

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: Data Matching and Merging: An Overview

Chapter Author: Benjamin Okner

Chapter URL: <http://www.nber.org/chapters/c10114>

Chapter pages in book: (p. 49 - 54)

DATA MATCHING AND MERGING: AN OVERVIEW

BY BENJAMIN A. OKNER*

This note surveys two papers which appear elsewhere in this issue of the journal. They were presented at a workshop on matching and merging of the bases, and the subsequent panel discussion is summarized here. It is clear that computer and analytical techniques are yielding promising approaches to the creation of "synthetic" data sets.

The data matching and merging workshop consisted of two sessions: (1) a joint meeting with the Conference on Price and Consumer Expenditure Data; and (2) a separate one-day meeting with informal presentations by a panel of persons working on match-merge projects, followed by general discussion among the panelists and other workshop participants.¹ Two of the papers (by Alter and Ruggles) given at the joint session and comments on them are presented elsewhere in this issue of the *Annals* (see pp. 353-437). Here I attempt to summarize the material presented by the panel members and highlight points raised by them and others during the day's discussion. This is not intended to be a transcript; rather, as indicated in the title, I try to provide a broad-brush summary.

Data matching projects fall into two broad categories: (1) exact matching or linking of microdata for identical units from two or more sources; and (2) synthetic or stochastic linking or data synthesizing. During the workshop we heard reports on both kinds of projects done in the past (or underway) plus plans for new work that is just getting started. The social security link and match projects described by Bridges and the linking of records for identical units over time in the Wisconsin Asset and Income Study (WAIS) presented by David are examples of exact matching. Reports on synthetic linking included Radner's (in collaboration with Edward Budd) on the 1964 file constructed at the Commerce Department's Office of Business Economics (now the Bureau of Economic Analysis, BEA);² Okner's on constructing the 1966 MERGE File at Brookings³; and Bristol and Turner's on work done at the Office of Tax Analysis. Although their work was not

* Senior Fellow, Economic Studies Division, The Brookings Institution. The views presented are those of the author and not necessarily those of the officers, trustees, or other staff members of the Brookings Institution.

¹ The workshop was held on May 4-5, 1973 at Williamsburg, Virginia. The panel members were Benjamin Bridges, Social Security Administration, U.S. Department of Health, Education, and Welfare; Ralph Bristol, Office of Tax Analysis, U.S. Treasury Department; Martin David, The University of Wisconsin; Benjamin A. Okner, The Brookings Institution; Jon K. Peck, Yale University; Daniel Radner, Bureau of Economic Analysis, U.S. Department of Commerce; and Scott Turner, Office of Tax Analysis, U.S. Treasury Department.

² A complete description of the Radner-Budd work is given in Edward C. Budd, Daniel B. Radner, and John C. Hinrichs, *Size Distribution of Family Personal Income: Methodology and Estimates for 1964*, U.S. Bureau of Economic Analysis, Staff Paper No. 21 (1973). A less detailed description appears in Edward C. Budd, "The Creation of Microdata File For Estimating the Size Distribution of Income," *The Review of Income and Wealth*, Series 17, No. 4 (December 1971), pp. 317-333.

³ See Benjamin A. Okner, "Constructing a New Data Base from Existing Microdata Sets: the 1966 MERGE File," *Annals of Economic and Social Measurement*, Vol. 1 (July 1972), pp. 325-342.

presented as part of the panel, the current Ruggles' NBER project and the work of Horst Alter at Statistics Canada are also examples of synthetic linking.

Regardless of whether exact or synthetic linking is used, there are several problems common to all work in this area. These include data comparability, missing data, specific techniques for linking and data manipulation, and the definition and evaluation of "goodness of a match." To some degree, these topics emerged with respect to all the projects discussed at the workshop. However, in this report, I emphasize particular problems in terms of examples relating to specific projects.

An important reason for linking or merging is to "enrich" or fill gaps in an existing data set. When viewed in this way, data linking or matching becomes a "missing information problem," and involves many of the same techniques (and problems) encountered in filling "missing" or "not ascertained" information in sample surveys.

The Treasury work described by Bristol and Turner provides a good example of filling missing information using synthetic linking. The problem they faced on the Internal Revenue Service (IRS) tax file was the lack of itemized deduction data for personal income tax returns with standard deductions. For the match, they assumed that most taxpayers who itemize are homeowners and that deductible expenditures, except those for mortgage interest and real property taxes, would be similar for homeowners and renters.⁴ They then used tax returns of homeowners who itemized only because of homeowner-related expenses—i.e., returns on which the deductions for other items were sufficiently small that the taxpayer otherwise would have used the standard deduction—as the basis for imputing nonhomeowner deductions to returns on which the standard deduction had been used. In other words they matched a subsample of itemized deduction returns with "similar" nonitemized ones to fill in the required information for nonitemizers. The method used to classify returns as being similar (on the basis of income and demographic characteristics) for matching was akin to that being used in the current Ruggles' work (see p. 353).

In the Treasury project, the problem of data comparability did not exist since income and other variables were defined in exactly the same way for both the itemizers' and nonitemizers' tax returns. However, in most linking projects involving data from two (or more) different files, such noncomparability of linking variables is a serious problem. Different data collection sources use different income concepts and population definitions (e.g., tax returns vs. Census families vs. special security earnings records). In addition, there are often different time-period problems. For example, Current Population Survey (CPS) data for the preceding income year are collected in March of the following year; however, some people will have died between the beginning of the preceding year and the survey date, or left the country, or become institutionalized. There is no way to "solve" the problem of data non-comparability: researcher must be aware of such differences and strive to make the data sets to be linked as consistent as possible.

⁴ Bristol noted that this assumption is subject to caveats. For example, homeowners may spend more on large appliances and therefore have higher sales tax and installment interest deductions than renters. Also, since contributions will "cost them less" because they itemize, homeowners' behavior in this area may differ from that of renters.

Another question regarding all the synthetic linking work was "how good" the resultant match is. In Alter's paper (p. 373), there are presented some statistics on this, but the question of useful criteria remained unresolved at the workshop. It is a difficult question because the notion of "good match" is undefined. Does good refer to each linked record or to the means, variances, and covariances in the distributions of all variables in the newly-created data set? Or should "goodness" be measured only for the "important" or "crucial" variables? What is the standard against which to measure "good?" What are the most appropriate statistical measures to present to users of the new file? Peck emphasized that the goodness of the result cannot be determined (by whatever method or standard) without prior knowledge of the ultimate research objective for which the new file is created. The new distributions of certain variables may be "poor" but that may be relatively unimportant for particular research purposes. On the other hand, there appeared to be agreement on the importance of giving users some indication of which variables are strong or weak so that they do not uncritically use a created data set for purposes where it is clearly inappropriate.

Obviously, if it were possible to do so, the best evaluation of a synthetic data set would be to compare it with similar information derived from a sample survey which includes all the variables. However, it is unlikely that this could ever be done: if all the information already existed in one data set, there would be no need to create the synthetic file. Alternatively, exact matches for earlier years can be useful for evaluating the quality of new synthetic matches.

In concept, exact matching such as was done in the Social Security Administration link projects for calendar years 1963 and 1964 would be expected to involve fewer problems than synthetic matching. Both of these projects involved matching data for identical units from the March CPS (including the February work-experience supplement and information from the Census control card for the household), Internal Revenue tax return data, and social security summary work history and benefit data. Link 1 involved about 7,000 individuals and Link 2 about 12,000; in both projects, social security numbers were used when available to aid in linking records from the three separate data sources.

In practice, exact matching involves a number of problems—perhaps most troublesome was being unable to find IRS tax returns for CPS units where they should have existed⁵. For 1963 about 80 percent of all tax returns were located and matched with CPS records. The corresponding match rate with social security earnings records was somewhat higher—about 90 percent were found and matched.⁶ The worst match rate with earnings records occurred among those with low earnings, i.e., the very young and the very old. Moreover, there was the difficult task of determining whether information that "should" agree differed in the various files because of different definitions or because incorrect or invalid data existed in one or more of the files.⁷

⁵ There was no search for returns of persons under age 14. However, for others a manual search in local Internal Revenue offices was undertaken.

⁶ The validity of social security numbers from the CPS control cards were first checked against social security master records. Sometimes these were missing and in other instances, the recorded numbers were invalid. When in error or missing, the numbers had to be searched for manually.

⁷ A series of reports on Link 1 will be published in 1973 and 1974; a monograph on Link 2 is expected in 1974.

The history of the link projects should teach us many things. First, while direct matching or linking is intuitively simple and appealing, it is not easy. Records that "should exist" often do not. Even when they exist, different records for presumably identical units may contain different data (even after allowing for conceptual differences). Arbitrary decisions had to be made when the CPS record, the tax record, and social security earnings record contained different (inconsistent) earnings amounts. Assuming that in fact these records were all for the same unit, some decision had to be made as to which one was correct and should be accepted for the final file. It seems clear that such problems inevitably require a large number of ad hoc human decisions in the case of exact matching. Given the quality of data collected by different agencies at different times, it is probably unwise to assume that one can merely assemble two or three sets of data to be sorted on the matching criteria and linked.

The construction of the WAIS longitudinal panel entailed many of the same exact linking problems noted in the discussion of the social security link projects. In the WAIS, the problems were multiplied because the file is very large and because of changes in marital status and household composition that occurred during the 18-year period for which data exist in the WAIS archive (1947-1964). The WAIS master file contains income tax, earnings, and social security benefit information for 21,000 individuals for one or more periods (which comprise a total of 200,000 records) as well as property tax data. This involved linking information from Wisconsin income tax records, federal social security earnings and benefit files, and data from local tax assessors. An interesting additional problem that has plagued the WAIS—an undoubtedly will be encountered in other studies in the future—is the fact that the researchers have had to deal with six different computers at the University of Wisconsin since the beginning of the project!

Again, the lesson to be learned is that exact matching is not easy. It is very costly and ironing out data inconsistencies and other problems tend to make such projects very time consuming. In addition, in any match that involves government data, there is need for a high degree of cooperation between the researcher and the agencies possessing the data (and information needed for clarifying them). Convincing agencies to cooperate is not always easy; but even after cooperation is secured, pressing day-to-day tasks may well prevent an agency from responding as quickly to data requests or answering questions about inconsistent (or invalid) information as the researcher would like.

Although work is still underway on many of the earlier projects, three new ones were also discussed at the workshop. These were plans for two more exact matches by the Social Security Administration and the Census Bureau (for 1969 and 1972) and new synthetic matches at the BEA and Brookings.

The Social Security-Census exact match for 1969 involves 10,000 households from the 1970 Decennial Census and is currently underway; the 1972 work is still in the planning stage. In the latter project, there is planned a direct linking of all 50,000 households in the 1972 CPS with a limited selection of tax return data from Internal Revenue and with social security earnings and benefit information. This is an extremely ambitious project. Because of the tentative nature of the plans, there was only brief discussion of the project. For similar reasons, the new syn-

thetic linking of IRS and CPS data contemplated at the BEA (which would be for 1970) was not discussed during the workshop.

The new Brookings undertaking will involve the synthetic linking of data from the 1971 CPS (for income year 1970) and the 1970 individual income tax file. There are two major objectives sought in the new project. The first is the construction of a more up-to-date base to be used for tax and income distribution analyses. The second is to develop a set of generally-applicable synthesizing techniques that can be utilized to construct similar files on a recurring basis over time. It was for the second reason that we decided to link CPS data, rather than Decennial Census information, with income tax records. Since the CPS and Internal Revenue data are both available annually, we hope eventually to construct a series of linked files with which it will be possible to examine tax and income distribution changes over time.⁸

The matching techniques to be employed in constructing the new Brookings file were described by Jon Peck who is working on the methodology. The procedure to be used involves evaluating distance (or criterion, or loss) functions involving variables available in the CPS and Tax File data to determine the quality of each proposed match. The algorithm will attempt to minimize the value of a criterion function that includes weighted values of existing variables available in the two files and derived ones to be added to the files. An example of a criterion function is the weighted sum of squares and cross products of the differences of variables in the two samples for pairs of observations. The estimation from extraneous sources and addition of such derived variables to the basic samples for matching is an important new feature of the procedure. It assumes that regression (or other techniques) can be utilized to estimate variables initially absent from one data set (but available in the other) in order to improve the quality of matches achieved. Such relationships might be developed from any number of outside (extraneous) data sources such as the University of Michigan longitudinal study, the Survey of Financial Characteristics of Consumers, or the Decennial Census of Population. Because these relationships are only estimates, they will enter the criterion function with lower weights than would be used if the data came from one of the original sources.

There were numerous technical points and questions raised with respect to all the projects; most of this is omitted because of space (and memory) limitations. However, a comment by Mollie Orshansky regarding the cost effectiveness of data matching and merging is a good note on which to conclude this report. In essence what she asked was: given the large amount of human and capital resources that now seem to be devoted to such projects, would it not be cheaper—and possibly more accurate—to get information directly from respondents through household sample surveys in which all needed data were simultaneously collected? As was the case with many of the important points raised at the workshop, it was impossible to proffer a good answer to the question at Williamsburg. However, the topic should be addressed systematically in a future meeting.

⁸ While it is theoretically possible to construct such files on an annual basis, detailed information on itemized deductions is provided by the Internal Revenue Service only every other year. If the work for 1970 is successful, the construction of future files will probably be done on a quadrennial basis. As yet, however, there are no definite plans for such future work.

Summing up, there were several useful results emanating from the workshop. First, it brought together a number of people engaged in similar work so that it was possible to assess the current state of the art. Second, based on the reports and discussion it seems clear that techniques for matching and merging microdata, while barely out of their infancy, are progressing rapidly along somewhat different lines. We simply do not yet know which of the possible techniques will be most fruitful and efficient. And third, there are obviously still many important, but unanswered questions in this area: these range from those having to do with rather narrow technical points concerning a given technique to whether data matching and merging is a worthwhile endeavor.

It seems likely that matching and merging projects will continue since the outlook for directly obtaining such data by another means is dim. Still, the cost effectiveness question does need study. Many of the more technical questions will be answered over time as work progresses on various projects. But again, certain of them—such as those concerning the “goodness” of a match—can and should be explored on a theoretical basis. It seems obvious that it would be useful to tackle such questions in a more comprehensive and systematic manner than was possible at the workshop and that it would also be desirable to get feedback in the future on the projects that were just getting underway. Another workshop or conference on data matching and merging to be held sometime in the future would therefore seem to be a fruitful follow-up to this meeting.

The Brookings Institution