Taylor & Francis
Taylor & Francis Group

©Journal of Business & Economic Statistics, Vol. 2, No. 1, January 1984

An Evaluation of Statistical Matching

Willard L. Rodgers
Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI 48106

The validity of findings based on statistically matched data sets depends on the accuracy of underlying assumptions about relationships between variables that are unique to each input file. Simulations of statistical matching procedures on samples from populations with known characteristics provide the basis for an evaluation of the usefulness of statistical matching, and for choosing among various matching techniques. In three such simulations frequent and substantial errors occur in estimates of bivariate and multivariate relationships between variables taken from two matched files. Alternative procedures for meeting the objectives of statistical matching with more solid theoretical justification and empirical support are proposed.

KEY WORDS: Statistical matching; Missing data; Imputation.

1. INTRODUCTION

Analyses of microdata often require data from units that are not available from a single source but are available from a set of sources. For example, suppose that one is interested in the relationships among two sets of variables: Perhaps one set consists of information about health care expenses incurred by individuals, and the other set consists of information about receipt of various types of welfare benefits. Suppose further that no existing data file contains all of the needed variables, but that two surveys have recently been conducted that, between them, contain all of the necessary variables for large samples of the target population. If mounting a new survey to obtain all of the needed variables from a single sample is not feasible, the only alternative to foregoing the analysis (and therefore, perhaps, developing policy based on a poor understanding of the empirical relationships) is to use information about relationships contained in the two separate data files.

Statistical matching is one method that is used to meet this objective. Statistical matching procedures were developed by analogy with exact matching procedures, whereby records from one source (perhaps a sample survey) are linked to records from a second source (perhaps administrative data) using identifiers such as social security numbers that permit information about the same individual unit to be identified in the two sources. Often, however, it is not possible to do an exact match of two data files, either because the cases in the two files have little or no overlap (e.g., if each consists of a relatively small sample of a large population), or because variables that would allow identification of individuals are not available in one or both of the input files. In statistical matching, each record from one of the data sources is matched with a record from the second source that generally does *not* represent the *same* unit, as in exact matching but does represent a similar unit. A description of the various procedures developed for statistical matching, and of existing statistical matches, may be found in Radner et al. (1980).

Statistical matching procedures have been developed without much theoretical foundation or empirical justification. More recently, efforts have been made to correct this deficiency. The purpose of this article is to review those efforts and to evaluate the potential usefulness of statistical matching. There are two parts to this evaluation. The first part is general and mathematical: We ask what types of assumptions underlie statistical matching techniques and the possible consequences of invalid assumptions. The second part is empirical and specific to particular data sets: In three separate studies, data from a single source were treated as though they were collected from two separate samples, and several statistical matches were performed. Information from those matched files was then compared with the information on the original, complete file. The quality of the files produced by different statistical matching techniques is compared both relative to one another and in absolute terms. We then try to answer the question, How often are misleading findings indicated by statistically matched files?

2. STATISTICAL MATCHING PROCEDURES

2.1 Notation

Let $\mathbf{S}$ be a vector of variables for each of n_A records on file A (sometimes referred to as the base file), and let $\mathbf{T}$ be a vector of variables for each of n_B records on

file B (the supplemental file), such that both vectors consist of indicators of a common set of characteristics of the analysis units. These common characteristics are referred to as **X** variables, $\mathbf{X} = (X_1, \ldots, X_p)$. The remaining variables in each of the files will be referred to as **Y** variables, $\mathbf{Y} = (Y_1, \ldots, Y_Q)$, in file A and as **Z** variables, $\mathbf{Z} = (Z_1, \ldots, Z_R)$, in file B. Observed or derived values of these variables are designated by lower-case letters (s, t, x, y, and z). In addition, a sampling weight, **w**, may be associated with each record in both files.

To illustrate the matching process with a miniature (and highly simplified) example, consider the two files shown in Tables 1a and 1b. There are eight records on file A and six records on file B. The **S** variables on base file A are labeled sex and year of birth, and the **T** variables on supplemental file B are labeled sex and age. These **S** and **T** variables are taken as indicators of two **X** characteristics, sex and age. No transformation is necessry to convert s_1 and t_1 to x_1 or t_2 to x_2, but s_2 must be subtracted from the year of the data collection (here taken to be 1980) to yield x_2. A single **Y** variable, ln(personal earnings), is shown for units in file A, and a single **Z** variable, ln(property income), is shown for units in file B. The units in both files are simple random samples drawn from a population of 24 total units. The weight assigned to each record is inversely proportional to the probability of selection.

The matching process requires the definition of a distance function, which permits the similarity of any pair of cases to be assessed in terms of a function, D_{ij}, of the **X** variables. Moreover, certain **X** variables may be treated as so-called cohort variables. Cohort variables establish subclasses of the records in each of the two files, with matching permitted only between pairs of cases in the same subclass. In our example, we define X_1, sex, as a cohort variable: A male can only be matched with another male, and a female with another female. The distance function is defined as the absolute difference in ages of two cases: $d_{ij} = |x_{2i} - x_{2j}|$. A weight is assigned to each record in the matched file: This weight, w_{ij}, may be equal to w_i, the weight associated with the input case from file A, or it may be modified depending on the matching technique and the need to align units in the two input files.

Table 1a. Simplified Example of Statistical Matching: File A

Case	$s_1 = x_1^A$ Sex	s_2 Year Born	x_2^A Age	y^A ln(E)	w_i^A Weight
A1	M	1938	42	9.156	3
A2	M	1945	35	9.149	3
A3	F	1917	63	9.287	3
A4	M	1925	55	9.512	3
A5	F	1952	28	8.494	3
A6	F	1927	53	8.891	3
A7	F	1958	22	8.425	3
A8	M	1955	25	8.867	3
Mean	.50		40.38	8.97	
SD	.53		15.32	.38	

Table 1b. File B

Case	$t_1 = x_a^B$ Sex	$t_2 = x_2^B$ Age	z^B ln(P)	w_j^B Weight
B1	F	33	6.932	4
B2	M	52	5.524	4
B3	M	28	4.223	4
B4	F	59	6.147	4
B5	M	41	7.243	4
B6	F	45	3.230	4
Mean	.50	43.00	5.55	
SD	.55	11.58	1.57	

2.2 Unconstrained and Constrained Matches

Within the realm of statistical matching, two basic types are recognized (Radner et al. 1980, p. 18). If there are no restrictions on the number of A records to which the values of the **Z** variables in supplemental file B can be imputed, then the statistical match is unconstrained (see, e.g., Okner 1972). This is illustrated in Table 1c: Each male in file A is matched with the male in file B whose age is closest to his own. Similarly, each female in file A is matched with a female in file B. Notice that three of the records in supplemental file B (B1, B4, and B5) are each matched to two records in base file A, while one B record (B6) is not attached to *any* A record. Moreover, both the mean and standard deviation of the **Z** variable in the matched file differ from the corresponding statistics in file B. Unconstrained statistical matching has the advantage of permitting the closest possible match for each A record, but at the cost of increasing the sample variance of estimators involving the **Z** variables. An unconstrained match amounts to taking a simple random sample, with replacement, of the records in file B. The distributions of the imputed **Z** variables added to file A, then, are distributions of the selected sample rather than the distributions as observed in file B. Despite this disadvantage, unconstrained matches have been used frequently.

An alternative type of statistical matching is called constrained matching (e.g., see Barr and Turner 1980).

Table 1c. Unconstrained Match

A Case	B Case	$x_1^A = x_1^B$	x_2^A	x_2^B	d_{ij}	y^A	z^B	w_{ij}
A1	B5	M	42	41	1	9.156	7.243	3
A2	B5	M	35	41	6	9.149	7.243	3
A3	B4	F	63	59	4	9.287	6.147	3
A4	B2	M	55	52	3	9.512	5.524	3
A5	B1	F	28	33	5	8.494	6.932	3
A6	B4	F	53	59	6	8.891	6.147	3
A7	B1	F	22	33	11	8.425	6.932	3
A8	B3	M	25	28	3	8.867	4.223	3
	Mean	.50	40.38	43.25	4.88	8.97	6.30	
	SD	.53	15.32	12.40	3.00	.38	1.06	

The conditions that must be met by a constrained match can be stated as

$$\sum_{j=1}^{m} w_{ij} = w_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{1}$$

and

$$\sum_{i=1}^{n} w_{ij} = w_j, \quad \text{for} \quad j = 1, \ldots, m. \tag{2}$$

Now, letting $d_{ij} \geq 0$ be the distance between cases i and j in files A and B, respectively, the objective of a constrained match is to minimize the following function:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (d_{ij} * w_{ij}), \tag{3}$$

subject to the conditions listed above and with $w_{ij} \geq 0$ for all i and j. This procedure is illustrated in Table 1d. It can be observed that the average distance between matched cases is greater for this constrained match than for the unconstrained match shown in Table 1c. It can also be observed, however, that the mean and standard deviation of the Z variable are identical in the matched file to their values in file B. The advantage of a constrained match relative to an unconstrained match is that the multivariate distribution of the Z variables as observed in file B is precisely replicated in the matched file. Disadvantages include the fact that the paired cases differ more with respect to the common (X) variables than is true for unconstrained matches, and the fact that for certain constrained matching procedures, which minimize the distances between paired cases across a large set of cases, the costs in computer time are considerable.

One final point—it should be noted that for the estimated standard deviation of Z in the constrained match to equal the estimate from file B, it is necessary to use the proper degrees of freedom (df) (5) from input file B rather than the 7 df for variables taken from file A or the 11 df that would be assumed by a standard analysis program that fails to take account of the origin of the data. The difference is trivial here, but this illustrates a hazard in the use of statistically matched files. The number of records in such a file may be quite different from the number of cases from which information about a particular variable or set of variables were actually collected. Failure to take this into account during the analysis could lead to highly misleading tests of statistical significance.

Table 1d. Constrained Match

A Case	B Case	$x_1^A = x_1^B$	x_2^A	x_2^B	d_{ij}	y^a	z^B	w_{ij}
A1	B2	M	42	52	10	9.156	5.524	1
A1	B5	M	42	41	1	9.156	7.243	2
A2	B3	M	35	28	7	9.149	4.223	1
A2	B5	M	35	41	6	9.149	7.243	2
A3	B4	F	63	59	4	9.287	6.147	3
A4	B2	M	55	52	3	9.512	5.524	3
A5	B1	F	28	33	5	8.494	6.932	3
A6	B4	F	53	59	6	8.891	6.147	1
A6	B6	F	53	45	8	8.891	3.230	2
A7	B1	F	22	33	11	8.425	6.932	1
A7	B6	F	22	45	23	8.425	3.230	2
A8	B3	M	25	28	3	8.867	4.223	3
	Mean	.50	40.38	43.00	6.46	8.97	5.55	
	SD	.53	15.32	11.58	5.81	.38	1.57	

3. THEORETICAL BASIS FOR STATISTICAL MATCHING

The inherent assumption in stastistical matching is that the random vector Y given X is independent of the random vector Z given X. For the particular case of multivariate normal distributions of the variables, this is equivalent to the assumption that the partial correlations among the Y and Z variables, controlling on the X variables, are all zero. This point was made by Sims (1972) and repeatedly by others since then.

The conditional independence assumption is a strong one for which little justification has generally been offered. It implies that Y's relationship to Z can be totally inferred from Y's relationship to X and Z's relationship to X. Occasionally, information about the relationship of a Y-Z pair is available from another source, permitting an improvement in the assumption of conditional independence. It should be clear, however, that this does not change the basic assertion that the statistical matching procedure does not generate new information about the conditional relationship of the Y-Z pair, but only reflects the assumptions used in creating the matched file (Kadane 1978, p. 166).

To illustrate the strength of the conditional independence assumption, it is useful to consider the total range of values that the correlation between a single observed Y variable and a single observed Z variable could have, given the constraints implied by the observed correlations of those variables with the X variables. Only if there is an extremely high multiple correlation between either a Y variable or a Z variable and the set of available X variables is the range of possible values for the correlation of the Y and Z variables at all narrow (cf. Wolff 1974; Rodgers and DeVol 1982a). Indeed, the multiple correlation must be so high that either the Y or the Z variable is close to being a simple linear combination of the X variables. For example, suppose that a linear combination of X variables correlates .80 with both a Y and a Z variable. The range of possible correlations between those Y and Z variables is from .28 to a perfect 1.0. If the combination of X variables correlates .8 with the Y variable but only .5 with the Z variable, the range of possible values for the Y-Z correlation is from −.12 to +.92.

The potential usefulness of statistical matching procedures depends not only on the observed correlations of the **X**, **Y**, and **Z** variables, but also on assumptions that the analyst is willing to make about the causal relationships among these variables. For example, suppose that one wishes to estimate a model that has the following form:

$$\mathbf{Z} = \mathbf{X}_1\beta + \mathbf{Y}\gamma + \epsilon, \tag{4}$$

where the dependent variable (**Z**) is measured for one sample of cases, along with some but not all of the predictors included in the model (i.e., the $\mathbf{X}_1$ variables); and that the other predictors (the **Y** variables) are measured for a second sample of the same population. Whether statistical matching will support the estimation of the parameters of such a linear model depends on the specific set of **X** variables, $\mathbf{X}_1$, included in expression (4) and the assumptions about the relationships of the **X** and **Y** variables. In particular, Expression (4) is not estimable unless for every **Y** variable included in that expression, at least one of the **X** variables used in the argument of the distance function is excluded from the set of $\mathbf{X}_1$ variables included in (4).

To show this, we summarize a demonstration given by Klevmarken (1982). Suppose that the entire set of **X**, **Y**, and **Z** variables is considered part of a system specified by the following expression:

$$\mathbf{Y}\mathbf{B} + \mathbf{Z}\Lambda + \mathbf{X}\Gamma = \mathbf{U}, \tag{5}$$

with $E(\mathbf{U}) = \mathbf{0}$, and with the **Y** and **Z** variables considered endogenous variables and the **X** variables as exogenous. Expression (4) can be considered one component of this system, and another component can be written for the **Y** variables:

$$\mathbf{Y} = \mathbf{X}\pi + \mathbf{V}. \tag{6}$$

It would be possible to estimate the parameters π from sample A, and then generate predicted values $\hat{\mathbf{Y}}^{\mathrm{B}}$ for cases in sample B, based on the observed values of the **X** variables. Under certain conditions, it would then be possible to estimate the parameters in expression (4) using sample B data:

$$\mathbf{Z}^{\mathrm{B}} = \mathbf{X}_1^{\mathrm{B}}\beta + \hat{\mathbf{Y}}^{\mathrm{B}}\gamma + \epsilon. \tag{7}$$

This expression can be rewritten as

$$\mathbf{Z}^{\mathrm{B}} = \mathbf{M}^{\mathrm{B}}\delta + \epsilon, \tag{8}$$

where $\mathbf{M}^{\mathrm{B}} = (\mathbf{X}_1^{\mathrm{B}} \mid \hat{\mathbf{Y}}^{\mathrm{B}})$ and $\delta = (\beta \mid \gamma)$. Then the parameters in δ can be estimated by ordinary least squares procedures:

$$\hat{\delta} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{Z}, \tag{9}$$

provided that the inverse matrix, $(\mathbf{M}'\mathbf{M})^{-1}$ is defined. Since the rank of $(\mathbf{M}'\mathbf{M})$ cannot exceed the number of **X** variables (P), the existence of this inverse matrix requires that at least as many of the **X** variables be omitted from expression (4) as the number of **Y** variables that are included in expression (4). Klevmarken (1982) shows that statistical matching is an alternative method to this two-step estimation procedure, albeit one in which the $\mathbf{X}_1^{\mathrm{B}}$ values are replaced by the matched $\mathbf{X}_1^{\mathrm{A}}$ values, since it is generally not possible to find perfect matches for all of the cases. In particular, the same conditions hold for the estimability of the parameters of expression (4).

We do not develop the argument for other types of multivariate analysis, but the logic is the same and could be applied, for example, to multivariate analyses of contingency tables. In short, one should exercise even greater caution before attempting to carry out multivariate analyses on statistically matched files than for estimating bivariate relationships.

4. ALTERNATIVES TO USUAL STATISTICAL MATCHING PROCEDURES

Statistical matching, although orders of magnitude less expensive than a special data collection, makes demands on professional and computer resources that are nonetheless considerable. It is pertinent to ask, therefore, whether there are alternatives to statistical matching that would either be less difficult and expensive, or more fully meet the objectives of statistical matching, or both.

For many types of analyses, the answer is unequivocably "yes." Any analysis that requires only the univariate distributions of variables does not require statistical matching. This does not necessarily mean that statistical matching provides incorrect information when the intention is only to provide better estimates of univariate distributions. It does appear, however, that statistical matching is an unwieldy and expensive method to apply for such a straightforward objective. For example, one objective in the creation of the 1972 Match File described by Radner (1983) was to correct for inaccuracies in reported income amounts in survey data from the Current Population Survey (CPS). Data from income tax returns are not subject to such large inaccuracies, but tax data are deficient in that not all individuals file tax returns. To provide better estimates of the distribution of income, then, it is reasonable to consider combining data from the two sources. The 1972 Match File was created through statistical matching, but if the only objective of this procedure had been that just described, alternative procedures would have been adequate and more straightforward to implement. For example, it would have been possible to compare the distributions of variables that are available on both of the input files with the 1972 Match File, and to supplement the tax records with cases from the CPS in appropriate numbers to match the distributions of the com-

mon variables in the combined file to the observed distributions in the CPS file (assuming that the CPS file is to be taken as representative of the target population).

Statistical matching is also unnecessary—indeed, it may be counterproductive—whenever a variance-covariance matrix is central to the analysis. Since an extremely wide variety of questions can be addressed through use of properly specified variance-covariance matrices, many types of analyses would never require files created through statistical matching. This assertion follows from consideration of the complete variance-covariance matrix, $V(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, representing the linear interrelationships of the **X**, **Y**, and **Z** variables:

$$V(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \begin{bmatrix} V(\mathbf{X}) & V(\mathbf{X}, \mathbf{Y}) & V(\mathbf{X}, \mathbf{Z}) \\ V'(\mathbf{X}, \mathbf{Y}) & V(\mathbf{Y}) & V(\mathbf{Y}, \mathbf{Z}) \\ V'(\mathbf{X}, \mathbf{Z}) & V'(\mathbf{Y}, \mathbf{Z}) & V(\mathbf{Z}) \end{bmatrix}. \tag{10}$$

The only submatrix of V that cannot be directly estimated from the input files is $V(\mathbf{Y}, \mathbf{Z})$. Statistical matching provides a method of estimating that submatrix, but only by making assumptions about the conditional covariances of the **Y** and **Z** variables controlling on the **X** variables. Those assumptions, whether they are the usual default of zero-conditional covariances or some set of alternative values based on outside information, are more efficiently introduced into the estimation of $V(\mathbf{Y}, \mathbf{Z})$ through matrix addition and multiplication than through statistical matching (cf. Anderson 1958):

$$V(\mathbf{Y}, \mathbf{Z}) = V(\mathbf{Y}, \mathbf{Z} \mid \mathbf{X}) + V'(\mathbf{X}, \mathbf{Y})V(\mathbf{X})^{-1}V(\mathbf{X}, \mathbf{Z}). \tag{11}$$

In particular, the assumption of the conditional independence of the **Y** and **Z** variables implies that

$$V(\mathbf{Y}, \mathbf{Z}) = V'(\mathbf{X}, \mathbf{Y})V(\mathbf{X})^{-1}V(\mathbf{X}, \mathbf{Z}), \tag{12}$$

where $V(\mathbf{X}, \mathbf{Y})$ is estimated from file A, $V(\mathbf{X}, \mathbf{Z})$ is estimated from file B, and $V(\mathbf{X})$ is estimated either from file A alone or from a concatenation of files A and B.

There are, of course, many types of analysis for which it is not sufficient to estimate the variance-covariance matrix, for example, for the analysis of polytomous dependent variables, or when there is concern about the validity of assumptions about the linearity or additivity of relationships, or the normality or homoscedasticity of the distributions of error terms. Some of these concerns can be alleviated by supplementing the variance-covariance matrix, of course, but in some cases it may be more appropriate to use techniques that require more than simply the variances and covariances.

The inappropriateness of the variance-covariance matrix as a basis for certain types of analysis does not necessarily imply the necessity for creating a microdata file consisting of complete records for each case. For example, one type of analysis in which it is not sufficient to estimate the variance-covariance matrix is one that requires producing bivariate and multivariate frequency tables involving both **Y** and **Z** variables. Such frequency tables, however, could be generated without statistical matching from the observed marginal and joint distributions, using iterative proportional fitting (cf. Bishop et al. 1975).

Nevertheless, despite the existence of alternative strategies that are available and probably preferable for many types of analysis, the creation of a file of microdata records with measures of all variables—**X**, **Y**, and **Z** variables—is attractive because of the flexibility it offers the analyst. This is particularly true if the data are to be used by many analysts for many different purposes—that is, if the data are to become a public resource.

Even in such circumstances, although statistical matching is one way to generate such a file, it is not the only way and may not be the best way. The basis for this assertion is the fact that commonly statistical matching procedures fail to preserve all of the information available in the input files. This is reflected in the fact that matched pairs do not agree completely on the **X** variables. Generally, only a subset of the potential **X** variables are used to define the distance function. (It should be noted that variables not used either as cohort variables or in the argument of the distance function, even though available on both input files, are equivalent to **Y** and **Z** variables; their values are simply transferred to the matched file.) Moreover, even for the subset of variables used either as cohort variables or in the argument of the distance function, it is not possible to obtain perfect agreement across all pairs of matched cases. This means that the relationships between the **X** variables (from input file A) and the **Z** variables (from input file B) do not coincide with those relationships as observed in file B. In a sense, then, statistical matching may introduce error into the values of the **X** variables vis-à-vis the **Z** variables. This source of error may often be minor compared with other sources of error, especially if the number of cases in the two input files is large so that fairly close matches can be found for most cases. It should be noted, however, that the error introduced by statistical matching is not uniform but tends to be larger in sparser regions of the multivariate distribution.

A related problem with statistical matching is that the distance function used to match cases in the input files may be defined in terms of a set of **X** variables that have a high explanatory power with respect to some of the **Y** and **Z** variables, but it is highly unlikely that any distance function is optimal with respect to the entire set of **Z** variables. This means that, at least for some of the **Z** variables, their relationships with **X** and **Y** variables estimated from a statistically matched file will be inconsistent with the relationships of the **X** and **Y** variables as estimated from file A and with the relation-

ships of the X and Z variables as estimated from file B.

In the case of a multivariate normal distribution for the entire set of variables, the variance-covariance matrix preserves all of the information in the input files. The effectiveness of this type of procedure with respect to a particular data set was tested in one of the simulations summarized in this article.

In addition to the direct estimation of the covariance matrix, alternative procedures have been proposed that would both preserve all of the information in the input files and support a full range of analysis procedures, including the production of bivariate and multivariate frequency tables. One such procedure was suggested by Sims (1978): Multivariate density functions involving X, Y, and Z variables would be estimated in accordance with the observed density distribution of X and Y variables in file A, and for X and Z variables in file B. A practical problem that often arises in the implementation of this procedure is that even with a moderate number of variables, the number of cases required to make reasonable estimates of the multivariate distribution may become excessive. A second alternative, which avoids this problem, is an imputation procedure proposed by Rubin (1980a).

5. SIMULATIONS OF STATISTICAL MATCHING PROCEDURES

The major conclusion of this article can already be stated, since it rests on the nature of statistical matching rather than empirical analysis: statistically matched files are a risky basis for any analyses that involve relationships between Y and Z variables. The separate files contain no information about the conditional relationships among the Y and Z variables. Statistical matching adds *no* information but only reflects the implicit or explicit assumptions made in the matching procedure. An important question, then, is how much confidence can be placed in the assumption of conditional independence. This question cannot be answered in general, of course, but it is possible to explore how often, and how well, such an assumption is met for a particular set of variables.

Empirical tests of statistically matched files may be helpful in providing guidelines for the use of this procedure. Such empirical tests may serve to indicate how often analyses based on statistically matched files lead to erroneous conclusions, as well as the magnitude of the errors introduced in the combined file. They can also serve to demonstrate the importance of such factors as the strength of associations among the X variables and the Y and Z variables; the number of cases in the component files; the nature of the distance function used to match cases; and the choice among alternative matching procedures.

The validity of the conditional independence assumption with respect to the Y and Z variables in any particular statistical match must in general remain untested. If information is available from another source concerning the relationship of the Y and Z variables, that information can and should be incorporated into the matching process. The resulting statistically matched file, however, provides no possibilities for testing the validity of the conditional independence assumption. Nevertheless, there have been efforts to assess the usefulness of statistical matching procedures by performing statistical matches on simulated data with known distributions. In addition, statistical matches have been performed on data sets that contain all three sets of variables (i.e., the X, Y, and Z variables), but ignore subsets of those variables in the matching process and then comparing the joint distributions of these variables in the statistically matched file to their observed joint distributions in the input files. It is important to note, however, that a file created in this fashion represents an idealized match—the populations from which the two sources of data are drawn, the definitions of the X variables used to match cases, and the operationalizations of those variables are all identical. Estimates from actual statistical matches can be expected to perform less well than such simulations to the extent that these conditions are not met.

5.1 Ruggles, Ruggles, and Wolff

Perhaps the first such empirical study was done by Ruggles, Ruggles, and Wolff (1977). In their study, these investigators used a matching method developed by Ruggles and Ruggles (1974) to match the 1970 Census 1/1,000 5% Public Use Sample (PUS) with the 1970 Census 1/1,000 15% PUS. More than 20 variables were measured for cases in both of these samples, but only a subset was used to match cases. Regression analyses were then done, using various combinations of variables from each of the two input files, and the statistical significance of differences in the estimates of regression coefficients based on the original and matched files. They report only 2 of 42 cases in which the differences in estimates were statistically significant, and on this basis conclude that their statistical matching technique can provide an adequate data source for many statistical applications. It should be noted, however, that most of these comparisons involved regressions that included only X and Y variables, so that for these regressions the differences in the estimated regression coefficients reflect only on the closeness of the matches, not on the validity of the conditional independence assumption.

5.2 Paass and Wauschkuhn

A more comprehensive evaluation of statistical matching procedures was carried out by Paass and Wauschkuhn (1980). For their simulations, these investigators used a sample of 10,000 cases included in a file of administrative records collected by the German Fed-

eral Ministry of Education and Science. Twelve of the variables were treated as **X** variables; these were demographic and income variables. Some of these were used as cohort variables to define six cells, with matches only allowed between cases falling in the same cell. The remaining variables were classified as either **Y** or **Z** variables. Four different matching procedures (and a total of five matches) were then used to create statistically matched files, and these files were then compared with one another and with the original, complete file. Specifically, the matching procedures that they implemented are those developed by Okner (1972); Alter (1974); Ruggles and Ruggles (1974); and Armington and Odle (1975, two different variations). All of these procedures are unconstrained.

The **X** and **Y** variables were transferred directly from the base file to the matched file in all four of these procedures, so their univariate and joint distributions are identical. The univariate distributions of the **Z** variables, and their joint distributions with one another and with the **X** and **Y** variables, may, on the other hand, be distorted by the statistical matching procedure, and it was on these distributions that Paass and Wauschkuhn (1980) focused their attention. They found that all four procedures produced univariate distributions of each of four **Z** variables that were very similar to their distributions in the original data. With respect to the bivariate distribution of **X** and **Z** variables, they found that in general the statistically matched files were quite similar to the original distributions, although the Alter method produced 3 out of 17 joint distributions that differed significantly ($p < .01$). Only one of those distributions differed significantly for the other three methods. With respect to the overall multivariate distribution of the **X** and **Z** variables, Paass and Wauschkuhn conclude that the methods of Ruggles and Ruggles and of Okner are satisfactory, but that the methods of Alter and Armington and Odle are unsatisfactory because of excessive reliance placed in those matches on a single **X** variable, income, to define the matches.

The most crucial test of the success of any statistical match is the accuracy with which the relationships of **Y** and **Z** variables are estimated. In this respect Paass and Wauschkuhn found that all four matching procedures produced joint distributions that in a majority of instances differed significantly from the distributions observed in the original data file. These distortions arose, in part, from imperfections in the matches between cases, but primarily because of the invalidity of the conditional independence assumption.

5.3 Barr, Stewart, and Turner

Barr, Stewart, and Turner (1982) performed a large number of statistical matches both on artificial and real sets of data. Each artificial data set consisted of 200 records with four variables: two **X** variables, one **Y** variable, and one **Z** variable. These records were generated from a multivariate normal population with means of 0 and variances of 1 for each variable, and various levels of covariances among the four variables. Altogether, these investigators generated 100 pairs of independent A ($\mathbf{X}_1$, $\mathbf{X}_2$, and **Y**) and B ($\mathbf{X}_1$, $\mathbf{X}_2$, and **Z**) files for each of 12 populations, where the populations differed with respect to the covariances of the variables. Then for each such pair of files, six statistical matches were performed: three constrained and three unconstrained matches, using three distance functions for each type of match. The first was a weighted sum of the absolute differences between the cases on the two **X** variables; the second the Mahalanobis distance, as suggested by Kadane (1978); and the third a modification of the Mahalanobis distance, also proposed by Kadane (1978), which expands the variance-covariance matrix of the **X** variables to include observed or predicted values of the **Y** and **Z** variables.

All three distance functions with the constrained matching procedure provided accurate estimates of the variance of the **Z** variable. This is to be expected, since constrained matching simply reproduces the distribution of the **Z** variable observed in the B sample. Unconstrained matches do not have this characteristic, however. Barr, Stewart, and Turner (1982) found that all three unconstrained matching procedures produced **Z** distributions that had significantly different means from the population value.

Barr, Stewart, and Turner report the estimated covariances of the **Z** variable with the two **X** variables only for two of the matching procedures, both of them constrained. They found that these covariances tended to be underestimated, especially when these variables were highly correlated; this may reflect the less than perfect agreement of the values of the **X** variables for many of the matched cases.

The findings of Barr, Stewart, and Turner, with respect to the most crucial question, the estimation of relationships between **Y** and **Z** variables, were unequivocal: if the conditional independence assumption was invalid, all statistical matching procedures provided estimates of the **Y**-**Z** covariance that were extremely poor. On the other hand, for the cases in which the conditional independence assumption was valid, all six procedures provided estimates of the **Y**-**Z** covariance that were generally quite accurate.

For their simulations with actual data, Barr, Stewart, and Turner used the 1975 Survey of Income and Education. Random samples of approximately 1,000 cases each were drawn from that file, and subsets of the variables were designated as **X**, **Y**, and **Z** variables in such a way as to provide a range of types of variables and covariances among them. Then a total of 50 statistical matches were performed on pairs of these files, all using the optimal-constrained procedure developed at

the U.S. Treasury Department, Office of Tax Analysis (Barr and Turner 1978). Six different distance functions were used in executing these matches: the Mahalanobis distance, using all of the **X** variables; and five weighted sums of the absolute differences of various subsets of the **X** variables. Again, the most important findings from this evaluation have to do with the estimation of the covariance of **Y** and **Z** variables, and again the findings are clear in this regard: Files created through statistical matching gave poor estimates of the **Y**-**Z** covariances if the conditional independence assumption was invalid. Most notably, the highest correlation of a **Y** and a **Z** variable in the original sample was .76 between family size and number of adults, but the average of the estimates of this correlation across the various procedures ranged from −.01 (for the matches using the Mahalanobis distance function) to .31.

Barr, Stewart, and Turner's simulations also indicated that the Mahalanobis distance function produced less accurate matching than subjectively weighted distance functions, and that distance functions with too few **X** variables included (specifically, three instead of five or six) produced less accurate matching. On the other hand, rather substantial infusions of noise and/or bias into one of the **X** variables used in the distance function had relatively little effect on the quality of the match.

5.4 Rodgers and DeVol

The objective of the simulation reported by Rodgers and DeVol (1982a, b) was to evaluate statistical matching techniques as they apply to data from the Income Survey Development Program (ISDP). In designating subsets of the available variables as **X**, **Y**, and **Z** variables, the objective was to define three sets of variables that were typical of the variables encountered in actual matches. That is, it was intended that the **X** variables that formed the basis for matching cases be typical of variables used as match variables in previous statistical matches and be available for matching ISDP data with other sample survey data. The sets of **Y** and **Z** variables were chosen because they were sets of variables that might reasonably be found in two separate surveys.

On the basis of their similarity on selected **Z** variables, the cases were first categorized into a set of cells, defined by a subset of the **X** variables (the cohort variables), and only cases that were in the same cell were eligible for matching. Two constrained matches were completed, using the same procedure as that used by Barr, Stewart, and Turner (1982) in their simulations. The two matches differed only with respect to the distance functions. The first match used a distance function based on regression analyses done separately for the cases in each of 13 cells. In contrast, the second match used a single distance function for the entire sample, with **X** variables and associated weights chosen arbitrarily. Similarly, three unconstrained matches were performed, one using the regression-based distance function, the other two using the second (arbitrary) distance function. The latter two matches differed with respect to the number of cases matched: One of the unconstrained matches was done on a quarter of the available cases; whereas, all of the other matches were done on the full set of about 16,000 cases.

For the purpose of comparison with these matching procedures, there was also a direct estimation of the variance-covariance matrix, $V(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, based on information in the input files A and B and a set of assumptions about the conditional covariances of the **Y** and **Z** variables. Rodgers and DeVol estimated the $V(\mathbf{Y}, \mathbf{Z})$ covariance matrix in the simplest possible manner: the **X**, **Y**, and **Z** variables were used as they were originally coded, and only additive and linear relationships were taken into account. In particular, unlike the statistical matches that were done separately for each of 13 subsamples as defined by the cohort variables, the direct estimation was based on the covariances estimated from a random quarter of the entire sample.

Rodgers and DeVol found that the means, variances, and covariances of some of the **Z** variables were severely distorted in the files created through unconstrained matching. These distortions can be partially attributed to the wide range of weights assigned to cases in the input files, which were not considered in the matching procedure. Because of this problem, the generalizability of the findings from the unconstrained matches is suspect, and findings from these matches will not be further described except to say that they provided considerably less accurate estimates than did the constrained matches and the direct estimation procedure.

The correlations between the 33 **X** and 18 **Z** variables as estimated from the matched files were compared with the correlations as estimated from the original file. The average discrepancy between these values was small, but the average value of the observed correlations was also small. The average value of these discrepancies was 29% and 22% of the average observed correlation for the first and second constrained matches, respectively. Moreover, there was an apparent bias to the estimates: the matches underestimated the absolute values of the **X**-**Z** correlations by an average of about 14% and 10% of their observed values, respectively.

With respect to the estimation of the relationships between the 25 **Y** and 18 **Z** variables, the constrained matches and the direct estimation procedure provided estimates of correlations that were generally quite close in absolute terms to the estimates from the original data, but quite poor if compared with the magnitude of the original correlations. For the first match, the average inaccuracy was only .029, and only about 2% of the correlations differed from the observed correlations by more than .10. On the other hand, most of the

observed correlations of these variables were also quite small: The average of their absolute values was only .074. From this perspective, the average discrepancy was about 40% as large as the average actual value. Moreover, the magnitudes of these correlations were underestimated by an average of almost a third of their actual values. The second match provided estimates of the Y-Z correlations that were somewhat more accurate than those from the first match (average absolute difference = .021), and with considerably less of a downward bias. The direct estimation of the covariance matrix produced estimates that were at least as accurate as either of the statistical matches. Since the discrepancies for this procedure are simply the partial correlations of the Y and Z variables controlling on the X variables, it happens that for this data set, and for these sets of variables, the conditional independence assumption is quite accurate for most Y-Z pairs.

Another set of comparisons made by Rodgers and DeVol (1982a) was concerned with estimation of parameters in multiple regression models. In each of a set of regression analyses, a single Y variable was used as a predictor along with several X variables. Comparisons between observed and matched data files were made with respect to standardized regression coefficients and the marginal predictive powers of the Y variables. With respect to the coefficients of the Y variables, some of the estimates from the constrained matches were fairly close to the actual values, but on the average the estimates from the first match were only 62% as large (in absolute value) as the estimates from the original data; the estimates from the second match were somewhat better, but still averaged only 74% of the original estimates. Some of the estimates were sufficiently inaccurate to give warning that one can not be certain about the accuracy of findings (or even the direction of the relationship) from statistically matched files when the predictors in the regression analysis include both X and Y variables. The direct estimation of the covariance matrix produced estimates of the regression coefficients that were not as good as those from the constrained matches. This may reflect the fact that the variance-covariance matrix was estimated from only a quarter of the cases on which the matches were performed, or the failure of this procedure as implemented to take into account possible interactions of the X and Z variables across the several cells. With respect to the marginal predictive powers of the Y variables, the findings were even more negative: the estimates from the constrained matches were an average of only 40% to 50% as large, and those from the direct estimation procedure only about a third as large as the estimates from the original data.

6. CONCLUSIONS

Statistical matching procedures have been implemented to meet a variety of objectives, but all (or at least all that would justify the effort of conducting a statistical match) center about the estimation of relationships between variables that are measured for two or more samples that either contain few overlapping cases or that cannot be linked through exact matching for practical or legal reasons.

The simulations of statistical matching procedures that have been summarized in this article provide insights into choices involved in the implementation of a statistical match. Some of these choices will be considered before turning to the more basic question of the usefulness of statistical matching techniques in general.

The first issue is the choice between constrained and unconstrained procedures. Unconstrained matching procedures have the advantage of relative simplicity and lower costs in terms of computer processing time and memory requirements, at least as compared with the constrained optimization matching procedure used in two of the simulations described in this study. The disadvantages of unconstrained procedures, however, are considerable. In the simulations described by Barr, Stewart, and Turner (1982) and by Rodgers and DeVol (1982a), substantial distortions were introduced into the univariate and joint distributions of the Z variables; such distortions are completely avoided by constrained matching. Moreover, the other comparisons, involving the covariances of X and Z variables, the covariances of Y and Z variables, and regression analyses involving all three sets of variables, all indicated that unconstrained matches introduce more error than do constrained matches.

Constrained matching procedures avoid the problems just described for unconstrained matching procedures and produced considerably more accurate estimates of covariances, regression coefficients, and of the proportion of explained variance in regression analyses. The costs of carrying out a constrained match may be considerable, however.

Another issue in the implementation of a statistical match is the choice of cohort variables and of X variables to include in the distance function. The evidence from the simulation studies is mixed with respect to this issue. Paass and Wauschkuhn (1980) found that statistical matches created using procedures proposed by Alter (1974) and by Armington and Odle (1975) were less satisfactory than those created using procedures proposed by Ruggles and Ruggles (1974) and by Okner (1972), and they attribute the poorer results of the former procedures to their excessive reliance on a single X variable in the distance function. Barr, Stewart, and Turner found that matches utilizing the Mahalanobis distance function were less accurate than matches based on subjectively weighted sums of the absolute differences between a set of X variables. Rodgers and DeVol (1982a) found that a distance function derived empirically on the basis of regression analyses to a set of Z variables produced matched files that were less

accurate than an arbitrarily defined distance function. There is little basis for making any reasonable recommendation on the basis of the available information, but neither is there any evidence that would suggest any alternative that would be consistently superior to a simple subjectively weighted sum of the absolute differences between values on the **X** variables. In particular, there is neither theoretical nor empirical justification for the use of any version of the Mahalanobis distance function.

A third issue in the implementation of statistical matching procedures is the minimum size of the input files required to perform a match. Existing implementations of statistical matching procedures have generally been done on large input samples, with 50,000 or more cases apiece. In contrast, the simulations of statistical matching have been performed on much smaller samples. The simulations by Paass and Wauschkuhn involved a sample of 10,000 cases; those by Barr, Stewart, and Turner (1982) involved samples of about 1,000 cases, and those by Rodgers and DeVol involved a sample of about 16,000 cases. Rodgers and DeVol also specifically tested for the importance of the sample size by performing one of three unconstrained matches on just a quarter of the cases (about 4,000 cases) and could detect no deterioration in the quality of the matched file compared with an identical match on the full sample.

To make sense of this finding, it is helpful to think of statistical matching as a procedure for imputing values of the **Z** variables to the base file, A; if the relationship between a **Z** variable, $\mathbf{Z}_i$, and the set of **X** variables in the population is expressed as $\mathbf{Z}_i = \mathbf{X}\beta + \epsilon$, then the lack of perfect matches on the **X** variables introduces error into the **X** values associated with each imputed value of $\mathbf{Z}_i$. The **X** variables, however, are often measured with considerable error even in the input files, and the **X** variables explain only part of the variance in the $\mathbf{Z}_i$ variable, so that the amount of error introduced because of the matching procedure is often probably a rather small proportion of the measurement and stochastic error. This argument is speculative, but it would explain why the quality of statistically matched files does not appear to be sensitive to the number of cases in the input files, despite the closer fit of matched cases that is obtained as the number of available cases increases.

A final issue is whether consideration should be given to alternatives to statistical matching. We have observed that constrained matching, and to an even greater extent unconstrained matching procedures, fail to take full account of the information in the input files. In particular, some distortion is introduced into the estimates of the covariances of **X** and **Z** variables. The direct estimation of the complete variance-covariance matrix for the **X**, **Y**, and **Z** variables is one alternative that may, if properly implemented, avoid the loss of certain information that characterizes statistical matching procedures. This procedure was also simulated in the study by Rodgers and DeVol, but not under ideal conditions, so the accuracy of the estimates should not be regarded as a good indication of its potential. The fact that the direct estimation of the covariance matrix did as well or better than the statistical matches on most of the comparisons, even under these less than ideal conditions, implies that further consideration should be given to this procedure. Moreover, the cost of implementing this procedure should be much less than that of implementing a statistical match.

The variance-covariance matrix is sufficient for many types of analyses, but not for all, and in particular not for multivariate contingency table analyses and for other types of nonparametric techniques. If this is the case, I suggest that an imputation procedure (Rubin 1980a) be seriously considered as an alternative to statistical matching. This procedure should produce a data file in which the univariate distributions of the **Z** variables, and the joint distributions of these variables with one another and with the **X** variables, are all very similar to the distributions observed in the input file containing actual observations on the **Z** variables.

The simulations of statistical matching procedures also suggest ways in which at least some aspects of the quality of a statistically matched file can, and should, be assessed. Certainly the univariate and joint distributions of the **Z** variables as estimated from a file created through an unconstrained procedure should be compared with the distributions of those variables in the input file. Instances of substantial distortions were found in the simulations by Rodgers and DeVol (1982a) and by Barr, Stewart, and Turner (1982). The former set of simulations indicates that the possibility of distortion may be especially important if the cases in the supplemental file have a wide range of weights, since unconstrained matching procedures do not necessarily take account of such weights.

Another check that should be performed on a statistically matched file is the accuracy with which the relationships between **X** and **Z** variables are estimated. Constrained as well as unconstrained matching procedures may introduce distortions into these distributions, and all three of the simulations resulted in instances where this in fact occurred.

In general, no direct check is possible on the estimates of the relationships among the **Y** and **Z** variables. Rubin (1980a), however, suggests a procedure that, although it provides no direct evidence on the accuracy of such estimates, at least offers the opportunity to assess the sensitivity of those estimates to changes in assumptions about their conditional relationships after controlling on the **X** variables. Rubin's proposal is to match each case in the base file A, not with a single case from file B, but with multiple cases. (See also Rubin 1980b, where he discusses the usefulness of multiple imputa-

tions for missing data.)

This brings us to the primary issue with respect to statistical matching. All the issues concerning the best way to implement a statistical match are subordinate to the underlying issue of whether *any* type of statistical match can provide useful information about the unobserved relationships of **Y** and **Z** variables. From a mathematical perspective, we have observed that statistical matching lacks a strong foundation as a procedure for estimating these **Y**-**Z** relationships. The relationships among **X** and **Y** variables, as observed in one input file and among **X** and **Z** variables as observed in a second file, provide only quite broad constraints on the possible values of the covariances between **Y** and **Z** variables. Nevertheless, it could be hoped that by judicious choice of the **X** variables used to match the two files, it would turn out that in practice most of the conditional relationships between **Y** and **Z** variables would indeed be close to zero, so that statistical matching would, in fact, provide acceptable estimates for those relationships.

The findings from the simulations summarized in this article (excluding the more limited simulation by Ruggles, Ruggles, and Wolff 1977) are consistent in showing that statistical matching often produces very poor estimates of these relationships among **Y** and **Z** variables. This conclusion holds for all four unconstrained matching procedures tested by Paass and Wauschkuhn (1980); for both the constrained and unconstrained matching procedures tested by Barr, Stewart, and Turner (1982); and for the constrained and unconstrained matching procedures and the direct estimation procedure tested by Rodgers and DeVol (1982a).

On the basis of these simulations, which confirm the caution arising from the absence of any mathematical justification for statistical matching, it seems clear that statistical matching may not in general be an acceptable procedure for estimating relationships between **Y** and **Z** variables, or for any type of multivariate analysis involving both **Y** and **Z** variables.

If statistical matching is not generally acceptable, is there any remaining choice for the analyst confronted with the need for conducting an analysis on a set of variables that are not all available in a single data file? In some circumstances, the answer is yes. Consider again the complete variance-covariance matrix of the **X**, **Y**, and **Z** variables (Expression 10): It was assumed that submatrices of this matrix can be estimated from two separate files, but that the $V(\mathbf{Y}, \mathbf{Z})$ component cannot be estimated from either of those files. It may be, however, that this set of relationships *is* estimable from a *third* source; perhaps not for a sample of exactly the population in which we are interested, or not for very recent data, but nevertheless a set of estimates we are willing to accept, if only because they are preferable to the alternative conditional independence assumption. In such a case, it is straightforward to carry out any analysis that can be based on a variance-covariance matrix.

If the analysis requires the generation, not of a variance-covariance matrix but of a multivariate contingency table, then iterative proportional fitting techniques can be used to generate such a table from the marginal distributions observed in the various input files. If it is necessary to generate a microdata file, modified statistical matching procedures have been proposed by Paass (1982) that combine the information from all sources in the creation of a microdata file that can then be used as the basis for any desired analysis. The procedure proposed and evaluated by Paass provides predictions of the **Z** values for input file A (which contains values for only **X** and **Y** variables), based on three sources of information: the relationships of the **X** and **Y** variables in file A; the relationships of the **X** and **Z** variables in file B; and finally, information about the relationships of the **Y** and **Z** variables from a third source. The generation of these predicted values employs the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977).

It should be remembered, however, that the information in a variance-covariance matrix, multivariate contingency table, or microdata file created through any of the three procedures just mentioned is only as good as the information in the various input files. For example, if the relationships among the **Y** and **Z** variables are estimated from a relatively small file, but a statistical match is implemented on a pair of large files, tests of statistical significance and confidence intervals based on the matched file may be highly misleading unless proper account is taken of the sources of information. If, moreover, the information about the **Y**-**Z** relationships is taken from a sample drawn at an earlier point in time than the other files, or from a sample of a different population, or from a sample of inferior quality, the interpretation of analyses based on such a matched file should be especially cautious.

It must also be kept in mind that all three of the methods just mentioned, while preferable to traditional matching procedures because they eschew unwarranted assumptions about the overall relationships between **Y** and **Z** variables, nevertheless resemble those traditional procedures in that they continue to rely on a critical set of assumptions that are in general not testable, assumptions that the relationships of the **Y** and **Z** variables are constant across levels of the **X** variable. That is, interactions involving the **X**, **Y**, and **Z** variables can only be detected from a file that contains observations on all three sets of variables. This suggests the advisability of obtaining information about all three sets of variables from at least a small sample of the target population, or carrying out an exact match on the two input files if there is a sufficient overlap of identical cases in the two files and if those overlapping cases are representative of

the target population.

It is possible that much of the interest in statistical matching procedures arises from an unfortunate analogy of statistical matching with exact matching. Historically, the roots of statistical matching are clearly related to exact matching. In exact matching, moreover, it is often necessary to deal with problems of errors in the link variable (say, in Social Security numbers as reported in a survey), and supplementary information may be used to increase the probability that matched records apply to the same individual, and to find the most probable match when unique identifiers disagree because of recording or reporting errors (Radner et al. 1980). It is a relatively small step, computationally, from such procedures for exact matching of identical individuals to statistical matching of similar individuals. A small step for the computer is in this case a giant step for the statistician—a step that should only be taken with full awareness of the importance of the implicit assumptions and the potential consequences of the incorrectness of those assumptions.

ACKNOWLEDGMENTS

The research reported in this article was supported by Contract HEW-100-79-0127 from the Income Survey Development Program to the Survey Development Research Center in Nonresponse and Imputation, for which Graham Kalton was Principal Investigator. An early draft of this article was done in collaboration with Edward DeVol. The author also wishes to acknowledge contributions to this research from Graham Kalton and Daniel Kasprzyk, and helpful comments on a previous version of this article from the associate editor and referees.

[*Received May 1982. Revised September 1983.*]

REFERENCES

ALTER, H. 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*, 3, 373–394.

ANDERSON, T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley.

ARMINGTON, C., and ODLE, M. (1975), "Creating the MERGE-70 File: Data Folding and Linking," Research on Microdata Files Based on Field Surveys and Tax Returns, Working Paper 1, Washington, D.C.: The Brookings Institution.

BARR, R. S., STEWART, W. H., and TURNER, J. S. (1982), "An Empirical Evaluation of Statistical Matching Methodologies," Unpublished mimeo, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Tx., January 1982.

BARR, R. S., and TURNER, J. S. (1978), "A New, Linear Programming Approach to Microdata File Merging," *1978 Compendium of Tax Research*, Office of the Treasury, Washington, D.C.: U.S. Government Printing Office, 131–149.

——— (1980), "Merging the 1977 Statistics of Income and the March 1978 Current Population Survey," Prepared for the Office of Tax Analysis, Washington, D.C.: U.S. Dept. of the Treasury.

BISHOP, Y. M. M., FEINBERG, S. E., and HOLLAND, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Mass.: The MIT Press.

DEMPSTER, A. P., LAIRD, N. M., and RUBIN, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

KADANE, J. B. (1978), "Some Statistical Problems in Merging Data Files," *1978 Compendium of Tax Research*, Office of Tax Analysis, Dept. of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159–171.

KLEVMARKEN, N. A. (1982), "Missing Variables and Two-Stage Least Squares Estimation From More Than One Data Set," *1981 Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 156–161.

OKNER, B. A. (1972), "Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement*, 1, 325–342.

PAASS, G. (1982), "Statistical Match With Additional Information," Internal Report IPES.82.0204, Gesellschaft Für Mathematik und Datenverarbeitung, Bonn, West Germany.

PAASS, G. and WAUSCHKUHN, U. (1980), "Experimentelle Erprobung und Vergleichende Bewertung Statistischer Matchverfahren," Internal Report IPES.80.201, Gesellschaft Für Mathematik und Datenverarbeitung, Bonn, West Germany.

RADNER, D. B. (1983), "Adjusted Estimates of the Size Distribution of Family Money Income," *Journal of Business & Economic Statistics*, 1, 136–146.

RADNER, D. B., ALLEN, R., GONZALEZ, M. E., JABINE, T. B., and MULLER, H. J. (1980), "Report on Exact and Statistical Matching Techniques," *Statistical Policy Working Paper 5*, U.S. Dept. of Commerce, Washington, D.C.: U.S. Government Printing Office.

RODGERS, W. L., and DEVOL, E. (1982a), "An Evaluation of Statistical Matching," unpublished report submitted to Income Survey Development Program, Dept. of Health and Human Services, Ann Arbor, Michigan: Institute for Social Research, The University of Michigan.

——— (1982b), "An Evaluation of Statistical Matching," *1981 Proceedings of the American Statistical Association, Section on Survey Research Methods*, 128–132.

RUBIN, D. B. (1980a), "File Concatenation With Adjusted Weights and Multiple Imputations: A Solution to the File Matching Problem Different in Principle From the Constrained Optimization Approach," Unpublished manuscript, 17 July 1980.

——— (1980b), *Handling Nonresponse in Sample Surveys by Multiple Imputations*, Washington, D.C.: U.S. Bureau of the Census.

RUGGLES, N., and RUGGLES, R. (1974), "A Strategy for Merging and Matching Microdata Sets," *Annals of Economic and Social Measurement*, 3, 353–371.

RUGGLES, N., RUGGLES, R., and WOLFF, E. (1977), "Merging Microdata: Rationale, Practice, and Testing," *Annals of Economic and Social Measurement*, 6, 407–428.

SIMS, C. A. (1972), "Comments" (on Okner 1972), *Annals of Economic and Social Measurement*, 1, 343–345.

——— (1978), "Comment" (on Kadane 1978), *1978 Compendium of Tax Research*, Office of Tax Analysis, Dept. of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172–177.

WOLFF, E. N. (1974), "The Goodness of Match," National Bureau of Economic Research Working Paper No. 72, December.