Journal of Official Statistics, Vol. 17, No. 3, 2001, pp. 423–433

Some Statistical Problems in Merging Data Files[1]

Joseph B. Kadane[2]

Suppose that two files are given with some overlapping variables and some variables unique to each of the two files. Notationally, let X represent the common variables, Y, the variables unique to the first file, and Z, the variables unique to the second file. Thus the basic data consist of a sample of pairs (X, Y) and a sample of pairs (X, Z).

Merging of such microdata files may occur in two contexts. In the first, the files are known to consist of the same objects or persons, although their identities may be obscured by measurement errors in the common variables X. In the other case, the two files are random samples from the same population, but only accidentally will the same object or person be on both lists.

To want to merge data files in the first context is a very natural impulse. A merged file permits statements about (Y, Z) cross-classifications that are unavailable without merging. If the measurement errors in the variables X are low (for instance, if X includes accurate social security numbers), the merging can be very accurate, and the meaning of an item in a merged file is clear. It represents the (X, Y, Z) information on the object or person in question.

Merging data files in the second context requires greater caution. Again, facts are sought about (Y, Z) cross-classifications, but the items in the merged file have no natural meaning. The information on the Z variables for persons in the first file and on the Y variables for persons in the second file are missing. A mechanical method of merging can be seductive in this context because it will produce a file of records with X, Y, and Z entries inviting treatment as if they refer to the same persons. Yet it is clear that information cannot be created by the merging process where none existed before. Great care must be exercised in the second context.

One important method, reported by Okner (1972a), sets up ''equivalence classes'' of X's and makes a random assignment of an (X, Y) with an (X, Z) among ''equivalent'' (X, Z)'s that achieve a minimum closeness score. Sims (1972a, 1972b) stresses the need for a theory of matching and criticizes the Okner procedure for making the implicit assumption that Y and Z, given X, are independent. Peck (1972) defends the assumption, while Okner (1972b) discusses the validity of the assumption in various cases. Budd (1972) compares Okner's procedure to one then being used in the Commerce Department.

A second round of discussion – Okner (1974), Ruggles and Ruggles (1974), and Alter

[1] This article is a reprint from 1978 Compendium of Tax Research, U.S. Department of the Treasury, 159–171. Reprinting permission has been given by the author.

[2] Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213-3890, U.S.A. E-mail: kadane@stat.cmu.edu

(1974) – shows some improvements in method but a continuing concentration on equivalence classes. Sims (1974) again stresses his belief that the methods proposed will not perform well in sparse X-regions.

The first section of this report considers the case in which the lists are known to consist of the same objects or persons, and the second section takes up the case in which the lists are unrelated random samples from the same population. Although the final section, ''Why Match?'', is obviously speculative, that term really describes all of the work in this article.

Files Consist of the Same Objects or Persons

A statistical model

We assume that originally there were true triples (X_i, Y_i, Z_i) that had a normal distribution with means (μ_X, μ_Y, μ_Z) and some covariance matrix. These were broken into two samples, (X_i, Y_i) and (X_i, Z_i), and then independent normal measurement error $(\epsilon_i^1, \epsilon_i^2)$ was added. Let

$$X_i^1 = X_i + \epsilon_i^1$$

and

$$X_i^2 = X_i + \epsilon_i^2$$

where $(\epsilon_i^1, \epsilon_i^2)$ has a normal distribution with zero mean. Suppose, also, ϵ_i^1 has covariance matrix Ω_{11} and ϵ_i^2 has covariance matrix Ω_{22}, and that ϵ_i^1 and ϵ_i^2 have covariance matrix Ω_{12}. Then we observe a permutation of the paired observations (X_i^1, Y_i) and (X_i^2, Z_i).

There are two ways in which the assumed joint normality of X, Y, and Z is restrictive. First, some of our data is binary or integer-valued. Second, this implies that all the regressions are linear, which is not likely to be the case, as pointed out by Sims (1972a, 1972b, 1974). One way around that problem might be to assume joint normality region-by-region in the X space. This thought is not pursued further here.

Let $T_i = (X_i^1, Y_i)$ and $U_i = (X_i^2, Z_i)$ be vectors of length k and l respectively, where without loss of generality we take $k \leq l$. Also without loss of generality, take $\mu_X = 0$, $\mu_Y = 0$, $\mu_Z = 0$. The covariance matrix of T and U can be written as

$$\Sigma = \left[\begin{array}{cc:cc} \Sigma_{XX} + \Omega_{11} & \Sigma_{XY} & \Sigma_{XX} + \Omega_{12} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YZ} \\ \hdashline \Sigma_{XX} + \Omega_{12} & \Sigma_{XY} & \Sigma_{XX} + \Omega_{22} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZX} & \Sigma_{ZZ} \end{array}\right] = \left[\begin{array}{c:c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right]$$

Let

$$\Sigma^{-1} = \left[\begin{array}{c:c} C_{11} & C_{12} \\ \hdashline C_{21} & C_{22} \end{array}\right]$$

so that, in particular, we have

$$C_{12} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$$

Note that all these covariances can be estimated easily except Σ_{YZ} and $\Sigma_{XX} + \Omega_{12}$. Treatment of them is deferred.

Now suppose that $v_1, \ldots, v_n$ is the random permutation of $T_1, \ldots, T_n$ which is observed, and $w_1, \ldots, w_n$ is the random permutation of $U_1, \ldots, U_n$ which is observed. Let $\phi = [\phi(1), \ldots, \phi(n)]$ be a permutation of the integers $1, \ldots, n$.

According to DeGroot and Goel (1976), the likelihood function of ϕ is

$$L(\phi) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} v_i' C_{12} w_{\phi(i)}\right\}$$

Thus the maximum likelihood ϕ minimizes

$$C(\phi) = \sum_{i=1}^{n} v_i' C_{12} w_{\phi(i)}$$

Let

$$p_{ij} = v_i' C_{12} w_j$$

Then minimizing $C(\phi)$ is equivalent to minimizing

$$C = \Sigma p_{ij} a_{ij}$$

subject to the conditions

$$\Sigma_i a_{ij} = 1$$

$$\Sigma_j a_{ij} = 1$$

and

$$a_{ij} = 0 \text{ or } 1$$

which is a linear assignment problem (Degroot and Goel 1976).

There may be cases in which v_i and w_j occur several times in the files and consequently are recorded together. In general, suppose that v_i occurs q_i times $(i = 1, \ldots, n)$ and w_j occurs y_j times $(j = 1, \ldots, m)$, where we assume

$$\sum_{i=1}^{n} q_i = \sum_{j=1}^{m} w_j$$

Then a simple transformation of $C(\phi)$ yields the minimization of

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} a_{ij}$$

subject to the conditions

$$\Sigma_i a_{ij} = y_j \text{ for } j = 1, \ldots, m$$

$$\Sigma_j a_{ij} = q_i \text{ for } i = 1, \ldots, n$$

and

$a_{ij} =$ nonnegative integers.

This minimization is in the form of a transportation problem. The matrix C_{12} appears to be a natural choice of a distance function in this context.

Information about Σ_{YZ}

One of the difficulties of this method is that it requires knowledge of Σ_{YZ}. There are several possible sources of such information. First, from a coarse but perfectly matched sample, certain elements of Σ_{YZ} may be known. If so, surely this information should be used. Second, the assumption may be made, as is customary in the literature on matching, that Y and Z are conditionally independent given the X's. That is,

$$f(Y, Z \mid X^1, X^2) = f(Y \mid X^1, X^2) f(Z \mid X^1, X^2)$$

The covariance matrix of $(Y, Z \mid X^1, X^2)$ is (Anderson 1958, pp. 28–29)

$$\begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix} - \begin{bmatrix} \Sigma_{YX^1} & \Sigma_{YX^2} \\ \Sigma_{ZX^1} & \Sigma_{ZX^2} \end{bmatrix} \begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} \begin{bmatrix} \Sigma_{X^1Y} & \Sigma_{X^1Z} \\ \Sigma_{X^2Y} & \Sigma_{X^2Z} \end{bmatrix}$$

Conditional independence occurs iff the upper-right partitioned submatrix is zero, i.e., iff

$$\Sigma_{YZ} - (\Sigma_{YX^1} \quad \Sigma_{YX^2}) \begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{X^1Z} \\ \Sigma_{X^2Z} \end{pmatrix} = 0$$

Thus this assumption gives a condition that uniquely defines Σ_{YZ} in terms of the other Σ's. Some simplification of this answer is possible. Using

$$\Sigma_{YX^1} = \Sigma_{YX^2} = \Sigma_{YX} \text{ and } \Sigma_{ZX^1} = \Sigma_{ZX^2} = \Sigma_{ZX}$$

we have

$$\Sigma_{YZ} = (\Sigma_{YX} \quad \Sigma_{YX}) \begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{pmatrix}$$

Suppose, without loss of generality, that

$$\begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} = \begin{pmatrix} R & S \\ S' & V \end{pmatrix}$$

Then

$$\begin{aligned} \Sigma_{YZ} &= (\Sigma_{YX} \quad \Sigma_{YX}) \begin{pmatrix} R & S \\ S' & V \end{pmatrix} \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{pmatrix} \\ &= (\Sigma_{YX}R + \Sigma_{YX}S' \quad \Sigma_{YX}S + \Sigma_{YX}V) \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{pmatrix} \\ &= \Sigma_{YX}R\Sigma_{XZ} + \Sigma_{YX}S'\Sigma_{XZ} + \Sigma_{YX}S\Sigma_{XZ} + \Sigma_{YX}V\Sigma_{XZ} \\ &= \Sigma_{YX}(R + S' + S + V)\Sigma_{XZ} \end{aligned}$$

A well-known fact about inverses of partitioned matrices (Rao 1965, p. 29) is

$$\begin{bmatrix} A & B \\ B' & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + F\,E^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{bmatrix}$$

where

$$E = D - B'A^{-1}B \text{ and } F = A^{-1}B$$

Then

$$\begin{aligned} R + S' + S + V &= A^{-1} + FE^{-1}F' - FE^{-1} - E^{-1}F' + E^{-1} \\ &= A^{-1} + (I - F)E^{-1}(I - F)' \\ &= A^{-1} + (I - A^{-1}B)E^{-1}(I - B'A^{-1}) \\ &= A^{-1}(A + (A - B)(D - B'A^{-1}B)^{-1}(A - B'))A^{-1} \end{aligned}$$

Thus, in our case,

$$\begin{aligned} \Sigma_{YZ} = {} & \Sigma_{YX}\Sigma_{X^1X^1}^{-1}(\Sigma_{X^1X^1} + (\Sigma_{X^1X^1} - \Sigma_{X^1X^2})(\Sigma_{X^2X^2} - \Sigma_{X^2X^1}\Sigma_{X^1X^1}^{-1}\Sigma_{X^1X^2})^{-1} \\ & \cdot (\Sigma_{X^1X^1} - \Sigma_{X^2X^1})\Sigma_{X^1X^1}^{-1}\Sigma_{XZ} \end{aligned}$$

Thus Σ_{YZ} is given by this equation as a function of Σ_{YX}, Σ_{XZ}, $\Sigma_{X^1X^1}$, $\Sigma_{X^2X^2}$, and $\Sigma_{X^2X^1}$. All of these can be estimated directly except the last, $\Sigma_{X^2X^1} = \Sigma_{XX} + \Omega_{12}$

Estimation of $\Sigma_{X^2X^1} = \Sigma_{XX} + \Omega_{12}$

There are really two topics in this section. First I consider the elicitation of the measurement error process variance-covariance matrix Ω. Then I consider how to use that with other information to obtain $\Sigma_{X^2X^1}$.

In the elicitation of Ω, I must first emphasize what it is *not*. It does not refer to the levels of the common variables X. That is, we are dealing only with the spread in measured X's caused by the measurement process. Second, it does not refer to any systematic bias there may be in the measurement error process, but refers only to variability around what would be expected, taking into account both the level of the X variable and the measurement bias, if any.

Begin, then, with the diagonal elements of Ω, which are variances. Each variance refers to a specific measurement error variable, that is, to a specific X-variable and the associated source (one of the two). Choose any value for the true underlying X variable, for instance x. Write down what you think the measurement bias b is. (This must be independent of the value you gave for the X-variable, x. While this is not exactly the case, take for b a typical value). Not everyone with this true value x will have a measured value $x + b$. Write down the number y such that only 33.3 percent of such people will lie below y and 66.7 percent, above. Write down the number w such that 66.7 percent will lie below w and 33.3 percent, above. These numbers should line up so that $y < x + b < w$. There are now two measures for the standard deviation: 2.17 $(w - x - b)$ and 2.17 $(x + b - y)$. These values should be close. The variance is then the square of the standard deviation. This variance should not, according to the model, depend on x, so try it for a number of x's and hope that the results are close. If they are, take the median as the best value. If they are not, the model is not a good representation of reality.

Now we turn to the off-diagonal elements of Ω, which have to do with the relationship between two variables. Suppose that those variables are A and B. Then the work above defines for us the following: x_A, b_A, σ_A, w_A, and y_A, and similarly, x_B, b_B, σ_B, w_B and y_B. We now are trying to capture the extent to which A and B affect one another. The characteristic we focus on is the proportion p of times a measurement error on A is smaller than w_A and, simultaneously, a measurement error on B is smaller than w_B. If A and B have nothing to do with one another, this proportion would be $2/3 \times 2/3 = 4/9 = .44$, slightly under 50 percent. However, if A and B are related to one another, this proportion p may vary from .44. Write down the number you think is correct, and then convert it into a correlation between A and B using Table 1.

This yields a ρ_{AB} for each pair of variables A and B. The proper element for Ω is then the covariance of A and B, which is $\sigma_A \sigma_B \rho_{AB}$.

Not every matrix formed in this way is positive definite, as a covariance matrix must be. Hence, additional checks must be made to ensure that the covariance matrix is positive definite. One convenient way to achieve this is to augment Ω one row and column at a time, making use of the following simple fact:

If A is positive definite, then

$$\begin{pmatrix} A & b \\ b' & c \end{pmatrix}$$

is positive definite, iff $c - b'A^{-1}b > 0$. The proof is simple (see Kadane et al. 1977.)

In this way, every element of Ω can be elicited. Now the sample also has some information about Ω, which can be used as a check on the process. The variance-covariance matrix of X^1 is $\Sigma_{X^1X^1} = \Sigma_{XX} + \Omega_{11}$ and of X^2, $\Sigma_{X^2X^2} = \Sigma_{XX} + \Omega_{22}$. This gives two independent estimates for Σ_{XX}, namely $\Sigma_{X^2X^2} - \Omega_{22}$ and $\Sigma_{X^1X^1} - \Omega_{11}$. These should be very close. I suggest rechecking the work if they are not. If they are, then an estimate for Σ_{XX} is at hand. Finally we obtain $\Sigma_{X^2X^1} = \Sigma_{XX} + \Omega_{12}$, for we now have estimates of both of the latter.

Some Concluding Remarks About This Case

The case in which the files are known to consist of the same objects or persons is not well understood. Recently DeGroot and Goel (1975) obtained the astonishing result that such matched samples contain information about Σ_{YZ}. Their results suggest that there may not be a lot of information, and we do not know whether the amount of information in some relevant sense increases or decreases (or stays constant) with n. In particular, we do not know if a consistent estimate of Σ_{XZ} can be found in this case, although this writer's intuition is that it cannot.

Another case, one in which the lists may or may not contain the same individuals, is

Table 1. Relation between p and ρ

p	.33	.35	.37	.40	.42	.44	.46	.48	.50	.54	.59
o	−.9	−.7	−.5	−.3	−.1	.0	.1	.3	.5	.7	.9

called record linkage. A few important papers in record linkage have been written by DuBois (1969), Fellegi and Sunter (1969), Newcombe and Kennedy (1962), and Tepping (1968).

Matching When the Files Are Random Samples from the Same Population

We assume here that there were true triples (X_k, Y_k, Z_k) that had a normal distribution with means (μ_X, μ_Y, μ_Z) and some covariance matrix. Suppose that in some of these triples the Z coordinates were lost, yielding a sample (X_j, Y_j), $(j = 1, \ldots, m)$, and that for others the Y coordinates were lost, yielding a sample (X_i, Z_i), $(i = 1, \ldots, n)$. The parameters μ_X, μ_Y, μ_Z, Σ_{XX}, Σ_{XY}, Σ_{XZ}, Σ_{YY}, and Σ_{ZZ} can all be estimated consistently, and so we will take them as known. However, the covariance matrix of Y and Z, Σ_{YZ}, cannot be consistently estimated from such data.

In fact, in the domain in which Σ_{YZ} is such that the matrix

$$\begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}$$

is positive semidefinite, nothing is learned from the data about Σ_{YZ}. In Bayesian terms, whatever our prior on Σ_{YZ} was, the posterior distribution will be the same (see Kadane, 1975 for other examples of this).

Hence we cannot hope to make realistic progress on this problem without a prior probability distribution on Σ_{YZ}. Our intention is to trace through the analysis using a particular value for Σ_{YZ}, for the purpose of obtaining results that would ultimately yield the expected value of some quantity – for instance, the expected amount of taxes a particular kind of tax schedule would raise. The taxes raised would then be a random variable, where the uncertainty would arise from the uncertainty about Σ_{YZ}. Hence we may assume that the distribution of Σ_{YZ} is known, and we may take values of Σ_{YZ} from the distribution, weighting the final results with the probability of that particular value of Σ_{YZ}. We proceed, then, with a value for Σ_{YZ} sampled in this way.

A natural first thing to do is to estimate the missing values, and the obvious way to do that is by the conditional expectation:

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

Let

$$\Sigma_{RS \cdot T} = \Sigma_{RS} - \Sigma_{RT} \Sigma_{TT}^{-1} \Sigma_{TS}$$

for any matrices R, S, and T.

Then

$$\begin{aligned} E(Z_j \mid X_j, Y_j) &= \mu_Z + (\Sigma_{ZX} \quad \Sigma_{ZY}) \begin{pmatrix} \Sigma_{XX \cdot Y}^{-1} & -\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY \cdot X}^{-1} \\ -\Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX \cdot Y}^{-1} & \Sigma_{YY \cdot X}^{-1} \end{pmatrix} \begin{pmatrix} X_j - \mu_X \\ Y_j - \mu_Y \end{pmatrix} \\ &= \mu_Z + \Sigma_{ZX \cdot Y} \Sigma_{XX \cdot Y}^{-1} (X_j - \mu_X) + \Sigma_{ZY \cdot X} \Sigma_{YY \cdot X}^{-1} (Y_j - \mu_Y) \end{aligned}$$

Similarly, we may predict missing Y_i with its conditional expectation

$$E(Y_i \mid X_i, Z_i) = \mu_Y + \Sigma_{YX \cdot Z}\Sigma^{-1}_{XX \cdot Z}(X_i - \mu_X) + \Sigma_{YZ \cdot X}\Sigma^{-1}_{ZZ \cdot X}(Z_i - \mu_Z)$$

Then the joint distribution of $(X_j, Y_j, \hat{Z}_j)$ is normal with mean vector (μ_X, μ_Y, μ_Z) and covariance matrix

$$S_1 = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} & T_1' \\ \Sigma_{YX} & \Sigma_{YY} & T_2' \\ T_1 & T_2 & T_3 \end{bmatrix}$$

where

$$T_1 = \Sigma_{ZX \cdot Y}\Sigma^{-1}_{XX \cdot Y}\Sigma_{XX} + \Sigma_{ZY \cdot X}\Sigma^{-1}_{YY \cdot X}\Sigma_{YX}$$

$$T_2 = \Sigma_{ZX \cdot Y}\Sigma^{-1}_{XX \cdot Y}\Sigma_{XY} + \Sigma_{ZY \cdot X}\Sigma^{-1}_{YY \cdot X}\Sigma_{YY}$$

and

$$\begin{aligned} T_3 = {} & \Sigma_{ZX \cdot Y}\Sigma^{-1}_{XX \cdot Y}\Sigma_{XX}\Sigma^{-1}_{XX \cdot Y}\Sigma_{XZ \cdot Y} + \Sigma_{ZY \cdot X}\Sigma^{-1}_{YY \cdot X}\Sigma_{YY}\Sigma^{-1}_{YY \cdot X}\Sigma_{YZ \cdot X} \\ & + \Sigma_{ZX \cdot Y}\Sigma^{-1}_{XX \cdot Y}\Sigma_{XY}\Sigma^{-1}_{YY \cdot X}\Sigma_{YZ \cdot X} + \Sigma_{ZY \cdot X}\Sigma^{-1}_{YY \cdot X}\Sigma_{YX}\Sigma^{-1}_{XX \cdot Y}\Sigma_{XZ \cdot Y} \end{aligned}$$

This is a singular distribution, of course, since $\hat{Z}_j$ is a linear function of X_j and Y_j.

Similarly, the joint distribution of $(X_i, \hat{Y}_i, Z_i)$ is normal with mean vector (μ_X, μ_Y, μ_Z) and covariance matrix

$$S_2 = \begin{bmatrix} \Sigma_{XX} & T_4' & \Sigma_{XZ} \\ T_4 & T_6 & T_5' \\ \Sigma_{ZX} & T_5 & \Sigma_{ZZ} \end{bmatrix}$$

where

$$T_4 = \Sigma_{YX \cdot Z}\Sigma^{-1}_{XX \cdot Z}\Sigma_{XX} + \Sigma_{YZ \cdot X}\Sigma^{-1}_{ZZ \cdot X}\Sigma_{ZX}$$

$$T_5 = \Sigma_{YX \cdot X}\Sigma^{-1}_{XX \cdot Z}\Sigma_{XZ} + \Sigma_{YZ \cdot X}\Sigma^{-1}_{ZZ \cdot X}\Sigma_{ZZ}$$

and

$$\begin{aligned} T_6 = {} & \Sigma_{YX \cdot Z}\Sigma^{-1}_{XX \cdot Z}\Sigma_{XX}\Sigma^{-1}_{XX \cdot Z}\Sigma_{XY \cdot Z} + \Sigma_{YZ \cdot X}\Sigma^{-1}_{ZZ \cdot X}\Sigma_{ZZ}\Sigma^{-1}_{ZZ \cdot X}\Sigma_{ZY \cdot X} \\ & + \Sigma_{YX \cdot Z}\Sigma^{-1}_{XX, Z}\Sigma_{XZ}\Sigma^{-1}_{ZZ \cdot X}\Sigma_{ZY \cdot X} + \Sigma_{YZ \cdot X}\Sigma^{-1}_{ZZ \cdot X}\Sigma_{ZX}\Sigma^{-1}_{XX \cdot Z}\Sigma_{XY \cdot Z} \end{aligned}$$

which again is a singular distribution. Now a natural impulse is to pool these two samples $w_j = (x_j, y_j, \hat{z}_j)$, $(j = 1, \ldots, m)$ and $v_i = (x_i, \hat{y}_i, z_i)$, $(i = 1, \ldots, n)$. However the covariance matrices S_1 and S_2 are not the same, and all such data would lie on two hyperplanes in (X, Y, Z) space. Another impulse is to match the data. Suppose now that $m = n$, so that simple matching has some hope of making sense.

Observe that $w_j - v_i$ has a normal distribution with mean of zero and covariance matrix $S_1 + S_2$, which is nonsingular.

Hence, using the Mahalanobis distance, we may define the distance from w_j to v_i to be d_{ij}, where

$$d_{ij} = (w_j - v_i)'(S_1 + S_2)^{-1}(w_j - v_i)$$

Thus a match would minimize

$$C' = \Sigma_{i,j} d_{ij} a_{ij}$$

over choices of a_{ij} subject to the conditions

$$\Sigma_i a_{ij} = 1$$

$$\Sigma_j a_{ij} = 1$$

and

$$a_{ij} = 0 \text{ or } 1$$

which again is a linear assignment problem. In the case in which the observations have weights, we relax the condition $n = m$ and suppose v_i has weight q_i $(i = 1, \dots, n)$ and w_j has weight y_j $(j = 1, \dots, m)$. The condition $n = m$ is replaced by the condition $\Sigma q_i = \Sigma y_j$. Then the natural generalization is to minimize

$$\Sigma_{i,j} d_{ij} a_{ij}$$

over choices of a_{ij} subject to the conditions

$$\Sigma_i a_{ij} = y_j$$

$$\Sigma_j a_{ij} = q_i$$

and

$$a_{ij} \geq 0$$

which is a transportation problem.

An interesting alternative to the matrix $S_1 + S_2$ to use in the Mahalanobis distance is the matrix

$$\begin{bmatrix} \Sigma_{XX} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

This alternative avoids ‘‘bias’’ that might be introduced by paired Y_j and Z_i, at the cost of not using some of the available information. I regard the relative benefits of these two methods as an open question.

Once the merging is complete, suppose – with slight abuse of notation – that w_j and v_i have been matched. Then it might be natural to take (x_j, y_j, z_i) and (x_i, y_j, y_i) as simulations of the underlying distributions.

Now the expected taxes can be computed. Again I stress that this is conditional on a value of Σ_{YZ}. Many such matchings and averagings should be done, to explore the sensitivity of the results to Σ_{YZ}.

Another aspect of this problem that is not well understood is the relation of matching to the prior reduction of the files (Turner and Gilliam 1975). Perhaps the two processes can be combined into one, or mutually rationalized.

Why Match?

At first, matching seems to be a peculiar way to treat data. If Σ_{YZ} were known in either framework, the complete joint distribution of the data would be consistently estimated, and any devised probabilities or expectations could in principle be calculated from that estimated jointly normal distribution or, if necessary, simulated on a computer. This approach is less than satisfactory because the variables are in truth not normally distributed. Hence we use the matched sample as if it were a sample from the true distribution and estimate, for instance, the expected value of some tax variable as if by simulation. The normality assumption is used to derive the matching methodology but need not be relied on for the rest of the estimation.

The soundness of this approach is very difficult to assess, and that question will not be settled in this article. It is clear that a matched sample cannot be treated uncritically as if it were a joint sample that had never been split nor had missing values. Thus the question is not the quality of the match itself, but rather the correct use and interpretation of statistics derived from the matched sample. Our understanding of this question is in its infancy.

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). Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons.

Budd, E.C. (1972). Comments. Annals of Economic and Social Measurement, 1, 349–354.

DeGroot, M.H. and Goel, P. (1975). Estimation of the Correlation Coefficient from a Broken Random Sample: Technical Report No. 105. Pittsburgh: Carnegie-Mellon University, Department of Statistics (Mimeo).

DeGroot, M.H. and Goel, P. (1976). The Matching Problem for Multivariate Normal Data. Sankyā, 38 (Series B, Part 1), 14–28.

DuBois, N.S. D'Andrea. (1969). A Solution to The Problem of Linking Multivariate Documents. Journal of the American Statistical Association, 69, 163–174.

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

Kadane, J.B. (1975). The Role of Identification in Bayesian Theory. In Studies in Bayesian Econometrics and Statistics, eds. S.E. Fienberg and A. Zellner, Amsterdam: North Holland Publishing Co., 175–191.

Kadane, J.B., Dickey, J.M., Winkler, R.L., Smith, W.S., and Peters, S.C. (1977). Interactive Elicitation of Opinion for a Normal Linear Model. Pittsburgh: Carnegie-Mellon University, June 8 (unpublished fifth draft).

National Bureau of Standards. (1959). Tables of the Bivariate Normal Distribution Function and Related Functions (Applied Mathematics Series, No. 50.

Newcombe, H.B. and Kennedy, J.M. (1962). Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. Communications of the Association for Computing Machinery, 5, 563–566.

Okner, B. (1972a). Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File. Annals of Economic and Social Measurement, 1, 325–342.

Okner, B. (1972b). Reply and Comments. Annals of Economic and Social Measurement, 1, 359–362.

Okner, B. (1974). Data Matching and Merging: An Overview. Annals of Economic and Social Measurement, 3, 347–352.

Peck, J.K. (1972). Comments. Annals of Economic and Social Measurement, 1, 347–348.

Rao, C.R. (1965). Linear Statistical Inference and Its Applications (1st ed.). New York: John Wiley and Sons.

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

Sims, C.A. (1972a). Comments. Annals of Economic and Social Measurement, 1, 343–345.

Sims, C.A. (1972b). Rejoinder. Annals of Economic and Social Measurement, 1, 355–357.

Sims, C.A. (1974). Comment. Annals of Economic and Social Measurement, 3, 395–397.

Tepping, B.J. (1968). A Model for Optimum Linkage of Records. Journal of the American Statistical Association, 63, 1321–1332.

Turner, J.S. and Gilliam, G.B. (1975). A Network Model to Reduce the Size of Microdata Files. Paper presented to ORSA Conference, Las Vegas, 1975. (Mimeo).