ALTERNATIVE TYPES OF RECORD MATCHING: COSTS AND BENEFITS

Daniel B. Radner, Social Security Administration
Hans J. Muller, Bureau of the Census

I. Introduction

This paper reports on work being done by a Subcommittee on Matching Techniques associated with the Federal Committee on Statistical Methodology.1/ Because the topic of record matching2/ is so broad, we can only give an overview. At a later date the Subcommittee will issue a final report which will expand upon the discussion presented here.

The matching of data files is a very useful technique for many purposes. In this paper, we are interested only in matching for research and statistical purposes. Matching for other purposes, e.g., administrative, will not be considered. In the matching considered here, identification of individuals, if needed at all, is only necessary to make the match. After matching, that identification can be removed.

When we are considering only the accuracy of the matched data, the preferred method of matching is ordinarily what is commonly called "exact matching,"3/ i.e., combining data for the same individuals from different data sources, usually by means of personal identifiers (e.g., name, address, Social Security Number).4/ The use of the term "exact" match is not meant to suggest that such matches are made without error; problems associated with exact matching are mentioned later.

In many cases, for technical or legal reasons, or both, exact matches cannot be carried out. For example, both files might be samples which have few persons in common; or, the information might not be sufficient to identify the individuals in both files. Legal restrictions on exact matching, which have existed for some time, have been increasing in recent years (e.g., the Privacy Act of 1974 and the Tax Reform Act of 1976). These limitations on the use of exact matching have led to interest in alternative methods of matching.

This paper focuses on one such alternative approach, what is commonly called "statistical matching."5/ In a statistical match, the information brought together from the different files (ordinarily) is not for the same person, but is for similar persons. The match is made on the basis of similar characteristics,6/ rather than personal identifying information, as in the usual exact match.

The distinction between exact and statistical matches is not always clear-cut. In this paper, matches in which the aim is to link data for the same person from two files are defined to be exact matches. As defined here, exact matches can be carried out using similar characteristics, but ordinarily personal identifiers are used. Matches in which the aim (for all or most records) is to link data of similar persons, rather than the same person, are defined to be statistical matches. In general, statistical matches have been carried out in situations in which an exact match was not possible.

II. Overview of Matching Applications

The Subcommittee has collected many examples of matching of data files, most by government agencies and most from the U.S. This overview is based upon the examples we have collected, only a few of which can be mentioned in this paper. We will separate the applications of matching, somewhat arbitrarily, into two broad types: (1) adding more variables or additional reports on the same variables; and (2) comparing the presence of units in two files. Within type (1), several different kinds of applications can be identified. One application is the addition of more variables to make possible analyses which otherwise could not be done or to enrich analyses with more variables. Both exact and statistical matching have been used in this application. A cross-section example of one such exact match is the addition of Social Security Administration (SSA) age, race, and sex data to federal individual income tax returns in order to provide better income and tax data by those characteristics. In another cross-section example, a statistical match was carried out between observations from a household survey and a sample of federal individual income tax returns in order to add more detailed and more accurate income information to the household survey data [8]. A longitudinal example of exact matching is the linkage of hospital admission and separation records into cumulative health histories [27].

Another kind of application within type (1) is the evaluation of data, in which initial variables are compared with added variables, or with additional reports on the same variables--from other existing sources or from special evaluation surveys. Evaluation of the accuracy of data was carried out using the 1973 Current Population Survey--Internal Revenue Service--SSA Exact Match Study. In that work, the income data from the different data sources were compared and response and reporting errors were analyzed (e.g. [3]). Definitional differences were examined in Sweden using exact matching. Two different definitions of unemployment--from a household survey and from the labor market board--were compared by matching survey responses and labor market board records [10].

In type (2), two different kinds of applications can be identified: evaluation of coverage and construction of more comprehensive lists. The Bureau of the Census has conducted numerous coverage evaluation studies in connection with the Decennial Censuses. For example, in connection with the 1960 Population Census, samples from 1950 Census records, registered births, and other sources were matched with 1960 Census records, and coverage was assessed [19]. In such matches, the emphasis is upon the presence of units in the files, rather than upon the relationships between data in the two files. In an example of list construction, the Statistical Reporting Service of the U.S. Department of Agriculture used exact matching in the construction of a master list sampling frame of farms in each state. This master list was constructed from several different lists, and exact matching was used to detect duplication between

(and within) the different lists [9]. Statistical matching has not been used in type (2) applications, and is not appropriate for such applications.

In most of the applications mentioned above, one possible effect of matching was a reduction of response "burden"--i.e., to get the same information without matching, a considerable amount of direct data collection would have been necessary. Also, in some of those applications, cost reduction was a beneficial effect--i.e., matching was less expensive than direct collection of the same combination of data would have been. The Office of Management and Budget recently has suggested the use of statistical matching to reduce response burden and cost by means of what are called "nested surveys." In such surveys, different samples from the same population are asked different sets of questions, with a core of questions in common. The data from these different samples can then be matched statistically to obtain relationships between the items not in the common core of questions [17].

III. Exact Matching

For exact matching it is necessary that all or most of the individuals in one file ("base file") be included in the other file ("reference file"). However, rarely do both source files include enough identifiers to allow unique identification of all individuals; the identifiers that are used are usually missing from some records and reported inaccurately or with variations in some other records; each file may--correctly or incorrectly--include some persons absent from the other file. As a consequence, an apparently matched pair of records with the same or very similar identifiers usually links the records of the same person in both files ("true match"), but it may link the records of two different persons ("false match" or "mismatch"). On the other hand, if a record in one file appears to have no match in the other file, that may be because there really is no record for that unit in the second file ("true nonmatch"), or there really may be records for the same person in both files but one or both records may include errors or spelling variations that prevent them from being recognized as a match ("false nonmatch").

In many cases the true match status could only be ascertained at great expense or not at all; generally, a matched file must be assumed to contain some errors. The relative importance of false matches and false nonmatches varies in accordance with the purpose of each project. Techniques have been developed for designing the matching process for a particular study in such a way that the type of error most harmful in the context of that study can be minimized and the remaining error can be estimated.

An exact matching procedure generally includes the following steps (although they may not always be clearly distinguishable).7/8/

1. Data preparation: Transfer to machine-readable form, resequencing, reformatting, elimination of out-of-scope records, and other editing steps. If one or both of the files do not already exist, this step includes data collection.

2. Selection of matching variables and tolerances: Ideally, the most accurately reported and the most discriminating variables are preferred, but these are often conflicting requirements. Confidentiality restrictions may interfere by making identifiers such as names or Social Security Numbers unavailable. Because of the inaccuracies in the source files, strict agreement on such variables as age or on name spelling cannot always be required. More or less elaborate techniques have been used for selecting, for a particular matching project, the combination of matching variables and tolerances that will keep the probability of matching errors as low as feasible [14, 33].

3. File blocking: In order to avoid having to compare each base file record with all reference file records, relatively small portions of both files are selected for intensive searching, (e.g., all records with addresses in the same city block, or all records with a certain group of last names, including variant spellings of the same name). Ideally, these "comparison classes" or "blocks" should be formed on the basis of characteristics that will virtually never disagree in the case of true matches, and will almost always disagree in the case of true nonmatches [32, 33].

4. Weights and thresholds: Since a block ("comparison class") may include several possible reference file matches with the same base record ("comparison pairs"), some rules are needed for deciding which pair--if any--is accepted as a match. Each pair contains a particular configuration of agreements and disagreements on the matching variables; explicitly or implicitly, the decision is based on the probability of that configuration occurring if the pair were truly matched, or truly not matched (paired at random).

The rules for making that decision need to take into account the fact that different variables contribute different amounts of relevant information. This is done by assigning different weights to various degrees of agreement or disagreement on each variable, and deriving a total weight for each comparison pair. For carrying this out in practice, a great variety of procedures have been used, ranging from the intuitive judgment of a researcher to mathematical models of the matching process that require a computer for their application. The weights can be based on external evidence or derived from special pilot studies or from thorough investigation of samples, or their derivation can be incorporated in the computer program that uses them.

Finally, once it has been determined how likely or unlikely it is that a particular comparison pair constitutes a true match and which of several possible pairs is the most likely match, it must be decided whether it is likely enough to be accepted as a match, taking into consideration the purpose of the project.

This final decision, explicitly or implicitly, takes the form of setting a threshold that divides the range of total weight scores into "matched" and "not matched". This is not an isolated decision; it is affected by the previous decisions on matching variables, tolerances, and weights. All of these decisions must be coordinated with the aim of achieving the

results that are optimal in terms of the purpose of the particular matching project [9, 11, 19, 27, 29].

5. Except for very small studies, it is practically impossible to clear up all doubts and avoid all matching errors. In well planned matching studies, the probable impact of such errors may be estimated and, if necessary, appropriate adjustments may be made in the results [16, 23, 25].

IV. Statistical Matching

To the best of our knowledge, the vast majority of the statistical matches and of the developmental work carried out has been in the field of economics.9/ The most common application has been to combine data from a household survey with data from income tax returns where there was little overlap between the two files. Early statistical matches were performed at the Bureau of Economic Analysis of the U.S. Department of Commerce in connection with estimates of the size distribution of family personal income [5, 6, 7] and the Brookings Institution in connection with analysis of the tax system [18]. More recent matching work has been done at Statistics Canada [1], Yale University [22], the Office of Tax Analysis of the U.S. Treasury Department [30], Brookings [4], and the Office of Research and Statistics of the Social Security Administration [21]. 10/

Because statistical matching is not a well-known technique, the theoretical steps involved in the most common case will be summarized.11/ We begin with two microdata sets of observations on variables for units in a universe, U; these sets, A and B, are the sets we want to match statistically. A and B are assumed to be probability samples from U. It is also assumed that very few units from U are in both A and B. For example, A might be the persons interviewed in a household sample survey, and B might be an independent sample of income tax returns. Some variables from U may be contained in both A and B, while at least some are contained in only one set.

It is assumed that at least some of the variables in A and B contain errors, while in U they do not. Because of different error components, a variable from U which appears in both A and B can have different values in the two sets for the same unit in U. For example, even if wage income were defined identically in the household survey and the tax return, the survey response might differ from the amount shown on the tax return.

We now define C, a hypothetical data set which represents the results of an exact match (carried out without error) between A and B, if the units in A were also in B. The set C is hypothetical because that exact match cannot be carried out, since very few of the units in A are also in B. By assumption, C contains all variables from A and all variables from B, including their error terms. Because a statistical match is an approximation of an exact match, C is the data set which we try to approximate when we perform a statistical match. In our example, for each unit in A, C contains the survey response given by that A unit and the data from the tax return filed by that A unit. As noted above, that tax return probably does not appear in B.

When we actually want to make a match, we do not know C. Therefore we make an estimate of C, called L, using whatever information is available. Estimated values (for the B information) might be obtained by assumption. For example, for a given A unit, it might be assumed that the value for a given B variable should be equal to the value for a given A variable. We could say that wage income in B should be identical to wage income in A. This would be valid if wage income were defined identically and had identical error patterns (e.g., response and reporting error) in A and B. Ordinarily, this is not the case. Estimated values can also be obtained by other means, for example, by regression techniques or by using cross-tabulations from an exact match between sets similar to A and B. In our example, for each unit in A, L contains that unit's survey response data and estimates of (some or all of) the variables in the tax return filed by that A unit.12/

We now introduce M, the result of statistically matching A and B (in some unspecified way). Using our example, for each unit in A, M contains that unit's survey response data and the tax return data from the B unit assigned to that A unit in the statistical match. It is not necessary that every B unit be used in the match solution; some B units can be used more than once in the solution.13/ It follows from the definition of a statistical match that the variables assigned to a given A unit in the match are all from one B unit.

In making a statistical match we choose among alternative solutions; each alternative solution is characterized by the particular set of B units assigned and the particular A unit(s) to which each is assigned. We choose the solution in which M approximates L as closely as possible, in terms of the variables and relationships of greatest importance in the results of the match. This approximation can be viewed in terms of a distance function which measures the distance of M from L. The distance is defined in a subjective way according to the purpose of the match. The statistical match solution which minimizes this distance is the optimal match result.14/

In practice, many different statistical matching methods have been used. In most cases the variables in both files were separated into "matching variables" (which were similar in the two files and were used to carry out the match) and "nonmatching variables" (which were the "added" variables). In most matches, both files were separated into comparable subsets of units. Within each subset, rules were specified for the choice of a record from the second file to be assigned to each record from the first (or "base") file. The selection of the record within the subset usually was based upon a distance function by which a distance was computed between a given base set record and each potential match in the other set. The distance was based upon differences between matching variables in the two files.15/ In some cases, these differences were weighted according to the relative importance of the variables and the comparability of the pairs of variables for which values were

compared. The potential match with the smallest distance ordinarily was chosen as the match; a maximum distance has been used to define a subset of potential matches from which a random choice was made. In some cases, subsets were defined so narrowly that most subsets contained only one record. In other cases, the choice within subsets was random.

Very little work on the reliability of statistical matching has been done.16/ Given this lack, we will merely attempt to identify several types of errors which can arise in statistical matching, assuming that the matching is done in an optimal way. "Error" is defined as the difference between the "true" joint distribution of A variables and B variables that would be obtained from an exact match (carried out without error) between A and B, if such a match were possible, and the estimated joint distribution of those variables obtained from a statistical match. The following three sources of error can be identified. First, because of lack of comparability between matching variables in the two sets (i.e., the variables are not defined identically and/or have different error patterns), we cannot know with certainty the values of the matching variables that we are searching for in B. Second, even if we knew those values with certainty, often we could not find a B record with such values because B is a sample which ordinarily does not contain the true match. Third, even if we could find a B record with such values (assuming it is not the true match), the values for nonmatching variables in B probably would differ from the true values because those nonmatching variables are not "completely explained" by the matching variables.

V. Summary of Costs and Benefits

In this section the costs and benefits (or advantages and disadvantages) of exact and statistical matching are summarized. Three aspects of this topic will be touched upon: (1) the reliability of the data resulting from the match; (2) the confidentiality problems involved; and (3) the resource cost of the match. Of course, it is very difficult to generalize, since matches vary widely in these aspects. But we feel that some general statements contrasting exact and statistical matching can be made.

Reliability--Error at a single record level will be discussed first; then error on an aggregate level will be mentioned. Initially it will be assumed that the same persons are in the two sets to be matched; therefore an exact match of all units in the base set is possible. Under this assumption we can compare sources of error for an exact match using personal identifying information and an exact match using characteristics (which is a statistical matching type of technique). In this case, error in the data used to match is the main source of error in the match result. In most cases, the personal identifying information has been more reliable than characteristics in making the match; thus we could generalize and say that, in this case, exact matching is more accurate than a statistical matching type of technique. It should be noted that we are considering not only whether the match for any given record is correct, but how far the values are from the true match values if a mismatch is made.

We will now assume that the persons in the two files are all different, and examine additional sources of error in statistical matching. In this case, statistical matching faces what might be called the "proxy" problem; that is, how good a proxy for the true match can be found. Even if we assume that the characteristics used to match on are defined identically and have identical error patterns, the proxy found is not likely to have values which are identical to the true match values. Even if it did have such values for the matching variables, the values for nonmatching variables probably would not be identical to the true match values.

On an aggregate level, it is difficult to identify generally applicable measures of accuracy. Unless the statistical match is constrained to use all non-base file records, the means of variables in the non-base set can be biased (e.g., because amounts are matched too low more often than too high, even though the best match for any base file record is chosen). Or, the variance of the values in the records chosen from the non-base set can be too low (e.g., if records with extreme values are not chosen often enough in the match). In exact matching, biases can arise from false matches and from false nonmatches. In general, the reliability of the results can be estimated in exact matching more easily than in statistical matching. With both methods, it may be necessary to adjust the matched file to a set of independently established control totals.

Confidentiality--The confidentiality problems clearly are much greater for exact matches than for statistical matches for two reasons. First, if personal identifiers are used (as they usually are in exact matching), persons must be identified, at least at some stage of the matching. Second, in an exact match (assuming that the true match is found), the matched file contains more information regarding the person than either of the original files matched. Thus, there is a greater risk of a record in the matched file being identifiable even after the removal of the personal identifiers. Protective measures against disclosure can be taken in both cases, but they usually entail greater expense and/or some loss of information. These problems ordinarily do not exist in the case of statistical matching.

Resource Costs--It is very difficult to generalize regarding cost differences between exact and statistical matches. Costs can vary for many reasons, depending upon, for example, the amount of computer time used, the amount of clerical time used, the lengths of the files, the complexity of the statistical matching program, and the amount of preliminary data analysis and reformatting that need to be carried out. Because it is so difficult to make meaningful comparisons, no generally valid conclusions regarding cost comparisons can be made here; the costs of possible alternative procedures must be evaluated specifically for each project.

In discussing the comparisons in this section, we have assumed situations in which either exact or statistical matching might be

useful. However, there are many situations in which statistical matching would not be useful. In addition to the type (2) applications (comparison of presence of units in two files) mentioned earlier, statistical matching also can be inappropriate for many type (1) applications. For example, if we want to compare the earnings of persons who have had a given training program with those who have not, we can use an exact match between a list of trainees and earnings records. However, a statistical match between those two data sets would not be useful unless the earnings observations could be separated into those who had been trained and those who had not.

VI. Summary and Conclusions

Exact matching is extremely useful in a variety of research and statistical applications. In many of those applications, statistical matching is not an acceptable alternative because the resulting data would not be useful. However, statistical matching has been useful in a few limited applications. When statistical matching can be used, the data obtained from the statistical match in general contain far more error than those from an exact match. Statistical matches can be as expensive as, or more expensive than, exact matches, but statistical matches do have the important advantage that they are carried out without the use of personal identifying information and that they ordinarily do not bring together information for the same person. Thus, statistical matching does not pose the same confidentiality difficulties that exact matching does.

The data which result from matched files should be used with caution because matching, exact or statistical, is not error free. This is particularly true for statistical matching. A substantial literature on exact matching and its nature and reliability exists. However, little has been written regarding the nature and reliability of statistical matching. A great deal of research into the reliability of statistical matching is needed; of particular importance is an examination of the effects of lack of comparability between matching variables. One possible approach which has been suggested is to compare the results of exact and statistical matching of the same files.

FOOTNOTES

1/ The authors are greatly indebted to the members of the Subcommittee, particularly the ex officio members, Maria Gonzalez, Thomas Jabine, and Tore Dalenius, for their many helpful comments.

2/ Other terms have also been used, e.g., "record linkage."

3/ Other terms have also been used, e.g., "actual" and "object" matching.

4/ Although most of the discussion in this paper is in terms of matching information for persons, the discussion also applies to other units for which confidentiality can be an issue (e.g., business firms, hospitals).

5/ Other terms have also been used, e.g., "attribute," "data," "stochastic," and "synthetic" matching.

6/ It is possible to match on characteristics which are not similar; the requirement is that for one or more variables in one set, corresponding values of one or more variables in the other can be identified.

7/ Some of these steps can be executed efficiently by computer. For some applications, a prepared program is available that works with user-specified variables, weights, and tolerances [31].

8/ In addition to the references cited for various steps, [15] includes a very comprehensive treatment of all aspects of exact matching. Brief overviews of exact matching procedures and problems are given in [12, 28].

9/ Related work on matching (or "pairing") samples to reduce extraneous variation has been done outside economics (e.g., [2]). Also, the imputation of values to nonrespondents in household surveys is a closely related technique.

10/ For several comments and replies on statistical matching and an overview article on matching, see the July 1972 and April 1974 issues of the Annals of Economic and Social Measurement. [13] and [34] are somewhat more theoretical papers on statistical matching.

11/ This formulation was suggested in [20].

12/ L can also include constructed variables for both A and B.

13/ Some matching methods do require that every B unit must be used in the match solution, and used only once [20, 30]. In some matching methods, more than one B unit can be assigned to an A unit.

14/ This is not meant to suggest that any given match should be carried out using a distance function, or that a distance function method is the best method in theory.

15/ The matching variables ordinarily were chosen partly because they were (thought to be) significantly correlated with important variables which could not be used to make the match. In exact matches, such a correlation has not been an important factor in the choice of information used to make the match.

16/ See [26] and [34] for examples of work that has been done.

REFERENCES

[1] Alter, Horst E. (1974). "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 (April) 2: 373-394.

[2] Althauser, Robert P., and Rubin, Donald (1969). "The Computerized Construction of a Matched Sample." American Journal of Sociology (September) 76: 325-46.

[3] Alvey, Wendy and Cobleigh, Cynthia (1975). "Exploration of Differences Between Linked Current Population Survey and Social Security Earnings Data for 1972," 1975 Proceedings of the ASA, Social Statistics Section, 121-28.

[4] Armington, Catherine, and Odle, Marjorie (1975). "Creating the MERGE-70 File: Data Folding and Linking." Research on Microdata Files Based on Field Surveys and Tax Returns, Working Paper I, The Brookings Institution (June). Mimeographed.

[5] Budd, Edward C. (1971). "The Creation of a Microdata File for Estimating the Size Distribution of Income." Review of Income and Wealth (December) 17: 317-33.

[6] Budd, Edward C., and Radner, Daniel B. (1969). "The OBE Size Distribution Series: Methods and Tentative Results for 1964." American Economic Review (May) LIX: 435-49.

[7] Budd, Edward C., and Radner, Daniel B. (1975). "The Bureau of Economic Analysis and Current Population Survey Size Distributions: Some Comparisons for 1964," in James D. Smith, ed., The Personal Distribution of Income and Wealth, Studies in Income and Wealth, 39: 449-558.

[8] Budd, Edward C.; Radner, Daniel B.; and Hinrichs, John C. (1973). "Size Distribution of Family Personal Income: Methodology and Estimates for 1964." Bureau of Economic Analysis Staff Paper No. 21. U.S. Department of Commerce (June).

[9] Coulter, Richard W. (1977). "An Application of a Theory for Record Linkage." Paper presented at the April 6 meeting of the Washington Statistical Society, Washington, D.C.

[10] Dalenius, Tore (1974). "Tva matare av arbetslosheten. En studie 1 svensk arbetsmarknadsstatistik." Report No. 81 of the research project "Errors in Surveys," Department of Statistics, University of Stockholm.

[11] Fellegi, Ivan P., and Sunter, Alan B. (1969). "A Theory for Record Linkage." JASA 64: 1183-1210.

[12] Hansen, Morris H. (1971). "The Role and Feasibility of a National Data Bank, based on Matched Records and Interviews." Report of the President's Commission on Federal Statistics 2: 1-63. Washington.

[13] Kadane, Joseph B. (1975). "Statistical Problems of Merged Data Files," OTA Paper 6, Office of Tax Analysis, U.S. Treasury Department (December 12).

[14] Madigan, Francis C., and Wells, H.B. (1976). "Report on Matching Procedures of a Dual Record System in the Southern Philippines." Demography (August) 13: 381-95.

[15] Marks, Eli S.; Seltzer, William; and Krotki, Karol J. (1974). Population Growth Estimation - A Handbook of Vital Statistics Measurement. The Population Council, New York.

[16] Neter, John; Maynes, E.S.; and Ramanathan, R. (1965). "The Effect of Mismatching on the Measurement of Response Errors." JASA 60: 1005-1027.

[17] Office of Management and Budget (1977). "Standards for Statistical Methodology." Statistical Reporter, No. 77-9 (June), pp. 423-24.

[18] Okner, Benjamin A. (1972). "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." Annals of Economic and Social Measurement (July) 1: 325-42.

[19] Perkins, Walter M., and Jones, Charles D. (1965). "Matching for Census Coverage Checks." 1965 Proceedings of the ASA, Social Statistics Section, 122-41.

[20] Radner, Daniel B. (1974). "The Statistical Matching of Microdata Sets: The Bureau of Economic Analysis 1964 Current Population Survey--Tax Model Match." Ph.D. dissertation, Department of Economics, Yale University. Microfilm.

[21] Radner, Daniel B. (1977). "Federal Income Taxes, Social Security Taxes, and the U.S. Distribution of Income, 1972." Paper prepared for the 15th General Conference of the International Association for Research in Income and Wealth, University of York, England, August 19-25.

[22] Ruggles, Nancy, and Ruggles, Richard (1974). "A Strategy for Merging and Matching Microdata Sets." Annals of Economic and Social Measurement (April) 2: 353-72.

[23] Scheuren, Fritz and Oh, H. Lock (1975). "Fiddling Around with Nonmatches and Mismatches." 1975 Proceedings of the ASA, Social Statistics Section, 627-33.

[24] "Selected Bibliography on the Matching of Person Records from Different Sources." 1974 Proceedings of the ASA, Social Statistics Section, 151-54.

[25] Seltzer, William and Adlakha, Arjun (1969). "On the Effect of Errors in the Application of the Chandrasekar-Deming Technique." (Reprinted as Laboratories for Population Statistics Reprint Series No. 14. Chapel Hill, 1974.)

[26] Sims, Christopher A. (1972). "Comments." Annals of Economic and Social Measurement (July) 1: 343-46.

[27] Smith, Martha E., and Newcombe, H.B. (1975). "Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories." Methods of Information in Medicine (July) 14: 118-25.

[28] Steinberg, Joseph, and Pritzker, Leon (1967). "Some Experiences with and Reflections on Data Linkage in the United States." Bulletin of the I.S.I. 42:786-805.

[29] Tepping, Benjamin J. (1968). "A Model for Optimum Linkage of Records." JASA 63: 1321-32.

[30] Turner, J. Scott, and Gilliam, Gary B. (1975). "Reducing and Merging Microdata Files," OTA Paper 7, Office of Tax Analysis, U.S. Treasury Department (October).

[31] "Unimatch 1 Users Manual--A Record Linkage System" (1974). Bureau of the Census, Census Use Study. Washington, March.

[32] U.S. Dept. of Agriculture, Statistical Reporting Service (1977). "Selection of a Surname Coding Procedure for the SRS Record Linkage System." (B.T. Lynch and W.L. Arends). Paper presented at the April 6 meeting of the Washington Statistical Society.

[33] U.S. Dept. of Commerce, National Bureau of Standards (1977). "Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers." NBS Special Publication 500-2.

[34] Wolff, Edward N. (1974). "The Goodness of Match," National Bureau of Economic Research Working Paper No. 72 (December).