

This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: *Annals of Economic and Social Measurement*, Volume 3, number 2

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher: NBER

Volume URL: <http://www.nber.org/books/aesm74-2>

Publication Date: April 1974

Chapter Title: A Strategy for Merging and Matching Microdata Sets

Chapter Author: Nancy Ruggles, Richard Ruggles

Chapter URL: <http://www.nber.org/chapters/c10115>

Chapter pages in book: (p. 55 - 74)

A STRATEGY FOR MERGING AND MATCHING MICRODATA SETS*

BY NANCY AND RICHARD RUGGLES

This paper reviews the problem of integrating microdata sets with each other and examines a number of alternative approaches which have been used. A strategy for merging and matching microdata sets based on the use of statistically derived hierarchical sort tags is described with reference to the 1970 Public Use Sample and the Social Security Longitudinal Employer-Employee Data File. The formatting of microdata sets for merging into single data sets is also discussed.

In the past ten years sets of data which are samples of information about individual households and persons have emerged as a major tool of economic analysis. These microdata sets can be thought of as alternative and supplementary to the national accounts. For example, recent work of the Bureau of Economic Analysis of the Department of Commerce [1] shows how microdata can be used to supplement the information in the national accounts in studying the distribution of income for the household sector. In a somewhat different way, the work of the Brookings Institution [2] on tax models shows how the analysis of appropriate microdata sets can provide answers to major questions which could not be obtained with macrodata alone. Other uses of microdata sets for analyzing income maintenance schemes [3], the distribution of income of the aged [4], and more recently simulations of the demographic and social characteristics of the population [5] have been undertaken with a considerable degree of success.

Unfortunately, no single microdata set contains all of the different kinds of information required for the problems which the economist wishes to analyze. Different microdata sets contain different kinds of information. Thus for example, the microdata set containing information on tax returns does not include the kind of household social and demographic information which is available in the Survey of Economic Opportunity sample. It was this fact which led the Brookings Institution to create a single microdata set merging these two types of information. Ideally, one would like to combine for a given household and even for individuals within the household the different types of data which are available in a wide variety of different sources. Thus, it would be desirable to assemble, for each household or individual, census records, tax records, and social security records. For the researcher outside of government, any such assembly of data would raise problems of confidentiality, since as the amount of information about an individual increases, identification of a specific case is more likely to be possible. Nevertheless, within the Federal Government, considerable effort is being channeled into making such exact matches for significant bodies of data.

* This work has been supported by National Science Foundation Grant Project GS-33956.

In many instances, however, exact matches may not be theoretically possible. A great deal of information is collected on a sample basis. Where two samples are involved the probability of the same individual appearing in both may be very small, so that exact matching is impossible. Other methods of combining the types of information contained in the two different samples into one microdata set will be required.

One of the traditional ways of transferring information between data sets is by the use of regression analysis. Information is imputed from one data set to another by setting up a multiple regression model to predict for each case in sample A an estimated value of a variable contained in sample B. For this method to be successful, it is of course necessary that the two samples contain common variables which can serve as the independent variables in the regression equation. Thus for example, if one sample showed the union status of wage earners and their characteristics in terms of age, sex, race, occupation, industry, and income, union status information might be imputed to each wage earner in another file containing the same age, sex, race, etc., characteristics. The validity of such an imputation would of course depend on how well the variable which is being imputed (union status) is explained by the variables (i.e., the characteristics) which are in common. For many analytical purposes it would not be necessary for the estimate to be accurate at the individual observation level. It is merely necessary that the estimate perform satisfactorily on average over the existing range of variation. If the regression fit is quite close, the substitution of the regression value for an actual value may not invalidate subsequent analysis.

The technique of imputation by regression is considerably less satisfactory in transferring complex sets of information. Thus for example, if budget information is to be imputed to a sample containing richer social and demographic information, a problem arises in that budget outlays are all highly interrelated. A separate estimate for each outlay would produce an inconsistent budget pattern for any specific individual. One of the major objectives of collecting budget information, furthermore, is the study of the interrelationships among budget items—interrelationships which would be lost if each budget 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 in the original sample, such a model would be highly complex, especially if the actual relationships were not well approximated by a linear or log linear additive model. A simpler and somewhat 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. Imputation by regression would normally result in assigning mean values, whereas the matching technique reproduces the distributions 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. The success of the matching technique depends on the data being quite dense, so that similar

cases can be found in both data sets. It should also be noted that for matching purposes no specific functional relationship need be determined in advance. 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. 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.

SPECIFICATION OF THE MATCHING PROBLEM

If two data sets are to be merged and the observations within them matched with each other, formal procedures should be set up so that there are objective and valid criteria for making matches. Consider for example two data sets: (A) the 1970 Public Use Sample (PUS), and (B) the Social Security Longitudinal Employer-Employee Data file (LEED) as candidates for merging. These will have certain variables $x_1 \dots x_n$ in common. There will be $y_1 \dots y_n$ variables in the Public Use Sample which are not available in the LEED file, and conversely there will be $z_1 \dots z_n$ variables available in the LEED file which are not available in the PUS file. Table 1 below indicates exactly what these variables are. For the matching to be valid, the common x variables must separate the observations into analytically meaningful groups. Trivial x variables which are unrelated to any of the y and z variables would merely result in a stochastic matching.

It may be that for some x variables a derived value will have to be created in one of the files. Thus for example, the year last worked is not explicitly given in the LEED file, but it can be derived from the longitudinal work history. There is also a very serious problem of alignment, in that an x variable in one data set may not correspond exactly to the corresponding x variable in the other data set. For example, the wage information collected by the Bureau of the Census will not correspond for both definitional and statistical reasons to the wage information reported to the Social Security Administration. On the one hand, the wage information in the Public Use Sample refers to all wages, whether or not covered by the social security system. On the other hand, the level of accuracy of the Social Security wage reporting, where given, is statistically better than the corresponding information in the PUS. Differences in definition can sometimes be taken into account. Thus for example if a person's occupation or type of employment as shown in the Public Use Sample is such that he is obviously not covered by the social security system, no attempt would be made to find a match in the LEED file. If, after adjusting for differences in coverage, the distributions of wages in the two files are still markedly different, a further statistical adjustment will be needed to align the two sets of information. In this particular case the alignment will probably involve adjusting the Public Use Sample wage data so that it more nearly conforms to the wage information in the LEED file.

TABLE 1

VARIABLES IN 1970 PUBLIC USE SAMPLE (PUS) AND THE LONGITUDINAL EMPLOYER-EMPLOYEE DATA (LEED) FILE

A. Public Use Sample		B. LEED File	
x variables			
x_1	Age	x_1	Age
x_2	Sex	x_2	Sex
x_3	Race	x_3	Race
x_4	State	x_4	State
x_5	Hours worked	x_5	(Hours worked—derived)
x_6	Year last worked	x_6	(Year last worked—derived)
x_7	Current industry	x_7	Current industry
x_8	Class of worker	x_8	Class of worker
x_9	Employment status	x_9	(Employment status—derived)
x_{10}	Worked last year	x_{10}	(Worked last year—derived)
x_{11}	Weeks worked	x_{11}	(Weeks worked—derived)
x_{12}	Wage	x_{12}	Wage
x_{13}	Work status 5 years ago	x_{13}	(Work status 5 years ago—derived)
x_{14}	State 5 years ago	x_{14}	(State 5 years ago—derived)
x_{15}	Industry 5 years ago	x_{15}	(Industry 5 years ago—derived)
y variables		z variables*	
y_1	Basic relationship in family	z_1	Employee identifier (scrambled)
y_2	Detailed relationship	z_2	Number of years employee in file
y_3	Subfamily number	z_3	Year of employment
y_4	Type of group quarters	z_4	Number of employers for z_3
y_5	Spanish surname	z_5	Employer identifier
y_6	Quarter of birth	z_6	Number of wage items
y_7	Marital status	z_7	Annual wages
y_8	Place of birth	z_8	First quarter wages
y_9	Highest grade attended	z_9	Second quarter wages
y_{10}	Finished grade	z_{10}	Third quarter wages
y_{11}	Children ever born	z_{11}	Fourth quarter wages
y_{12}	Current occupation	z_{12}	Total estimated wages
y_{13}	In armed forces 5 years ago		
y_{14}	In college 5 years ago		
y_{15}	Business earnings		
y_{16}	Farm earnings		
y_{17}	Social Security income		
y_{18}	Welfare income		
y_{19}	Other income		
y_{20}	Persons total income		
y_{21}	Poverty status		
y_{22}	Persons in family		
y_{23}	Subfamily relationship		
y_{24}	Family unit membership		
y_{25}	Spanish descent		
y_{26}	Citizenship		
y_{27}	Year of immigration		
y_{28}	Times married		
y_{29}	Age at first marriage		
y_{30}	Quarter of first marriage		
y_{31}	Vocational training		
y_{32}	Field of training		
y_{33}	Disability		
y_{34}	Occupation 5 years ago		

* Available on each individual quarterly by each employer paying FICA tax for a 12-year period.

The problems of definition and alignment of the x variables for matching purposes are extremely important and may consume a large part of the energy of a matching effort. Certainly the quality of the ultimate match will depend on how thoroughly the definitional adjustment and alignment of the x variables in the different data sets has been carried out. This topic deserves a paper in its own right, but it is not the function of this paper to cover it. Rather, the remainder of this paper focuses on an examination of different strategies of merging and matching microdata sets which do contain already-aligned x variables.

The process of matching involves comparing values of the x variables in one data set to the values of x variables in another data set in order to bring together observations from the two data sets. The central question in this process resolves itself into the choice of criteria to determine a match. Where the values of the x variables in sample A precisely match the values of the x variables in sample B there is no problem. In such an instance the observations in files A and B having identical values for the x variables can be matched on a stochastic basis. In the absence of additional matching information it is not possible to do better than this. The real problem arises when the values of the x variables in the two data sets differ somewhat, and it becomes necessary to decide which combination of x values is most satisfactory for determining the match.

Conceptually, a distance function could be constructed to express the difference between the values of all the x 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, an analytic measurement of what is meant by the difference between the values of the x variables is required.

In principle, the x variables are intermediate in the sense that their function is to bring the y and z variables together synthetically. Although it is true that there is no information in either data set about the joint distribution of the y and z variables conditional on x , information is available on the joint distributions of the x and y variables and of the x and z variables, and this information is relevant to the creation of a satisfactory distance function. If outside information on the joint distribution of y and z conditional on x is available it could be introduced as part of the matching criteria; this possibility is not being considered here. If the matching is undertaken for a specific analytic purpose, certain y and z 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. For such a purpose the y and z variables themselves can be used as general criteria for determining whether two observations are similar.

METHODS OF DETERMINING MATCHES

One approach to developing distance functions is to use multivariate regression analysis, in which the dependent variables are the y and z non-matching

variables, and the independent variables are the x variables, to determine the weights to be attached to each of the x variables to get the best explanation of the y and z variables. From such information a distance function can be constructed. The paper by Horst Adler illustrates the use of such a procedure by Statistics Canada [6].

The work by Okner in merging the Survey of Economic Opportunity files with the tax model files 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 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 at BEA on merging the Current Population Survey files and the tax model files 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 rank classes, and within these, by self employment income and property income. The actual 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 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.

A somewhat different approach was developed by Richard Rockwell to match the 1970 Public Use Sample with the Survey of Economic Opportunity file. In this match five variables classified into quite broad intervals were used to cross-classify the data into 288 cells. Within these cells matches were achieved by using three additional variables successively to arrive at a final match. The Rockwell result could actually have been achieved by a pure sort and merge process, since the cross tabulation cell matches are based on sequential ordering of the three additional variables.

ELABORATION OF CROSS TABULATION TECHNIQUES OF MATCHING

The matching process could in fact be carried out by means of an n -dimensional cross tabulation using all of the x variables, with matches being made stochastically among observations falling in the same cell. This process would produce results different from those obtained by the use of a distance function, since it is quite possible that two observations lying at the opposite boundaries of a cell would be matched with each other, whereas if a distance function had been used observations lying near the boundary of one cell would be matched to observations near the boundary of an adjacent cell. Another disadvantage of the cross tabulation technique is that for any given cross tabulation the density of observations in some cells may be quite high so that closer matching could have been achieved either by use of a distance function or by finer cross classification. Furthermore, cross

tabulation may result in cells which contain one or more observations of sample A and no observations of sample B, and vice versa.

These difficulties could be resolved by using extremely fine cross classification grids to begin with, matching those cases which can be matched, and gradually on an iterative basis increasing the cell size until a complete matching of all observations is achieved. Such a procedure does face exactly the same basic problem as other techniques: some objective criterion will be needed to determine the intervals of the x variables to be used to develop the cross tabulations. The intervals of the x variables, furthermore, will depend not only upon the relationship of the x variables to the y and z variables, but also upon the density of the observations over the variable space. Finer cross classifications are appropriate and possible for large samples, and higher quality matches can be achieved without excessive cost.

This matter of cost is of considerable importance, since if matching techniques are employed which require the comparison of the observations in the two files to determine the best match, the cost of handling very large data sets becomes prohibitive. With large samples, therefore, some adaptation of the cross tabulation technique of matching becomes quite attractive.

THE SORT-MERGE STRATEGY FOR MATCHING

It is quite possible to accomplish the same result as iterative cross tabulation processing by a single sort of the files which will yield a hierarchical nesting of cross tabulated cells. Sorting is in fact the traditional method of producing cross tabulations. In order to create a hierarchical nesting of cells, a series of sets of sort tags, each representing one level of the hierarchical nesting, is attached to each observation. The first (left-most) set of sort tags determines the broadest cells which are to be used. To create this first set of sort tags each x variable is partitioned, on some basis, into broad intervals. These interval specifications for all variables constitute the set of sort tags which define the cell boundaries for the cross-tabulation. This procedure is somewhat similar to the equivalence class concept used by Okner. Within the initial broad cells, a second set of sort tags is then created to introduce finer classifications. This is accomplished by partitioning each of the x variables into somewhat narrower intervals. This process can be repeated until, if desired, the raw values of the x variables are reached. The process, in other words, is one of taking a fairly broad cell and breaking it up into smaller cells, and then taking each of these smaller cells and breaking it up into even smaller cells, the process continuing until extremely small cell sizes are reached. Sorting the two data sets according to these nested sets of sort tags, and then merging the two data sets, will yield a merged data set in which the observations which are closest to each other will by definition fall within a common cell at some level of the hierarchy, as long as there is at least one A and one B observation in the first level of the hierarchy (i.e., the broadest cell): thus a match will be assured at some cell size. The size of the cell at which the match occurs will of course depend on the density of the observations in the two samples. In very large samples, quite fine classifications of the x variables can be used at the most detailed level of the hierarchy since a substantial number of matches may be expected to occur at that level of specification. In smaller samples where matches are less likely, broader classifications will have to

be used. This is another way of saying that higher density samples can produce better matches.

THE STATISTICAL MEASUREMENT OF DIFFERENCES BETWEEN INTERVALS OF THE MATCHING VARIABLES

The determination of the intervals of the x variables which are to be used as cell boundaries is central to the problem of matching. Ideally, one would like to have the assurance that within a specified interval of a given x variable the distributions of the y and z variables are invariant. In other words, it should only be necessary to distinguish between one interval of x and another if doing so results in significantly different distributions of some y or z variable.

To test this, the observations falling into two different specified intervals of an x variable can be treated as different samples. If the probability that these samples come from different universes is low, this means that there is no statistical basis for maintaining the distinction between these intervals for matching purposes. Conversely, if the probability is high that the samples for the specified intervals come from different universes, it will be important to utilize this information in developing matching criteria.

The chi-square test can be applied to the y and z distributions for different intervals of an x variable to determine whether the observed differences are significant. Where the number of observations is small, it may not be possible to detect differences between intervals of an x variable even where such differences may actually exist. On the other hand, where the number of observations is very large, even relatively small differences in the y and z distributions of observations for different intervals will result in highly significant chi-square values. To the extent possible large samples should be used to determine the significance of the observed differences: in some cases this may mean that stratified samples should be sought so that an adequate number of observations will be available for each value of the x variables.

Where significant differences are found in the y or z distributions for different intervals of x , it will then be necessary to make a further evaluation of the relative importance of these differences, in order to provide the basis for the hierarchical nesting of cells based on different intervals of the x variables. This can be done by measuring how closely the percentage distributions for the y and z variables are correlated for any two specific intervals of x . If the two percentage distributions are the same, they would lie on a 45-degree regression line, and the correlation coefficient would be 1.00. If the two percentage distributions differ, the correlation coefficient will indicate the size of this difference. Where the correlation is high for specified x intervals, collapsing these intervals to a single interval for matching purposes will result in less distortion in the y and z variables than it would if a low correlation exists. What is being asked is whether the combined interval of x is a good proxy for either one alone. If the correlation coefficient indicates that one x interval will produce about the same distribution for the y and z variables as the other x interval, a combined interval will be a satisfactory proxy. This statistical measure makes it possible to specify the hierarchical levels of the sort tag in terms of different levels of the correlation coefficient.

Thus two criteria have been introduced. The chi-square criterion is intended to determine whether the distributions for the y and z variables accompanying two intervals of an x variable are significantly different from each other, based upon both sample size and the observed differences in the distributions. In those instances where no significant difference is found, intervals can be combined without doing violence to the match. Where significant differences are found, the importance of these differences needs to be evaluated. The correlation measure asks *how* different the distributions are, in terms of how much of the total variance in the distributions of the y and z variables is explained. Where the unexplained variance is very small (i.e., where the correlation is high), the two intervals of the x variable may be combined without significantly altering the distribution for the y and z variables in question. Both measures, chi-square and correlation, are necessary to provide valid and meaningful distinctions. On the one hand, with very large samples, chi-square may be large, but the correlation coefficient may also be large. On the other hand, with small samples, a low chi-square may accompany a low correlation coefficient. In the first instance, there is a statistically significant difference between the distributions but the difference is trivial, so that combining the intervals will do no violence to the matching process. In the second instance, there is a large difference between the distributions but it is not statistically reliable, and so should not be used as a matching criterion. Only when a relatively high chi-square is combined with a relatively low correlation is maintenance of the distinction between two intervals desirable.

Specific examples of how the chi-square and correlation measures are applied may help to clarify the analysis. Table 2 shows how two intervals of the x variable "work status" are related to the y variable "size of family." The question which is posed is whether the distinction between the interval "employed at work" and the

TABLE 2
DISTRIBUTION OF FAMILY SIZE BY WORK STATUS OF EMPLOYED WORKERS

y variable Size of Family (Number of Persons)	x variable: Employment Status			
	Employed at Work		Employed Not at Work	
	Number of Observations	Percent	Number of Observations	Percent
1	973	11.9	16	13.2
2	1602	19.5	26	21.5
3	2487	30.3	31	25.6
4	1740	21.2	29	24.0
5	846	10.3	13	10.7
6	329	4.0	5	4.1
7	135	1.6	1	0.8
8	52	0.6		
9	19	0.2		
10 and over	12	0.1		
TOTAL	8195	100.0	121	100.0

Comparison between distributions:

Chi square probability 0.0086 (based on distributions of number of observations)

Correlation coefficient 0.9852 (based on percentage distributions).

interval "employed not at work" results in significantly different family size distributions. The chi-square test gives a very low probability that the observed difference in the distributions is significant. For the y variable "size of family," therefore, it can be determined that there is no statistical reason not to combine the two intervals of work status into one for matching purposes.

In Table 3, the x variable is "class of worker," and the y variable is "business income." Chi-square is 1.000, indicating that the difference between the distributions of business income for "employed" and "self-employed" is statistically significant. The low correlation coefficient indicates that the difference is important. It is therefore important to maintain the distinction between employees and self-employed as a matching criterion, if business income is one of the y variables.

TABLE 3
DISTRIBUTION OF BUSINESS INCOME FOR EMPLOYEES AND SELF-EMPLOYED

y variable	x variable: Class of Worker			
	Employee		Self-employed	
	Number of Observations	Percent	Number of Observations	Percent
-9900 - 100	19	4.4	7	0.7
0-200	74	17.3	32	3.2
201-600	80	18.7	52	5.2
601-1,000	37	8.6	52	5.2
1,001-1,300	10	2.3	40	4.0
1,301-2,000	45	10.5	116	11.5
2,001-2,500	23	5.4	64	6.4
2,501-3,200	26	6.1	87	8.6
3,201-4,100	23	5.4	129	12.8
4,101-5,000	25	5.8	108	10.7
5,001-7,600	40	9.3	152	15.1
7,601-15,500	23	5.4	146	14.5
15,501-24,500	2	0.5	16	1.6
24,501 and over	1	0.2	6	0.6
TOTAL	428	100.0	1007	100.0

Comparisons between distributions:

Chi square probability 1.000 (based on distributions of number of observations)

Correlation coefficient 0.1479 (based on percentage distributions).

In Table 4, the x variable is "class of worker," and the y variable is "size of family." The chi-square of 0.9536 indicates a strong probability that the observed difference between the two distributions of size of family is statistically significant. However, the correlation coefficient is also high, indicating that in terms of total variance the differences between the two distributions are small. Keeping government employees and private employees in separate intervals for matching purposes would therefore not appreciably improve the attribution of family size.

THE PARTITIONING OF A MATCHING VARIABLE INTO INTERVALS

Application of the chi-square and correlation measures as criteria for partitioning x variables requires the development of suitable algorithms which can be

TABLE 4
DISTRIBUTION OF FAMILY SIZE FOR PRIVATE AND GOVERNMENT EMPLOYEES

x variable: Class of Worker				
y variable Size of Family (Number of Persons)	Private Company Employee		Government Employee	
	Number of Observations	Percent	Number of Observations	Percent
1	869	12.4	186	13.6
2	1394	19.9	279	20.4
3	2075	29.6	439	32.1
4	1445	20.6	288	21.1
5	728	10.4	115	8.4
6	289	4.1	38	2.8
7	124	1.8	13	1.0
8	50	0.7	6	0.4
9 or more	17	0.2	3	0.2
Total cases	8537	100.0	1707	100.0

Comparison between distributions:

Chi Square Probability 0.9536 (based on distributions of number of observations)

Correlation Coefficient 0.9966 (based on percentage distributions).

embodied in computer programs to process the data and report out the results in an intelligible form. Different algorithms will be required depending on whether the x variables are (1) well ordered, or (2) non-ordered or partially ordered. Wage income is an example of a well-ordered variable. Race and class of worker are non-ordered, and such variables as industries or regions and states are partially ordered into hierarchical sets.

For a well-ordered variable with a relatively small number of raw values and a large number of observations for each raw value, the procedure is quite straightforward. The distributions of y and z variables for adjacent intervals of the raw values of the x variable are compared and the chi-square and correlation measures computed. If no significant difference is found or if the size of the difference is below a given level, the raw values are combined. A comparison can then be made between the newly combined interval and other intervals adjacent to it. In this way the x variable can be partitioned into a set of intervals based on specified levels of chi-square and correlation coefficients.

In some cases a well-ordered x variable may have an inconveniently large number of raw values. Thus the variable "wages" in the Public Use Sample consists of 250 intervals of \$100, and the LEED file reports wages in \$1 units. Instead of comparing each raw value, a different procedure is used. The x variable is arbitrarily partitioned into a relatively small number of intervals which are then compared. Where significant differences are found, each of these intervals is split into two intervals, and these are compared. This process continues until either no significant differences are found between intervals or raw values are reached. Various techniques could be used to partition the x variables into broad intervals, but the one which has been adopted is based on ordering the sample on the x variable and

dividing it into eight major segments, each of which has the same number of observations. This approach assures that the resulting intervals will contain an adequate number of observations to provide reliable comparisons, and that optimal use can be made of the sample size.

The only difference in the procedures for analyzing well-ordered x variables with few raw values and well-ordered x variables with many raw values is that in the former case smaller intervals are aggregated into larger intervals whereas in the latter case large intervals are disaggregated into smaller intervals.

For non-ordered x variables, the concept of adjacent intervals is not meaningful. It will therefore be necessary to make all possible pairwise comparisons between intervals in order to determine which can be combined. For partially ordered or hierarchical variables, the comparisons are first made at the broadest group level (e.g., major industry or region). For these groups all possible pairwise comparisons would be made. Where separate groups are identified, pairwise comparisons would be carried out for sub-groups within the major group. This process would be continued until the hierarchical ordering is exhausted.

It should be apparent that the specification of the chi-square and correlation criteria for combining intervals will determine the number of intervals in the partitioning. If even a small difference between intervals is considered statistically significant and important then there will be more intervals. If large differences are tolerated then the number of intervals into which the x variable is partitioned will be reduced. Thus by using different levels of chi-square and/or correlation coefficients as criteria, different levels of partitioning will be generated, yielding a hierarchical set of intervals.

An x variable is generally analyzed in terms of more than one y or z variable. It is therefore necessary to consider how a generalized partitioning is to be derived from the individual partitionings resulting from individual y , z variables. Two different rules could be applied. First, it would be possible to construct the generalized partitioning so that it would reflect the most detailed intervals represented in the individual partitionings. Second, it would be possible to pool the percentage distributions for all the y , z variables and compute the correlation coefficient on the basis of these pooled distributions.

An example of an x variable (wages) which has been partitioned into three nesting sets of wage intervals is shown in Table 5. The raw wage values consisted of 250 wage classes of \$100 each ranging from \$1-99 to \$25,000 or more. In making the interval analysis 27 y variables were used. At the most detailed hierarchical level (level 3) only those wage classes were combined where the chi-square measure of the difference between intervals for every y distribution was less than 0.95. This criterion resulted in 21 intervals, ranging in size from \$100 to \$13,200 and including from 0.7 to 13.1 percent of the observations. It should be pointed out that the wide wage class for the 21st interval (i.e., 11,800-25,000 and over) is due in large part to the relatively small number of observations in this range. The sample on which these runs were made contained about 20,000 observations; this means that about 300 observations were in the 21st interval. An increase in sample size and/or the use of stratified sampling would probably have resulted in the 21st interval being broken down into additional intervals. In terms of the matching process, such finer intervals would improve the matching for only 1.7 percent of the data to be

TABLE 5
PARTITIONING OF WAGE CLASSES INTO INTERVALS

Wage Classes (Dollars)	Hierarchical Levels					
	Level 1		Level 2		Level 3	
	Interval Number	% of Observations	Interval Number	% of Observations	Interval Number	% of Observations
1- 99	1	31.7	1	31.7	1	3.3
100- 499	↓	↓	↓	↓	2	9.8
400- 599					3	2.1
600- 799	↓	↓	↓	↓	4	3.4
800- 1,799					5	13.1
1,800- 2,299	2	68.3	2	39.6	6	6.9
2,300- 2,799	↓	↓	↓	↓	7	6.0
2,800- 3,499					8	9.1
3,500- 3,899	↓	↓	↓	↓	9	5.1
3,900- 4,299					10	6.0
4,300- 4,499	↓	↓	↓	↓	11	1.7
4,500- 4,899					12	4.8
4,900- 5,299	↓	↓	3	14.7	13	6.3
5,300- 5,499			4	5.8		
5,500- 6,299	↓	↓	↓	↓	14	1.4
6,300- 7,499					5	3.0
7,500- 8,499	↓	↓	↓	↓	15	7.0
8,500- 9,099					6	2.0
9,100- 9,799	↓	↓	↓	↓	16	5.8
9,800-11,799					7	1.5
11,800-25,000*	8	1.7	8	1.7	17	3.0
					18	1.3
					19	0.7
					20	1.5
					21	1.7

* Top income class is \$25,000 and over.

Specifications for combining intervals:

If chi-square is in the range between 0.00 and 0.94 intervals will be combined irrespective of correlation coefficient.

If chi-square is in the range between 0.95 and 1.00, intervals will be combined if the correlation coefficient is above the levels shown below for the different hierarchical levels:

Hierarchical Level	Correlation coefficient
1	0.70
2	0.90
3	1.00

matched, but for research where analysis of the highest wage classes is important, however, special attention might well be directed to improving matching in these wage classes. For level 2, the criterion for combining intervals used for level 3 was relaxed to combine, in addition, intervals where chi-square was more than 0.95 but the correlation coefficient exceeded 0.90. This reduced the number of intervals to eight, with a minimum wage class size of \$1000 and minimum coverage of 1.5 percent of the observations. It is interesting to note that four of the 21 intervals specified at level 3 of the hierarchy were carried over unchanged at level 2 of the hierarchy. Finally by relaxing the correlation coefficient criterion to 0.70, the eight intervals at level 2 of the hierarchy collapse to two intervals for level 1. At this level the two income classes distinguished are \$1-1,799 and \$1,800 and above. The first interval contains 32 percent of the observations.

It is of course possible to generate as many hierarchical levels as desired. For some of the x variables, however, it may be decided that exact matching is needed. This would be somewhat similar to defining equivalence classes within which all matching is required to take place: observations in different equivalence classes would never be matched with each other. Three possible candidates for such an exact match are age, sex, and race. Exact matching on these variables would have the advantage that specific age, race, and sex cohorts would be recognized in both files, and the mean values and distributions of the y , z variables for these cohorts would not be affected by the matching process.

THE OPERATIONAL PROCESS OF MERGING AND MATCHING DATA FILES

Once the concepts have been developed for establishing hierarchical levels of sort tags based upon intervals of x variables derived from the comparison of distributions of the non-matching y and z variables, the foundation is laid for matching and merging any two data files with each other. The validity of such a match will depend on (1) the adequacy of the x variables as the basis for the match, (2) the correspondence of the different concepts of the x variables in the two samples and their alignment, and (3) the density of the observations in the two samples which are being matched. Unless all these conditions are adequately met, the matching process will not be satisfactory, and the merged body of data will probably not be very useful for any kind of analysis.

To some extent, the importance of these various conditions can be tested experimentally by splitting a large data set in half, and then carrying out the process of matching the two halves with (a) different combinations of matching variables, (b) stochastic or systematic biases which have purposely been introduced into specific variables, and (c) varying sample densities. Since matching a sample against itself can provide information on how the relationships resulting from the match correspond to the actual relationships, some measure of the adequacy of the matching process under various experimental conditions can be obtained. The NBER matching project is now carrying out such experiments, to determine the sensitivity of the matching process to different kinds of limitations.

The actual process of merging and matching data sets breaks down into a number of different steps. *First*, the two data sets which are to be merged and matched must be formatted in such a way that observations can be uniquely identified. *Second*, the interval of the x variables which are appropriate for each level of the hierarchical sort tags must be derived. *Third*, a new file for each data set must be created, containing only the identifier for each observation and a hierarchical sort tag based upon its values of the x variables. *Fourth*, the new tagged but unmatched files must be sorted in the same order, producing another matched file in which specific observations in data sets A and B are linked. *Finally*, the linkage between matched observations must be introduced into one of the data sets and the data set sorted in such a manner that the full sets of information for the matched observations can be brought together by merging the two data sets.

Formatting the data sets. It is unfortunately true that data sets which are to be merged and matched usually have quite different formats. Often there are no unique identifiers for the different observations, and this information will need to

be added so that a specific observation from one file can be linked with a specific observation in the second file. It is also important, in any merged file, to be able to identify which information came from what source. Finally, it should be possible to introduce new kinds of information into a file without disturbing existing information. To meet these needs, a special 80 character record has been created consisting of a 20 character information tag and 60 characters of data. The information tag serves the function of uniquely identifying the observation, indicating the source of the data, and providing information about the format of the data. The contents of the information tag are as follows :

<i>Information Tag</i>	<i>No. of characters</i>
Identifier for observation	10
Source	2
Information type	2
Item	2
Line format	2
Sequence number	2
Total	20

The serial identifier provides an identification for each observation. In the Public Use Sample the identifier is broken down as follows :

<i>Identifier</i>	<i>No. of characters</i>
Household serial number	6
Type of unit in household	2
Serial number of unit in household	2
Total	<u>10</u>

Type of unit is used to differentiate between records referring to (1) the household, (2) the family, (3) the sub-family or (4) the person. In the LEED file, the identifier for individuals is assigned on a sequential serial basis (7 characters), followed by a work history identifier (3 characters).

The tag for source identifies the origin of the data in the 60 character data portion of the record. By using alphanumeric source references the two character source tag permits approximately 1,300 sources to be identified. The information type, item, and line tags are used to designate the format of the data record itself. The item tag within this set permits keeping track of multiple sets of data which have precisely the same format. The continuation tag allows the 80-character record to be extended by additional 80-character records as supplements. Such a device permits text material such as comments, footnotes, etc., to be introduced at a specific point in a file without affecting the data. In other words what the information tag accomplishes is (1) identifying a specific observation, (2) indicating the source of the information, and (3) specifying the format in which the data is classified. The system is open ended in that different kinds of data from additional sources of information can be added at any time without disturbing the existing record. Programs which are designed to run on the original file will continue to operate on augmented files.

The Public Use Sample household and person records are each 120 characters long. These were split into two household records, each containing 60 characters of data, and two person records, each also containing 60 characters of data. The conversion to tagged records did not increase the size of the file, since only one 80-character person record was required for individuals of 14 years of age or less. The second record was not required, since it contains only information which is not applicable to individuals 14 years and under—information such as times married, veteran status, and employment history.

In the case of the LEED file, the original data came in variable records, from 92 characters to 32 thousand characters in length. In reformatting this file, one type of record was created for employee information, and another for employer work history information. A given individual would have one basic employee record and as many employer records as required to cover his work history.

Recasting the different data sets into compatible formats makes it possible not only to use common programs for handling and processing the different files, but to merge the data sets after linkages have been made between specific observations. The new merged file will then contain data for the linked observations from both sources, in such a way that both the source and the format of the data are easily identified.

Derivation of the intervals for matching variables. The derivation of the proper intervals of each x variable for each hierarchical level of the sort tag constitutes the heart of the matching process. A program named $I(x)$ has been developed which will for any given x variable create sort tags based upon chi square and correlation criteria applied to the distributions of specified y or z variables. The conceptual basis of this derivation has already been discussed for well-ordered variables, non-ordered, and hierarchical variables. This program can be run on samples of the data sets rather than the full data sets which are to be matched, or if desired can be run within age and sex cohorts, in order to determine whether different intervals should be used in the matching of different age and sex groups. The input required for this program includes the distributions of the y, z variables for each possible interval of the x variables. The $I(x)$ program also requires as input the chi square and correlation criteria which are to be used in determining the intervals of x for each hierarchical level in the sort tag. These criteria can easily be altered so that the program can generate different sets of hierarchical sort tags which are based on different criteria.

Creation of sort tags. Given the output from the $I(x)$ program for each x variable, the next task is to apply the $I(x)$ criteria to each observation in both data sets, and attach to every observation the hierarchical sort tags required for the matching process. For this purpose a tagging program examines the values of the x variables for each observation and produces a new file containing an appropriate set of hierarchical sort tags attached to the identifier for the observation.

The linking of observations. The linking of observations is achieved by sorting the file of tagged identifiers for each data set in the order of the hierarchical sort tag. It is then possible to process the two sorted tagged files to find for each observation in data set A the closest match in data set B. The closest match in this sense means the observation for which the sort tag matches at the lowest (most detailed) possible hierarchical level. Since the data are fed from the sorted tagged files

sequentially, this comparison can be made simply and at low cost. It should be noted that what is being done in this process is that each observation in data set A is being matched with the best possible choice in data set B in accordance with the $I(x)$ hierarchical sort tags. If a match of data set B with data set A is wanted, it is merely necessary to alter the program, so that for each observation in B the best match from data set A is chosen. From an analytic point of view, it may in fact be desirable to generate a single data set in which the best possible matches of both A with B and B with A are represented.

An example of a portion of merged tagged files used to link observations is given in Table 6. In this example each observation is identified by a person serial number and by a source number which indicates which file (A or B) the observation comes from. Exact matching is done on age, sex, and race (30 year old, white, males) and 10 other x variables ($a-j$). Six hierarchical levels specified by the $I(x)$ program are used. For each observation a set of sort tags was generated for each hierarchical level and both files were sorted and merged on the basis of the sort tags. The objective of this matching was to find for each observation in file A (source 31) the closest observation from file B (source 32). The underlined sort tags indicate the level at which the match is made.

It is obvious that the specification of the hierarchical levels will determine the level at which matches take place. If the specification is such that almost all matches take place at the most detailed level, the quality of the match could be improved by introducing stricter chi-square and correlation criteria to increase the number of intervals in the sort tag. If almost all the matches occur at the broadest level of hierarchical sort tags, this would mean that the more detailed intervals are not useful, given the sample size, and the efficiency of the matching process could be improved by somewhat relaxing the chi-square and correlation criteria. The exact calibration of the chi-square and correlation criteria thus depends on the matching process itself: experimental runs with the sets of data to be matched can be used to provide the necessary calibration.

Merging the basic data. Once the identifiers in data set A have been linked with identifiers in data set B, the problem resolves itself into purely a sort-merge process. Probably the simplest way to accomplish the sort and merge of the two files is to sort the linked identifiers in the order of the identifiers for file B, and assign the file A identifier as a sort tag to the identifier in file B. Where an observation in file B is used more than once, it will be necessary to replicate the data accordingly. It is then only necessary to sort the records in data set B in the order of the identifiers for file A, and merge the resulting file with file A. The merged file will then contain the final results of the matching process.

SUMMARY

The strategy of merging and matching data which has been outlined here was designed primarily to provide for systematic processing of information based upon objective rules and criteria. An attempt was made to make maximum use of the information contained in the data sets about the relationships between the matching and non-matching variables. The explicit utilization of a distance function was rejected not only because it was difficult to design conceptually but also

TABLE 6
MERGING AND MATCHING MICRODATA
An Example of Merged Tagged Files

Person Serial Number	Source	Age	Sex	Race	Tag Level 1 abcdefg h i j	Tag Level 2 abcdefg h i j	Tag Level 3 abcdefg h i j	Tag Level 4 abcdefg h i j	Tag Level 5 abcdefg h i j	Tag Level 6 abcdefg h i j
35068201	31	30	0	0	1111110101001	11121110101012	13121122001015	12221143003016	12226243804015	71216133934010
45554201	32	30	0	0	1111110101001	11121110101012	13121122001015	32121153003016	32126353806015	11316343942010
76429601	31	30	0	0	1111110101001	11121110101012	13121122001015	12221152903014	12226253706014	71216143840110
57084401	32	30	0	0	1111110101001	11121110101012	13123142001015	12224132802016	12225233602015	71214123728010
64372701	32	30	0	0	1111110101001	11121110101012	13123142001015	12224133003016	12225233802015	71214123920010
23954701	31	30	0	0	1111110101001	11121110101014	13121342001017	12221233003017	12226433805117	71216423937130
44695901	32	30	0	0	1111110101001	11121110101014	13121342001017	12221233003017	12226433806117	71216423943130
13223501	31	30	0	0	1111110101001	11121110101015	1312142001099	12221133003105	12226133806109	71216223942119
84284601	32	30	0	0	1111110101001	11121110101015	1312142001099	12221133003105	12226133806109	71216223942119
33173501	31	30	0	0	1111110101001	11121110101015	1312142001099	12221133003105	12226133806109	71216223942119
08000701	32	30	0	0	1111110101001	11121110101015	1312142001099	12221133003105	12226133806109	71216223942119
18082301	31	30	0	0	1111110101001	11121110101015	1312142001098	12221132703104	12226133504108	71216223634118
45299401	32	30	0	0	1111110101001	11121110101052	1312142101090	12221132702096	12226233503100	71216123627108
					1111110101001	11121110101055	13121122001093	12221143003099	12226243806103	72316133945113

Source: 31 is "File A" and 32 is "File B", that is, records of source 32 are matched onto those of source 31. A source 32 record may be used once, several times, or never. Every source 31 record is matched.

Matching Variables: a Year last worked
b Class of worker

- c Employment status record
- d Worked last year
- f Year moved into unit
- g Residence five years ago
- h State of residence
- i Wage or salary earnings
- j Current industry.

because the comparison of observations to arrive at minimal distance measures would consume excessive computer time if used for merging and matching large data sets. The utilization of hierarchical sort tags based upon the $I(x)$ technique was developed primarily because the sort-merge process is relatively economical of computer time and can be implemented for large data sets. Since large data sets do provide closer matching because of the higher density of observations, it can be expected that a simple technique applied to large data sets will yield better results than more complicated procedures which try to find good matches in small data sets where no satisfactory matches exist. This suggests strongly that in order to develop a well matched data set it may be desirable to use large samples even when this sample size is not required for the end purpose. Thus, the two million observations in the Public Use Sample may profitably be matched with the two million cases in the social security files, even if the final sample size which one is aiming at may be only 50,000 cases. Once the larger matched data set is created, it is a simple matter to select a smaller sample from it.

When one of the two data sets to be matched is small, it is still true that a high-quality match may be obtained if the second data set is large. But where both data sets are small, it is quite possible that the resulting match will not be highly significant when done by any method, and under such circumstances other multivariate techniques may be preferable.

*National Bureau of Economic Research
Yale University*

REFERENCES

- [1] 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. 317-334.
- [2] 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. 325-342.
- [3] Nelson McClung, John Moeller, and Eduardo Siguel. "Transfer Income Program Evaluation." The Urban Institute, Washington, D.C., Working Paper 950-3, September 2, 1970.
- [4] James H. Schulz. "Comparative Simulation Analysis of Social Security Systems." *Annals of Economic and Social Measurement*, Vol. 1, No. 2, April 1972, pp. 109-128.
- [5] Harold W. Guthrie, Guy H. Orcutt, Steven Caldwell, Gerald E. Peabody, and George Sadowsky. "Microanalytic Simulation of Household Behavior." *Annals of Economic and Social Measurement*, Vol. 1, No. 2, April 1972, pp. 141-170.
- [6] Horst Adler. "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey, 1970".

