Office of Tax Analysis
U.S. Treasury Department
Washington, D.C. 20220
Issued: July, 1976

Statistical Problems of Merged Data Files

Joseph B. Kadane

OTA Paper 6 December 12, 1975

TABLE OF CONTENTS

		page
I.	Introduction	1
II.	A Statistical Model	3
	A. Assumptions	3
	B. Complications	3
	C. The Model and Its Application to the Transportation Problem	4
III.	Assumption of Conditional Independence	7
IV.	Interpretation of Matched Sampling	10
V.	Unresolved Questions About Matching	11

I. INTRODUCTION

Merging micro data files is a common occurrence. There are two aspects to the statistics of merged files that deserve attention: the proper procedure of merging, and the correct interpretation of the results. This report will concentrate on the first question, although some thoughts on the second are offered in Section IV.

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 consists of a sample of pairs (X,Y) and a sample of pairs (X,Z). (Later the possibility of weighted samples, important to the Treasury application, will be considered. Weighted sampling does not radically complicate the analysis.)

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 which achieve a minimum closeness score. Sims, in his comment [1972a] and rejoinder [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.

In a second round of discussion, Okner [1974], Ruggles and Ruggles [1974], Alter [1974] and Budd [1974] improve the method, but continue to

concentrate on equivalence classes. Sims' [1974] comment again stresses his belief that the methods proposed will not perform well in sparse X-regions.

Section II of this report gives a model for matching and derives the maximum likelihood match. This leads to a distance function as sought by the Treasury Department. Section III gives some thoughts about Σ_{YZ} and conditional independence. Section IV discusses the question of interpretation of calculations from a matched sample, and Section V concludes with some unanswered questions.

II. A STATISTICAL MODEL

A. Assumptions. 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 Σ. These were broken into two samples, (X_i, Y_i) and (X_i, Z_i), and then independent normal measurement errors (ξ_i) were added.

Let

$$X_i^1 = X_i + \xi_i^1$$

$$X_i^2 = X_i + \xi_i^2$$

where (ξ_i^1, ξ_i^2) each has a normal distribution with zero mean. Suppose also that ξ_i has covariance matrix Ω_1 and ξ_i^2 has covariance matrix Ω_2, and that ξ_i^1 and ξ_i^2 are independent for all i. We may then observe a permutation of the paired observations (X_i^1, Y_i) and (X_i^2, Z_i).

B. Complications. There are two aspects of unrealism in this model. First, we assume that the two samples represent the same individuals, and the X values have been distorted only by some measurement errors. We know that in general, this is not the case. Nonetheless, we use this assumption not only because it gives a reasonable answer, but also because it is the only assumption available. Second, we assume joint normality of X, Y, and Z. This is untrue of our data in at least two important respects. First, some of our data is binary or integer-valued. Second,

joint normality implies that all the regressions are linear, which is not likely to be the case, as pointed out by Sims [1972a,b, 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.

C. The Model and Its Application to the Transportation Problem. Let $T_i = (X_i^1, Y_i)$ and $U_i = (X_i^2, Z_i)$ be vectors of length k and 1 respectively, where without loss of generality we take $k \leq 1$. 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

$$\Sigma = \left[\begin{array}{cc:cc} \Sigma_{XX}+\Omega_1 & \Sigma_{XY} & \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YZ} \\ \hdashline \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XX}+\Omega_2 & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZX} & \Sigma_{ZZ} \end{array}\right] = \left[\begin{array}{c:c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right]$$

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 covariances above can be estimated easily except Σ_{YZ}. Treatment of Σ_{YZ} is deferred to Section III.

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 [1975a], the likelihood function of ϕ is

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

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

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

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

where $a_{ij} = 0$ or 1.

This is a linear assignment problem (DeGroot and Goel, 1975a).

In the case of observations with weights, suppose v_i has weight $x_i (i=1,\ldots,n)$ and w_j has weight $y_j (j=1,\ldots m)$ where we assume

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{m} y_i.$$

Then the natural generalization is to minimize

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

subject to the conditions

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

$$\sum_j a_{ij} = x_i \text{ for } i=1,\ldots,n.$$

This is a transportation problem, which happens to be what the Treasury Department is now programming. Thus the matrix C_{12} above is my answer to the question of what distance function to use in the transportation problem.

III. ASSUMPTION OF CONDITIONAL INDEPENDENCE

One of the difficulties of the preceding 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 matching literature, that Y and Z are conditionally independent, given the X's. That is

$$f(Y,Z|X^1,X^2) = f(Y|X^1,X^2)f(Z|X^1,X^2).$$

The covariance matrix of $(Y,Z|X^1,X^2)$ is (see Anderson pp. 28, 29)

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

Conditional independence occurs if and only if the upper-right partitioned submatrix is zero, i.e., if and only if

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

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

Using

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

we have

$$\Sigma_{YZ} = (\Sigma_{YX} \; \Sigma_{YX}) \begin{bmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{bmatrix} .$$

Suppose, without loss of generality, that

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

Then

$$\Sigma_{YZ} = (\Sigma_{YX} \; \Sigma_{YX}) \begin{bmatrix} R & S \\ S' & V \end{bmatrix} \begin{bmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{bmatrix}$$

$$= (\Sigma_{YX}R+\Sigma_{YX}S' \quad \Sigma_{YX}S+\Sigma_{YX}V) \begin{bmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{bmatrix}$$

$$= \Sigma_{YX}R\Sigma_{XZ} + \Sigma_{YX}S'\Sigma_{XZ} + \Sigma_{YX}S\Sigma_{XZ} + \Sigma_{YX}V\Sigma_{XZ}$$

$$= \Sigma_{YX}(R + S' + S + V)\Sigma_{XZ} .$$

A well-known fact about inverses of partitioned matrices (see 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 $E = D - B'A^{-1}B$ and $F = A^{-1}B$.

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

Hence in our case,

$$\Sigma_{YZ} = \Sigma_{YX}\Sigma^{-1}_{X^1X^1}(\Sigma_{X^1X^1} + (\Sigma_{X^1X^1} - \Sigma_{X^1X^2})(\Sigma_{X^2X^2} - \Sigma_{X^2X^1}\Sigma^{-1}_{X^1X^1}\Sigma_{X^1X^2})^{-1}$$
$$(\Sigma_{X^1X^1} - \Sigma_{X^2X^1}))\,\Sigma^{-1}_{X^1X^1}\Sigma_{XZ}$$

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 directly estimated except the last, $\Sigma_{X^2X^1}$.

One way to obtain an estimate for $\Sigma_{X^2X^1}$ is to formulate an opinion about the covariance matrix of the error process, that is, to take Ω_1 and Ω_2 as known. Knowing one implies an estimate for the other (under the assumption of independence between ξ^1_i and ξ^2_i), so this can be used as a check on the procedure.

IV. INTERPRETATION OF MATCHED SAMPLING

If Σ_{YZ} were known, the matching procedure of Section II could be used without further assumption. And yet it would not be necessary to do matching at all, since then the joint distribution of (X^1,Y,X^2,Z) would be known. Any probability desired could in principle be calculated from this distribution, or, if necessary, simulated directly. Therefore, the value of matching, at least in this jointly normal world, depends on a situation in which Σ_{YZ} is not known. In this case, some assumption must be made about Σ_{YZ} in order to use the procedure of Section II.

It is my judgment that this line of criticism is not as damaging as it might first appear. If Y is well predicted by X^1 and Z is well predicted by X^2, and if X^1 and X^2 are close, the conditional independence assumption is not bad because the conditional variability will be low. So while covariances might be well estimated, correlations might be poorly estimated. Yet recent work of DeGroot and Goel [1975b] suggests that even some information about the correlation can be squeezed out of a matched sample, although not very much. What is clear is that a matched sample cannot be treated uncritically as though it were a joint sample that had never been split and reunited. Thus the right 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.

V. UNRESOLVED QUESTIONS ABOUT MATCHING

The foregoing discussion raises a number of unresolved questions. The last three questions listed below concern the relation of merging to file reduction (Turner and Gilliam, 1975).

(i) Is there a simple form for C_{12}? Under what conditions is it true that

$$C_{12} = \begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}$$

for some W? (In this case C_{12} would be a distance matrix just on X^1 and X^2.)

(ii) The discussion of Section IV indicates that there might be a theorem characterizing those functions of the parameters that have consistent estimates. Possibly the conditions specified in Section IV are necessary for certain functions to be consistently estimated.

(iii) Can a method be found for estimating X, the common value of X^1 and X^2? When the time comes to use the matched sample, some values must be taken for the X variables.

(iv) How should the theory be extended to deal with non-normal variables? This question is particularly important for binary variables.

(v) Can a more realistic model for matching be found, one which does not assume the same people in both populations? What is the meaning of the estimates derived from such a model?

(vi) Can a similar distance function be found for file reduction?

(vii) Can the common variables to which two or more files are reduced be estimated, perhaps similarly to (iii) above?

(viii) Can a common model, theory, and procedure be found for simultaneous file reduction and merging?

REFERENCES

Alter, H. E., "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, 1974, pp. 373-394.

Anderson, T. W., Introduction to Multivariate Statistical Analysis, J. Wiley and Sons, New York, 1958.

Budd, E. C., "Comments", Annals of Economic and Social Measurement 1, 1972, pp. 349-354.

DeGroot, M. and P. Goel, "The Matching Problem for Multivariate Normal Data", Technical Report No. 94, Department of Statistics, Carnegie-Mellon University mimeo, 1975a.

DeGroot, M. and P. Goel, "Estimation of the Correlation Coefficient from a Broken Random Sample", Technical Report No. 105, Department of Statistics, Carnegie-Mellon University mimeo, 1975b.

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

Okner, B., "Reply and Comments", Annals of Economic and Social Measurement 1, 1972b, pp. 359-362.

Okner, B., "Data Matching and Merging: An Overview", Annals of Economic and Social Measurement 3, 1974, pp. 347-352.

Peck, J. K., "Comments", Annals of Economic and Social Measurement 1, 1972, pp. 347-348.

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

Ruggles, N. and R. Ruggles, "A Strategy for Merging and Matching Microdata Sets", Annals of Economic and Social Measurement 3, pp. 353-371.

Sims, C., "Comments", Annals of Economic and Social Measurement 1, 1972a, pp. 343-345.

Sims, C. A., "Rejoinder", Annals of Economic and Social Measurement 1, 1972b, pp. 355-357.

Sims, C., "Comment", Annals of Economic and Social Measurement 3, 1974, pp. 395-397.

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