different levels of significance, so that a hierarchy of intervals is obtained; a nesting system can be developed which classifies a given matching variable from broad classes into fine classes by applying stricter and stricter standards of significance. This technique is then used to produce sort tags to order the microdata sets which are to be synthetically matched. The NBER research project has developed specialized computer programs which analyze large samples of microdata and generate the hierarchical sort tags required for the synthetic matching process. Once the sort tags have been developed, the actual matching process is accomplished simply by sorting the two data sets and merging them. Adjacent cases will then automatically constitute the best matches.

One of the major advantages of this technique is that it easily handles very large data sets. One of the data sets involved in the NBER project, for example, is the one percent Census Public Use Sample, which contains approximately two million cases. Because modern computer technology makes the sorting of even large files economically feasible, this technique is practical where more elaborate techniques involving comparison of actual cases in the two files would be prohibitively expensive.

Summary and Conclusions

Microunit data are essential for the design and evaluation of a wide range of government programs. Only by its use is it possible to evaluate the distributive impact of various government programs on enterprises and households and to analyze the specific and detailed interactions among them. It is only on the microunit level, further, that nontransactions information of a social and demographic nature such as age, race, and sex can be embodied in the data, and the needed locational and industrial dimensions of economic activity specified.

Although the national accounts are a useful framework for analyzing many problems relating to the economic system as a whole and the interactions among sectors, they are not by themselves sufficient. Even the extensive disaggregation recommended by the new United Nations System of National Accounts and System of Social and Demographic Statistics does not accomplish this, since these present data based on the disaggregation of economic constructs and totals rather than data relating to individual microunit observations.

From the point of view of the statistical system, a broader and more systematic framework is needed to make the integration of microdata sets possible. There are two reasons for this. First, existing microdata sets originate from a large variety of sources, many of which are administrative in nature. The lack of standardized classification procedures and common definitions and differences in coverage result in statistical chaos. Apparently conflicting conclusions can be reached through the use of different microdata sets, since the conceptual and statistical differences among them are not easily understood. Second, just as the integration of the different sectors in the national accounts is achieved by articulating the intersectoral flows, thus permitting the analysis of the interactions among different sectors, it is essential for the same analytic reasons to be able to interrelate the microdata sets for the different sectors of the economy, and it is important therefore to develop procedures to integrate them with the national accounts and with each other.

Integration of microdata sets with the national accounts is possible if the national accounting system is sectored according to decision-making units and the macro constructs are developed to reflect aggregations of microdata. The developments of the last 30 years in national accounting have been in this direction, and only minor modifications are now required to meet this criterion.

Satisfactory microdata sets for specific sectors of the economy do not currently exist, but it would be possible to develop synthetic microdata sets which are aligned with all of the information that is available for each sector of the economy. The techniques for doing this are not conceptually different from those which the national accountant already uses in the construction of the national accounts. Although such synthetic microdata sets will not represent any actual microunits in the system, they would, if properly constructed, have the same statistical characteristics as the actual microunits. For many countries, there are already in existence various microdata sets which provide many different kinds of useful information. Economists are now utilizing a variety of techniques for aligning, imputing, and merging and matching microdata from different sources to create new sets of microdata for particular analytic purposes. These same techniques can be used to provide general-purpose microdata sets which are integrated with the national accounts.

From a methodological point of view there is increasing uneasiness among economists about the failure of economics as a science to unite economic theory and empirical analysis. Wassily Leontief in his presidential address to the American Economic Association in 1970 stressed that a major gap exists between theoretical formulations and empirical research. In particular he argued for "an iterative process in which improved theoretical formulation raises new empirical questions, and the answers to these questions in their turn lead to new theoretical insights."[8] Nicholas Kaldor, in his article on "The Irrelevance of Equilibrium Economics", goes further and says that "The powerful attraction of the habits of thought engendered by 'equilibrium economics' has become a major obstacle to the development of economics as a science—meaning by the term 'science' a body of theorems based on assumptions that are empirically derived (from observation) and which embody hypotheses that are capable of verification both in regard to the assumptions and predictions."[9] Martin Shubik, in his perceptive article, "A

[8] *American Economic Review*, Vol. LXI, No. 1, March 1971, p. 5.

[9] *Economic Journal*, Vol. 82, p. 127.

Curmudgeon's Guide to Microeconomics"[10], charges that "the very power and elegance" of economic theory may have set the subject back as far as it set it forward, since it made it appear that the "abstraction was somehow central, universal, and of broad application."

One of the major problems of microeconomics has been that its basic concepts are not empirically operational. Not only are utility functions and indifference curves not measurable, but seemingly more concrete concepts such as production functions, factors of production, monopoly, and even the apparently very real concepts of price and output, are not really possible to measure in meaningful operational terms. In contrast, macroeconomics as it has been related to the national accounts does deal with operational empirically measurable concepts. Thus for example, consumer expenditures, government revenue, and gross fixed capital formation are all definable and measurable. The accounting approach to the classification of transactions makes it possible for the national accountant to provide empirical measurements. One may quarrel with the definitions, and even decide to change them, but there is a correspondence between the accounting definitions and the data. Yet macroeconomic analysis still cannot adequately analyze behavior, since the observed aggregates reflect both the behavior of the decision-making units at the micro level and structural changes in the composition of the microunits. A microanalytic framework that employs operational concepts and can be related to macroeconomic analysis is badly needed.

Sets of microunit accounts which are nested within the national accounts would go far toward accomplishing such an integration. The concepts used at the micro level should, when aggregated, correspond to the economic constructs at the macro level. This would permit the interaction between theory and empirical analysis that Leontief asks for. Present definitions of concepts at both the micro and macro level are not optimal from the point of view of developing a theoretical structure, but their shortcomings will only become evident as theoretical microanalytic models are developed to explain the behavior and variance in behavior observed at the microunit level. One of the advantages of microunit analysis is that the data immediately suggest the need for a probabilistic approach. Contrary to the assumptions of received economic doctrine, it is not true that each decision-making unit operates in an optimal manner with full information. Analysis in terms of "representative" units is likely to be quite misleading. The variation in behavior among microunits needs to be taken into account at the theoretical level as well as in the empirical analysis.

New types of analytic techniques also becomes feasible once microdata sets that are integrated with one another and with the aggregate accounts are developed. For instance, microanalytic simulation as outlined by Guy Orcutt[11] becomes a powerful tool for testing hypotheses about different types of behavior. Just as regression analysis tests to see how much of the total variation in a

[10] *Journal of Economic Literature*, Vol. 8, No. 2, June 1970, p. 413.

[11] Guy Orcutt, "Microanalytic Simulation of Households in the U.S. Economy", paper prepared for The Second Latin American Conference of the International Association for Research in Income and Wealth, Rio de Janeiro, 1974.

dependent variable can be explained by its functional relationships with independent variables, so the success of a microanalytic simulation in tracing out the pattern of events and in producing cross-sectional distributions can throw considerable light on whether the theoretical model on which it is based does adequately explain the behavior and structure of the economic system.

It is the development of the computer over the last several decades that has made the integration of microdata sets and national accounts possible. Without the computer, the microdata sets would not exist. Before its development, data-collecting agencies reduced raw data to a cross-tabulated form as the first step in data processing, because of the difficulties and high costs of handling raw data. Aggregation was thus a form of data reduction intended to make information manageable, and any subsequent data processing was confined to manipulation of the aggregated data. In recent years, the trend of data processing by governments has been in the opposite direction. Maintaining the raw data in its original form makes it easier to produce the wide variety of aggregations and cross tabulations required for different purposes. Such considerations have led to a data revolution, where basic data files are computerized and maintained in their microunit form.[12]

The computer has also played a major role in permitting the analyst to carry through computations at the microunit level on large numbers of observations at a relatively low cost. This means that a large variety of different microdata sets can be synthesized, aligned, merged, and matched with one another, and integrated with macroeconomic data. It also means that once a data base has been created it can be used at the microunit level for a wide variety of simulations or other types of analysis where a detailed case-by-case processing is required.

[12] Ivan Fellegi and S. A. Goldberg, "The Computer and Government Statistics", in N. Ruggles, ed., *The Role of the Computer in Economic and Social Research in Latin America,* National Bureau of Economic Research (New York, 1974).