Annals of Economic and Social Measurement, 6/4, 1977

MERGING MICRODATA: RATIONALE, PRACTICE AND TESTING

BY NANCY RUGGLES, RICHARD RUGGLES, EDWARD WOLFF

In the first section of the paper, we argue for the need for the statistical matching of microdata sets as a way of reconciling diverse bodies of data. As in the construction of national (macro) accounts, the objective is to map information from different sources onto a common framework to resolve conflicts among data drawn from different samples. The end result is the creation of a sample replication of the micro information for each sector of the economy aligned to the macroeconomic account for that sector. In the next section of the paper, one particular matching technique, developed at the National Bureau of Economic Research, is discussed in some detail. The method is illustrated with results from a statistical match carried out between the 1970 Census Public Use Sample and the 1969 Internal Revenue Service Tax Model. In the third section of the paper, we perform several econometric tests to evaluate the reliability of our matching technique. We do this by matching two samples drawn from the same population, with the same sampling frame, and with almost the same set of variables (the 1970 Census 5% and 15% Public Use Samples). We then substitute imputed values obtained from the match for the actual values to determine whether the imputed joint distributions differ statistically from the actual joing distributions. The results show that in 95 percent of the cases there is no statistical difference between the imputed and actual joint distributions.

I. THE RATIONALE FOR MERGING MICRODATA

A. *The Relation of Macro- and Microdata*

One of the major achievements in economics over the past forty years has been the development of systematic macroeconomic data. Out of the combined efforts of academic and government economists and statisticians there has emerged a macroeconomic data base including the national income accounts, input-output tables, the flow-of-funds system, and more recently national balance sheets, which together provide an overview of the operation of the economic system. The macroeconomic data base is a general purpose tool, which can be used for planning, monitoring, and study of fiscal and other general policy questions, and for the construction of many different kinds of aggregative economic models designed for forecasting and analysis. Its focus is on the functioning of the economic system as a whole: problems relating to output, employment and economic growth; the relationships among the major sectors the government, households, business, the foreign sector; and monitoring the performance of the economic system in terms of the behavior of such key elements as consumer spending, gross capital formation, the government surplus or deficit, the balance of payments. It is an essential characteristic of the macroeconomic data that it forms an integrated system. All of its

parts fit together, so that it gives a unified picture of the behavior of the economy.

In contrast, the situation with regard to microdata – that is, data relating to individual firms, establishments, governmental units, households, and persons – is much different. These data are gathered primarily for use in regulatory activities and the administration of government programs, or as input in constructing the macroeconomic statistics. When administrative and regulatory agencies collect information to carry out their own specific operating functions, it is the operating function, not the statistical by-product, that determines the nature of the data collected. Thus the Internal Revenue Service collects microdata in the course of processing tax forms; the Social Security Administration collects information from employers and beneficiaries; and the SEC and FTC require financial and line of business reports from businesses. In the Federal statistical agencies, large quantities of microdata are collected, but their primary use is still viewed as provision of the basis for compiling general purpose aggregative data. For the most part, the microdata collecting activities of even the statistical agencies are conducted as many separate inquiries. Thus, although the Census Bureau conducts censuses and surveys yielding microdata on households and businesses, the Bureau of Labor Statistics collects microdata on prices, wages, and employment, other agencies regularly collect information on agriculture, health, and many other fields, and special surveys produce information on consumer expenditures, crime, and a wealth of other social and demographic questions, it is difficult to relate these separate inquiries either to one another or to the macrodata.

Increasingly, this situation has come to be viewed as unsatisfactory. Broad social programs involving interrelationships between governments and households, the distribution of income, the position of specific social and demographic groups, and distributions by region and type of community call for new kinds of data, as does the setting of goals and the measurement of performance for social programs in such areas as manpower training, education, and health care. Macroeconomic data systems do not provide either the distributional information or the social and demographic information which is needed, but the existing microdata are often partial, biased, internally inconsistent, and impossible to reconcile with the macrodata.

The response of the traditional national accountant and macroeconomic analyst to the increasing demand for such new kinds of information has been to increase the amount of detail in the macrodata by disaggregating the major transactions flows, and by providing supplementary information that can in some degree be related to the aggregate data. More detailed data are provided for the government sector, linking the

revenue and outlays in the government budget with the national income accounts so that the aggregate impact of various government programs can be traced. More detailed industrial breakdowns are provided for the enterprise sector: this is of interest especially to those who are following specific industries and who wish to be able to predict changes in the level and composition of industrial activity. Work is progressing also on regional breakdowns. It was this approach - disaggregation - that was followed in the latest revision of the United Nations System of National Accounts (SNA) [8], where the disaggregations called for have been pushed well beyond the capacity of most countries to supply the data.

An alternative approach relies upon detailed cross tabulations of social and demographic data. As Richard Stone [7] has demonstrated, 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 multidimensional cross tabulations of the information germane to a specific type of analysis (as for instance the educational process). The System of Social and Demographic Statistics of the United Nations [9] has gone in this direction.

However, neither disaggregation of the macroeconomic acounts like that in SNA nor the more detailed and elaborate social and demographic statistics of SSDS can provide the kind of detailed information required in many instances for the design and evaluation of specific programs and policies. Unfortunately, cross-tabulation rapidly becomes explosive. If for example one wishes to study the interaction of 10 variables in a socio-demographic matrix, and each of these variables contains 10 categories (both very small numbers to characterize a socio-demographic system), the number of cells in the matrix is 10^{10}, or 10 billion. Given the size of the populations of most countries and the natural clustering, most of the cells in the matrix would of course be empty, but it is still true that the data would be spread over so many cells as to be essentially unmanageable. In this connection, it is interesting to note that the cross-tabulations of the population census regularly produced by the Bureau of the Census fill many more computer tapes than do the original individual census reports, despite the fact that the data cross-tabulated seldom exceed three or four variables.

Furthermore, the microunit data needed for the kinds of uses mentioned above is quite different in nature from data obtained by disaggregation. In samples of microdata all of the information relevant to a specific microunit is available as a separate and distinguishable set, but in disaggregated cross-tabulations individual microunits cannot be observed as separate entities.

It is possible, however, to envisage a fundamentally different kind of approach, based on systematization of microdata far beyond what now

exists. This possibility has been opened up by the revolution in data processing brought about by the increasing capacity of computers, which in turn has led to a very great change in the technology of statistical activities. The traditional use of data aggregation as a technique for reducing processing requirements to manageable proportions is no longer required. Before the computer was developed to its present level, government agencies conceived of their statistical output as the provision of specific tabulations, and subsequent data processing was confined to manipulation of the tabulated data. Now, the emphasis has shifted from tabulations to the processing and editing of the primary data. It is increasingly clear that data are most efficiently stored in the form of microunit records relating to each separate reporting unit. In some cases, samples of such microunit data may be released to other users with appropriate confidentiality protection, but the desirability of preserving the microdata records does not depend on whether or not there is any intention of releasing individual records. This form of storage not only permits far greater flexibility in generating the wide variety of aggregations and cross-tabulations which may be required for different purposes, but it also compresses the data storage space required. This change in methodology permits the analyst effective access to large bodies of information, at relatively low cost, and it means that the possibility of relating microdata sets more directly both to other microdata and to macrodata has become real (see [5]).

B. *Problems Inherent in Using Microdata*

It has long been recognized that different microdata sets may contain conflicting or inconsistent information, but the magnitude of this problem was not generally appreciated until the microdata began to be used themselves for analysis. As long as microdata were used only as the basis for aggregative tabulations, it was customary to make corrections and adjustments at the aggregate level rather than to carry them back to the microdata records. Once the microdata began to be used directly for analysis, however, the problems of editing and cleaning the individual records became central. Glaring inconsistencies or impossible values which would have been undetected in aggregate data -- for example, a seven year old girl with ten children -- became apparent in the microrecord form. The problem of missing values also had to be faced, and techniques were developed like the "hot deck" employed by the Bureau of the Census to impute reasonable values for missing data.

Even after editing for inconsistencies and allocations for nonresponse, however, the data in a single microdata set often are very different from data derived from other independent sources. Such differences may arise because of underreporting, or from differences in classification

TABLE 1

COMPARISON OF SURVEY OF INCOME AND EDUCATION AND CURRENT POPULATION SURVEY AGGREGATE MONEY INCOME IN 1975 BY TYPE OF INCOME, REPORTED AND ALLOCATED, AND WITH INDEPENDENT ESTIMATES

Source of Income	Independent Sources		SIE as a Percent of Independent Sources		CPS as a Percent of Independent Sources	
	Billions of Dollars	Percent	Reported	Allocated	Reported	Allocated
	(1)	(2)	(3)	(4)	(5)	(6)
Total Income	(NA)	(X)	(X)	(X)	(X)	(X)
Total income, independent estimates available	1,115.0	100.0	82.4	11.3	72.1	18.1
Sources with Independent Estimates						
Wage or salary income	788.2	70.7	89.3	10.9	79.4	18.0
Nonfarm self-employment income	63.4	5.7	76.7	22.1	63.9	33.1
Farm self-employment	20.9	1.9	56.0	9.1	43.1	13.9
Social Security and Railroad Retirement	65.0	5.8	83.8	8.5	72.5	18.6
Supplemental Security income	5.6	0.5	67.9	1.8	55.4	8.9
Aid to Families with Dependent Children and other public assistance	10.2	0.9	75.5	3.9	67.6	9.8
Interest and dividends	82.0	7.3	39.9	14.1	30.7	13.9
Net rental income and royalties	11.2	1.0	75.0	14.3	56.2	17.0
Veteran's payments	12.0	1.1	66.7	5.0	56.7	10.0
Unemployment compensation	18.3	1.6	63.9	3.8	53.0	10.4
Workmen's compensation	5.3	0.5	43.4	1.9	35.8	7.5
Private pensions and annuities	13.8	1.2	74.6	13.0	61.6	19.6
Federal government and military retirement	13.5	1.2	80.7	11.1	74.1	19.3
State and local government employee retirement	6.2	0.6	72.6	6.5	56.5	19.4

Bureau of the Census, Background Paper for Census Advisory Committee, April 1977.

or coverage. Table 1 below gives an example of the magnitude of such differences. It compares income data reported and allocated in the 1976 Survey of Income and Education (SIE) and the Current Population Survey (CPS) with estimates from independent sources. It is apparent from this comparison that serious discrepancies exist.

When different sources give different results, a careful evaluation is needed to determine which source is superior. Thus for example, the payments of social security benefits to households as recorded by the Social Security Administration are likely to be more accurate than receipts of social security benefits reported in a household survey. Similarly, interest and dividends reported to the Internal Revenue Service by the payers are likely to be superior to interest and dividends reported as received by individuals in a household survey. For analytic use, an effort needs to be made to correct the biases in each microdata set to align it with the sources judged to be most accurate. Special resurveys, audits, and small exact matches of records may be found useful in some cases in analyzing the types of bias involved in particular microdata sets and in suggesting techniques for introducing appropriate adjustments. Thus the audits carried out by the Internal Revenue Service have been found to be very useful in assessing the quality of the different items of information in the income tax file, and suggesting kinds of adjustments that would be appropriate. In some cases, adjustments can be based on internal relationships in the data itself. Thus for example, the number of individuals who are receiving social security benefits can be brought into line with the totals reported by the Social Security Administration on the basis of the age, sex, and employment status of the individuals in the survey. Such adjustments can be introduced as additional information rather than as alternatives in the original entries contained in the microrecords, thus adding to the information which can be utilized by the analyst rather than limiting it.

Quite apart from these considerations, another class of difficulty arises because different microdata sets do, of course, cover different reporting units and contain different information. A single survey yielding all of the desired kinds of information relating to any given group of respondents is impractical. Any given survey is limited both by cost and by the reporting burden on individual respondents. There is necessarily some trade-off between sample size and questionnaire size. In instances like the population census where exhaustive coverage of the population is desired, the number of items that can be obtained from every respondent is limited; more detailed questions are asked only from samples of the population. Even these samples, however, are very large when compared with those used for surveys which seek to obtain extremely detailed

information about each respondent, such as the survey of consumer expenditures.

All of these problems pose barriers to the generalization of information contained in microdata sets, but the goal is nevertheless worth pursuing. As long as the microdata were collected only to produce aggregated totals or cross-tabulations, it was considered sufficient to relate the information from different microdata sets at these aggregate levels, and unfortunately it is still true that this is the most generally used method of relating information from different sources. But this method both hides what may be important inconsistencies and differences among data sources and sharply reduces the use that can be made of their information content. The information which the microdata contain on the joint distributions of different variables becomes lost in the process of aggregation.

C. *The Construction of Integrated Microdata Sets*

As any general equilibrium economist will quickly point out, all macrodata are in effect aggregated microdata. The national income estimates compiled by the Bureau of Economic Analysis depend on the tabulations produced by the Internal Revenue Service, the Social Security Administration, and the Bureau of the Census, all of which are based upon their microrecords relating to firms, establishments, governmental units, households, and individuals. Maintaining microdata files at the level of the individual reporting unit makes it possible to preserve both the basic interrelations of data within the reporting unit and the distribution among reporting units, but it would be very useful also to be able to bring the diverse types of information which are available at the micro level from many different sources together in a cohesive form.

This could be accomplished by constructing microdata files that are composites of the known information about various types of reporting units. Construction of such composite microdata sets would be directly analogous to what is done in constructing the macroeconomic accounts. As in the macro accounts, the objective would be to map information from different sources onto a common framework so that conflicts among data from different sources could be examined and resolved in the light of the best available information. The end result would be a general purpose sample replication of the micro information for each sector of the economy aligned to the macroeconomic account for that sector. Such microdata sets could then be used to generate any specific tabulation desired, at any level of disaggregation of the macrodata, and they could also be used as a vehicle for microanalytic simulation.

Conceptually, one can imagine one or more microdata sets underlying the data for any specific sector of the economy. Thus for example, sets of accounts relating to individual firms and their establishments can be considered to underlie the macroeconomic data relating to the enterprise sector. Such accounts are the basis of IRS tax records, FTC and SEC records, and the Censuses of Manufactures and Business. In a well-organized central statistical office, it is not inconceivable that such information could be integrated into a cohesive whole. Even outside the government, companies such as McGraw-Hill and Dun and Bradstreet assemble information on specific firms and their establishments. These data sources are generally open, and academic economists are rapidly beginning to use such sources of open data. For the government sector, the Census of Governments contains the accounts of some 75,000 individual governmental units at the federal, state, and local levels. For the household sector, a wide variety of data sources exists, both from surveys and censuses and from administrative records maintained by government agencies. The development of integrated microdata sets for the major macroeconomic sectors is of course a long term objective, but some of the steps that could be taken are clear.

With respect to firms and establishments, industrial concentration makes it almost imperative that exact matching be used at least for the larger firms, since each one is fairly unique. In open data sets, exact matching poses no serious problem for large companies since their identity is hard to conceal. For small companies and establishments exact matching may in some cases be quite difficult, and other expedients may be required. Considerable analytic problems are also caused, of course, by births, deaths, and mergers of companies, but this kind of information constitutes a significant part of the data base and is a topic of much analysis. Data inconsistencies between different sources will continue to exist, and the analyst will be forced to choose among the available sources. Some of the government's data sources on firms and establishments are open (i.e., SEC, FTC), but others are confidential (such as IRS and census returns). The disclosure rules, however, do not prevent the Bureau of the Census from bringing together information from a variety of different sources, so long as the published results do not disclose confidential information. Thus the published enterprise statistics are based upon use of both IRS and census records by the Bureau of the Census. Similarly, County Business Patterns makes use of Social Security data. It is certainly not beyond the bounds of possibility for the Bureau of the Census to develop a comprehensive microdata file for all the larger firms and establishments in the economy. Although such a microdata file could not be released to the public, it would be an extremely useful tool in the development of the statistical system.

For the government sector, data sources are open, and the microdata tapes of the accounts of the individual governmental units do provide a valuable data source to the analyst. Work on linking these accounts to the macro accounts for the government sector in the national income accounts is going forward. Here, problems of definition, classification, timing, and inaccuracies in reporting all arise, but such problems can be overcome with improved reporting.

With respect to households, the problem of relating microdata records from different sources poses a different sort of difficulty. Analytically, it would be desirable to match different files on a person-by-person basis, and some such exact matching has in fact been carried out. Files of tax returns, social security records, and the Current Population Survey have been linked with each other by matching the social security numbers which were reported in all three files. However, there were a substantial number of non-matches or mismatches due to non-reporting or errors in reporting of the social security number. Attempts to match files by using names and addresses of the respondents meet with much greater difficulties due to the variation in names recorded in different files, the existence of duplicate names, changes in addresses, and even changes in names, i.e., due to marriage. Thus, even in those instances where it is technically feasible, exact matching is costly to carry out. Furthermore, it is subject to the objection that the complete identification of individuals constitutes an invasion of privacy. In instances where the files to be matched are samples of populations, exact matching is of course not possible since generally different samples will contain different individuals. For these reasons, although exact matching can be useful in special instances, it cannot be relied on as the basic method for integrating microdata records from different sources into a composite for the household sector.

In some cases it may be possible to use regression analysis to impute variables contained in one file to another file. For this method to be successful, it is of course necessary that both files contain variables which are closely related to the variable which is being imputed. Such imputation may be satisfactory for many purposes. It should be recognized, however, that using the regression value for imputation entails that the joint distribution of the imputed variables with other variables may not be correctly measured. Where a substantial number of imputations are required and the joint distribution among the imputed variables is important, regression methods may not be appropriate and other techniques must be found.

The concept of a statistical match, which relates a set of data about an individual or household in one file to a record for a similar individual or household in another file, is intuitively appealing. Thus for example, at

an aggregate level, governments often publish a special consumer price index which is said to represent the prices paid by a household of average income with two children living in an urban area. In effect, what is being said in this instance is that the general pattern of goods purchased by a family of this type and level of income is very similar to that of any other family in the same circumstances. The same approach can be extended to provide the basis for statistical matching of household microdata files. For a satisfactory statistical match, it is necessary to be able to select households whose characteristics are similar enough so that the merged information is consistent with all the known information in both files. For example, if both a household survey and a sample of tax returns contain fairly complete information on the composition of households and their sources and levels of income, it would be possible to select for any given household in the household survey an actual tax return from the tax sample which would be representative of the tax return which that household did in fact file.

In recent years there have been a number of efforts directed at the statistical matching of files. In a pioneering study, the Bureau of Economic Analysis of the Department of Commerce undertook the statistical matching of microdata from a sample of individual tax returns, the Census Current Population Survey, and the Federal Reserve Board's Survey of Financial Characteristics of Consumers [2]. The purpose of this effort was to develop improved estimates of the size distribution of income related to the social and demographic characteristics of income recipients. At about the same time, the Brookings Institution undertook a statistical match of a sample of income tax records with the Survey of Economic Opportunity sample in order to analyze the impact of proposed changes in tax laws on the tax payments of individuals [3]. This effort at tax modelling based upon the integration of different bodies of data has continued to be used by both Brookings and the Treasury Department. Using a somewhat different approach, Statistics Canada carried out a statistical match between the 1970 Canadian Survey of Consumer Finances and the 1970 Family Expenditure Survey, which made it possible to relate information on balance sheets and consumer expenditure patterns [1].

At the National Bureau of Economic Research, a research project on the measurement of economic and social performance[1] has been working on developing the methodology of statistical matches of household microdata sets [4]. The methodology developed employs a technique based upon sorting and merging of microdata files. The variables which are common to the two files to be matched are used to develop a hier-

[1]NSF Grant No. Soc74-21391.

archical ordering. Both files are then sorted in the same way according to this hierarchical order, and merged. The cases which are adjacent in the merged file can then be considered to be suitable for matching. Statistical techniques have been developed to determine the matching intervals which should be specified for each variable for any given level of probability. By specifying different levels of probability, a nested set of matching intervals can be developed which results in a sorting order of each file appropriate for statistical matching. The matching procedures are described in more detail in section II of this paper.

It should be recognized that such statistical matching is only valid in fairly dense data sets. Where these are only a few cases within broad matching intervals, the possibility of mismatching is obvious. For this reason, this matching technique is not generally applicable to records contained in small samples, or to those records in large samples which have unusual or extreme characteristics. It is also apparent that although the matching technique takes into account the relation between the matching variables and the remaining variables in each data set, it can say nothing about the conditional joint distributions of the non-matching variables in the two data sets. The assumption is made that such conditional joint distributions are stochastic. Nevertheless, to the extent that the non-matching variables are correlated with the matching variables, the raw joint distributions will be correctly reflected.

A test of the accuracy of matching which was outlined in the earlier piece [4] was to split a large sample into two halves and match one half against the other One could then examine whether the imputed values obtained by the match can satisfactorily be substituted for the actual values. An alternative to this is to match two samples drawn from the same population and with the same sampling frame, which have an almost identical list of variables. One could then substitute imputed variables for actual ones to determine the reliability of the match. We did this using the 1970 Census 5% and 15% Public Use Samples. The results of the test are shown in Section III of the paper.

II. Specification of the Matching Procedure

Since the sort-merge procedure for matching microdatasets was first outlined [4], we have executed three full-scale matches at the National Bureau of Economic Research. The procedures we have used follow very closely the original description of the method. In this section we will present a more technical description of the matching procedure and indicate any modifications to the original strategy. Moreover, in way of illustration we will present some results from the match we executed between

the 1970 Census 15% 1/1000 Public Use Sample (PUS) and the 1969 Internal Revenue Service Tax Model (IRS).[2]

A. *The Direction of the Match.* In our matching procedure one file, the *B* file, is matched to the other file, the *A* file. This means, in effect, that information from the *B* file is transferred to each record of the *A* file. In the PUS-IRS match, for example, we decided to match the IRS file to the PUS file. The reason was that the PUS file is a random (representative) sample of the U.S. population, while the IRS file is a stratified sample with upper income groups over-represented. By matching the IRS file to the PUS file, we could assure that the tax information would be given its appropriate population weight.

B. *The Matching Unit.* Microdatasets have different units of observation. For matching purposes, it is necessary to select a common unit between the two files. Sometimes this entails the creation of a corresponding unit in one of the files. In the PUS file, for example, the basic unit is the household, but the household is broken down into family and individual observations. In the IRS file, the basic unit is the return – that is, a single or joint return. By assuming that all married couples file joint returns, we constructed single and joint tax return units from the individuals in the PUS file and matched the two files on the tax unit.

C. *The Selection of Matching Variables.* In the matching procedure, there are four kinds of variables in each file. The first kind is the cohort variable. These are variables common to both microdatasets which are matched on the basis of *exact* values. In the PUS-IRS match, we selected the type of tax return, the sex of the respondent in the case of single returns, and the age and race of the head of household in the case of joint returns and of the respondent in the case of single returns (see Table 2).[3] Cohort variables are ones we consider too important to match with approximate values from the other file.

The second kind is the *X* variable. These are the remaining variables common to both files but are less important than cohort variables. These variables are matched on the basis of *approximate* values or matching intervals (see below). In the PUS-IRS match, the *X* variables were the number of children, house ownership, wage and salary earnings, business earnings, farm income and total income. Since *X* variables sometimes differ in concept and distribution between the *A* and *B* file, we designate them X_a and X_b, respectively (see below).

The third kind is the *Y* variable. These are non-common variables which are used to construct the matching intervals (see below). In prin-

[2]We used the 1969 IRS file in the match, because earnings and income information in the PUS file is for calendar year 1969.

[3]Age and race information was added to our IRS file by a special run by the Social Security Administration using the actual social security numbers on the tax returns.

TABLE 2

STRUCTURE OF THE 1970 PUBLIC USE SAMPLE–1969 INTERNAL REVENUE SERVICE TAX MODEL MATCH

A. Cohort Variables

1. Type of tax return
2. Sex of respondent (single returns)
3. Race of head of household
4. Age of head of household

B. X Variables

1. Number of children
2. Owner-occupied home or rental unit
3. Wage and salary earnings
4. Business earnings
5. Farm income
6. Total income

C. Y Variables[a]

1. Education
2. Birthplace
3. Occupation
4. Industry of employment
5. Class of worker
6. Years married (married couples only)
7. Number of years at current address
8. Value of property (homeowners only)
9. Number of automobiles in household

[a]PUS file only.

ciple one set of matching intervals should be generated for each file, and the two sets interwoven to form an integrated set of matching intervals. In this case, there would be a separate set of Y variables for each file—Y_a and Y_b. In practice, however, we have generated the matching intervals only from the A file because of the enormous computer cost of the process. The Y variables in the PUS-IRS match are listed in Table 2.

The remaining set of variables in each file are designated Z_a and Z_b, respectively. These are also non-common variables, but are one which we have little interest in or which appear, on the surface, unrelated to the X variables. Quarter of birth, veteran status, transportation to work and language spoken in the home are examples of Z variables in the PUS file.

D. *The Construction of the Matching Intervals.* Since it is very unlikely to find records in the A and B file with identical values for the X variables, it is necessary to match the X variables on close or approximate values. The range of values of the X variable which can be considered "sufficiently close" forms the "matching interval." We construct these ranges by analyzing the sensitivity of the conditional distribution of Y_a on X_a ($f(Y_a \mid X_a)$) to X_a. Using a Chi-square or a correlation test preset at a given significance level we determine among which values of X_a the con-

TABLE 3

NESTED INTERVAL STRUCTURE FOR THE WAGE EARNINGS VARIABLE

Earnings (In Dollars)	Chi Sq (.99)	Correl (.97)	Correl (.90)	Correl (.80)	Correl (.70)	Correl (.50)
		Interval Number by Matching Level				
$ 0 200	1	1	1	1	1	1
201 300	2	—	—	—	—	—
301 400	3	2				
401 500	4	3				
601 700	5		2			
701 800	6	4				
801 900	7	—	—	2	2	2
901 1,400	8	5				
1,501 1,700	9					
1,701 1,800	10	6	3			
1,801 2,000	11					
2,001 2,200	12	—	—	—	—	—
2,201 2,500	13	7				
2,501 2,800	14	—				
2,801 2,900	15	8	4	3		
2,901 3,100	16	—				
3,101 3,400	17	9	—	—	3	
3,401 3,800	18	10				
3,801 4,100	19	—	5	4		
4,101 4,300	20	11				
4,301 4,800	21	—	—	—	—	
4,801 4,900	22					
4,901 5,100	23	12	6			3
5,101 5,400	24					
5,401 5,900	25					
5,901 6,400	26	—	—			
6,401 7,100	27	13				
7,101 7,500	28	14	7			
7,501 8,000	29	—	—			
8,001 8,700	30	15		5	4	
8,701 9,700	31	16				
9,701 13,600	32	17	—			
13,601 15,600	33	18	9			
15,601 18,600	34	19				
18,601 25,500	35	20				
25,501 50,000+	36	—	—	—	—	—

ditional distribution $f(Y_a \mid X_a)$ is statistically different. Values of X_a for which the conditional distribution is statistically different are placed in different matching intervals, and values for which the conditional distribution is not statistically different are placed in the same matching interval. Moreover, by varying the Chi-square and correlation levels, we can generate different sets of matching intervals at different matching levels. In fact, by continually relaxing the criterion for a significant difference, we are able to create a "nested" set of matching intervals. The set of matching intervals for the X variable Earnings and the matching levels used in the PUS-IRS match are shown in Table 3. For example, at the Chi-square (.99) level, IRS earnings of $3500 would be considered a suitable match for PUS earnings of $3700 but not for PUS earnings of $3900. At the correlation (.97) level, any IRS earnings in the range of $3401 to $4100 would be considered a suitable match for PUS earnings in that range. As is apparent from Table 3, the range of the matching intervals widens and the number of matching intervals decreases between the first and last matching intervals. This is true of the other matching variables in the PUS-IRS match, as can be seen in Table 4. From this table it is also apparent that the number of matching intervals and the rate of "collapse" differ substantially among the X variables. (In addition, see [4] for a more complete discussion of the procedures used in constructing the matching intervals and the rationale for it.)

E. *The Alignment of the X Variables.* It often happens that an X variables differs somewhat in concept or sampling distribution between the A and B files. Before the two files can be matched, it is necessary to reconcile or "align" the X_a and X_b variables. In the first case, where two variables differ in concept, it is often possible to transform one concept

TABLE 4

THE NUMBER OF MATCHING INTERVALS BY MATCHING LEVEL IN THE PUS-IRS MATCH

	Matching Level					
	6 Correlation (.50)	5 Correlation (.70)	4 Correlation (.80)	3 Correlation (.90)	2 Correlation (.97)	1 Chi Sq (.99)
X Variable						
1. Number of children	1	1	1	1	4	8
2. Homeowner status	1	1	1	2	2	2
3. Wage earnings	3	4	5	9	20	36
4. Business earnings	1	1	1	1	1	13
5. Farm income	1	1	1	1	1	2
6. Total income	2	2	3	6	16	36

into the other if other information is present. In the PUS-IRS match, for example, adjusted gross income (AGI) in the IRS file was matched to total personal income in the PUS file. Because of the other information present in the IRS file, it was possible to add dividend exclusions and other adjustments back in to AGI to obtain personal gross income. The two concepts were still not identical, since gross income in the IRS file still excluded social security income but included capital gains, whereas total income in the PUS file included social security income but excluded capital gains. To align the two concepts, it was, in addition, necessary to subtract capital gains from gross income in the IRS file and subtract social security income from total income in the PUS file. In the other case, where X_a differs from X_b in sampling distribution, either because of differences in sampling frame or because of differences in reporting errors, it becomes necessary to align the distributions of the two variables. We did not encounter this problem in the PUS-IRS match.[4]

F. *The Sort-Merge Match and Calibration.* The match itself is executed in the following steps: First, the X variables in each record in the A file and in the B file are recoded into matching intervals. Second, the records in each file are sorted on the basis of their cohort values and, within cohort, on the basis of their matching interval values. Third, for each record in the A file, a search is made for a B record with identical matching interval values as that of the A record at the first (most detailed) matching level. If this fails, a search is made for a B record with identical matching interval values at the second matching level; if this fails, a candidate is searched for at the third matching level, and so on, to the cohort level. Once the matching level is established, the matching B record is randomly selected from all the B records that match the A record at this level.[5] The selected B record is then merged with the A record. Fourth, the distribution of the match by matching level is calibrated. If the distribution is uneven, new matching intervals are generated with a new set of probability levels and the match repeated.[6] This process is continued until the resulting calibration is relatively even. Three iterations were

[4]One possible fix-up procedure is to align the distributions of X_a and X_b on the basis of their rank order or percentile distribution. In effect the nth percentile value in the B file would be treated as equivalent to the nth percentile value in the A file. Before matching intervals for the B file are generated, the X_b values would be transformed to their equivalent values in terms of the X_a variable.

[5]This is actually true only if the B file is a random sample. If the B file is a stratified sample, the B record is chosen on a probability basis on the basis of the sample weights.

[6]The reason for this is that the match can be *improved* by re-specifying the matching levels. For example, suppose that 50 percent of the matches occur at level 4, correlation (.80), and that level 3 is correlation (.90). This indicates that a large proportion of the matches would likely occur at some correlation level between (0.80 and 0.90) – say, 0.85. Since a correlation level of 0.85 yields more narrow matching intervals than a correlation level of 0.80, the matches between the A and B records would occur at closer X values and the match thereby improved.

TABLE 5

CALIBRATION OF THE PUS-IRS MATCH

Matching Level	Percentage of Matches
1. Chi Sq (.99)	16.0%
2. Correl (.97)	18.8
3. Correl (.90)	30.6
4. Correl (.80)	14.3
5. Correl (.70)	12.2
6. Correl (.50)	6.2
7. Cohort	3.0
	100.0

necessary in the PUS-IRS match, and the final calibration is shown in Table 5.

III. An Empirical Test of the Matching Procedure

Though it is true, as Sims [6] points out, that we can say nothing about the joint distribution of Y_a and Y_b conditional on X, we can nevertheless say something about the (raw) joint distribution of Y_a and Y_b.[7] As shown previously [10], the stronger the correlation between Y_a and X, and Y_b and X, the closer the imputed joint distribution of Y_a and Y_b will be to the actual (unknown) joint distribution. In the case where Y_a and X, and Y_b and X are perfectly correlated, the imputed joint distribution will exactly replicate the actual joint distribution. However, the more the correlation coefficients deviate from positive or negative 1.0, the greater the likely error between the imputed and actual distributions.

In this section, we provide a statistical test of the matching procedure. The test is performed on the match of the 1970 Census 1/1000 5% Public Use Sample (PUS) to the 1970 Census 1/1000 15% PUS (see Table 6). The two datasets are random samples, of the same size, and identical in their variable list except for about a dozen variables.[8] As a result, almost all the Y and Z variables will be the same in the two datasets.[9] This thus allows a comparison of the imputed joint distribution of Y_a and Y_b with the actual *known* joint distribution. Moreover, the imputed joint distribution of Y_a and Z_b can be compared with the actual.

[7] We can say nothing about the conditional joint distribution, since this is precisely the information missing. As in almost all imputation procedures, we assume the relation is stochastic (see Section I).

[8] The designations "5%" and "15%" refer to the percentage of the population receiving the respective questionnaires. We performed this match to transfer information on consumer durable holdings for construction of household balance sheets.

[9] In fact, all the Y variables are the same.

TABLE 6

STRUCTURE OF THE 1970 PUBLIC USE SAMPLE 5% - 15% MATCH

A. *Cohort Variables*
1. Marital status
2. Age of head of household
3. Sex of head of household
4. Race of head of household
5. Owner-occupied home or rental unit

B. *X Variables*
1. Number of children in household
2. Value of property or gross monthly rental
3. Wage earnings of head of household
4. Wage earnings of spouse of head of household (if married)
5. Total family income

C. *Y Variables*[a]
1. Education of head of household
2. Education of spouse of head of household (if married)
3. Industry of employment of head of household
4. Occupation of head of household
5. Place of birth of head of household
6. Farm income
7. Professional income
8. Social Security income
9. Welfare income

D. *Z Variables*[b]
1. Hours worked per week by head of household
2. Weeks worked per year by head of household
3. Year last worked by head of household

E. *Matching Levels (and Calibration)*[c]
1. Chi-square (.99): 18.3%
2. Correlation (.98): 19.4%
3. Correlation (.97): 25.8%
4. Correlation (.93): 17.6%
5. Correlation (.90): 13.5%
6. Correlation (.80): 3.9%
7. Cohort : 1.5%

[a]All *Y* variables are common to both the 5% and 15% samples.

[b]Partial list of those common to the 5% and 15% samples.

[c]The calibration is in terms of the percentage of the total number of households matched at the indicated level.

and the imputed joint distribution of Y_a, Y_b, and Z with the actual for most Z variables.

We used a "Chow test" to compare the two joint distributions. A Chow test is a regression technique where the coefficients of an equation estimated using one sample are compared with the coefficients of the same equation estimated on a different sample. The test determines whether the full set of coefficients or any subset is statistically different in the two estimations. In our application we will use the Chow test to compare co-

efficient estimates of a regression on the original sample with one where one or more imputed (matched) variables has been substituted for the original variables. If the two sets of coefficients are statistically different, then the indication is that the imputed joint distribution is not a good replication of the actual joint distribution. If they are not statistically different, then the replication is good according to this criterion. This criterion, it should be noted, has limitations characteristic of regression techniques. In particular, a regression is a summary statistic, capturing only the first and second moments of the joint distribution of the regression variables (that is, the means and the covariance matrix).[10]

We ran the Chow test on a variety of combinations of X, Y, Z, and cohort variables to obtain a comprehensive picture of the relation between actual and imputed joint distributions. In all we ran six sets of regressions. In each set we ran four equations. In the first, variables from the PUS 15% sample were used; in the second, the left-side variable was drawn from the 15% sample and the right-side variables from the matched record of the 5% sample; in the third, the left-side variable was drawn from the 5% sample and the right-side variables from the matched 15% record; in the fourth, all variables were drawn from the 5% sample. We then ran a Chow test on each pair of equations, resulting in six separate Chow tests. The number of observations in each regression was 6341.

We used equations that are commonly found in the labor economics literature. The first equation was a regression of the logarithm of earnings (Log E) on years of schooling (S). Earnings is an X variable (in matching nomenclature) and schooling a Y variable. The resulting F-statistics of the Chow tests are shown in Table 7. The upper left-hand entry of the Table shows the results of comparing the regression of earnings on schooling from the actual 15% sample with a regression of the same equation where earnings is an original variable and schooling the imputed variable. The F-statistic indicates no significant difference in the coefficients. The next entry shows the results of comparing the actual 15% PUS regression with the regression of imputed earnings on actual schooling. Again, there is no statistical difference in the set of coefficients. The third entry on the first line indicates no significant difference in the coefficient estimates when the sample is drawn from the PUS 15% sample and when it is drawn from the PUS 5% sample. The other three entries indicate no significant difference in the coefficients from the remaining regression pairs. In the second equation we substituted income – another X variable – for earnings and found no significant difference in coefficient estimates between regression pairs. In the third equation, we added age (A), a cohort

[10]There is, of course, the added possibility that the actual *sample* distribution will differ from the *population* distribution. In our test we are interested only in the relation of the actual sample distribution and the imputed sample distribution.

TABLE 7

CHOW TEST F-STATISTICS FROM REGRESSIONS RUN ON THE PUBLIC USE SAMPLE 5% 15% MATCHED SAMPLE

Equation 1: $\text{Log}(E) = \beta_0 + \beta_1 S + u \quad (E > 0)$

	(E_{15}, S_{15})	(E_{15}, S_5)	(E_5, S_{15})	(E_5, S_5)
(E_{15}, S_{15})		2.140	1.027	2.515
(E_{15}, S_5)			0.449	0.081
(E_5, S_{15})				0.468

Equation 2: $\text{Log}(Y) = \beta_0 + \beta_1 S + u \quad (Y > 0)$

	(Y_{15}, S_{15})	(Y_{15}, S_5)	(Y_5, S_{15})	(Y_5, S_5)
(Y_{15}, S_{15})		0.180	0.923	1.930
(Y_{15}, S_5)			1.238	1.545
(Y_5, S_{15})				0.774

Equation 3: $\text{Log}(E) = \beta_0 + \beta_1 A + \beta_2 S + u \quad (E > 0)$

	(E_{15}, S_{15})	(E_{15}, S_5)	(E_5, S_{15})	(E_5, S_5)
(E_5, S_{15})		1.738	0.823	2.154
(E_{15}, S_5)			0.327	0.078
(E_5, S_{15})				0.388

Equation 4: $\text{Log}(Y) = \beta_0 + \beta_1 A + \beta_2 S + u \quad (Y > 0)$

	(Y_{15}, S_{15})	(Y_{15}, S_5)	(Y_5, S_{15})	(Y_5, S_5)
(Y_{15}, S_{15})		0.947	0.202	1.031
(Y_{15}, S_5)		--	1.612	1.172
(Y_5, S_{15})				0.634

Equation 5: $\text{Log}(E) = \beta_0 + \beta_1 A + \beta_2 S + \beta_3 H + \beta_4 W + \beta_5 R + \beta_6 M + u \quad (E > 0)$

A. Chow test on full set of coefficients

	$(E_{15}, (S, H, W)_{15})$	$(E_{15}, (S, H, W)_5)$	$(E_5, (S, H, W)_{15})$	$(E_5, (S, H, W)_5)$
$(E_{15}, (S, H, W)_{15})$		0.669	1.419	2.059
$(E_{15}, (S, H, W)_5)$			1.933	1.975
$(E_5, (S, H, W)_{15})$				2.556*

B. Chow test on (S, H, W)

	$(E_{15}, (S, H, W)_{15})$	$(E_{15}, (S, H, W)_5)$	$(E_5, (S, H, W)_{15})$	$(E_5, (S, H, W)_5)$
$(E_{15}, (S, H, W)_{15})$		0.551	1.047	1.238
$(E_{15}, (S, H, W)_5)$			1.901	1.484
$(E_5, (S, H, W)_{15})$				1.192

TABLE 7 (*continued*)

Equation 6: $W = \beta_0 + \beta_1 A + \beta_2 S + \beta_3 M \quad (W > 0)$

	(W_{15}, S_{15})	(W_{15}, S_5)	(W_5, S_{15})	(W_5, S_5)
(W_{15}, S_{15})	0.977	0.268		1.242
(W_{15}, S_5)		0.482		0.882
(W_5, S_{15})				2.627*

Key:
Log: logarithm
E: earnings
Y: income
A: age
H: hours worked
W: weeks worked
R: race
M: marital status
u: random error term
*Significantly different at the 5% significance level.
**Significantly different at the 1% significance level.

variable (and thus with identical values in the 15% and 5% sample records), to the first equation and again found no significant differences. In the fourth equation, age was added to the second equation with similar results.

In the fifth equation, the logarithm of earnings was regressed on schooling; age, race (R), and marital status (M), which are all cohort variables; and hours worked per week (H) and weeks worked per year (W), which are Z variables. In the first specification, earnings, schooling, hours worked, and weeks worked were drawn from the 15% sample; in the second specification, the earnings variable was drawn from the 15% record and the other variable from the 5% record; in the third, earnings was drawn from the 5% record and the others from the 15% record; in the fourth, all variables were drawn from the 5% record. (The remaining three variables—age, race, and marital status—are cohort variables, with identical values in the two files.) Chow tests on the equality of all coefficients indicated only one instance where the coefficients were significantly different at the five percent significance level. Chow tests on the equality of the coefficients of schooling, hours worked, and weeks worked showed no instances. In the sixth equation weeks worked, a Z variable, was regressed on two cohort variables and a Y variable. There was only one instance where the coefficients were significantly different.

These statistical results provide strong support that the imputed joint distributions resulting from the matching procedure are good replications of the actual joint distributions. In only 2 of the 42 Chow tests we performed were there significant differences in estimated coefficients be-

tween regressions involving original sample variables and regressions involving both sample and imputed variables. This test thus indicates that the sort-merge matching procedure can provide reliable synthetic data sources for many kinds of statistical applications.

United Nations Statistical Office
Yale University
New York University

REFERENCES

[1] Alter, Horst, "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.

[2] Budd, E. C., "The Creation of a Microdata File for Estimating the Size Distribution of Income," *The Review of Income and Wealth*, Series 17, No. 4, December 1971.

[3] Okner, Benjamin, "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.

[4] Ruggles, Nancy, and Richard Ruggles, "A Strategy for Merging and Matching Microdata Sets," *Annals of Economic and Social Measurement*, Vol. 3, No. 2, April 1974.

[5] Ruggles, Richard and Nancy D. Ruggles, "The Role of Microdata in the National Economic and Social Accounts," *The Review of Income and Wealth*, Series 21, No. 2, June 1975.

[6] Sims, Christopher A., "Comments" to Okner, op. cit., *Annals of Economic and Social Measurement*, July 1972.

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

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

[9] United Nations, *Towards a System of Social and Demographic Statistics*, Studies in Methods, Series F, No. 18 (New York, 1975, Sales No. E.74.XVII.8)

[10] Wolff, Edward N., "The Goodness of Match," NBER Working Paper No. 72, December 1974.