Journal of Official Statistics, Vol. 17, No. 3, 2001, pp. 407–422

Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure

Chris Moriarty[1] *and Fritz Scheuren*[2]

Statistical matching has been widely used by practitioners without always adequate theoretical underpinnings. The work of Kadane (1978) has been a notable exception and the present article extends his insights. Kadane's 1978 article is reprinted in this JOS issue. Modern computing can make possible, under techniques described here, a real advance in the application of statistical matching.

Key words: Multivariate normal; complex survey designs; robustness; resampling; variance-covariance structures; and application suggestions.

1. Introduction

Many government policy questions, whether on the expenditure or tax side, lend themselves to microsimulation modeling, where ''what if'' analyses of alternative policy options are carried out (e.g., Citro and Hanushek 1991). Often, the starting point for such models, in an attempt to achieve a degree of verisimilitude, is to employ information contained in several survey microdata files. Typically, not all the variables wanted for the modeling have been collected together from a single individual or family. However, the separate survey files may have many demographic and other control variables in common. The idea arose, then, of matching the separate files on these common variables and thus creating a composite file for analysis.

''Statistical matching,'' as the technique began to be called, has been more or less widely practiced since the advent of public use files in the 1960's. Arguably, the desire to employ statistical matching was even an impetus for the release of several of the early public use files, including those involving U.S. tax and census data (e.g., Okner 1972).

Statistical matching always has had an ad hoc flavor (Scheuren 1989), although parts of the subject have been examined with care (e.g., Cohen 1991; Rodgers 1984; Sims 1972). In this article we return to one of the important attempts to underpin practice with theory. This is the work of Joseph Kadane (1978), which is now over 20 years old (reprinted in this JOS issue).

[1] 200 Spring Avenue, Takoma Park, MD 20912. Email: chrismor@cpcug.org
[2] The Urban Institute, 2100 M Street. NW, Washington, DC 20037, U.S.A. Email: FScheure@ui.urban.org

Acknowledgments: This article is based in large part on the unpublished doctoral dissertation cited as Moriarty (2001). Results sketched here are fully developed in that reference. Thanks are due to Tapan Nayak, Reza Modarres, and Hubert Lilliefors of The George Washington University Department of Statistics for useful discussions. The authors also thank the associate editor and referees for thoughtful comments that improved this article.

We begin by describing Kadane's contribution. Needed refinements are then made which go hand-in-hand with advances in computing in the two decades since his article was written. To frame the results presented, after this introduction (Section 1) we include a section (Section 2) entitled, ''What is Statistical Matching?'' This is followed by a restatement of the original results by Kadane (Section 3 and the Appendix). Then, in Section 4, improvements of Kadane's procedure are presented and their properties developed. Section 5 provides some simulation results that illustrate the new approach. Section 6 discusses generalizations of the approach. In the last section (Section 7) an application is sketched as part of a brief discussion of implementation issues.

2. What is Statistical Matching?

Perhaps the best description to date of statistical matching was given by Rodgers (1984). Other good descriptions are given by Cohen (1991) and Radner et al. (1980). A summary of the method is provided here.

Suppose there are two sample files, File A and File B, taken from two different surveys. Suppose further that File A contains potentially vector-valued variables (X, Y), while File B contains potentially vector-valued variables (X, Z). The objective of statistical matching is to combine these two files to obtain at least one file containing (X, Y, Z).

In contrast to record linkage, or exact matching (e.g., Fellegi and Sunter 1969; Scheuren and Winkler 1993 and 1997), the two files to be combined are *not* assumed to have records for the same entities. In statistical matching the files are assumed to have little or no overlap; hence, records for *similar* entities are combined, rather than records for the *same* entities. For example, we might want to match individuals who are similar on characteristics like sex, age, poverty status, health status, etc.

All statistical matches described in the literature have used the X variables in the two files as a bridge to create a single file containing (X, Y, Z). To illustrate, suppose File A consisted, in part, of records

$X_1, \quad Y_1$

$X_2, \quad Y_2$

$X_3, \quad Y_3$

while File B has records of the form

$X_1, \quad Z_1$

$X_3, \quad Z_3$

$X_4, \quad Z_4$

$X_5, \quad Z_5$

The matching methodologies employed almost always have made the assumption that (Y, Z) are conditionally independent, given X, as pointed out initially by Sims (1972). From this it would be immediate that we could create

$X_1, \quad Y_1, \quad Z_1$

$X_3, \quad Y_3, \quad Z_3$

Notice that matching on X_1 and X_3 (where X is, say, age) in no way implies that these are the same entities.

What to do with the remaining records is less clear and techniques vary. We could stop where we are, only partially matching the two files, but this option is not usually taken.

Broadly, the various strategies employed for statistical matching can be grouped into two general categories: ''constrained'' and ''unconstrained.'' Constrained statistical matching requires the use of all records in the two files and basically preserves the marginal Y and Z distributions (e.g., Barr, Stewart, and Turner 1982). In the above (toy) example, for a constrained match we would have to end up with a combined file that also had additional records that used the remaining unmatched File A record (X_2, Y_2) and the two unmatched File B records (X_4, Z_4) and (X_5, Z_5). In other words, all of the records on both files get used. Notice that, as would generally be the case, we could not preserve the role of X in the matching as one where only identical X's were matched. We would have to settle on matching X's that were only close (similar) to one another.

Unconstrained matching does not have the requirement that all records are used. Indeed, we might stop after creating (X_1, Y_1, Z_1) and (X_3, Y_3, Z_3). Usually in an unconstrained match, though, all the records from one of the files (say File A) would be used (matched) to ''similar'' records on the second file. Some of the records on the second file may be employed more than once, or not at all. Hence in the unconstrained case the remaining unmatched record on File A, the observation (X_2, Y_2), would be matched to make the combined record (X_2, Y_2, $Z_{??}$). The observations (X_4, Z_4) and (X_5, Z_5) from File B might or might not be included. Exactly how we defined ''similar'' would, of course, determine the values of the variables without specific subscripts. To go further into this now, however, would be to get ahead of ourselves.

A number of practical issues, not part of our present scope, need to be addressed in statistical matching; for example, alignment of universes (i.e., agreement of the weighted sums of the data files) and alignment of units of analysis (i.e., individual records represent the same units). Also, X variables can have different measurement or nonsampling properties in the two files. See Cohen (1991) and Ingram et al. (2000) for further details. Statistical matching is by no means the only way to combine information from two files. Sims (1978), for instance, described alternative methodologies to statistical matching that could be employed under conditional independence.

Other authors (e.g., Singh et al. 1993; Paass 1986 and 1989) have described methodologies for statistical matching if auxiliary information about the (Y, Z) relationship is available. While an important special case, this option is seldom available (Ingram et al. 2000). See also National Research Council (1992), where the subject of combining information has been taken up quite generally.

Rodgers (1984) includes a more detailed example of combining two files, using both constrained and unconstrained matching, than the example we have provided here. We encourage the interested reader to consult that reference for an illustration of how sample weights are used in the matching process.

3. Kadane's Procedure for Statistical Matching

In the setting described above, Kadane (1978) sets out a methodology for statistical matching where the vector (X, Y, Z) is assumed to have a nonsingular multivariate normal distribution with covariance matrix

$$\sum = \begin{pmatrix} \sum_{XX} & \sum_{XY} & \sum_{XZ} \\ \sum_{YX} & \sum_{YY} & \sum_{YZ} \\ \sum_{ZX} & \sum_{ZY} & \sum_{ZZ} \end{pmatrix}$$

Note that all elements of $\sum$ can be estimated from File A or File B except $\sum_{YZ}$ and its transpose, $\sum_{ZY}$.

Kadane's procedure begins by selecting an admissible value of $\sum_{YZ}$ (Cov(Y, Z) or Corr(Y, Z) in the univariate case). (Henceforth in this article, we usually will use $\sum_{YZ}$ generically for both the multivariate case and the univariate case.) What we mean by ‘‘an admissible value of $\sum_{YZ}$’’ is a value of $\sum_{YZ}$ that would give a positive definite $\sum$.

The given value of $\sum_{YZ}$ is used in regressions done on both files to produce augmented files containing (X, Y, $\hat{Z}$) (File A) and (X, $\hat{Y}$, Z) (File B). The files then are matched using a Mahalanobis distance (as defined in Section 3.3.), and Y and Z values are exchanged in matched records to obtain the augmented records (X_j, Y_j, Z_i) (File A) and (X_i, Y_j, Z_i) (File B), where the jth record of File A was matched to the ith record of File B. The matching methodology prescribed by Kadane is a constrained match, so all records in the two files are used in the matching process. The end result is two files containing records of the form (X_j, Y_j, Z_i) (File A) and (X_i, Y_j, Z_i) (File B), where the jth record of File A was matched to the ith record of File B.

Kadane recommended that this procedure be repeated for many values of $\sum_{YZ}$, thereby obtaining a range of synthetic datasets (X, Y, Z) under various assumptions on the value of $\sum_{YZ}$. Kadane's procedure is one of only two procedures described in the statistical matching literature for assessing the effect of alternative assumptions of the inestimable value $\sum_{YZ}$. Rubin (1986) discusses the other procedure.

3.1. Specification of $\sum_{YZ}$

Kadane's procedure requires the specification of $\sum_{YZ}$. Kadane correctly states that nothing can be learned about $\sum_{YZ}$ from File A and File B, beyond the assumed nonsingularity of the distribution of (X, Y, Z).

More specifically, in the case of univariate (X, Y, Z), one can start from several sources, e.g., the definition of the partial correlation of (Y, Z), given X, or the requirement that the correlation matrix of (X, Y, Z) must be positive definite, to show that Corr(Y, Z) must lie in the interval

$$Corr(X,Y) * Corr(X,Z) \pm \sqrt{[1-(Corr(X,Y))^2] * [1-(Corr(X,Z))^2]}$$

This provides a bound on Corr(Y, Z) in terms of Corr(X, Y) and Corr(X, Z) (which can be estimated using File A and File B, respectively).

Note that Corr(Y, Z) is equal to Corr(X, Y)$*$Corr(X, Z) in the special case of conditional independence of (Y, Z), given X. This special case has been discussed extensively in the statistical matching literature. However, here it is seen that the ‘‘conditional independence value’’ is the midpoint of a range of admissible values of Corr(Y, Z).

In general, the bounds given above are wide, even for large values of Corr(X, Y) and Corr(X, Z). For example, when both Corr(X, Y) and Corr(X, Z) are equal to 0.75 (or both are equal to -0.75), the admissible range of Corr(Y, Z) values is (0.125, 1). In most cross-sectional surveys, finding correlations larger than 0.4 in absolute value has been extremely rare in the experience of the authors. If Corr(X, Y) and Corr(X, Z) both equal 0.4, the range of admissible Corr(Y, Z) values is (-0.68, 1), virtually useless as a guide to practice. This is one of the reasons why practitioners commonly have invoked the conditional independence assumption, whether or not it could be justified.

3.2. *Regression step*

As Kadane discusses, for a given value of $\sum_{YZ}$, the procedure begins by estimating the missing values in the two sample files using conditional expectation (i.e., regression). For example, if Z is missing and all needed quantities are known, then (e.g., Anderson 1984, p. 36)

$$E(Z_j|X_j, Y_j) = \mu_Z + \left(\sum_{ZX}\sum_{ZY}\right)\begin{pmatrix}\sum_{XX} & \sum_{XY}\\ \sum_{YX} & \sum_{YY}\end{pmatrix}^{-1}\begin{pmatrix}X_j-\mu_X\\ Y_j-\mu_Y\end{pmatrix}$$

In this application, all quantities other than $\sum_{ZY}$ can be estimated using one or both sample files. For the given value of $\sum_{ZY}$, this procedure is followed for File A, and a similar procedure is followed for missing Y in File B.

As stated by Kadane, it can be shown that the joint distribution of (X_j, Y_j, $\hat{Z}_j$) is normal with mean (μ_X, μ_Y, μ_Z) and (singular) covariance matrix

$$S_1 = \begin{pmatrix}\sum_{XX} & \sum_{XY} & \Phi_1'\\ \sum_{YX} & \sum_{YY} & \Phi_2'\\ \Phi_1 & \Phi_2 & \Phi_3\end{pmatrix}$$

Kadane gave an analogous expression for the joint distribution of (X_i, $\hat{Y}_i$, Z_i):

$$S_2 = \begin{pmatrix}\sum_{XX} & \Phi_4' & \sum_{XZ}\\ \Phi_4 & \Phi_6 & \Phi_5'\\ \sum_{ZX} & \Phi_5 & \sum_{ZZ}\end{pmatrix}$$

When rederiving these expressions, we realized that the formulas Kadane gave for the Φ_i, $i = 1$ to 6 (refer to the appendix), were more complicated than they needed to be. It can be shown (Moriarity 2001) that

$$(\Phi_1\Phi_2) = \left(\sum_{ZX}\sum_{ZY}\right)$$

and

$$\begin{aligned}\Phi_3 &= \left(\sum_{ZX}\sum_{ZY}\right)\begin{pmatrix}\sum_{XX} & \sum_{XY}\\ \sum_{YX} & \sum_{YY}\end{pmatrix}^{-1}\begin{pmatrix}\sum_{XZ}\\ \sum_{YZ}\end{pmatrix}\\ &= \sum_{ZX.Y}\left(\sum_{XX.Y}\right)^{-1}\sum_{XZ} + \sum_{ZY.X}\left(\sum_{YY.X}\right)^{-1}\sum_{YZ}\end{aligned}$$

In a similar manner, analogous results can be derived for Φ_4, Φ_5, and Φ_6 (Moriarity 2001). These simplifications were useful to us in the evaluation of Kadane's procedure and the development of an improved methodology.

3.3. Matching step

Letting $W_j = (X_j, Y_j, \hat{Z}_j)$ from File A and $V_i = (X_i, \hat{Y}_i, Z_i)$ from File B, the mean of $W_j - V_i$ is the zero vector. Kadane states that the covariance matrix of $W_j - V_i$, which is the sum of the two singular covariance matrices S_1 and S_2, is nonsingular. (Moriarity (2001) provides a proof of this result.)

Given the nonsingularity of $S_1 + S_2$, Kadane suggested using the Mahalanobis distance in (X, Y, Z) to do a constrained match of File A and File B. In matrix algebra notation, the Mahalanobis distance being minimized in the matching would be of the form

$$(W_j - V_i)'(S_1 + S_2)^{-1}(W_j - V_i)$$

Alternatively, Kadane suggested using a Mahalanobis distance involving only X. (For this alternative, it would not be necessary to carry out the regression step, as the regressed values would not be in the distance computation.)

As noted by Kadane, using either of these suggested matching procedures while requiring use of a constrained match can be shown to be equivalent to solving a ''transportation problem,'' a type of linear programming problem (see, e.g., Barr and Turner 1978; Bertsekas 1991). That is, it is known that an algorithm exists to carry out these matching procedures.

4. Preservation of $\sum_{YZ}$ During the Procedure

The derivations sketched in Section 3.2. show that the specified value of $\sum_{YZ}$ is preserved during the regression step. However, an extensive simulation (see Section 5 for details of the simulation methodology) of Kadane's suggested matching procedure showed that the specified value of Corr(Y, Z) is not preserved during the matching step. This was true regardless of whether matching was performed using (X, Y, Z) or X, the two methods suggested by Kadane. Hence, a revision of Kadane's methodology was needed. This section describes our proposed revision.

4.1. Additional alternatives

Matching on (X, Y, Z) or X, as Kadane did, does not exhaust the possibilities. Why not match on (Y, Z)? This is the very relationship we are interested in. Matching using (Y, Z) might do the best job of preserving (Y, Z) relationships. Indeed, simulations showed that matching using (Y, Z) was far more successful, on average, in retaining the specified value of Corr(Y, Z), than the two methods Kadane examined. Even so, this method did not always give good results, particularly (refer to Table 1) for values of Corr(Y, Z) that were far from the conditional independence value of Corr(X, Z)$*$Corr(X, Y).

Referring to results sketched in Section 3.2., S_1 and S_2 are given in simplified form by

$$S_1 = \begin{pmatrix} \sum_{XX} & \sum_{XY} & \sum_{XZ} \\ \sum_{YX} & \sum_{YY} & \sum_{YZ} \\ \sum_{ZX} & \sum_{ZY} & \Phi_3 \end{pmatrix}$$

and

$$S_2 = \begin{pmatrix} \sum_{XX} & \sum_{XY} & \sum_{XZ} \\ \sum_{YX} & \Phi_6 & \sum_{YZ} \\ \sum_{ZX} & \sum_{ZY} & \sum_{ZZ} \end{pmatrix}$$

with Φ_3 as given in Section 3.2. It can be shown (Moriarity 2001) that

$$\begin{aligned} \Phi_6 &= \left(\sum_{YX} \sum_{YZ}\right) \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{ZX} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{XY} \\ \sum_{ZY} \end{pmatrix} \\ &= \sum_{YX.Z}\left(\sum_{XX.Z}\right)^{-1}\sum_{XY} + \sum_{YZ.X}\left(\sum_{ZZ.X}\right)^{-1}\sum_{ZY} \end{aligned}$$

Now, $\sum_{ZZ} - \Phi_3$ and $\sum_{YY} - \Phi_6$, the variances of the residuals from the regressions, can be characterized as variances of random variables with certain *conditional distributions* (e.g., Anderson 1984, p. 37). Hence, the two covariance matrices can be made equal by imputing independently drawn normally distributed random residuals with mean zero and the specified variances, and adding these residuals to $\hat{Z}_j$ and $\hat{Y}_i$. In addition to making the two covariance matrices equal (both to each other and to the covariance matrix of the common parent distribution), this approach also has the benefit of converting singular distributions to nonsingular distributions. Given this observation, imputation of random residuals prior to matching appeared to be a promising innovation to explore.

4.2. *Success of alternatives*

In our simulations of matching using (Y, Z), after imputation of random residuals, we found our new method to be very successful in retaining the specified value of Corr(Y, Z), always giving good results.

Furthermore, as shown in Table 1, the method is *robust* for values of Corr(Y, Z) close to, and far from, the conditional independence value of Corr(X, Z)*Corr(X, Y). This robustness is a feature not found in other methods employed in statistical matching (as noted by Rodgers 1984).

In summary, it appears to us that in order to preserve the value of $\sum_{YZ}$ during the matching step, it is critical to align the covariance matrices by imputing residuals *prior* to the matching step, rather than doing the match first and then, in essence, imputing residuals *afterwards*.

5. Simulations Conducted

In order to assess the performance of Kadane's original procedure, and our variants, simulations were carried out. For simplicity, univariate X, Y, and Z were used, and (X, Y, Z) were constructed to have a multivariate normal distribution. Also, for convenience and without loss of generality, (X, Y, Z) were assumed to have zero means and unit variances.

5.1. *Basic simulation setup*

Corr(X, Y) was allowed to vary from 0 to 0.95 in increments of 0.05. For a given value of Corr(X, Y), Corr(X, Z) was allowed to vary from Corr(X, Y) to 0.95 in increments of 0.05.

Given the symmetry of Y, Z in Corr(Y, Z) and the fact that Corr(−A, B) = Corr(A, −B) =− Corr(A, B), the effect of allowing Corr(X, Y) and Corr(X, Z) to range from −0.95 to 0.95 can be inferred. For given values of Corr(X, Y) and Corr(X, Z), Corr(Y, Z) was allowed to take four to ten different values within the range of admissible values specified in Section 3.1. (The number of values of Corr(Y, Z) depended on the length of the interval of admissible values of Corr(Y, Z).)

For given values of Corr(X, Y), Corr(X, Z), and Corr(Y, Z), two independent samples of size 1,000 were drawn from the specified multivariate normal distribution. We felt that using a sample size of 1,000 was a reasonable compromise to simulate a dataset of realistic size with minimal sampling validity, while avoiding excessive computational burden.

The regression step was carried out as described in Section 3.2. Note that although it is possible to pool the X values from both files to estimate Var(X), this can and does lead to occasional problems of nonpositive definite covariance matrix estimates in the regression step (Moriarity 2001). To avoid any possibility of a nonpositive definite covariance matrix estimate, X values only from File A were used to estimate Var(X) when predicting Z from X and Y for File A, and X values only from File B were used to estimate Var(X) when predicting Y from X and Z for File B.

For the matching procedures that included imputation of residuals, random residuals with mean zero and specified variances were added to the $\hat{Z}_j$ and the $\hat{Y}_i$ after the regression step. $\sum_{ZZ} - \Phi_3$ and $\sum_{YY} - \Phi_6$ were estimated using the data. In a small proportion (about 8 percent) of the simulations, one of these estimates was negative; when a negative estimate occurred, no random residuals were added to the respective variable. In our simulations, there were no occurrences where both estimates were negative (although such an occurrence is possible). Any such occurrence in practice could be expected to give results consistent with the results presented in Table 1 for matching on (Y, Z) (without imputing residuals to the $\hat{Z}_j$ and the $\hat{Y}_i$ after the regression step); that is, better than other suggested methods in the existing literature, but not ideal.

The constrained match was carried out using RELAX-IV software (Bertsekas 1991, Bertsekas and Tseng 1994), which is written in FORTRAN and is in the public domain. The RELAX-IV software was executed from within SAS macro code that included extensive use of SAS/IML. Note, however, that it was not necessary to use the RELAX-IV software for conducting the match using univariate X. As shown by Goel and Ramalingam (1989), it suffices to match the sorted values of X in the two files – a major saving in computational effort.

All of the simulation work was carried out on a Sun Ultra 60 workstation, with a 360 MHz Sparc-II CPU and 512 MB RAM. A set of 1,873 simulations typically took about a week of continuous computer processing to complete.

5.2. Results obtained

Table 1 summarizes five sets of results: matching on (X, Y, Z) and matching on X (the two procedures suggested by Kadane), plus three more variants: matching on (Y, Z), and matching on (X, Y, Z) and (Y, Z) after residuals were added to $\hat{Y}$ and $\hat{Z}$.

A comparison of the first two columns within a given row shows the relative performance of a matching procedure for values ‘‘near’’ conditional independence versus values

Table 1. Summary of simulation results for Kadane's method and related methods

Matching procedure	Average absolute difference between specified value of Corr(Y, Z) and value computed from matched (Y, Z) pairs		Performance reproducing specified values of Corr(X, Z) in File A and Corr(X, Y) in File B
	Corr(Y, Z) values near conditional independence value 1,049 simulations	Corr(Y, Z) values away from conditional independence value 824 simulations	
match on (X, Y, Z)	.07	.49	bad
match on X	.07	.52	good
match on (Y, Z)	.04	.08	bad
match on (X, Y, Z) after adding residuals to $\hat{Y}$ and $\hat{Z}$	.01	.01	usually OK, but a few bad results
match on (Y, Z) after adding residuals to $\hat{Y}$ and $\hat{Z}$	.01	.01	good

''far'' from conditional independence. In general, performance was worse for values ''far'' from conditional independence; for some procedures, dramatically worse. Matching on (X, Y, Z) and (Y, Z) after adding residuals were the only procedures with robust performance.

For the first two columns, a comparison of the rows within a given column illustrates the relative ability of different procedures to maintain the specified value of Corr(Y, Z) during the matching step. The differences in the procedures are particularly easy to see in the column corresponding to values ''far'' from conditional independence. Matching on (X, Y, Z) or on (Y, Z) after adding residuals performed better than the other procedures, both for values ''near'' conditional independence and for values ''far'' from conditional independence.

The last column of Table 1 provides summary-level information about each procedure's ability to accurately reproduce other correlation estimates, such as Corr(X, Z) in File A and Corr(X, Y) in File B, over the set of simulations. (Note that since constrained matching was used, means and variances automatically were preserved.) We examined both the overall average absolute deviations and individual occurrences of large absolute deviations (e.g., larger than .10). Procedures with ''good'' performance had low averages (e.g., less than .05) with no occurrences of large deviations, while procedures with ''bad'' performance had both high averages and occurrences of large deviations. Matching on (X, Y, Z) after adding residuals had low averages, but some occurrences of large deviations (as high as .18).

The occurrences of some large deviations when matching on (*X*, *Y*, *Z*) after adding residuals fell among the small proportion of simulations, mentioned earlier, when residuals were not imputed for one of the variables. No large deviations occurred for this (adverse) situation when matching on (*Y*, *Z*) after adding residuals. These results suggest that matching on (*Y*, *Z*) after imputing residuals gives more robust performance in this specific situation. However, as our research did not focus on this area, we think that more examination is warranted before reaching such a conclusion.

It is useful to compare and contrast the performance of the two methods suggested by Kadane and the method of matching on (*Y*, *Z*) (without adding residuals). We found that matching on *X* consistently led to the creation of synthetic files with Corr(*Y*, *Z*) values very close to the ''conditional independence'' value of Corr(*X*, *Y*)*Corr(*X*, *Z*), as expected. Matching on (*X*, *Y*, *Z*) had some effect on moving the synthetic estimates of Corr(*Y*, *Z*) away from the conditional independence value, but at the price of introducing distortions into the synthetic estimates of Corr(*X*, *Z*) in File A and Corr(*X*, *Y*) in File B. Matching on (*Y*, *Z*) gave great improvement in terms of allowing the synthetic estimates of Corr(*Y*, *Z*) to do a better job of reproducing the specified values of Corr(*Y*, *Z*); however, as was seen for matching on (*X*, *Y*, *Z*), distortions were introduced into the synthetic estimates of Corr(*X*, *Z*) in File A and Corr(*X*, *Y*) in File B. It is not surprising to see these distortions when the matching process involves regression estimates of *Y* and *Z* as linear functions of *X*.

5.3. Simulation summary

Several conclusions can be drawn from the results summarized in Table 1, and the discussion in Section 5.2:

(1) Matching on *X* alone should be avoided, as this procedure can be expected to lead to ''conditional independence'' synthetic estimates of Corr(*Y*, *Z*) (Corr(*X*, *Y*)* Corr(*X*, *Z*)).

(2) Matching with variables that are linear combinations of other variables (e.g., regression estimates) can be expected to introduce distortions into the synthetic estimates of Corr(*X*, *Z*) in File A and Corr(*X*, *Y*) in File B.

(3) Matching on (*Y*, *Z*) after adding residuals performed, overall, better than any of the other procedures that we examined. It always did a good job of accurately reproducing the other correlation estimates. For both theoretical and operational reasons, we recommend it to our fellow practitioners.

5.4. Additional research on proposed new procedure

We conducted additional research on the proposed new procedure of matching on (*Y*, *Z*) after adding residuals. Specifically, we were interested to see how the procedure would work for specified values of Corr(*X*, *Y*), Corr(*X*, *Z*), and Corr(*Y*, *Z*), and what would happen when an incorrect (but admissible) value of Corr(*Y*, *Z*) was postulated and used in the regression process.

In particular, we were curious to see if our recommended procedure could produce matched datasets with accurate synthetic estimates of Corr(X, Z) in File A and Corr(X, Y) in File B, while also reproducing the postulated (and incorrect) value of Corr(Y, Z). Based on extensive simulation results (Moriarity 2001), the answers to these questions all appear to be ‘‘yes.’’ This is to be expected, as per the following quote from Kadane’s article:

> ‘‘In the domain in which $\sum_{YZ}$ is such that the matrix
>
> $$\begin{pmatrix} \sum_{YY} & \sum_{YZ} \\ \sum_{ZY} & \sum_{ZZ} \end{pmatrix}$$
>
> is positive semidefinite, nothing is learned from the data about $\sum_{YZ}$. In Bayesian terms, whatever our prior on $\sum_{YZ}$ was, the posterior distribution will be the same.’’

Note that for the situation where an auxiliary source provides information about the value of Corr(Y, Z), the results discussed in this section suggest that our recommended procedure could produce matched datasets with accurate synthetic estimates of Corr(X, Z) in File A and Corr(X, Y) in File B, while also reproducing the value of Corr(Y, Z) obtained from the auxiliary source. That is, our recommended procedure might be an alternative to other procedures previously described for this situation (e.g., Singh et al. 1993; Paass 1986 and 1989).

6. Generalizations

We have outlined a procedure that can be used to explore variability due to alternative assumptions made during statistical matching. The procedure gave robust performance in our simulations, and we would expect robust performance from the procedure in other applications as well. We have provided explicit guidance for implementing the procedure for univariate (X, Y, Z), where the variables can be taken to be normally distributed (or approximately so).

As might be expected, the procedure is more complicated when the dimension of (X, Y, Z) exceeds three. However, one of the many strengths of Kadane’s procedure is that it is formulated in general terms, and provides a framework for the more general case. It should be noted that the specification of an admissible value of $\sum_{YZ}$ requires some effort.

One possible strategy is to begin with the ‘‘conditional independence value’’ $\sum_{YX}(\sum_{XX})^{-1}\sum_{XZ}$, which always is an admissible value for $\sum_{YZ}$. This provides a starting point for generating perturbations that would then need to be checked for admissibility.

For multivariate X and univariate (Y, Z), Rodgers and DeVol (1982) give a bound for Corr(Y, Z) that is a generalization of the bound given in Section 3.1. Let the dimension of X be P, and let $C^{i,j}$ denote the (i, j)th element in the inverse of the correlation matrix of X. Then Corr(Y, Z) must lie in the interval

$$C \pm \sqrt{D}$$

where

$$C = \sum_{i=1}^{p}\sum_{j=1}^{p} Corr(Y, X_i) * C^{i,j} * Corr(X_j, Z)$$

and

$$D = \left[1 - \sum_{i=1}^{p}\sum_{j=1}^{p} Corr(Y, X_i) * C^{i,j} * Corr(X_j, Y)\right]$$

$$* \left[1 - \sum_{i=1}^{p}\sum_{j=1}^{p} Corr(Z, X_i) * C^{i,j} * Corr(X_j, Z)\right]$$

This can be seen to reduce to the bound given in Section 3.1. if the dimension of X is 1.

In the more general case, the following recursion formula for partial correlations (e.g., Anderson 1984, p. 43) can be used to generate admissible values of $\sum_{YZ}$:

$$\rho_{ij\cdot q+1,\ldots,p} = \frac{\rho_{ij\cdot q+2,\ldots,p} - \rho_{i,q+1\cdot q+2,\ldots,p}\rho_{j,q+1\cdot q+2,\ldots,p}}{\sqrt{1-\rho^2_{i,q+1\cdot q+2,\ldots,p}}\sqrt{1-\rho^2_{j,q+1\cdot q+2,\ldots,p}}}$$

For example, in the simplest case of univariate (X, Y, Z), this formula gives

$$\rho_{YZ\cdot X} = \frac{\rho_{YZ} - \rho_{YX}\rho_{ZX}}{\sqrt{1-\rho^2_{YX}}\sqrt{1-\rho^2_{ZX}}}$$

from which the bound given in Section 3.1. can be derived by allowing $\rho_{YZ\cdot X}$ to vary from -1 to 1.

For (X, Y, Z_1, Z_2), a value for ρ_{YZ_1} could be specified in accordance with a bound analogous to what is given in Section 3.1.; i.e., in the interval,

$$\rho_{YX}\rho_{Z_1X} \pm \sqrt{1-\rho^2_{YX}}\sqrt{1-\rho^2_{Z_1X}}$$

This value of ρ_{YZ_1} then determines $\rho_{YZ_1\cdot X}$ via the equation

$$\rho_{YZ_1\cdot X} = \frac{\rho_{YZ_1} - \rho_{YX}\rho_{Z_1X}}{\sqrt{1-\rho^2_{YX}}\sqrt{1-\rho^2_{Z_1X}}}$$

The recursion formula then could be used to obtain the relation

$$\rho_{YZ_2\cdot X,Z_1} = \frac{\rho_{YZ_2\cdot X} - \rho_{YZ_1\cdot X}\rho_{Z_2Z_1\cdot X}}{\sqrt{1-\rho^2_{YZ_1\cdot X}}\sqrt{1-\rho^2_{Z_2Z_1\cdot X}}}$$

which specifies $\rho_{YZ_2\cdot X,Z_1}$ in terms of $\rho_{YZ_1\cdot X}$, $\rho_{Z_2Z_1\cdot X}$, and $\rho_{YZ_2\cdot X}$. $\rho_{Z_2Z_1\cdot X}$ can be estimated from File B, and then by allowing $\rho_{YZ_2\cdot X,Z_1}$ to vary from -1 to 1, the allowable range of values for $\rho_{YZ_2\cdot X}$ can be determined. Using these bounds, the allowable range of values for ρ_{YZ_2} can be determined that correspond to the given value of ρ_{YZ_1}.

7. Summary and Application Issues

Kadane's article presented pioneering work in the area of statistical matching. Clearly, Kadane was the first to propose a methodology to evaluate the variability of alternative assumptions in statistical matching; we are surprised that he has not gotten due recognition in the existing literature.

7.1. Conclusion

The discussion in this note has shown that most of Kadane's procedure rests on a firm theoretical basis. There is an important qualification, though. Due to the loss of the specified value of $\sum_{YZ}$ during the matching step, Kadane's procedure is not feasible as originally described. However, an innovation (adding residuals to the regression estimates prior to matching) makes the procedure feasible. The end result is a collection of datasets formed from various assumed values of $\sum_{YZ}$, where analyses can be repeated over the collection and the results can be averaged (or summarized in some other meaningful way) to assess the variability due to alternative assumptions about the value of $\sum_{YZ}$. (For a recent reference on ways for averaging the resulting values, see Hoeting et al. 1999. The methods described in Rubin (1987) might also be employed.)

7.2. Application

In this article we have made the assumption that the variables to be statistically matched come from multivariate normal distributions. This does not really fit most practice, where the variables come from complex survey designs and do not have a standard theoretical distribution, let alone a normal one.

While beyond our scope here to develop a complete paradigm, there are some suggestions that we would make:

(1) Constrained matching is a good starting point. It is expensive but affordable now with recent advances in computing. Unconstrained matching may also be used but this topic is taken up elsewhere (Moriarity and Scheuren 2000).

(2) Applications that match files as large as 1,000 (the sample size we simulated) would be unusual. Even in large-scale projects like matching the full Current Population Survey (CPS) with the Survey of Income and Program Participation (SIPP), the matching would be done separately in modest-sized demographic subsets defined by categorical variables such as sex, race, etc.

We do want to mention here that during the course of our research, we noticed that file sizes of 100 appeared to be too small (too much ''noise''), and file sizes of 500 appeared to be adequate. However, we did not conduct extensive research in this particular area.

(3) We believe that the general robustness of normal methods can be appealed to, even when the individual observations are not normal. While not necessarily optimal, the statistics calculated from the resulting combined file will be approximately normal because of the central limit theorem.

(4) Naturally, we recommend that Kadane's procedure, as modified here, be used. Whether one chooses to postulate a prior distribution for $\sum_{YZ}$ and then make random draws from that distribution, as Kadane suggested, or to specify values deterministically from the range of admissible values (our approach), distributional information will be created as a result. This can be used to check on robustness and should be.

(5) As samplers we are more worried about the assumption of independent identically distributed (IID) observations that exist in our modified method. With unequal

weights and clustering, common in many surveys (including the CPS and SIPP, for example), the matching results could still be unsafe.

(6) Resampling of the original sample, prior to employing the techniques in this article, could help here to expose potential lack of robustness to failures in the IID assumption. One such technique is found in the article by Hinkins, Oh, and Scheuren (1997), albeit it can be computationally expensive depending on the sample designs of the two files being matched.

(7) Often researchers that do statistical matching do not bring in the survey designers and at a minimum this is needed. The use of sample replication if possible, even approximately, is one way that designers can help matchers.

(8) Deep subject-matter knowledge is required too, in order to deal with differences across files in the measurement error and other nonsampling concerns (e.g., edit and imputation issues) that arise and which could even be dominant as a limitation to statistical matching.

7.3. *Final observation*

In this article we have reviewed and strengthened the theoretical underpinnings of one form of statistical matching, that proposed by Kadane. There has also been a brief discussion of applications issues. The application suggestions, however, are clearly not enough for a novice to begin a statistical matching exercise unaided. In all applications, no matter what the experience level of the matcher, caution would recommend that with a new problem, simulations should always be done and a small prototype involving real data should be conducted before beginning on a large scale. No decision on how or even whether to do a statistical match should be made until these steps have been taken.

Appendix

Formulas given by Kadane for the Φ_i, $i = 1$ to 6 (refer to Section 3.2.)

$$\Phi_1 = \sum_{ZX\cdot Y}\left(\sum_{XX\cdot Y}\right)^{-1}\sum_{XX} + \sum_{ZY\cdot X}\left(\sum_{YY\cdot X}\right)^{-1}\sum_{YX}$$

$$\Phi_2 = \sum_{ZX\cdot Y}\left(\sum_{XX\cdot Y}\right)^{-1}\sum_{XY} + \sum_{ZY\cdot X}\left(\sum_{YY\cdot X}\right)^{-1}\sum_{YY}$$

$$\begin{aligned}\Phi_3 = &\sum_{ZX\cdot Y}\left(\sum_{XX\cdot Y}\right)^{-1}\sum_{XX}\left(\sum_{XX\cdot Y}\right)^{-1}\sum_{XZ\cdot Y}\\ &-\sum_{ZY\cdot X}\left(\sum_{YY\cdot X}\right)^{-1}\sum_{YY}\left(\sum_{YY\cdot X}\right)^{-1}\sum_{YZ\cdot X}\\ &-\sum XZ\cdot Y\left(\sum_{XX\cdot Y}\right)^{-1}\sum_{XY}\left(\sum_{YY\cdot X}\right)^{-1}\sum_{YZ\cdot X}\\ &-\sum_{ZY\cdot X}\left(\sum_{YY\cdot X}\right)^{-1}\sum_{YX}\left(\sum_{XX\cdot Y}\right)^{-1}\sum_{XZ\cdot Y}\end{aligned}$$

$$\Phi_4 = \sum_{YX\cdot Z}\left(\sum_{XX\cdot Z}\right)^{-1}\sum_{XX} + \sum_{YZ\cdot X}\left(\sum_{ZZ\cdot X}\right)^{-1}\sum_{ZX}$$

$$\Phi_5' = \sum_{YX\cdot Z}\left(\sum_{XX\cdot Z}\right)^{-1}\sum_{XZ} + \sum_{YZ\cdot X}\left(\sum_{ZZ\cdot X}\right)^{-1}\sum_{ZZ}$$

and

$$\begin{aligned}\Phi_6 = {} & \sum_{YX\cdot Z}\left(\sum_{XX\cdot Z}\right)^{-1}\sum_{XX}\left(\sum_{XX\cdot Z}\right)^{-1}\sum_{XY\cdot Z} \\ & + \sum_{YZ\cdot X}\left(\sum_{ZZ\cdot X}\right)^{-1}\sum_{ZZ}\left(\sum_{ZZ\cdot X}\right)^{-1}\sum_{ZY\cdot X} \\ & + \sum_{YX\cdot Z}\left(\sum_{XX\cdot Z}\right)^{-1}\sum_{XZ}\left(\sum_{ZZ\cdot X}\right)^{-1}\sum_{ZY\cdot X} \\ & + \sum_{YZ\cdot X}\left(\sum_{ZZ\cdot X}\right)^{-1}\sum_{ZX}\left(\sum_{XX\cdot Z}\right)^{-1}\sum_{XY\cdot Z},\end{aligned}$$

where $\sum_{AB\cdot C} = \sum_{AB} - \sum_{AC}(\sum_{CC})^{-1}\sum_{CB}$.

8. References

Anderson, T.W. (1984). An Introduction to Multivariate Statistical Analysis. Second Edition. New York: John Wiley.

Barr, R.S. and Turner, J.S. (1978). A New, Linear Programming Approach to Microdata File Merging. 1978 Compendium of Tax Research, U.S. Department of the Treasury, 131–149.

Barr, R.S., Stewart, W.H., and Turner, J.S. (1982). An Empirical Evaluation of Statistical Matching Strategies. Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.

Bertsekas, D.P. (1991). Linear Network Optimization: Algorithms and Codes. Cambridge, Massachusetts: MIT Press.

Bertsekas, D.P. and Tseng, P. (1994). RELAX-IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems. Available on the Internet at http://web.mit.edu/dimitrib/www/home.html

Citro, C.F. and Hanushek, E.A. (eds.) (1991). Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling, Volume I: Review and Recommendations. Washington, D.C.: National Academy Press.

Cohen, M.L. (1991). Statistical Matching and Microsimulation Models. In Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling, Volume II: Technical Papers, Citro, C.F. and Hanushek, E.A. (eds.). Washington, D.C.: National Academy Press, 62–88.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183–1210.

Goel, P.K. and Ramalingam, T. (1989). The Matching Methodology: Some Statistical Properties. Springer-Verlag Lecture Notes in Statistics, Vol. 52. New York: Springer-Verlag.

Hinkins, S., Oh, H.L., and Scheuren, F. (1997). Inverse Sampling Design Algorithms. Survey Methodology, 23, 11–21.

Hoeting, J.A., Madigan, D., Raftery, A., and Volinsky, C.T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, 14, 382–417.

Ingram, D., O'Hare, J., Scheuren, F., and Turek, J. (2000). Statistical Matching: A New Validation Case Study. To appear in the 2000 Proceedings of the Survey Research Methods Section, American Statistical Association.

Kadane, J.B. (1978). Some Statistical Problems in Merging Data Files. 1978 Compendium of Tax Research, U.S. Department of the Treasury, 159–171. (Reprinted in Journal of Official Statistics, 17, 3, 423–433.)

Moriarity, C. (2001). Unpublished doctoral dissertation submitted to The George Washington University, Washington, D.C.

Moriarity, C. and Scheuren, F. (2000). A Note on Rubin's 'Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations'. Forthcoming.

National Research Council (1992). Combining Information: Statistical Issues and Opportunities for Research. Washington, D.C.: National Academy Press

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. (1986). Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information. In Microanalytic Simulation Models to Support Social and Fiscal Policy, Orcutt, G.H., Merz, J., and Quinke, H. (eds.). North-Holland, Amsterdam, 401–420.

Paass, G. (1989). Stochastic Generation of a Synthetic Sample from Marginal Information. Fifth Annual Research Conference Proceedings, U.S. Bureau of the Census, 431–445.

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. Department of Commerce. Washington, D.C.: Government Printing Office.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91–102.

Rodgers, W.L. and DeVol, E.B. (1982). An Evaluation of Statistical Matching. Report submitted to the Income Survey Development Program, Department of Health and Human Services, by the Institute for Social Research, University of Michigan.

Rubin, D.B. (1986). Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics, 4, 87–94.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Scheuren, F. (1989). A Comment on ''The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration. Survey of Current Business, 40–41.

Scheuren, F. and Winkler, W.E. (1993). Regression Analysis of Data Files that Are Computer Matched. Survey Methodology, 19, 39–58.

Scheuren, F. and Winkler, W.E. (1997). Regression Analysis of Data Files that Are Computer Matched – Part II. Survey Methodology, 23, 157–165.

Sims, C.A. (1972). Comments on ''Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File'', by B.A. Okner. Annals of Economic and Social Measurement, 1, 343–345.

Sims, C.A. (1978). Comments on ''Some Statistical Problems in Merging Data Files'', by J.B. Kadane. 1978 Compendium of Tax Research, U.S. Department of the Treasury, 172–177.

Singh, A.C., Mantel, H.J., Hinack, M.D., and Rowe, G. (1993). Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. Survey Methodology, 19, 59–79.

Received April 2000
Revised April 2001