Technical Report 88-OR-15

QUALITY ISSUES AND EVIDENCE IN STATISTICAL FILE MERGING

Richard S. Barr[1]
J. Scott Turner[2]

Prepared for the
Survey Data Quality Control Workshop
Oak Ridge National Laboratory
April 21-22, 1988

[1]Department of Operations Research and Engineering Management
Southern Methodist University
Dallas, Texas 75275

[2]College of Business Administration
Oklahoma State University
Stillwater, Oklahoma 74074

TABLE OF CONTENTS

ABSTRACT

In developing microsimulation models or research databases, it is common to discover that the desired data is not available from a single source. In such cases, practitioners can merge a pair of sample survey files to form a composite microdata file by linking record pairs. Statistical merging is a widely–used class of techniques to link records of sample units which have similar data attributes, but are not necessarily the same person or household.

This paper describes a computational study undertaken to investigate empirically the impacts of merging scheme, distance function, and data measurement error on the statistical characteristics of the resultant merge file. Merges of national data sets were performed to test the procedures' ability (or lack thereof) to create composite files which replicate an actual sample drawn from the original population.

The results indicate specific instances where merging works well and other cases in which it does not. The optimal–constrained merge technique with an absolute difference distance function appears to be the best of the methodologies in current use. Other distance functions proposed in the literature yielded extremely poor matches when applied to sample survey data. The robustness of merge techniques when bias and noise are present is clearly demonstrated as is the need for a reasonable number of variables in the distance function.

In addition, the need for modifications to existing merge procedures which address their shortcomings is discussed and easily–implementable improvements described.

ACKNOWLEDGMENTS

We wish to acknowledge the strong support and cooperation of the following organizations: the Office of Evaluation and Technical Analysis, Department of Health and Human Services; the Office of Research and Statistics, Social Security Administration; the Internal Revenue Service and Office of Tax Analysis, Department of the Treasury; and our home universities. This research was also supported by the Department of Defense under Contract Number MDA 903-86-C0182, the Air Force Office of Scientific Research under Contract Numbers AFOSR 83-0278 and AFOSR 87-0199, the Office of Naval Research under Contract Number N00014-87-K-0223, and ROME Air Development Center under Contract Number SCEEE PDP/86-75.

We especially thank Nelson McClung, Fritz Scheuren, Al Schwartz, Joan Turek-Brezina, and David Zalkind for lending their resources, expertise, and encouragement to this project. Their many contributions and suggestions greatly enhanced the research and this report.

PART 1. BACKGROUND AND OVERVIEW

The concept of microanalytic simulation models was developed by Guy Orcutt in the mid–1950's [24]. Today, these models abound in governmental agencies and research organizations and are used widely for policy analysis and projection of program needs. Examples include the various versions of the Transfer Income Model (TRIM), the dynamic, demographic simulation model DYNASIM, and the tax policy simulations at the U.S. Department of the Treasury, Brooking Institute and Statistics Canada.

At the heart of these models are sample survey files, or microdata. These files consist of data records for a representative set of decision units (individuals, households, taxpayers, firms, etc.) which are processed by the simulator individually with data collected to identify aggregates, distributions, and interactions. By working at the record level, this modeling technique is very flexible and can accommodate as much detail as desired.

Microdata matching is useful when the data required for a model is located in two or more microdata sets, as is often the case with governmental tax– or welfare–system models. Means for performing such matchings are the subject of this study. At present, all known applications for microdata matching are in the public sector, primarily at the national level. However, these same methodologies could be used in business applications such as the construction of marketing research databases from a set of separately–compiled samples.

1.1. Microdata Files

While the recording unit may vary, microdata files usually represent the national population or a major subset such as taxpayers or Social Security system participants. Various sampling schemes are used in collecting the data, hence each record includes a weight indicating the number of population units it represents. These weights often differ among records in a given file.

Microdata files are created as byproducts of ongoing governmental programs, from legislative mandate, or as special commissioned studies. For example, both I.R.S.'s Statistics of Income (SOI) and Social Security Earnings (SSA) files are drawn from data collected in the process of program implementation and control. The U.S. Constitution mandates the taking of a decennial census, subsets of which are used as microdata, and the Current Population Survey (CPS) is performed monthly to determine the unemployment rate, as required by law. The Survey of Income and Education (SIE) was a special study, as are numerous university–based surveys.

For the model designer and user, there are several pertinent characteristics of microdata files. First, they are expensive to create, on the order of $10 millions each. Hence, their construction is not a trivial undertaking. Second, several versions are often created through editing procedures to "correct" the data for underreporting, sample bias, etc. Third, a variety of sampling designs may be used, including stratified, clustered, and simple randomized, in order to combine information richness with brevity.

Fourth, the end product of these sometimes elaborate machinations is a multi–attribute representation of the underlying population, including all interactions and distributions of the reported data items. The distributional and interaction details are especially important for microanalytic models since they operate at the record level and base their computations on combinations of item values. Finally, by virtue of taking a middle ground between a census and population aggregates, these files are efficient from both a computational and information–content standpoint.

1.2. Limitations of Individual Samples

As illustrated by files such as the SOI and CPS, microdata are often collected primarily for the construction of aggregates or for program implementation, analysis, and control. Their use as general research data bases or in microanalytic simulation models is of secondary concern in the sample survey designs, an aspect which creates problems for these applications.

As models are built and policy proposals are analyzed, data are often required which (a) are not part of the current program, study, or system, as when new tax deductions are considered, or (b) are of superior quality since sample survey items are deemed to be unreliable, such as business income on the CPS.

The model user has four choices available: (1) commission a new study, at great expense and investment of time, (2) ignore the variables in question, and jeopardize the validity of the model's results, (3) impute the missing or unreliable items into an existing file, using methods which often ignore the distributional and interaction characteristics of the variables in question, or (4) merge a pair of microdata files to combine the information from two surveys. This last option, file merging, is currently in widespread use and is investigated in this paper.

1.3. Microdata File Merging

The basic idea behind file merging, or matching, is to combine one file A with another file B to form a composite file C with all data items from the two original files. This is accomplished by selecting pairs of records to match based on data items which are common to both files. The schemes for performing the matching process fall into two general categories: exact and statistical matching.

Exact matching uses unique–valued common items to mate records for the same individual in both files. By using a unique identifier, such as a social security number, the matching process is theoretically a simple sort and merge operation. Problems with this approach include: insignificant overlapping of samples causing few records to be matched, absence of or error in the "unique" identifiers, confidentiality restrictions which preclude legal linking of records, and the expense of handling a large number of exceptions.

Statistical merging (also called synthetic, stochastic, or attribute matching or merging) mates *similar* records using several common items with non–unique values. By matching like records, file C contains records which may be composites of two different persons, but whose attributes are similar enough for research purposes. There are a variety of statistical merging schemes in use today, as discussed below.

In choosing a methodology, exact matching is obviously preferable. But where such a match is not possible, statistical merging is often employed.

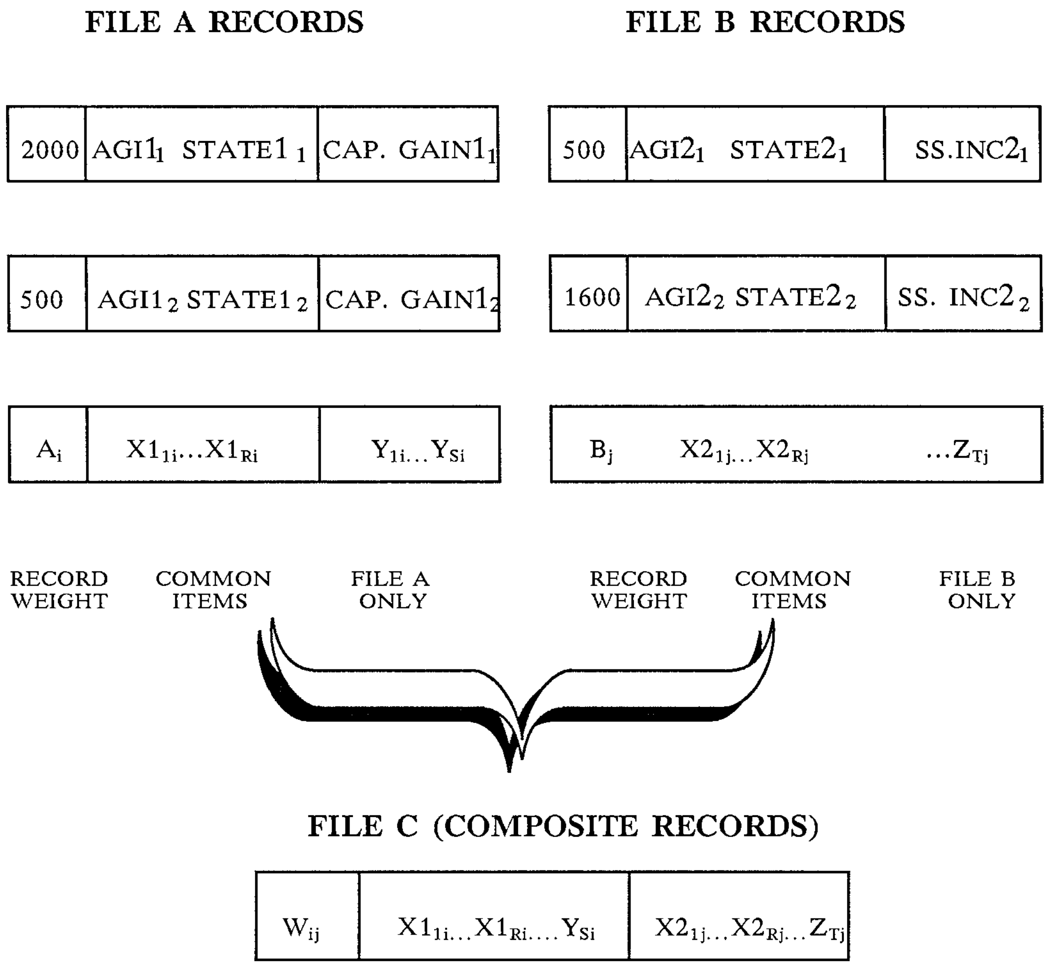
1.4. Statistical Merging

A pictorial description of statistical merging is presented in Figure 1.1. In this drawing, a_i represents the weight of the i–th record in file A and b_j the weights of the j–th record in file B. The merged file, C, contains composite records formed by matching a record in file A with a record in file B, and assigning a merge record weight of w_{ij} . An interrecord dissimilarity measure d_{ij} , or distance function, is used to choose matched record pairs. The "distance" between a pair of records is usually determined from a user–defined function which compares corresponding common items and assigns a penalty value for each item pair which differs significantly. These penalties are summed to create a measure of dissimilarity, with a zero distance meaning all common items are identical or "close enough."

There are two general categories of statistical merges: Unconstrained and constrained. In an unconstrained merge, file A is designated the base file and file B the augmentation file. Each base file record is matched with the most similar record in the augmentation file; the selected file B record is appended to the base file record and the base record's weight is used for w_{ij} . This is, in essence, sampling with replacement since some augmentation file records may not be matched while others may be used repeatedly. This is a very popular technique as evidenced by its use by Ruggles and Ruggles of Yale and NBER [41], Radner of the Social Security Administration [37], Okner and Minarik at Brooking Institute [29,32], Statistics Canada [20], and the Bureau of the Census.

In contrast, a constrained merge uses matching without replacement. The merging algorithm enforces constraints on the record weights in both files to ensure that each record is neither under– nor over–

Figure 1.1.
Statistical File Merging


FILE A RECORDS
FILE B RECORDS
2000 | AGI1_1 STATE1_1 | CAP. GAIN1_1
500 | AGI2_1 STATE2_1 | SS.INC2_1
500 | AGI1_2 STATE1_2 | CAP. GAIN1_2
1600 | AGI2_2 STATE2_2 | SS. INC2_2
A_i | $X1_{1i}...X1_{Ri}$ | $Y_{1i}...Y_{Si}$
B_j | $X2_{1j}...X2_{Rj}$ | $...Z_{Tj}$
RECORD WEIGHT
COMMON ITEMS
FILE A ONLY
RECORD WEIGHT
COMMON ITEMS
FILE B ONLY
FILE C (COMPOSITE RECORDS)
W_{ij} | $X1_{1i}...X1_{Ri}....Y_{Si}$ | $X2_{1j}...X2_{Rj}...Z_{Tj}$

INTERRECORD DISSIMILARITY MEASURE (DISTANCE FUNCTION):

$$D_{ij} = F(X1_{1i}, \ldots, Y_{Si}, X2_{1j}, \ldots, Z_{Tj})$$

matched relative to the number of population units represented. Mathematically, the constrained merge model is as follows.

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j \tag{1.1}$$

$$\sum_{j=1}^{n} w_{ij} = a_i, \quad i = 1, \ldots, m, \tag{1.2}$$

$$\sum_{i=1}^{m} w_{ij} = b_j, \quad j = 1, \ldots, n, \tag{1.3}$$

$$w_{ij} \geq 0, \quad \text{for all } i \text{ and } j. \tag{1.4}$$

Constraint (1.1) reflects the assumed equivalent underlying population sizes for the two files, although files A and B have m and n records, respectively. Some minor adjustments may be needed to accomplish this in practice. Again, w_{ij} is the merged record weight for matching record i in file A with record j in file B, and the records are not matched if $w_{ij} = 0$. Constraints (1.2) and (1.3) allow any record to be matched one or more times but such that the merge file weights must sum to the original record weights. Negative weights are precluded by (1.4). This merging algorithm is currently used by Mathematics Policy Research.

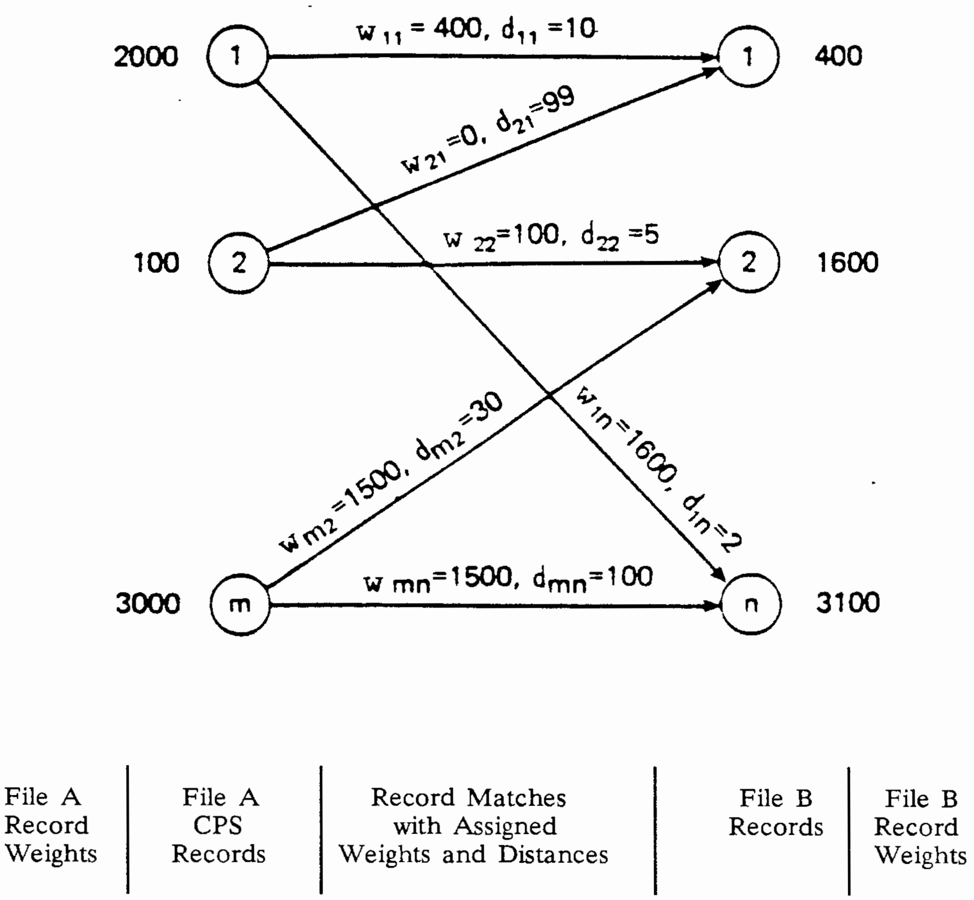
Pictorially, the constrained merge process is depicted in Figure 1.2 where the leftmost set of circles, or nodes, represent file A records with their respective weights, the rightmost nodes the file B records and weights, and the connecting arcs the possible record matches. A set of w_{ij} merge record weights are shown which meet constraints (1.1)–(1.4).

This merge technique can be further refined by requiring the procedure to

$$\text{minimize} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} w_{ij} \tag{1.5}$$

subject to (1.1)–(1.4). This model, originally proposed by Turner and Gilliam [48] and later derived by Kadane [25], seeks to find the best constrained match, the one with the minimum aggregate distance between matched records. This optimal constrained merge procedure requires the solution of a linear programming problem of extremely large dimensions, and is currently used by the U.S. Department of the Treasury [5, 10, 13, 14].

Figure 1.2
Constrained Merge Model


2000
1
100
2
3000
m
$w_{11} = 400, d_{11} = 10$
$w_{21} = 0, d_{21} = 99$
$w_{22} = 100, d_{22} = 5$
$w_{m2} = 1500, d_{m2} = 30$
$w_{1n} = 1600, d_{1n} = 2$
$w_{mn} = 1500, d_{mn} = 100$
1
400
2
1600
n
3100
File A Record Weights
File A CPS Records
Record Matches with Assigned Weights and Distances
File B Records
File B Record Weights

1.5. Statistical Aspects of Merging Techniques

Unconstrained procedures utilize (1.5), subject to (1.1), (1.2), and (1.4); thus, by dropping constraint set (1.3), the composite records match up well at the record level. However, file B item statistics in the composite file are distorted by their implicit reweighting of the augmentation records through over– and under–matching. This reweighting has a strong impact on important extreme values, variances, covariances, and other distributional aspects of the file B items, as shown below.

While the constrained procedures may not match up as well at the record level as unconstrained procedures, their merged files contain all of the information from the original files and preserve all statistical properties of the A and B data items. Further, if optimization is applied, the best overall constrained match is insured. Appendix A details, using two small example files, the effect of various merging schemes on the mathematical structure of the resultant composite file.

All of these aspects influence the results of the microanalytic models and research studies which use the merge file.

1.6. Underlying Merge Rationale

When two files are merged, we assume that two files (X1, Y) and (X2, Z) are drawn from the same population, where X1 and X2 are the sets of common items in file A and B, respectively, and Y and Z the sets of items unique to file A and B, respectively (the alignment assumption). The objective of merging is to form a file (X1, X2, Y, Z) which corresponds statistically to a sample of (X, Y, Z) taken from the same population. We do this in order to make inferences about (Y, Z) and (Y, Z|X) relationships, since we can already make (X, X), (X, Y), and (X, Z) inferences from the two original files.

1.7. Quality Considerations

A strong theoretical justification for merging or an explanation of exactly what is being accomplished by a merge is not available in the literature. What is needed is both a measure of the "accuracy" of (X1, X2, Y, Z) in replicating (X, Y, Z), and a means for making decisions such as: are the two files mergable? Is the composite file acceptable?

Typical reported measures of match accuracy are: counts of X1–X2 item agreements, item means, and percentage agreement by common item. The notion is that a file which matches well on the X–items matches well on the Y–Z relationships. Rarely reported are (1) comparisons of covariances, such as cov(Y) in unconstrained matches and cov(Y, Z) versus expected cov(Y, Z), (2) conditional and joint frequencies for augmentation variables, and (3) other Y–Z studies. Ruggles, Ruggles, and Wolff [42] are the only contributors in this area. Moreover, the following empirical data directs attention to potential problems measured by the above mentioned statistical measures.

1.8. Preliminary Empirical Data

In a recent set of experiments to investigate the effect of merging techniques on resultant file quality, subsets of the 1975 SOI and 1975 SIE were chosen, based on a nine–state geographic region. There were 7144 SOI records and 6283 SIE records, representing 12.7 million tax filling units, or approximately 15 percent of the population. The X–variables used in the distance function were: age, race, sex, marital status, family size, wage income, business income, property income, spouse's income, and adjusted gross income. The two files were merged three ways: unconstrained with SOI as base file, unconstrained with SIE as the base file, and optimally–constrained, all using the same distance function.

In Table 1.1, the distribution of SOI wages and business income is shown for both the original file and the unconstrained merge file using the SIE as the base. Not only are the means not in agreement but the distributions are altered, and dramatically in the case of business income. Of course, in the constrained merge, the distributions were identical to the originals.

To evaluate the unconstrained procedure's effect on covariance structure, the variance–covariance matrices of several common items were compared with the original matrices. The median percentage differences, by item, are shown in Table 1.2. In some cases, the median error is as small as 7 percent, but in others these second–order statistics differ greatly. Analysis of the constrained merge verified the expectation of zero error.

1.9. Research Questions

Despite the widespread use of merging as a data enrichment technique, there is a paucity of much–needed research in this area. Consider the following questions.

1.9.1. Constrained Versus Unconstrained Techniques

When does either procedure create a match file which is statistically equivalent to a valid (X, Y, Z) sample drawn from the population? A goodness–of–match criterion is needed not only to answer this question but to compare alternative matching algorithms.

Table 1.1
SOI Item Distributions

Income Class ($000's)	Total SOI Wages ($ Millions)		Total SOI Business Income ($ Millions)	
	Original	Unconstrained	Original	Unconstrained
< 0	0	0	- 712	- 86
1-5	12,218	11,699	857	607
5-10	22,535	21,124	836	1,194
10-15	24,745	26,639	617	1,497
15-20	10,326	23,882	677	909
20-30	21,133	24,930	808	1,402
30-50	9,597	10,141	785	964
50-100	3,371	2,741	777	1,108
100-200	1,010	136	298	608
> 200	244	0	111	0
Total	115,784	121,291	5,055	8,187
Mean	$9,108	$9,542	$398	$644

Table 1.2

Variance-Covariance Differences

Common Variable	Median Variance-Covariance Error Relative to Original Data	
	Unconstrained SIE	Unconstrained SOI
Age	31.2%	26.4%
Family size	35.5	17.4
Wages	7.3	23.9
Business income	72.1	38.4
Farm income	97.7	88.8
Property income*	78.5	850.4
Spouse income	73.5	31.0
Adjusted gross income	9.9	24.4

*Interest + dividend + rental income.

1.9.2. Covariance of (Y, Z|X)

What is the effect of omitting or including cov(Y, Z|X) in the matching methodology? Do correlated X–variables carry along their correlated Y and Z variables properly?

1.9.3. Distance Functions

How do the various dissimilarity measures affect the resultant merge file? What is a "correct" distance function? (See [25].) In practice, distance functions usually reflect the data aspects of greatest importance in the target microanalytic model or database.

For other research question and issues surrounding merging activities, see [47] by the Federal Committee on Statistical Methodology.

1.10. Experimentation Overview

In order to benchmark the various merging schemes and study the statistical aspects of the merge process, a series of experiments were performed with public–use national data sets. The 1975 SIE file was designated to be a test population from which a selected series of samples were drawn. Each record item was declared to be in set X, Y, or Z based on data type and correlations with other data items. The resultant set of files were merged pairwise in various combinations using a variety of distance functions, merge schemes, and levels of data bias and error. By designating the full SIE file to be the population, the actual (X, Y, Z) is known, unlike the usual case in practice. This availability of the complete population provides an accurate standard for comparison with any merge file.

The experimental design was structured to study the effect, if any, of the above parameters on Y–Z relationships, standard statistical tests, and measures of "goodness" of the match. The study also investigates the sensitivity of the various merge algorithms to the distance function used and to the introduction of bias and error.

PART 2. STATISTICAL FRAMEWORK AND EXPERIMENTAL DESIGN

2.1 Notation and Overview of the Study

Statistical matching methods have been developed for the purpose of combining the information from two microdata files, each collected from a separate sample survey, into a single composite file. The objective of statistical matching is to create a single file which is "equivalent" to a valid sample taken from the population of interest.

The two input files, A and B, are of the form (X1, Y) and (X2, Z), respectively, where (X1, Y) is a sample with multivariate observations (x_i, y_i) on each sampling unit, while (X2, Z) is another independent sample with multivariate observations (x_j, z_j) on each sampling unit. Note that sets X1 and X2 are measured on the same data items (e.g., wages, interest income, family size) but the observation sets X1 and X2 are measured on different sampling units arising from the two different surveys. The data items (X1, Y) and (X2, Z) are obtained from either stratified or probability samples taken from the national population. The number of observations is typically large, e.g., in excess of 50,000 records.

From such files a statistical match would create a single file of the form (X12, Y, Z) where set X12 is a composite of X1 and X2. Presumably such a file would in practice be used as if it were a valid random sample taken from the population of (x, y, z) measurements. Statistical inferences would be made with standard methods developed to account for sampling variability of such random samples.

A fundamental question to be addressed is: when do matched files really contain the same sampling variability as ordinary random samples? It is the goal of this project to empirically investigate the performance of some known matching methodologies from this point of view. The experimentation took the general form of (1) creating file A and file B from known populations, (2) statistically matching the two files, and (3) calculating a statistical summary of the matched files. By repeating these steps many times for each matching methodology, the empirical sampling results of the matched files may be compared with the known sampling properties of valid random samples.

2.2 Statistical Matching Issues to be Addressed

2.2.1. Statistical Inference with Matched Files

One objective of this work is to identify conditions under which the matching methodologies will perform well. If a matching technique produces matched (X12, Y, Z) files which behave like random samples (X, Y, Z), then the technique would be totally successful. This, however, may be too strict a requirement to reasonably expect. A weaker condition for judging a matching technique as acceptable would be to require that point estimates of the (x, y, z) population parameters be unbiased or consistent. Matched files which provided accurate estimates would be of great value, even if the precisions of such estimators were difficult to access.

By conducting many replications of the matching process for known populations, it is possible to statistically study the properties of matched file estimators of key population parameters, such as cov(Y, Z). Since the success of a matching technique will quite possibly depend on the properties of the sampled population, the experiments were conducted for a variety of theoretical populations which are expected to affect the matched file estimators in different ways.

2.2.2. Constrained Versus Unconstrained Procedures

As described previously, matching procedures may be divided into two principal types, "constrained" and "unconstrained." In each case the (X1, Y) file is linked, record–by–record, to the (X2, Z) file to form the (X12, Y, Z) composite file. In unconstrained matching, each record in the (X1, Y) file is matched with the single closest record in the (X2, Z) file. The composite (x12, y, z) record weight is the weight of the (x1, y) record. In constrained matching, each record in each file may be matched one or more times; however, for a given record, the sum of the linked record weights must equal the original weight.

The desirable statistical characteristics of unconstrained matching is that the degree of association between X1 and X2 is closer at the unit level than the unit level association of X1 and X2 in a constrained match. The potential disadvantage of unconstrained matches is that the statistical characteristics of (X2, Z) can be altered in the matching process (assuming the (X2, Z) file is the one that is unconstrained). The advantage of constrained matching is that all statistical properties of (X1, Y) and (X2, Z) are preserved in the matching process. It must be noted that the statistical characteristics of (X1, Z) might not be the same as for (X2, Z) even though the data item X1 can be accepted as statistically equivalent to X2. The disadvantage of constrained matching is that unit level associations between X1 and X2 are not as close as can be obtained using unconstrained matching. However, for both constrained and unconstrained matching the ultimate test is whether or not the matched file generated is statistically equivalent to a valid sample of (X, Y, Z) drawn from the population of such data items. The question becomes one of identifying the conditions under which either constrained or unconstrained matches produce valid results.

2.2.3. Cov(Y, Z|X)

If Y and Z are uncorrelated for given levels of X, i.e., cov(Y, Z|X=x)=0 for all values x, then non-matching methods distribution could be used to estimate the joint of (X, Y, Z) from information obtained from the unmatched files alone. As pointed out by Sims [46] if, under conditional independence, the joint distributions of (X, Y, Z) admit probability density functions, then

$$f_{XYZ}(x, y, z) = f_{XY}(x, y) \cdot f_{XZ}(x, z)/f_X(x). \tag{2.1}$$

The probability density functions on the right hand side of the above equation could all be estimated from the separate (X1, Y) and (X2, Z) files, and from this the joint distribution of (X, Y, Z) could be estimated. Then any population parameters of interest could be estimated using this estimated joint distribution. However, when the data files are large the computational effort for this approach may be as great or greater than that for matching techniques. Furthermore, statistical properties for this estimation approach might require unusual methods not available in standard statistical computing packages.

Current matching techniques usually accomplish the match by aligning X1 and X2 values which are close by some distance function criterion. (See distance function discussion below.) Since information about Y and Z is not used in the matching criteria, it would seem that the created matched files will likely have sample cov(Y, Z|X=x) close to 0. This is because when several records have exactly the same x information the matching is accomplished within these records by arbitrary or random selection. However, this might not be a problem if in fact there are only a few records with the same set of x values. Most matching projects have not included Y–Z relationships in the matching methodology even though it is not assumed that cov(Y, Z|X=x)=0. One of the major objectives of this project is to examine the results of matched files which do not use Y–Z relationships in the matching when in fact cov(Y, Z|X=x) $\neq$ 0.

2.3. Distance Functions and Matching Methodologies

The matching methodologies considered here all proceed by defining a distance function which measures the dissimilarity between a pair of records. This function assigns a value d_{ij} to any pair of records (x_i, y_i) and (x_j, z_j) from files A and B, respectively. For a given match, say, M, of the two files an overall distance D_M is defined as a weighted sum of the distances of all matched record pairs as follows:

$$D_M = \sum_{(i, j) \in M} w_{ij} d_{ij}. \tag{2.2}$$

In unconstrained and constrained matches, the final matched file is chosen as the match M* which minimizes D_M over whatever class of possible matches is being considered.

This study focuses on four types of distance functions. There are two major groups: weighted absolute difference methods and Mahalanobis distances. In addition, each of these types may be applied to the X items alone, or expanded to include all X, Y, and Z items. Each of these four types of distance functions is used in conjunction with both an unconstrained and the constrained–optimal matching scheme, thus giving eight primary matching methodologies for study.

2.3.1 Weighted Absolute Difference Measures

Distance functions in this category are of the type used by the U.S. Department of the Treasury's Office of Tax Analysis [13] and the Social Security Administration's Office of Research and Statistics [37]. This function uses subjective weights, reflecting the relative importance of each data item, which are multiplied by the absolute differences of values of the corresponding items in the pair of records under consideration. Specifically, when only the X items are being considered, the distance between record i in file A and record j in file B is defined as:

$$d_{ij} = \sum_{k=1}^{r} s_k \cdot |x1_{ik} - x2_{jk}|, \qquad (2.3)$$

where $x1_{ik}$ and $x2_{jk}$ denote the kth data items in the respective files, r is the number of data items in X, and s_k is the subjective weight for data item k.

This procedure can be expanded to include additional (X12, Y, Z) relationships by adding other difference terms to the function definition. For example, to include some information about the relation of data item k of X and data item λ of Y, an additional component of the distance function could be $s_{\lambda k} \cdot |y1_{i\lambda} - x2_{jk}|$.The weight s_k would be determined subjectively with the sign of the term corresponding to the sign of $\mathrm{cov}(X_k, Y_k)$. Similarly, relationships between the various X items themselves, X and Z items, and even Y and Z items could be included in the matching criteria. Of course, the choice of the subjective weights is an important one since they will have a strong impact on the matches obtained.

2.3.2. Mahalanobis Distance Metrics

The other category of distance functions studied was proposed by Kadane [26]. One procedure, which uses only the X items, defines

$$d_{ij} = (s1_i - x2_j)'(\Sigma_{XX})^{-1}(x1_i - x2_j), \qquad (2.4)$$

where Σ_{XX} is the covariance matrix of the X variables. This is the Mahalanobis distance between two x values and, using only the X information, it arises as the maximum likelihood solution for exact matching of normal random variables. It seems quite plausible that it will also perform well in statistical matching.

Kadane also suggests a procedure which employs full (X, Y, Z) information. In this instance, file A is expanded from (X, Y) to $V \equiv (X1, Y, \hat{Z})$, where $\hat{z}_i \equiv E(Z|X = x1_i, Y = y_i)$ is the regression prediction for the missing Z data of record i. Likewise, file B is expanded to $U \equiv (X2, \hat{Y}, Z)$ where $\hat{y}_j \equiv E(Y|X = x2_j, Z = z_j)$. S_1 and S_2, the covariance matrices of (X1, Y, Z) and (X2, Y, Z), respectively, may be formed from Σ, the covariance matrix of (X, Y, Z). The match is performed using

$$d_{ij} \equiv (v_i - u_j)'(S_1 + S_2)^{-1}(v_i - u_j). \qquad (2.5)$$

A difficulty here is in obtaining accurate Σ_{XX} entries used in calculating S_1 and S_2. These must come from a source outside of the two files being matched or, if $\mathrm{cov}(Y, Z|X=x) = 0$ is assumed, can be calculated as

$$\Sigma_{YZ} \equiv \Sigma_{YX}(\Sigma_{XX})^{-1}\Sigma_{XZ}. \qquad (2.6)$$

2.4. Experimental Design

To gain insight into the impacts of merging scheme, distance functions, and measurement error in the data, a set of experiments were performed using national datasets.

The 1975 Survey of Income and Education was treated as a population and, from this file, five randomly-drawn samples of 1000 records each were drawn. The records of each file were divided into two new data sets by designating each record item to be in set X, Y, or Z and forming an (X, Y) file and an (X, Z) file. These files were merged using various methodologies, and examined using the evaluation design in the section that follows.

The data items designated for sets X, Y, and Z were selected to include each of the various types of data available on the file and different levels of correlation. Also, the files were merged using different numbers of X-variables. The performance of the matching methodologies can be simultaneously evaluated, as in the Monte Carlo simulations, by comparison with known characteristics of the original records.

The existence of measurement error is simulated by adding bias, unbiased noise, and biased noise to subsets of the X variables prior to merging. The sensitivity of the distance function definitions and merging schemes to such error are then evaluated both at the record level and in the aggregate using the statistical evaluation design.

2.5. Statistical Evaluation of the Merged Files

2.5.1. Bias for Matched File Estimators

Does a matching methodology induce a bias in the estimation of important population parameters? To answer this question, the key parameters of interest will be cov(Y, Z) and cov(X, Z) with μ_Z and Var(Z) of particular interest for unconstrained matching. Let Θ be a parameter of interest. Note that Θ will be known exactly, since we know the population from which we have samples. Also let T_i be the estimate of Θ derived from the ith matched file generated with some particular methodology. Now let μ_T be the mean of all T which could have possible been obtained as estimates of Θ from matched files. Then the matching procedure produces unbiased estimates of Θ if $\mu_T = \Theta$.

The question of biased estimators of Θ can then be addressed by comparing the observed T_i with Θ . A simple test for $\mu_T = \Theta$ can be made by using

$$z = \frac{\sqrt{k}\,(\overline{T} - \Theta)}{S}, \tag{2.7}$$

where

$$\overline{T} \equiv \frac{\sum_{i=1}^{k} T_i}{k}, \tag{2.8}$$

and

$$S^2 \equiv \frac{\sum_{i=1}^{k} (T_i - \overline{T})^2}{(k-1).} \tag{2.9}$$

For large values of k (including this study's k=100), z will have an approximate standard normal distribution if $\mu_T = \Theta$.

As noted earlier, a randomized block design analysis would be available for comparing the $T_i - \Theta$ biases of several methodologies simultaneously.

2.5.2. Matched Samples versus Valid Random Samples

Does the sampling variability of matched files resemble the variability in random samples? Here the major concern is with the overall sampling distribution of the (X, Y, Z) merged files or with simply the bivariate sampling distribution of (Y, Z) obtained from merged files.

An excellent way to examine a multivariate sample is to divide each variable into classes and form a multiway contingency table of the sample. For example, for a (Y, Z) sample and selected values of a, b, c, d, e, and g, we can analyze the table below where f_{ij} is the observed frequency of data in cell (i, j).

	$z \leq d$	$d < z \leq e$	$e < z \leq g$	$g < z$
$y \leq a$	f_{11}	f_{12}	f_{13}	f_{14}
$a < y \leq b$	f_{21}	f_{22}	f_{23}	f_{24}
$b < y \leq c$	f_{31}	f_{32}	f_{33}	f_{34}
$c < y$	f_{41}	f_{42}	f_{43}	f_{44}

Since the (Y, Z) population is known, the expected value for each f_{ij} may be calculated for random sampling. It is well know that the Pearson Chi-square statistic

$$\chi^2 \equiv \sum \frac{[f_{ij} - E(f_{ij})]^2}{E(f_{ij})},$$

has an approximate Chi-square distribution with 15 degrees of freedom when valid random sampling is done. This suggests that we calculate χ^2 for each matched file we obtain using a given methodology; if the matched files are equivalent to random sampling, then these k χ^2 values should follow the appropriate Chi-square distribution. The Kolmogorov test can be used to test this, or we can simply describe χ^2 values with a histogram. A similar procedure can be used to compare the sampling variability of whole matched files with that of valid random sampling.

If we wish to bypass the file-by-file comparison of matched file distribution and random sampling, all matched file data can be pooled for a given methodology into a single table and the pooled distribution compared with the expected values calculated from the known population. To test the hypothesis that the matched file samples were simple random samples, a single Pearson Chi-square statistic can be used.

PART 3. ANALYSIS BASED ON A NATIONAL MICRODATA SET

The objective of this section is to empirically examine the statistical characteristics of selected matched files, given that the population characteristics of the data are known. These comparisons will be used to evaluate the matching techniques.

The primary statistical procedure employed in these comparisons is a simple χ^2 goodness-of-fit test using percentage distributions and prespecified categories. In addition, matched file correlation matrices are compared directly with the population correlation matrix. No specific test for equality of correlation matrices was applied since direct observation of the correlation matrices indicated that for certain data items there existed very large deviations.

An absolute difference distance function was tested under conditions of noise, bias, and combined noise and bias. The same function was studied for sensitivity to the number of common variables used and the Mahalanobis distance function was also tested. Also investigated were the consequences of matching a sample with itself under a wide variety of circumstances, and the effect of using an unconstrained merging procedure instead of a constrained model.

This work focuses upon the Y-Z distributions in the matched files and their relationship to the population Y-Z distributions, since the implied motivation for most statistical matching is the construction of Y-Z distributions. Prior to this study, published merge analyses paid almost exclusive attention to the relationship between X1 and X2 in the matched file and to the vector of Z means. However, in these previous studies, data limitations have been such that *ex-post* testing of Y-Z distributions has not been possible.

Fifty matched files were generated and solved with the transportation model for empirical analysis of the issues mentioned above. These constrained matches employed five samples of approximately 1000 records each, selected from the Survey of Income and Education family file. In addition to the constrained matches, a number of unconstrained matches were generated for comparison purposes.

3.1. Population, Samples, and Data Item Descriptions

In this phase of the study, an extract of the Survey of Income and Education family file, resident in the Office of Tax Analysis' data library, was designated to be a base population. This particular file has 54,034 records representing a full population of approximately 78 million families. The data items selected for inclusion in the extract file were: family wage or salary income, interest income, countable assets, age of the family head, highest grade of school completed by the family head, sex of the family head, total annual family income, social security income, number of adults in the family, dividend income, family size, and race of the family head.

Five subsamples of approximately 1000 records were randomly selected for matching purposes. In addition, one of the subsamples (SIE5) was used to form "other samples" with the data items perturbed to simulate noise and bias.

The five subsamples' identifiers and their respective sizes are: SIE1, 938 records; SIE2, 943 records; SIE3, 942 records; SIE4, 991 records; and SIE5, 951 records. The weights of each file summed to approximately 1.4 million. The subsamples created from SIE5 and their characteristics are described in Table 3.1.

Table 3.1.

Files Created from Perturbations of SIE5

Name	Modification(s) Made to Record Item(s)
SIE6A	Asset value reduced by 20% for simulated bias.
SIE6B	Asset value reduced by 10% for simulated bias.
SIE7A	25% of the records have asset value multiplied by a random number between .75 and 1.25 for simulated noise.
SIE7B	All asset values were multiplied by a random number between .75 and 1.25 for simulated noise.
SIE7C	25% of the records have asset value multiplied by a random number between .9 and 1.1 for simulated noise.
SIE7D	All asset values were multiplied by a random number between .9 and 1.1 for simulated noise.
SIE8A	25% of the records have asset value multiplied by a random number between .7 and 1.0 for simulated noise and downward bias.
SIE8B	All asset values were multiplied by a random number between .7 and 1.0 for simulated downward bias.

NOTE: Samples SIE6A–8B are otherwise the same as SIE5, and different random numbers were used for each randomly-perturbed record.

Six items were designated as X, or common, variables: wages and salaries, interest income, assets, age of family head, highest grade of head, and sex of family head. Total income, social security income, and number of adults were chosen to be the Y variables (i.e., the variables unique to the first matching file, A). The variables selected to be set Z (i.e., the variables unique to the second matching file, B) were dividend income, family size, and race.

The Y–Z correlation matrix for the full population, and the differences between each subsample's Y–Z correlation matrix and the population correlations are given in Tables 3.2A–F. In addition, the population percentage frequency counts for all Y–Z item pairs are given in Tables 3.3A–I.

Table 3.2A
Full SIE Population Y–Z Correlation Matrix

Total Income	1.00					
Social Security	–.14	1.00				
Number of Adults	.44	–.04	1.00			
Dividends	.33	.06	.02	1.00		
Family Size	.34	–.17	.76	–.01	1.00	
Race	–.12	–.15	–.02	–.04	.05	1.00

Table 3.2B
Difference Between SIE1 Correlation Matrix
and Population Correlation Matrix

Total Income	0				
Social Security	.05	0			
Number of Adults	.06	.01	0		
Dividends	–.03	.11	.05	0	
Family Size	.03	–.17	.02	–.01	0
Race	–.12	.08	.02	–.04	.05

Table 3.2C
Difference Between SIE2 Correlation Matrix and Population Correlation Matrix

Total Income	0					
Social Security	−.02	0				
Number of Adults	−.01	−.03	0			
Dividends	−.14	−.01	−.03	0		
Family Size	−.04	.03	−.04	−.01	0	
Race	−.04	−.01	−.02	−.04	.02	0

Table 3.2D
Difference Between SIE3 Correlation Matrix and Population Correlation Matrix

Total Income	0					
Social Security	.01	0				
Number of Adults	.01	−.05	0			
Dividends	−.05	.14	.02	0		
Family Size	.05	−.04	.03	−.01	0	
Race	−.08	−.07	.09	−.01	.11	0

Table 3.2E
Difference Between SIE4 Correlation Matrix and Population Correlation Matrix

Total Income	0					
Social Security	.03	0				
Number of Adults	−.03	−.05	0			
Dividends	.06	−.01	−.03	0		
Family Size	−.07	−.03	−.01	−.02	0	
Race	−.03	.06	−.06	.00	−.03	0

Table 3.2F
Difference Between SIE5 Correlation Matrix and Population Correlation Matrix

Total Income	0					
Social Security	−.02	0				
Number of Adults	−.01	.02	0			
Dividends	.02	−.03	−.02	0		
Family Size	.03	−.01	.02	−.01	0	
Race	−.08	.01	.00	−.01	−.03	0

Table 3.3A
Population
Total Income and Dividend Joint Distribution (Percentage Counts)

	Dividends		
Total Income	$0	$1 under $1,000	$1,000 Plus
Under $5,000	21.01	1.18	.16
$5,000 under $10,000	20.20	1.85	.58
$10,000 under $15,000	16.75	2.24	.55
$15,000 under $20,000	11.61	2.22	.56
$20,000 under $25,000	6.74	2.00	.54
$25,000 Plus	6.84	3.09	1.78

Table 3.3B
Population
Total Income and Family Size Joint Distribution (Percentage Counts)

	Family Size			
Total Income	1	2	3	4 Plus
Under $5,000	13.48	4.85	1.93	2.09
$5,000 under $10,000	7.69	7.71	2.97	4.27
$10,000 under $15,000	3.57	6.23	3.72	6.05
$15,000 under $20,000	1.20	4.06	2.91	6.22
$20,000 under $25,000	.45	2.65	1.92	4.29
$25,000 Plus	.46	2.89	2.35	6.06

Table 3.3C
Population
Total Income and Race Joint Distribution (Percentage Counts)

	Race	
Total Income	White	Nonwhite
Under $5,000	17.82	4.53
$5,000 under $10,000	19.52	3.11
$10,000 under $15,000	17.62	1.94
$15,000 under $20,000	13.25	1.13
$20,000 under $25,000	8.70	.61
$25,000 Plus	11.08	.67

Table 3.3D
Population
Social Security and Dividend Joint Distribution (Percentage Counts)

		Dividends	
Social Security	$0	$1 under $1,000	$1,000 Plus
$0	62.33	9.44	2.38
$1 under $3,000	12.92	1.59	.79
$3,000 Plus	7.9	1.55	1.01

Table 3.3E
Population
Social Security and Family Size Joint Distribution (Percentage Counts)

		Family Size		
Social Security	1	2	3	4
$0	17.26	17.42	13.35	26.20
$1 under $3,000	7.83	4.60	1.31	1.56
$3,000 Plus	1.77	6.35	1.14	1.21

Table 3.3F
Population
Social Security and Race Joint Distribution (Percentage Counts)

	Race	
Social Security	White	Nonwhite
$0	64.94	9.28
$1 under $3,000	13.36	1.94
$3,000 Plus	9.69	.77

Table 3.3G
Population
Number of Adults and Dividend Joint Distribution (Percentage Counts)

	Dividends		
Number of Adults	$0	$1 under $1,000	$1,000 Plus
1	27.35	2.64	1.02
2	37.38	6.28	2.06
3 Plus	18.18	3.66	1.09

Table 3.3H
Population
Number of Adults and Family Size Joint Distribution (Percentage Counts)

	Family Size			
Number of Adults	1	2	3	4 Plus
1	26.62	2.05	1.41	.95
2	0	26.33	7.43	12.02
3 Plus	0	0	6.96	16.00

Table 3.3I
Population
Number of Adults and Race Joint Distribution (Percentage Counts)

	Race	
Number of Adults	White	Nonwhite
1	26.05	4.96
2	41.55	4.23
3 Plus	20.22	2.74

3.2. Matched Files Generated Using the Transportation Model

A total of 50 matched files were generated using the Office of Tax Analysis' optimal-constrained merge system (see [14] for description). The sub-samples designated in the previous section were selectively matched pairwise using six different distance functions. These six distance models use the X vector of common data items.

3.2.1. Weighted Absolute Differences Model

Model 1 is an absolute difference distance function where for record i from the first file and record j from the second file:

$$d_{ij} = C_1 + C_2 + C_3 + C_4 + C_5 + C_6$$

where the six components are calculated as follows, the first component in the distance function for any given record match is

$$C_1 = \min\{400,\ \ 100 \cdot \frac{|(\text{File A wage} \ - \ \text{File B wage})|}{\text{File A wage}}\}$$

The index C_1 is the absolute value of difference in wages and salaries for any pair of A and B records divided by the File A wage, but constrained not to exceed 400. For example, if the File A wages are \$25,000 and the File B wages are \$25,596,

$$C_1 = \min\{400,\ \ 100 \cdot |(25,000 \ - \ 25,596)|/\ 25,000\} = 2.4.$$

In this example the index C_1 denotes the fact that the given B record has a wage which differs from the A by 2.4%. The upper limit of C_1 = 400 is arbitrary, but is intended to not allow differences in wages alone to determine a match for situations with large total distance, i.e., in excess of 400.

The record distance function component, C_2, is a penalty assessed for differences in countable assets and follows the same formula as C_1.

The index C_3 denotes an index for differences in interest income between a pair of A and B records. Interval categories are used for the calculation of C_3 , as defined in Table 3.4. The index C_3 has an upper limit of 52 which means that the greatest difference in property incomes has a distance function penalty equivalent to a 52% difference in wages. Hence, the matching algorithm will try to maintain compatibility between the broad categories of property income, but the penalty for noncompatibility is never very large. In the lower segment of the income distribution, the impact of C_3 is to match records with zero property income together, whereas in the upper range of the income distribution the index C_3 will keep records with large amounts of property income together, all else equal.

Demographic factors included in the distance function are age, sex, and highest grade attained by head of household. The age penalty is defined by the variable C_4 which is described in Table 3.5. The age penalty is based upon the age of the first person in the tax record.

Table 3.4
C_2 = Interest Income Difference Index

File A Interest			File B Interest Income		
Income	$0	$1-1000	$1001-10000	$10001-1000001	$100001 or more
0	0	13	26	39	52
$ 1 - $ 1000	13	0	13	26	39
$ 1001 - $ 10000	26	13	0	13	26
$10001 - $100000	39	26	13	0	13
$100001 and above	52	39	26	13	0

Table 3.5
C_4 = Penalty Index for Difference in Ages

			File B Age		
File A Age	≤ 17	18≤ 22	23≤ 61	62≤ 65	65 and over
≤ 17	0	12	32	80	80
18 ≤ 22	12	0	24	80	80
23 ≤ 61	32	24	0	64	80
62 ≤ 65	80	80	64	0	40
66 and over	80	80	80	40	0

The penalty for age difference is never greater than an 80% difference in wages. The broad age categories are defined to represent school age and retirement age. For example, the age breakpoint of 62 represents early retirement and 66 denotes regular retirement. The age "17 and less" represents children living at home, and the age interval "18–22" represents college or beginning employment age. For the objectives of this matched file, persons with ages between 23 and 61 are not considered to be different, if all other factors are the same. However, a large penalty is imposed if a person 61 or younger is matched with a person 62 or older in order to differentiate persons eligible for Social Security income from those who are not.

The penalty index for differences in highest grade attained by head-of-household is calculated as follows:

$$C_5 = 16 \cdot \mathrm{INT}\,(|\text{Grade of A} - \text{Grade of B}|/3)$$

where INT(x) is a function whose value is the smallest integer less than or equal to x. This value never exceeds a 100% difference in wages and represents a graduated penalty for increased differences in highest grade attained. Note that there is no penalty for a difference of under three years, a penalty of 16 for a three to five year difference, and so on.

The last penalty included in the distance function is the index C_6 for differences in sex of head of household. If the A record and the B record have different sex codes then the index C_6 is set equal to 100, which has the same impact as a 100% difference in wages.

$$C_6 = \begin{cases} 0 \text{ if A and B have the same sex code} \\ 100 \text{ if A and B have different sex codes} \end{cases}$$

The distance function value for a given potential record match is the sum of variables C_k . More precisely, the notation for the variable C_k discussed above should be C_{ijk} where i denotes the ith A record, j denotes the jth B record, and k denotes the index for income and demographic characteristics. Hence,

$$d_{ij} = \text{distance function value for the ith A record and the jth B record.}$$

$$= \sum_{k=1}^{6} C_{ijk}.$$

The objective of the distance function is to try to force matches within the intervals defined for interest income, age, highest grade, and sex, and to try to obtain very close absolute agreement based upon wages and assets.

3.2.2. Mahalanobis Distance Model

Model 2 is the Mahalanobis distance function value for record i from the first file (A) and record j from the second file (B) is defined in (2.4) as:

$$di_j = (x1_i - x2_j)'(\sum_{XX})^{-1}(x1_i - x2_j)$$

where

$x1_i$ is the vector of common data items from record i of file A,

$x2_j$ is the vector of common data items from record j of file B, and

$\sum_{XX}$ is the covariance matrix of the X variable from the population file.

3.2.3 Other Constrained Models

Model 3 is Model 1 without "assets" in the distance function, and *Model 4* is Model 1 without "assets," " age," and "sex" in the distance function. *Model 5* is an absolute value percentage difference distance function using only "wages and salaries," i.e., using only C_1 from Model 1. *Model 6* uses only the "age " and "highest grade attended," i.e., C_4 and C_5, of Model 1.

3.2.4. Matched Files Created

The specifications of the 50 generated matched files using the transportation algorithm are given in Table 3.17. For matching purposes, the Z elements of the file A samples and the Y elements of the file B samples were ignored.

These 50 matched files are in the following test classifications.

Matched files 1–10:	pairwise matching of all samples using an absolute difference distance function (model 1).
Matched files 11–20:	pairwise matching of all samples using the Mahalanobis distance function (model 2).
Matched files 23–30:	matching the sample SIE5 with itself under conditions of noise, bias, and combined noise and bias (model 1).
Matched file 21:	matching SIE5 with itself using the absolute difference distance function and six common variables (model 1).

Matched file 22: matching SIE5 with itself using the Mahalanobis distance function (model 2).

Matched files 31–38: matching samples SIE1 with SIE5 under conditions of noise, bias, and combined noise and bias (model 1).

Matched files 40–43: pairwise matching of sample SIE5 with sample SIE1, SIE2, SIE3, and SIE4 using absolute difference distance function with five common variables (model 3).

Matched files 44–47: pairwise matching of sample SIE5 with samples SIE1, SIE2, SIE3, and SIE4 using the absolute difference distance function with only three common variables (model 4).

Matched files 39: Sample SIE5 matched with itself using the absolute difference distance function with five common variables (model 3).

Matched file 48: sample SIE5 matched with itself using the absolute difference distance function with three common variables (model 4).

Matched file 49: sample SIE5 matched with itself using the absolute difference distance function with only the common variable wages and salaries (model 5).

Matched file 50: sample SIE5 matched with itself using the absolute difference distance function with only the two common variables age and highest grade attained (model 6).

3.2.5. Tests Used to Compare Matched File Distributions with the Population Distributions

Two tests were selected for comparing the Y–Z distributions in a matched file to calculate a simple χ^2 statistic for each Y–Z pair, based upon cross-tabulated, percentage for the categories specified in Table 3.3–A–I, and using the population percentage counts from these tables as the expected values. The χ^2 statistic is calculated in the following manner. For a given cell K in a Y–Z table,

Table 3.6
Matched Files Created Using the Transportation Matching Algorithm*

FILE B

FILE A	SIE1	SIE2	SIE3	SIE4	SIE5	SIE6A	SIE6B	SIE7A	SIE7B	SIE7C	SIE7D	SIE8A	SIE8B
SIE1		#1[M1] #11[M2]	#5[M1] #15[M2]	#8[M1] #18[M2]	#10[M1] #20[M2]	#31[M1]	#32[M1]	#33[M1]	#34[M1]	#35[M1]	#36[M1]	#37[M1]	#38[M1]
SIE2	•		•	•	•								
SIE3	•	#2[M1] #12[M2]		•	•								
SIE4	•	#3[M1] #13[M2]	#6[M1] #16[M2]		•								
SIE5	#43[M3] #44[M4]	#4[M1] #14[M2] #42[M3] #45[M4]	#7[M1] #17[M2] #41[M3] #46[M4]	#9[M1] #19[M2] #40[M3] #47[M4]	#21[M1] #22[M2] #39[M3] #48[M4] #49[M5] #50[M6]	#23[MI]	#24[MI]	#25[MI]	#26[MI]	#27[MI]	#28[MI]	#29[MI]	#30[MI]

Legend: #i [Mj] ≡ Matched file number i using model j, with File A and File B defined from row and column, respectively.

*Note: With this procedure, the designation of A and B is arbitrary; see B vs. A for missing A vs. B combinations.

$$\chi_k^2 = \frac{(fm - fp)^2}{fp}$$

where fm = weighted percentage of matched records in the k–th cell,

f_p = weighted percentage of population records in the k–th cell taken from Table 3.3A through 3.3I,

$$\chi^2 = \sum_{k-1}^{N} \chi_k^2,$$

N = number of cells in the Y–Z table, and

degrees of freedom = (number of rows) (number of columns)–1.

Also for simplicity, cell counts with less than one percent of the cumulation frequency were set equal to one, and all frequency counts were rounded to the nearest whole percent. This statistic was selected to neutralize the effect of having weighted samples with enormous cell values, where the slightest percentage difference will generate very large χ^2 figures. For example, a weighted sample with a weighted cell count of one million deviating by 1% from the population cell count would result in a cell χ^2 of 100, which would not pass a goodness–of–fit test with degrees of freedom less than or equal to 20.

Implied in the selection of a χ^2 goodness–of–fit test based upon percentage distributions is the assumption that percentage counts are sufficient to represent the data. That is, for most applications using microdata, a cell percentage count of 20.5% is just as useful as knowing that the actual weighted frequency count is, for example, 287,000. Another important assumption for the goodness–of–fit test is that the appropriate cell–defining categories have been selected. For instance, if the categories for dividend income specified in Table 3.3A, 3.3D, and 3.3G are sufficient for any use of dividend income, then χ^2 figures based upon these categories are meaningful. However, it must be stated that for the purposes of this study, the categories were selected to have relevance with the restriction that low–count cells were avoided by aggregation. Consequently, for a small–frequency data item (such as "social security" used in Tables 3.3D–F), the categories (zero, $1 to $2999, and $3000+) were selected so that cross tabulated counts using the other variables, categories are reasonable.

The second test used in the study is the direct comparison of the Y–Z correlation matrix of a matched file with the population Y–Z correlation matrix. The comparison is displayed by subtracting the population correlation matrix from the matched file correlation matrix. Ideally, the matrix obtained would be zero or have all elements very close to zero, hence the matched-file-generated Y–Z distribution is statistically the same as the corresponding population Y–Z distribution. As will be presented later in the report, a direct test for equality of correlation matrices is not necessary because of large differences observed between the matched file and population statistics for "dividends" and "total income," and "family size" and "number of adults."

It must be noted that the matched file will be different to a certain extent from the population file since the samples which are used for matching are slightly different from the population.

3.3. Comparison of Absolute Difference and Mahalanobis Distance Functions

Matched files 1–20 identified in Table 3.6 can be used to compare matched files generated by an absolute difference distance function and a Mahalanobis distance function. Matched files 1–10 represent all pairwise matching of the five subsamples SIE1 through SIE5 selected from the population file using an absolute difference distance function. Matched files 11–20 represent all pairwise matching of the five subsamples SIE1 through SIE5 using a Mahalanobis distance function.

Only the Y–Z distributions will be examined since the transportation model leaves all original distributions in their original form. For example, the covariance matrix for X1–Y in the matched file is the same as the corresponding matrix in file A, and the covariance matrix for X2–Z in the matched file is identical to the corresponding matrix in file B.

Table 3.8 summarizes this χ^2 statistic for each of the nine frequency count tables representing all Y–Z distributions. The rows in Table 3.8 were arranged to allow a direct comparison of Models 1 and 2 for the same input data files. For example, matched files 1 and 11 given in the first two rows of the table are for input data files SIE1 and SIE2 where matched file 1 uses Model 1 and matched file 2 uses Model 2. The row averages are for the average χ^2 for a given matched file for the nine Y–Z frequency tables. In all cases the average for the matched file using Model 1 is less than the matched file using Model 2 generated from the same input data files.

The following table is given to illustrate one of the χ^2calculations in Table 3.8. Table 3.7 gives the percentage counts of records for the Y–Z distribution "total income" and "dividend income" in matched file 1. An interesting feature of constrained matching models is that the marginal distributions in Table 3.7 are identical to the marginal distributions of the original files. For example, the marginal distribution of the Y variable "total income" in this table is identical to the marginal Y distribution in sample SIE1,

Table 3.8
Contingency Table χ^2 Values Based on Population Percentages as Expected Values

Matched File	Total Income & Div.	Total Income & Fa.Size	Total Income & Race	S.S. & Div.	S.S. & Family Size	S.S. & Race	No. of Adults & Div.	No. of Adults & Fa. Size*	No. of Adults & Race	Row Average
1	1.3	4.0	2.2	1.5	2.4	.23	.78	37	.58	1.6
11	8.8	38.6	4.6	.7	13.8	1.0	13.0	306	3.75	10.5
2	5.3	7.3	3.0	1.2	6.8	.72	.81	165	.56	3.2
12	10.2	48.3	6.1	1.9	18.8	1.6	8.25	376	5.1	12.5
3	2.8	7.5	1.5	1.7	3.0	1.1	1.31	133	1.03	2.5
13	13.4	45.9	5.2	2.3	9.1	1.96	8.5	268	3.63	11.2
4	4.5	4.3	1.8	.8	2.9	.56	1.14	121	.86	2.1
14	12.0	48.6	3.7	2.2	16.1	1.5	10.0	336	2.74	11.8
5	1.3	6.3	3.2	1.8	3.4	.80	2.04	127	2.67	2.7
15	8.7	70.1	5.2	3.26	8.2	.58	7.67	387	3.78	13.4
6	2.3	6.2	7.4	3.5	2.3	1.7	2.67	154	1.93	3.5
16	8.8	75.8	4.0	2.83	11.0	1.96	5.97	331	3.46	14.2
7	1.4	8.2	3.2	1.9	3.2	2.0	1.72	127	1.03	2.8
17	6.2	67.3	4.2	3.6	14.6	2.4	6.75	405	1.87	13.4
8	2.9	9.3	.5	1.2	4.2	.13	1.97	134	1.43	2.7
18	11.2	66.0	3.6	2.67	20.5	.23	11.4	339	1.38	14.6
9	1.4	10.0	1.1	1.3	3.9	1.90	1.7	122	.71	2.8
19	8.4	66.2	3.6	3.6	16.7	1.80	6.32	348	2.81	13.7
10	1.8	5.8	.9	.81	1.3	.21	1.06	131	.58	1.6
20	14.2	63.3	2.6	1.7	14.9	1.58	7.34	378	3.7	13.7
avg.for 1-10	2.5 σ=1.4	6.9 σ=2.0	2.5 σ=2.0	1.6 σ=.8	3.3 σ=1.5	.9 σ=.7	1.5 σ=.6	125 σ=34	1.1 σ=.7	
avg.for 11-20	10.2 σ=2.5	59.0 σ=12.5	4.3 σ=1.0	2.5 σ=.9	14.4 σ=4.0	1.5 σ=.7	8.5 σ -2.3	347 σ=41	3.2 σ=1.1	
Degrees of Freedom	17	23	11	8	11	5	8	11	5	

*Number of Adults and Familv Size Omitted and the marginal distribution of the Z variable "dividend," in this table is identical to the marginal Z distribution in sample SIE2.

Table 3.7
Matched File 1
Total Income and Dividend Joint Distribution (Percentage Counts)

	Dividends		
Total Income	**$ 0**	**$1 - under $1,000**	**$1,000 Plus**
Under $5,000	20.5	1.42	.39
$5,000 under $10,000	17.27	2.31	.73
$10,000 under $15,000	17.47	2.34	.43
$15,000 under $20,000	11.91	2.40	.72
$20,000 under $25,000	7.30	2.32	.43
$25,000 Plus	6.79	4.07	.61

The population percentages for the corresponding Y-Z distribution for Table 3.7 are given in Table 3.3A and using the χ^2 figure previously defined, the resulted χ^2 is 1.3. With (6)(3)-1=17 degrees of freedom, a χ^2 of 1.3 indicates that the distribution in Table 3.7 is, for all practical purposes, the same distribution in Table 3.3A, and consequently for this Y-Z distribution the matched file is the same as the population file. However, this result is only true if the relevant categories are those in Tables 3.3A and 3.7, and percentage distributions are sufficient for the data being represented.

The degrees of freedom for the Y-Z distributions are given in the bottom row of Table 3.8. It is observed in the column for "total income" and "race," that the average χ^2 for matched files 1 through 10 is 2.5 with a standard deviation of 2.0. A rough interpretation of these figures is that the mean χ^2 plus two standard deviations = 2.5 + 4 = 6.5, which is an acceptable χ^2 figure given DF = 11. In fact for matched files 1-10, the mean χ^2 plus two standard deviations yields an acceptable χ^2 for all Y-Z distributions with the exception of "number of adults" and "family size."

The average χ^2 for the Mahalanobis distance function for each Y-Z distribution is given in the row averages for matched files 11 through 20. If two standard deviations are added to the mean χ^2 figures the resultant sum is an acceptable χ^2 in only five of the Y-Z tables.

In summary, it is observed from Table 3.8 using the χ^2 test that the absolute difference distance function is much better then than the Mahalanobis distance function. It is also observed that at the 5%

level of acceptance, that all but one of the absolute value distance function Y–Z distribution are acceptable.

Another way to compare matched files with the population file, and to compare one matching model with another is to observe the difference between the correlation matrix of a matched file and the correlation matrix of the population file. Table 3.9A gives the result of subtracting the population correlation matrix given in Table 3.1 from the average correlation matrix obtained from matched files 1–10.

The blocked–in portion of Table 3.8 represents the Y–Z distribution where for the ideal match all entries should be zero or close to zero. It is observed from Table 3.9A that these are significant differences from zero where the big differences are for the correlations between "total income" and "dividends," and between "number of adults" and "family size." The difference between the population and the matched file distribution for "family size" and "number of adults" were also very evident from Table 3.8. However, the difference between the population and matched file distributions for "total income" and "dividends" was probably masked by classifying all dividends over $1,000 in the class "$1,000 Plus." Another feature of the blocked portion of Table 3.9A is that six of the differences are negative and only one is positive, reflecting the fact that the matched file correlations are, on the average, smaller in absolute value than the population correlations.

Table 3.9A
Average Correlation Matrix for Matched Files 1–10 Minus the Population Correlation Matrix (Absolute Difference Distance Function)

Total Income	0					
Social Security	.02	0				
Number of Adults	.01	–.01	0			
Dividends	–.27	0	–.04	0		
Family Size	–.08	–.04	–.47	–.01	0	
Race	.02	0	–.01	–.01	.03	0

Table 3.9B gives the difference between the average correlation matrix for matched files 11–20 and the population correlation matrix given in Table 3.1.

From Table 3.9B it is observed that the Mahalanobis distance function produces larger deviations from the population than the absolute difference distance function represented in Table 3.9A. The blocked-in portion of Table 3.9B represents the Y–Z distributions and it is observed that correlation between "family size" and "number of adults," between "dividends" and "total income," and between "family size" and "total income" are very different from the population correlations. As observed in Table 3.9A, there is a strong tendency for the matched file Y–Z correlations to be smaller in absolute value than the population correlations. It is also observed from Tables 3.9A and 3.9B that the absolute value distance function is better than the Mahalanobis distance function.

Table 3.9B
Average Correlation Matrix for Matched Files 11–20 Minus the Population Correlation Matrix (Mahalanobis Distance Function)

Total Income	0					
Social Security	.02	0				
Number of Adults	.01	–.01	0			
Dividends	–.32	–.08	–.05	0		
Family Size	–.40	.12	–.77	.03	0	
Race	.12	.07	.08	–.01	.03	0

3.4 Comparison of Matched Files Generated with an Absolute Value Distance Function Using a Range of Common Variables

Earlier in this chapter matching Models 1, 3, and 4 were specified. Essentially Model 3 is the same as Model 1 with the data item "assets" left out. In the population file "assets" is strongly correlated (.57) with the common variable "interest," moderately correlated (.38) with the y variable "total income," and highly correlated (.70) with the z variable "dividends." Model 4 is the same as Model 3 with the common data items "age" and "sex" left out. In the population file age is moderately correlated (–.22) with the common variable "wages and salaries," moderately correlated (–.34) with the common variable "highest grade of family head," strongly correlated (.58) with the Y variable "social security," and moderately correlated (–.20) with the Z variable "family size."

The objective of this section is to compare Models 1, 3, and 4 consequently to investigate the effect of altering the number of variables in the distance function. To achieve this objective matched files 9, 40, and 47 are grouped together representing samples SIE5 and SIE4 matched respectively with matching Models 1, 3, and 4. Matched files 4, 42, and 45 are grouped together representing samples SIE5 and SIE2 matched respectively with Models 1, 3 and 4. Matched files 10, 43, and 44 are grouped together representing samples SIE1 and SIE5 matched respectively using matching Models 1, 3, and 4. Also matched files 7, 41, and 46 representing files SIE5 and SIE3 matched respectively using matching Models 1, 3, and 4.

Table 3.10 displays χ^2 statistics as defined previously to compare Models 1, 3, and 4 using pairwise matching of sample SIE5 with samples SIE1, SIE2, SIE3, and SIE4. From the table it is observed from the row average column that Model 4 yields the largest average χ^2 statistics in three of the four groupings. It is also observed from Table 3.10 that Model 4 has the largest column average in six of the nine frequency tables. Models 1 and 3 appear to generate matched files with the same overall differences from the population file.

Once again it is very obvious that the Y–Z distribution for "number of adults" and "family size" is very poor, but the other distributions are reasonable given that the samples are different from the population.

Models 1, 3, and 4 can also be examined using the average correlation matrices for matched files using the different models. The correlation results for Model 1 were given in the previous section in Table 3.8. The differences between the average correlation matrix using Model 3 and the population correlation matrix is given in Table 3.11. The difference between the average correlation matrix using Model 4 and the population correlation matrix is given in Table 3.12.

Table 3.10
Contingency Table χ^2 Values Based on Population Percentages as Base

Matched File	Total Income & Div.	Total Income & Fa.Size	Total Income & Race	S.S. & Div.	S.S. & Family Size	S.S. & Race	No. of Adults & Div.	No. of Adults & Fa. Size	No. of Adults & Race	Row Average
9	1.4	10.0	1.1	1.3	3.9	1.9	1.7	122	.71	2.8
40	4.3	12.9	1.1	.66	5.1	1.8	1.6	133	.83	3.5
47	3.3	14.9	.91	.89	14.8	1.0	.62	232	1.0	4.7
4	4.5	4.3	1.8	.80	2.9	.56	1.1	121	.86	2.1
42	3.8	4.6	1.3	.67	1.1	.45	2.0	88	1.2	1.9
45	11.8	7.3	2.5	.90	6.1	.83	1.2	214	.7	3.9
10	1.8	5.8	.9	.81	1.3	.21	1.1	131	.58	4.2
43	7.5	6.2	1.6	1.40	1.7	.43	2.1	114	.45	2.7
44	4.0	11.9	.72	1.83	7.4	2.3	1.1	223	1.1	3.8
7	1.4	8.2	3.2	1.9	3.2	2.0	1.7	127	1.0	2.8
41	2.3	4.3	4.7	1.8	2.6	2.2	.56	106	1.0	2.4
46	4.1	11.4	4.4	2.3	6.5	.93	2.1	196	.4	4.0
avg.for 9,4,10&7	2.3 σ=1.5	7.1 σ=2.5	1.8 σ=1.0	1.2 σ=.5	2.8 σ=1.1	1.2 σ=.92	1.4 σ=.35	125 σ=46	.8 σ=.2	
avg.for 40 42,43&44	4.5 σ=2.0	7.0 σ =4	2.2 σ=1.7	1.1 σ=.6	2.6 σ=1.8	1.2 σ=.9	1.6 σ =.7	110 σ=19	.9 σ =.3	
avg.for 47 45,44&46	5.8 σ=4.0	11.4 σ =3.1	2.1 σ=1.7	1.5 σ=.7	8.7 σ=4.1	1.3 σ=.7	1.3 σ =.6	216 σ=15	.8 σ =.3	
Degrees of Freedom	17	23	11	8	11	5	8	11	5	

*Number of Adults and Family Size Omitted

Table 3.11
Average Correlation Matrix for Matched Files 40–43 Minus the Population Correlation Matrix (Model 3 Distance Function)

Total Income	0					
Social Security	−.02	0				
Number of Adults	−.01	.02	0			
Dividends	−.27	.02	.02	0		
Family Size	−.07	.04	.45	0	0	
Race	.02	.02	.05	.01	.04	0

Table 3.12
Average Correlation Matrix for Matched Files 44–47 Minus the Population Correlation Matrix (Model 4 Distance Function)

Total Income	0					
Social Security	−.02	0				
Number of Adults	−.01	.02	0			
Dividends	−.25	.01	−.03	0		
Family Size	−.11	−.01	−.58	0	0	
Race	.04	.04	.01	−.01	.04	0

The blocked–in portions of Tables 3.11 and 3.12 reflect the difference between the Y–Z distributions in the matched files using Models 3 and 4 and the population Y–Z distributions. As mentioned in the previous section, the ideal match would have zero or near zero entries. However, as in the case for Model 1 observed in Table 3.8, there are significant differences in the matched correlations for dividends and total income, and for family size and number of adults. These large differences are a result of the tendency for the matched file correlations to have smaller absolute values than the absolute values of the corresponding correlations in the full population. The population correlation for "dividends" and "family income" is .33 as opposed to the average Model 3 corresponding correlation of .06, and the corresponding correlation from the average results from Model 4 of .05. For "family size" and "number of adults" the population correlation is .76, the average Model 3 correlation is .31, and the average Model 3 correlation is .18.

It should be noted that the entries in Tables 3.8, 3.11, and 3.12 outside the blocked–in section are for the correlations within the Y's (given above the blocked–in portion), and for the correlations within the

Z's (given to the right of the blocked-in portion). Any nonzero entries outside the blocked-in portions are a consequence of differences between the samples and the population, since the transportation algorithm forces the within-Y and within-Z correlations to be the same as the sample values.

In summary, the results of this section indicate that Models 1 and 3 are better than Model 4. The implication for matching is that it is possible to have too few common variables in the distance function. However, because of the mixed results from Models 1 and 3, it cannot be stated that too many common variables can degrade the accuracy of a generated matched file.

3.5 Matching Under Conditions of Noise and Bias

Samples SIE1 and SIE5 were matched using Model 1 and the results of this match are designated as matched file 10 in Table 3.7. Earlier in this chapter, samples 6A–8B were identified as versions of SIE5 with noise and bias injected into the X variable assets. The purpose of this section is to compare matched file 10 with matched files 31–38 which are identified in Table 3.6. In all cases, matching Model 1 is used.

Table 3.13 displays the χ^2 statistic as defined in the previous two sections for comparing sample SIE1 matched with SIE5 when bias and noise are injected into SIE5. Match file 10 is for the unaltered sample SIE5, matched file 31 has "assets" in SIE5 reduced by 20%, matched file 32 has "assets" in SIE5 reduced by 10%, matched file 33 has a 25% noise factor in "assets" in 25% of the records in SIE5, matched file 34 has a 25% noise factor in "assets" in all records in SIE5, matched file 35 has a 10% noise factor in "assets" in 25% of the records in SIE5, matched file 36 has a 10% noise factor in "assets" in all SIE5 records, matched file 37 has an average 15% downward bias and noise factor in "assets" in 25% of the records in SIE5, and matched file 38 has an average 15% downward bias and noise factor in "assets" for all records in SIE5.

In Table 3.13 it is observed that the row average for matched file 10 is slightly better than the row averages for matched files 31–38. It is also observed from the column averages of matched files 31–38 when compared with the row entries for matched file 10 that, on the average, χ^2 statistics for matched file 10 are better than the average tables for the matched files 31–38.

Once again, as in the two previous sections, the Y–Z distribution for "number of adults" and "family size" are very poor, and most of the other distributions are reasonable. The empirical result taken from Table 3.13 is that moderate amounts of noise, bias, and combined noise and bias do not greatly affect the Y–Z distributions in matched files.

Table 3.13
χ^2 Statistics for Bias and Noise Tests

Matched File	Total Income & Div.	Total Income & Fa.Size	Total Income & Race	S.S. & Div.	S.S. & Family Size	S.S. & Race	No. of Adults & Div.	No. of Adults & Fa. Size	No. of Adults & Race	Row Average*
10	1.8	5.8	.9	.81	1.3	.21	1.1	131	.58	1.6
31	2.3	6.4	1.1	.35	1.8	.21	.73	133	.51	1.7
32	1.5	8.4	1.6	.33	1.8	.21	1.1	133	.76	2.0
33	2.8	6.1	1.3	.75	1.3	.21	.73	137	.58	1.7
34	1.6	7.9	1.2	.25	1.5	.21	.81	112	.75	1.8
35	2.3	7.1	1.3	.75	1.3	.21	1.2	132	.58	1.8
36	2.8	6.9	1.2	.33	1.3	.21	1.6	138	.78	1.9
37	1.1	6.0	1.2	1..6	1.8	.21	.73	117	.58	1.7
38	2.1	7.9	.78	.25	1.8	.21	.73	134	.76	1.8
avg.for 31–38	2.1 σ =.62	7.1 σ =.9	1.2 σ =.23	.58 σ=.46	1.6 σ =.25	.21 σ=0	.95 σ=.32	130 σ =9.6	.66 σ=.11	
Degrees of Freedom	17	23	11	8	11	5	8	11	5	

*Number of Adults and Family Size Omitted

3.6 Results of Matching A Sample With Itself

The sample SIE5 was matched with itself under a variety of conditions, i.e., using all of the matching models and using the conditions of bias, noise, and combined noise and bias. The files of particular interest are described in Table 3.6 as matched file 21 (Model 1), 22 (Model 2), 39 (Model 3), 48 (Model 4), 49 (Model 5) and 50 (Model 6) and matched files 23–30 (noise and bias tests).

Matched files 39, 48, 49, and 50 (generated under conditions of a reduced set of common variables) are unique in that they are identical with sample SIE5; that is, each sample record is matched with itself and with the matched weights equal to the record weights. There are 951 records in sample SIE5 and consequently there are 951 records in matched files 39, 48, 49, and 50. The correlation matrix for each of these files is identical to correlation matrix for SIE5.

Matched file 21 is slightly different from sample SIE5. However, this matched file has 974 records which implies that all except 24 records have been matched with themselves. These 24 exceptions have been split and cross–matched with each other. The consequences of the "mismatching" of 24 records can be observed in Table 3.14 where it is seen that the correlations are almost identical with the exception of "race" and "family size" which has an approximate difference of –.01. However, the percentage frequency in the table for "race" and "family size" are different between SIE5 and the matched file by less than .1%.

Thus the results of matching a file with itself are perfect using Models 3, 4, 5, and 6 and near–perfect using model 1. However, the results obtained using model 2, the Mahalanobis distance function, are very poor by comparison. The difference between the correlation matrix from matched file 22 and SIE5 is given in Table 3.15.

Table 3.14
Correlation Difference Matrix for Matched File 21 Minus SIE5 (Model 1)

Total Income	0					
Social Security	0	0				
Number of Adults	0	0	0			
Dividends	0	0	0	0		
Family Size	0	−.001	−.004	0	0	
Race	−.001	.004	−.009	0	0	0

From Table 3.15 it is observed that matched file 22 is very different from SIE5. These differences are probably due to the non-normal and discrete data distributions in the sample.

Table 3.15
Correlation Difference Matrix for Matched File 22 Minus SIE5

Total Income	0					
Social Security	0	0				
Number of Adults	0	0	0			
Dividends	.365	−.071	.039	0		
Family Size	−.49	.094	−.891	0	0	
Race	.173	.079	.071	0	0	0

The impact of bias and noise on sample data can be studied by comparing the characteristics of matched files 23–30. In all cases, the results are nearly identical to matched file 21, with correlation differences from matched file 21 less than .003. Each of these output files only differ from each other by less than 7 of the 951 records in SIE5 with the number of matched records ranging from 967 to 974. The conclusion is that limited amounts of bias, noise, and combined bias and noise do not affect a matched file.

3.7 Analysis of Unconstrained Procedures

To investigate the impact of unconstrained procedures on the resultant composite file, a single sample file, SIE2, was merged with each of the four remaining sample files using an unconstrained method with the absolute difference distance function of Section 3.2.1. In each case, the weights for SIE2 are not constrained and the other file is used as the base file. By observing the distributional statistics of a Z-variable, the effects of dropping the weight constraints are demonstrated.

The means and standard deviations of the Z-variable dividend income are shown in Table 3.16A for all five sample files. Table 3.16B shows the same statistics for SIE2 when used as file B. Not only do the means vary, depending on the base file, but the standard deviations are also distorted, as much as ± 38 percent from the original values.

Noise and bias factors also influence the pattern of data in the X2 and Z variables as illustrated in Table 3.16C. Files SIE6A through SIE8B are identical to file SIE5 but with the X2 variables perturbed with noise or bias as described in Table 3.1. Table 3.16C demonstrates the impact of the X-data perturbations on the same fundamental statistics. The presence of bias or noise tends to decrease the Z-variance and distort even the means either upwards or downwards.

Of course, all of these statistical changes are a result of the implicit modification of the weights on file B by the unconstrained merge process. Such variations are in contrast with constrained procedures which, as an integral part of the merge process, maintain the original (or equivalent) record weights and hence preserve all of the X2 and Z data items and their interrelationships.

Table 3.16A
Dividend Income Statistics for Sample Files

	Dividend Income	
Sample File	Mean	Standard Deviation
SIE1	$227	$1,471
SIE2	194	1,717
SIE3	350	1,942
SIE4	353	2,929
SIE5	292	1,990

Table 3.16B
SIE2 Dividend Income Statistics
After Unconstrained Merging

		Dividend Income from SIE2			
File A	File B	Mean	(Deviation*)	Standard Deviation	(Deviation*)
SIE1	SIE2	$176	(-9.2%)	$1,057	(-38.4%)
SIE3	SIE2	268	(+38.1%)	1,922	(+11.9%)
SIE4	SIE2	146	(-24.7%)	1,454	(-15.3%)
SIE5	SIE2	186	(-4.1%)	1,210	(-29.5%)

*Percent deviations from original SIE2 values per Table 4.16A

Table 3.16C
Unconstrained Merges: Noise and Bias Tests

		Dividends of File B	
File A	File B	Mean	Standard Deviation
SIE1	SIE5	$220	$1,713
SIE1	SIE6A	317	1,557
SIE1	SIE6B	294	1,433
SIE1	SIE7A	283	1,418
SIE1	SIE7B	216	1,028
SIE1	SIE7C	258	1,371
SIE1	SIE7D	286	1,475
SIE1	SIE8A	246	1,149
SIE1	SIE8B	295	1,992
SIE5 Original File		**292**	**1,990**

In summary, this information seems to suggest not only that unconstrained approaches have difficulty maintaining the basic descriptive statistics of Z-variables, but are also influenced by bias and noise in the X-variables, two problems not encountered in constrained procedures.

3.8 Summary and Results of the Real Data Empirical Investigation

All of the issues outlined in the introduction of this chapter were addressed using the fifty constrained matched files defined in Section 3.2.4 and the unconstrained matched files discussed in Section 3.7. The pattern of the results obtained indicate that sufficient observations have been generated for some general conclusions.

These results are:

1. Absolute difference distance function yields significantly better results than the Mahalanobis distance function.
2. Noise and bias have a nominal effect on constrained matched files.
3. When a file is matched with itself using the transportation model with an absolute differences distance function, the desired matching of records is produced even under conditions of bias, noise, and combined noise and bias.
4. All original statistical content in the input files is preserved with constrained matching. However, there is a tendency for the absolute value of correlations between the Y-Z items to be reduced from the population values.
5. The quality of a match is reduced if too few common variables are used in the distance function.

6. The absolute difference distance function generated acceptable Y–Z distributions in seven of the nine Y–Z distributions specified, using the correlations in Table 3.9A and the percentage distribution functions and categories specified in Tables 3.3A–I.

It is extremely interesting to note that the absolute difference distance function used with the transportation model generated the unacceptable Y–Z distributions when cov(YZ|X) was clearly non–zero.

Perhaps the most important conclusion is that the next applied research topic in this area should be the identification and development of a matching criterion which has the "known" Y–Z relationship included. That is, the matching function should exclude non–valid Y–Z patterns while encouraging the valid ones. This conclusion is based upon the empirical evidence that the transportation model using an absolute difference distance function produces acceptable Y–Z distributions in most situations, but not in all. Also it is reasonable to predict that in a proper environment with the necessary matching software and data that a matching function could be developed which would generate acceptable distributions for all Y–Z pairs.

The long–run implication of this conclusion is that statistical matching would be a very useful tool for data preparation in situations where, for example, every five years population Y–Z characteristics are observed from a sample, and during the intervening years file matching is done when the data is available only in X1–Y and X2–Z distributions. In this situation the population Y–Z characteristics are used to match X1–Y and X2–Z files such that the matched file Y–Z distribution conform to expected patterns.

PART 4.
SUMMARY AND CONCLUSIONS

The objective of this study was to empirically measure the quality of statistical matching. To achieve this result many statistical matches were generated and the properties of these matches were objerved and analyzed.

Statistical matching creates a composite microdata file from two original microdata files A and B. The records in the composite field are formed by appending the records from file B onto the recordds in file A. Statistical matching algorithms identify the record weight to be associated with the record formed by appending the j–th record from file B onto the i–th record in file A. The statistical matching procedures currently used in large–scale matching projects are described in Section 1 fo this report. The statistical framework and the experimental design are given in Section 2 and the statistical results in Section 3.

This study has been presented one of the first research studies of its kind in this area: an in–depth computational work designed to achieve a foothold of understanding into the statistical nature of files formed by state-of-the-art merging techniques. The study views from dual aspects — theoretical and direct-empirical — an important set of questions unanswered by the literature. An enormous body of data has been created and analyzed; in the process, over fifty linear programming problems were solved with dimensions of up to 2,000 constraints and 1 million variables.

Details of this research effort have been documented in the previous pages and, in this concluding section, several general conclusions suggested by these results are presented. These interrelated summary and conclusions are organized under the following headings: (1) the viability of merging, (2) choosing a merge technique and distance function, (3) the effects of data perturbations, (4) improvements needed in existing methods, and (5) future research directions.

4.1 The Viability of Merging

There are several instances suggested by this study in which specific statistical merging techniques perform well but others where merging to accomplish certain goals is perhaps not advisable. The study focused on the various merge techniques' abilities (or lack thereof) to preserve known relationships between data items that came from and were unique to separate files. These relationships were expressed in the form of correlation and covariance statistics and cross-tabulations of pairs of such items. These are the so-called Y-Z relationships.

There is evidence to suggest that applications requiring that Y-Z relationships be preserved in "modestly broad" categories can be obtained with generally good results from a merge file created by the transportation model and the weighted absolute differences distance function described in Section 3.2.1. While data categories such as "wages between $5,000 and $10,000" and "age between 20 and 30 years" would be considered "modestly broad," the categories "wages income between $5,000 and $5,100" and "age of 21 years" would not. Therefore many existing microsimulation models, such as the Treasury's Individual Tax Model, which do not have extremely strict requirements in this area are well-suited to the use of merged files.

This is not to suggest that there cannot be any problems with using such files or that improvements cannot be made in their construction. The empirical study using SIE data demonstrated the ability of merged files to create acceptable Y-Z data relationships in most cases, but also provided an excellent example of the creation of illogical relationships. Specifically, several records were created to form single-person families containing two adults. While such spurious results lead to reasonable concerns about merging, it is clear that these erroneous record matches could have been easily avoided by adding a penalty to the process's distance function for each illogical match pair. The extension of this notion to less clear-cut cases is discussed below.

Another application of merge files is for correcting or expanding an existing file's item to account for underreporting. When, for example, items in a given file are not deemed sufficiently trustworthy, that file might be merged with another primarily to upgrade those particular items. If the files have many items in common, this use might be focused on either X1–X2 relationships or X1–Z relationships. Such relationships seem to be retained by merging.

In the instance where there is no relationship between the non-common items, for a given set of values for the common items [i.e. $\mathrm{cov}(Y, Z|X)=0$], the Y–Z distribution is replicated well by merge techniques. Of course this situation is not as useful as where there is such a relationship [$\mathrm{cov}(Y, Z|X) \neq 0$], a case current merging techniques have difficulty replicating. While it might be thought that all nonzero relationships are forced to zero by the merge techniques studied, the testing based on SIE data indicated that such relationships were only softened, not eliminated. This means that relationships which are not an explicit part of the merge procedure (e.g., in the distance function) tend to be attenuated through random pairings but on average hold to a certain, although lesser, degree. Hence, the user of merge files should be cautioned about heavy reliance on them if a high degree of accuracy is required.

4.2 Choice of Merging Model

In summary, the study indicated that the best results can be obtained by applying an optimal-constrained merge model with an absolute difference distance function. Testing with "real" data files verified the superiority of the constrained approach but found the Mahalanobis distance function to yield extremely poor results, most likely a consequence of the presence of non-normally distributed data.

The number of common variables used in the distance function was also shown to have a strong effect on the representativeness of matched files. As expected, more variables seemed better than fewer, perhaps due to the procedures' inability to distinguish between records when only a few variables are uses.

4.3 Effects of Data Perturbations

A notable finding from the SIE data study was the robustness of the constrained merge techniques when the variables used in a distance function are subjected to noise, bias, and both noise and bias. When the transportation model with an absolute differences distance function was used to match a sample file with itself, 99 percent of all records were matched correctly, even under varying levels and types of noise and bias. This lack of sensitivity to such prevalent conditions of sample survey data is a very positive result that enhances the attractiveness of merging schemes in general.

4.4 Improvements Needed in Existing Methods

It appears likely that current techniques, including the transportation model, could be improved with a modicum of additional research. For example, distance functions should be designed to account for known relationships among non-common variables and at the very least to rule out illogical matches.

It is very clear from many aspects of the research that for merging methods to perform well in maintaining Y–Z relationships, the procedures must inject some measure of control over those relationships. In the simplest case, distance functions should associate heavy penalties with record matches that are illogical not only in terms of X1–X2 data configurations but for Y–Z combinations as well.

In the more general case, the higher–order statistical relationships between X, Y, and Z items should be incorporated into the merging procedures. Research to identify more sophisticated distance functions or matching schemes which directly address data characteristics such as non–normality and $\mathrm{cov}(Y, Z|X) \neq 0$ should be undertaken to counteract shortcomings inherent in procedures which are quite robust along other dimensions.

From a procedural point of view, it would be extremely useful for all merge file users if aggregate statistics for non–common variable pairs could be collected at regular (five– or ten–year) intervals in order to calibrate on–going merge models. Such information could be collected piecemeal and at various points in time for subsequent construction and maintenance of these statistical mosaics. For example, the correlations between some item pairs probably would not change dramatically from year–to–year and would need to be updated or verified only over large time intervals. However, if such statistics were available, new merging schemes could likely be designed to incorporate them and perhaps eliminate the problems associated with $\mathrm{cov}(Y, Z|X)$ significantly different from zero.

4.5 Further Research Directions

In addition to the research topics described above, it is felt that this work is only a starting point for research into the theory and practice of microdata file merging. In general, this line of research should be continued (1) to devise merge methodologies which are better able to capture the Y–Z relationships, (2) to identify criteria to determine when a pair of files can be said to be "mergeable," and (3) to study the impact of merge technique at the model output level, as opposed to the data input level, to gauge models sensitivities to data perturbations from this source.

A very different line of research would be to explore the applicability of new "learning" techniques to merging and imputation problems. For example, neural network models can be "trained," through exposure to a large set of examples, to identify highly nonlinear relationships between variables. A typical neural network model takes a set of input values and performs weighted computations to create a set of output values. The network is trained by repeatedly comparing desired outputs with computed ones and, if there is a difference, adjusting the model's internal weights. This is repeated until the model can adequately reproduce the desired outputs from a given set of input values. The ability of such a model to handle "noisy" data has made it useful in image processing and pattern recognition applications.

Such a model might be used to learn the interrelationships among variables from a sample, for use in massive imputation. The model might be trained with (X1, Y) data, using the X1's as input variables and the Y's as desired output variables, so as to automatically uncover the X–Y interrelationships. The model could then process a second file of (X2, Z) values, using the X2's as inputs, and use the computed–Y outputs to for (X2, Z, Y) records. The research would involve not only model construction and training, but a statistical analysis of the resulting imputed file, and a comparison with merged model results.

BIBLIOGRAPHY AND REFERENCES

[1] Alter, Horst E. (1974). "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970." *Annals of Economic and Social Measurement* (April)2: 373–394.

[2] Anderson, Erling B. (1980). *Discrete Statistical Models with Social Science Application*, North–Holland, Amsterdam.

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

[4] Barr, Richard S. (1981). "Design of Experiments to Investigate Joint Distributions in Micro–analytic Simulations," in *Proceedings of the 13th Conference on the Interface of Computer Science and Statistics*, Springer–Verlag, New York.

[5] Barr, Richard S. (in press). "Solution Strategies and Algorithm Behavior in Large–Scale Network Codes," in J. Mulvey, ed. *Testing and Validating Mathematical Programming Algorithms and Software*, Springer–Verlag, New York.

[6] Barr, Richard S., F. Glover, and D. Klingman (1977). "The Alternating Basis Algorithm for Assignment Problems." *Mathematical Programming*, 13: 1–13.

[7] Barr, Richard S., F. Glover, and D. Klingman (1978) "A New Alternating Basis Algorithm for Semi–Assignment Networks." In W.W. White, ed., *Computers and Mathematical Programming*, U.S. Government Printing Office, Washington, D.C.

[8] Barr, Richard S., F. Glover, and D. Klingman (1978). "The Generalized Alternating Path Algorithm for Transportation Problems." *European Journal of Operational Research*, 2: 137–144.

[9] Barr, Richard S., F. Glover, and D. Klingman (1979). "Enhancements to Spanning Tree Labeling Procedures for Network Optimization." *INFOR*, 17, 1: 16–34.

[10] Barr, Richard S. and J. Scott Turner (1978). "A New Linear Programming Approach to Microdata File Merging." In *1978 Compendium of Tax Research* sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. (Barr and Turner's reply to Goldman also appears in that volume.)

[11] Barr, Richard S. and J. Scott Turner (1978). "New Techniques for Statistical Merging of Microdata Files." Paper prepared for the Conference on Microeconomics Simulation Models for the Analysis of Public Policy, National Academy of Sciences, March.

[12] Barr, Richard S. and J. Turner (1980). "New Techniques for Statistical Merging of Microdata Files." In R. Haveman and K. Hollenbeck, eds., *Microeconomic Simulation Models for the Analysis of Public Policy*, Academic Press.

[13] Barr, Richard S. and J. Scott Turner (1980). "Merging the 1977 Statistics of Income and the March 1978 Current Population Survey." prepared for the Office of Tax Analysis, U.S. Department of the Treasury.

[14] Barr, Richard S. and J. Scott Turner (1981). "Microdata File Merging Through Large–Scale Network Technology." *Mathematical Programming Studies*, 5: 1–22.

[15] Budd, Edward C. (1971). "The Creation of a Microdata File for Estimating the Size Distribution of Income." *Review of Income and Wealth* (December) 17: 317–333.

[16] Budd, Edward C. (1972). "Comments." *Annals of Economic and Social Measurement* (July) 1: 349–354.

[17] Budd, Edward C. and Daniel B. Radner (1969). "The OBE Size Distribution Series: Methods and Tentative Results for 1964." *American Economic Review* (May) LIX: 435–449.

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

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

[20] Colledge, M.J., J.H. Johnson, R. Pare, and I.G. Sande, "Large Scale Imputation of Survey Data," *1978 Proceedings of the American Statistical Association, Survey Research Methods Section*, (1979) 431–436.

[21] Conover, W.J. (1971). *Practical Nonparametric Statistics*, John Wiley and Sons.

[22] Goldman, Alan J. (1978). "Comment." In *1978 Compendium of Tax Research* sponsored by the Office of Tax Analysis, U.S. Department of the Treasury.

[23] Green, Paul E. (1978). *Analyzing Multivariate Data*, Dryden Press.

[24] Greenberger, Martin, Matthew Crenson, Brian Crissey, *Models in the Policy Process*, Russell Sage Foundation, New York (1976).

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

[26] Kadane, Joseph B. (1978). "Some Statistical Problems in Merging Data Files." In *1978 Compendium of Tax Research* sponsored by the Office of Tax Analysis, U.S. Department of the Treasury. (Kadane's reply to Sims also appears in that volume.)

[27] Kilss, Beth and Fritz Scheuren (1978). "The 1973 CPS–IRS–SSA Exact Match Study: Past, Present, Future." Paper presented at the NBER Workshop on the Uses of Social Security Research Files, March 15–17.

[28] Makesh, S. and J. Scott Turner (1980). "Statistical Analysis of Oklahoma Microdata Files," working paper, Office of Business and Economic Research, Oklahoma State University.

[29] Minarik, Joseph J., "The MERGE 1973 Data File," in R.H. Haveman and K. Hollenbeck, *Microeconomic Simulation Models for Public Policy Analysis*, Academic Press (1980).

[30] Mulvey, John M. (1980). "Reducing the U.S. Treasury's Taxpayer Data Base by Optimization." *Interfaces* (October) 10: 101–112.

[31] Okner, Benjamin A. (1972). "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File." *Annals of Economic and Social Measurement* (July) 1: 325–342. (Okner's reply to comments also appear in that issue.)

[32] Okner, Benjamin A. (1974). "Data Matching and Merging: An Overview." *Annals of Economic and Social Measurement* (April) 2: 347–352.

[33] Peck, Jon K. (1972). "Comments." *Annals of Economic and Social Measurement* (July) 1: 347–348.

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

[35] Radner, Daniel B. (1978). "Age and Family Income." Paper presented at the NBER Workshop on Policy Analysis with Social Security Research Files, Williamsburg, Virginia, March 15–17. Mimeographed.

[36] Radner, Daniel B. (1978). "The Development of Statistical Matching in Economics." *Proceedings 1978 American Statistical Association, Social Science Section*, San Diego, August 16.

[37] Radner, Daniel B. (1980). "An Example of the Use of Statistical Matching in the Estimation and Analysis of the Size Distribution of Income," working paper, Office of Research and Statistics, Social Security Administration.

[38] Radner, Daniel B. and Hans J. Muller (1978). "Alternative Types of Record Matchings: Costs and Benefits." *1977 Proceedings of the ASA, Social Statistics Section.*

[39] Rao, C. Radhakrishna (1952). *Advanced Statistical Methods in Biometric Research*, John Wiley, New York.

[40] "Report on Exact and Statistical Matching Techniques," (1980). Statistical Policy Working Paper 5, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce.

[41] Ruggles, Nancy and Richard Ruggles (1974). "A Strategy for Merging and Matching Microdata Sets." *Annals of Economic and Social Measurement* (April) 2: 535–372.

[42] Ruggles, Nancy, Richard Ruggles, and Edward Wolff (1977). "Merging Microdata: Rationale, Practice and Testing." *Annals of Economic and Social Measurement* (Fall) 6: 429–444.

[43] Siegel, Sidney (1956). *Nonparametric Statistics for the Behavioral Sciences*, McGraw–Hill, New York.

[44] Sims, Christopher A. (1972). "Comments." *Annals of Economic and Social Measurement* (July) 1: 343–346. (Sims' "Rejoinder" also appears in that issue.)

[45] Sims, Christopher A. (1974). "Comments." *Annals of Economic and Social Measurement* (April) 2: 395–398.

[46] Sims, Christopher A. (1978). "Comments on Kadane's Work on Matching to Create Synthetic Data." In *1978 Compendium of Tax Research* sponsored by the Office of the Tax Analysis, U.S. Department of the Treasury.

[47] Subcommittee on Matching Techniques, Federal Committee on Statistical Methodology, "Report on Exact and Statistical Matching Techniques," Statistical Policy Working Paper 5, U.S. Department of Commerce (1980).

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

[49] Upton, Graham J.G. (1978). *The Analysis of Cross–tabulated Data*, John Wiley, New York.

[50] Wolff, Edward N. (1977). "Estimates of the 1969 Size Distribution of Household Wealth in the U.S. from a Synthetic Database." Paper presented at the Conference on Research in Income and Wealth, Williamsburg, Virginia, December.

[51] Yamane, Taro (1967). *Statistics, an Introductory Analysis*, Harper and Row.

APPENDIX A

Detailed Description of a Small Matching Problem Using Constrained and Unconstrained Methodologies

It can be shown algebraically and understood intuitively that the statistical merging technique chosen directly affects the statistical structure of the resultant composite matched file. The objective of this appendix is to illustrate these effects using two small hypothetical data files.

An Example Matching Problem

The data files, called A and B, that we will use in our examples have three records and four records respectively. These records are completely described in Figure A.1. Note that files A and B have some items in common and some that are not. The objective of merging would be to form a file of composite records, each containing items from both files, as depicted in Figure 1.1. As in all merging and matching techniques, the common items are used for identifying records with like attributes for matching purposes.

The tabulations given in Table A.1 show that the weight totals for each file are equal, indicating identically-sized sample populations. If we let a_i be the weight of the i-th record in file A and b_j be the weight of the j-th record in file B, this property can be expressed mathematically as:

$$a_1 + a_2 + a_3 = b_1 + b_2 + b_3 + b_4. \qquad (A.1)$$

Table A.1 also indicates that the weighted item sums are slightly different, indicating reporting or sample variations.

Overview of the Matching Problem

We shall let w_{ij} represent the weight assigned to the composite record formed by merging record i of file A with record j of file B, with a zero value indicating that the records are not matched. Microdata file merging may be viewed as a problem of finding a set of nonnegative values for all w_{ij}'s.

In order to guide the merge process to matching similar records a *distance function*, d, is used to measure the extent to which the attributes in any one record differ from the same attributes in another record. Intuitively, the parameter d_{ij} can be viewed as the "distance" between record i of file A and record j of file B, as illustrated in Figure A.2 below. In this example, file A record 2 (shown as point A2) is considered to be closer to file B record 1 (B1) than to file B record 2 (B2), that is $d_{21} < d_{22}$, since the schedule codes and AGI values are in closer agreement.

A simplistic distance function will be used for illustrative purposes. (The effects of different dissimilarity metrics could also be studied using this example.) In this model, the interrecord distance will be

Figure A.1. Example Files A and B

FILE A RECORDS:

Record Number	Record Weight	Schedule Code	Reported Adjusted Gross Income	Reported Deductions
1	1000	1	16,000	3,200
2	2000	1	12,000	2,300
3	500	2	20,000	4,000

Common Items | *Non-common Items*

FILE B RECORDS:

Record Number	Record Weight	Schedule Code	Reported Adjusted Gross Income	Family Size	Transfer Income
1	1400	1	14,000	2	500
2	400	2	19,500	4	0
3	1500	1	11,000	3	3,000
4	200	2	17,000	2	0

Table A.1

Item Tabulations for Example Files
(Weighted)

Description	File A	File B
Population size	3,500	3,500
Schedule code = 1	3,000	2,900
Schedule code = 2	500	600
AGI, Total ($000s)	50,000	47,300
Reported Deductions, Total	9,800	n.a.
Family size, avgerage	n.a.	2.65
Transfer income ($000s)	n.a.	5,200

defined as

$$d_{ij} = \frac{|(\text{File A AGI}_i) - (\text{File B AGI}_j)|}{100} + \begin{cases} 0, \text{ if schedule codes agree} \\ 25, \text{ if schedule codes differ.} \end{cases}$$

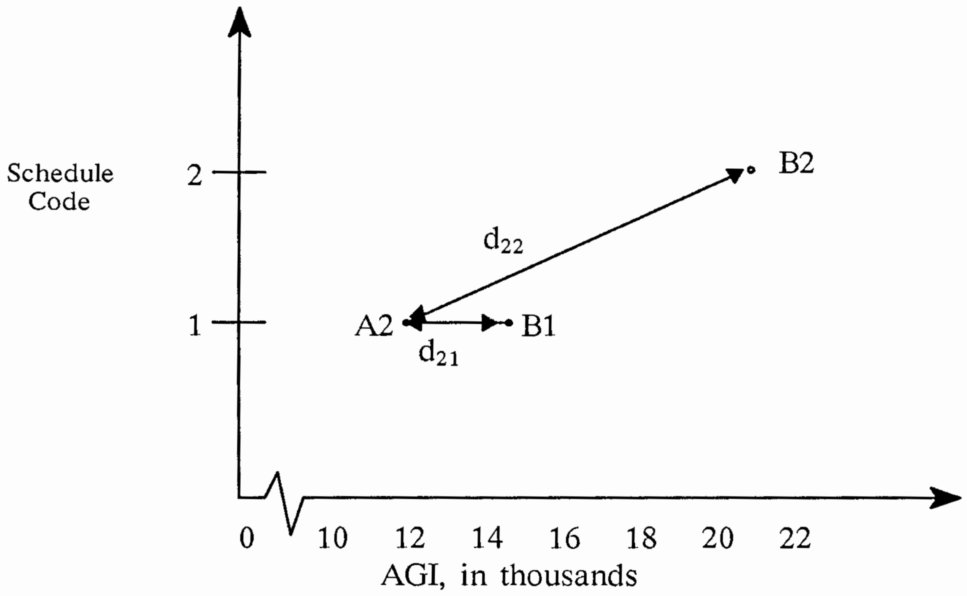


Figure A.2. Scatter Diagram of Selected Records

The following tableau can be used to summarize the matching problem. With a row for each record i in file A and a column j for each record in file B, a tableau cell (i, j) corresponds to a match possibility and an w_{ij} value. We will indicate a record match by including the composite record weight in a cell ($w_{ij} > 0$) and use a blank cell to mean that the records are not matched ($w_{ij} = 0$) . The box inset in each cell contains d_{ij} , the distance function value. Row and column totals reflect the weights associated with the record.

Problem Constraints

Of course any values could be assigned to the w_{ij} variables. However, since the record weights are an integral part of any computations made with the data items, these composite record weights directly affect the merge file's numerical structure. For this reason, we may wish that the sum of the w_{ij} values for any record in file A to equal the original record weight, thereby not overmatching or undermatching that record and preserving that record's intrinsic data structure. In our example, this translates to the following set of constraints that we may wish to be in force in our solution to the merge problem:

Tableau 0. Sample Tableau for Example Merge Problem

FILE		FILE B RECORD			
A RECORD	1	2	3	4	a_i
1	20 w_{11}	50 w_{12}	50 w_{13}	35 w_{14}	1000
2	20 w_{21}	100 w_{22}	10 w_{23}	75 w_{24}	2000
3	85 w_{31}	5 w_{32}	115 w_{33}	30 w_{34}	500
b_j	1400	400	1500	200	

$$\begin{aligned} w_{11} + w_{12} + w_{13} + w_{14} &= a_1 \\ w_{21} + w_{22} + w_{23} + w_{24} &= a_2 \\ w_{31} + w_{32} + w_{33} + w_{34} &= a_3 \end{aligned} \qquad (A.2)$$

If, in addition, we wish to place the same conditions on the file B weights, we could also require:

$$\sum_{i=1}^{3} w_{ij} = b_j, \quad \text{for } j = 1,\ 2,\ 3,\ \text{and } 4. \qquad (A.3)$$

Since we assume that negative weights are not permitted, we always require the constraints:

$$w_{ij} \geq 0, \quad i = 1,\ 2,\ 3 \ \text{and}\ j = 1,\ 2,\ 3,\ 4 \qquad (A.4)$$

Also, we may wish to use the distance function values to achieve a best overall solution, so that our objective would be to require that the merge process:

$$\text{minimize} \quad d_{11}w_{11} + d_{12}w_{12} + \ldots + d_{34}w_{34}. \qquad (A.5)$$

In so doing, we minimize the aggregate interrecord distance for the entire file.

Merging Techniques

Three statistical merging approaches will be considered in this study: unconstrained, constrained, and constrained–optimal. Each of these can be described in terms of some or all of the expressions (A.1) – (A.5).

The first, *unconstrained* matching uses one file as a base and matches each record with the minimum–distance record in the other file. The merged records use the weights from the base file records. This problem can be described mathematically by the set of expressions (A.1), (A.3), (A.4), (A.5) or (A.1), (A.2), (A.4), (A.5). In either case, one set of weight constraints is dropped from the problem.

Using our example files, one unconstrained match would be to drop the file B constraints (A.3) and use file A as the base file. The solution which minimizes the total distance (A.5) is found by matching each file A record with the minimum–distance file B record. This solution is shown in Tableau 1, where $w_{11} = 1000$, $w_{23} = 2000$, $w_{32} = 500$, and the remaining variables are zero. The result is a match with a low aggregate distance (42,500) and strong match statistics, but distortions in file B data. By applying these record weights to the file B data, as shown in the tabulations, note that not only are schedule code and AGI tabulations different, but the aggregate transfer income has increased by $1,300,000 and the average family size has grown from 2.65 to 2.94. Because the weights on file A records are maintained by constraints (A.2), the tabulated values of these record items do not change.

Tableau 2 illustrates the case in which file A weight constraints (A.2) are ignored but file B weight constraints (A.3) are enforced. The match solution for this situation is found by matching each file B record with the closest record in file A. By using the file B weights for merged records, the column totals are maintained, but the row weight totals are altered.

The result is, again, good match statistics and aggregate distance but distorted data values, this time in the file A items. Specifically, for file A, the schedule code tabulations are changed, total AGI has increased $2,400,000 and total deductions increased $530,000.

Tableau 1. Unconstrained File B (Ignore B Weights)

FILE A RECORD	FILE B RECORD 1	2	3	4	
1	20 1000	50	50	35	1000
2	20	100	10 2000	75	2000
3	85	5 500	115	30	500
	1400	400	1500	200	

Total solution distance = 42,500

WEIGHTED TABULATIONS

Description	This Merged File	Original Value
File A Record Data:		
Schedule code =1	3,000	3,000
Schedule code =2	500	500
Total AGI (000s)	50,000	50,000
Total Deductions (000s)	9,800	9,800
File B Record Data:		
Schedule code =1	3,000	2,900
Schedule code =2	400	600
Total AGI (000s)	45,750	47,300
Transfer income (000s)	6,500	5,200
Avg. Family Size	2.94	2.65
Match Statistics (Wtd.)		
% Agreement on Schedule Code	100%	n.a.
Average Absolute AGE Difference	1,214	n.a.

Tableau 2. Unconstrained File A (Ignore A Weights)

FILE A RECORD	FILE B RECORD 1	2	3	4	
1	20 1400	50	50	35	1000
2	20	100	10 1500	75	2000
3	85	5 400	115	30 200	500
	1400	400	1500	200	

Total solution distance = 51,000

WEIGHTED TABULATIONS

Description	This Merged File	Original Value
File A Record Data:		
Schedule code =1	2,900	3,000
Schedule code =2	600	500
Total AGI (000s)	52,400	50,000
Total Deductions (000s)	10,330	9,800
File B Record Data:		
Schedule code =1	2,900	2,900
Schedule code =2	600	600
Total AGI (000s)	47,300	47,300
Transfer income (000s)	5,200	5,200
Avg. Family Size	2.65	2.65
Match Statistics (Wtd.)		
% Agreement on Schedule Code	100%	n.a.
Average Absolute AGE Difference	1,457	n.a.

Tableau 3. Constrained Match

FILE A RECORD	FILE B RECORD 1	2	3	4	
1	$\boxed{20}$ 1000	$\boxed{50}$	$\boxed{50}$	$\boxed{35}$	1000
2	$\boxed{20}$ 400	$\boxed{100}$ 400	$\boxed{10}$ 1200	$\boxed{75}$	2000
3	$\boxed{85}$	$\boxed{5}$	$\boxed{115}$ 300	$\boxed{30}$ 200	500
	1400	400	1500	200	

Total solution distance = 120,500

WEIGHTED TABULATIONS

Description	This Merged File	Original Value
File A Record Data:		
Schedule code =1		3,000
Schedule code =2	Same	500
Total AGI (000s)		50,000
Total Deductions (000s)	Values	9,800
File B Record Data:		
Schedule code =1	As	2,900
Schedule code =2		600
Total AGI (000s)	Original	47,300
Transfer income (000s)		5,200
Avg. Family Size		2.65
Match Statistics (Wtd.)		
% Agreement on Schedule Code	80%	n.a.
Average Absolute AGE Difference	2,942	n.a.

Therefore, in either case, unconstrained matching can drastically distort the values associated with one file or the other. This is of particular concern in the case of the non-common data. The purpose of matching is to be able to draw inferences regarding relationships between one data file and the non-common items in another file. When the values from one file are distorted, the reliability of such inferences is lessened and, thus, the objective of matching is being defeated.

An attempt to remedy this distortion problem is to include both sets of weight constraints, (A.2) and (A.3). This is called *constrained* matching and is described mathematically by expressions (A.1) – (A.4), with the imposition of (A.5) being the special case of *constrained–optimal* matching.

Tableau 3 depicts a constrained non–optimal match. Note that all match weights in a row sum to the row total (original file A record weight) and column sums are similarly kept. By so doing, the original data structures are maintained and all tabulations are the same as those for the original files.

This improvement is not without its costs. The trade–off is in terms of total solution distance and poorer match statistics. The aggregate distance and average AGI discrepancy have more than doubled plus schedule code agreement has dropped 20 percent, relative to the unconstrained matches.

To improve this solution to the greatest extent possible, expression (A.5) can be included and a constrained–optimal match sought. This optimization problem can be solved iteratively by devising a series of improved matches, each of which merges a new pair of records and drops an existing record match, while simultaneously maintaining the weight totals. Tableaus 3 through 3b illustrate this process.

Tableau 3a. Constrained Match, Improved Solution 1

	FILE B RECORD				
FILE A RECORD	1	2	3	4	
1	20 1000	50	50	35	1000
2	20 400	100 100	10 1500	75	2000
3	85	5 300	115	30 200	500
	1400	400	1500	200	

Total solution distance = 60,500

WEIGHTED TABULATIONS

Description	This Merged File	Original Value
File A Record Data:		
Schedule code =1		3,000
Schedule code =2	Same	500
Total AGI (000s)		50,000
Total Deductions (000s)	Values	9,800
File B Record Data:		
Schedule code =1	As	2,900
Schedule code =2		600
Total AGI (000s)	Original	47,300
Transfer income (000s)		5,200
Avg. Family Size		2.65
Match Statistics (Wtd.)		
% Agreement on Schedule Code	97%	n.a.
Average Absolute AGE Difference	1,657	n.a.

Tableau 3b. Constrained Match, Improved Solution 2

FILE A RECORD	FILE B RECORD 1	2	3	4	
1	$\boxed{20}$ 900	$\boxed{50}$ 100	$\boxed{50}$	$\boxed{35}$	1000
2	$\boxed{20}$ 500	$\boxed{100}$	$\boxed{10}$ 1500	$\boxed{75}$	2000
3	$\boxed{85}$	$\boxed{5}$ 300	$\boxed{115}$	$\boxed{30}$ 200	500
	1400	400	1500	200	

Total solution distance = 55,500

WEIGHTED TABULATIONS

Description	This Merged File	Original Value
File A Record Data:		
Schedule code =1		3,000
Schedule code =2	Same	500
Total AGI (000s)		50,000
Total Deductions (000s)	Values	9,800
File B Record Data:		
Schedule code =1	As	2,900
Schedule code =2		600
Total AGI (000s)	Original	47,300
Transfer income (000s)		5,200
Avg. Family Size		2.65
Match Statistics (Wtd.)		
% Agreement on Schedule Code	97%	n.a.
Average Absolute AGE Difference	1,542	n.a.

Tableau 4. Optimal Constrained Match

FILE A RECORD	FILE B RECORD				
	1	2	3	4	
1	20 900	50	50	35 100	1000
2	20 500	100	10 1500	75	2000
3	85	5 400	115	30 100	500
	1400	400	1500	200	

Total solution distance = 51,500

WEIGHTED TABULATIONS

Description	This Merged File	Original Value
File A Record Data:		
Schedule code =1		3,000
Schedule code =2	Same	500
Total AGI (000s)		50,000
Total Deductions (000s)	Values	9,800
File B Record Data:		
Schedule code =1	As	2,900
Schedule code =2		600
Total AGI (000s)	Original	47,300
Transfer income (000s)		5,200
Avg. Family Size		2.65
Match Statistics (Wtd.)		
% Agreement on Schedule Code	97%	n.a.
Average Absolute AGE Difference	1,400	n.a.