Southern Methodist University
SMU Scholar

Historical Working Papers

Cox School of Business

1982

An Empirical Evaluation of Statistical Matching Methodologies

Richard A. Barry
Southern Methodist University

William H. Stewart
Oklahoma State University - Main Campus

J. Scott Turner
Oklahoma State University - Main Campus

Follow this and additional works at: https://scholar.smu.edu/business_workingpapers

 Part of the Business Commons

This document is brought to you for free and open access by the Cox School of Business at SMU Scholar. It has been accepted for inclusion in Historical Working Papers by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.

Research and Development
School of Business Administration
Southern Methodist University
Dallas, Texas 75275

AN EMPIRICAL EVALUATION OF STATISTICAL MATCHING METHODOLOGIES

Working Paper 82-804*

by

Richard A. Barr

William H. Stewart

J. Scott Turner

Richard A. Barr
Edwin L. Cox School of Business
Southern Methodist University
Dallas, Texas 75275

William H. Stewart
Oklahoma State University
Stillwater, Oklahoma

J. Scott Turner
Oklahoma State University
Stillwater, Oklahoma

*This paper represents a draft of work in progress by the authors and is being sent to you for information and review. Responsibility for the contents rests solely with the authors. This working paper may not be reproduced or distributed without the written consent of the authors. Please address all correspondence to Richard A. Barr.

ACKNOWLEDGEMENTS

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.

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.

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 which link records of sample units which have similar data attributes, but are not necessarily the same person or household.

This paper describes a massive 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. Over 7,000 merges of both randomly-generated files and 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.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF TABLEAUX

PREFACE

1. BACKGROUND AND OVERVIEW 1

1.1 Microdata files 2
1.2 Limitations of individual samples 3
1.3 Microdata file merging 4
1.4 Statistical merging 5
1.5 Statistical aspects of merging techniques 10
1.6 Underlying merge rationale 10
1.7 Quality considerations 11
1.8 Preliminary empirical data 11
1.9 Research questions 12
1.9.1 Constrained versus unconstrained techniques 15
1.9.2 Covariance of $(Y,Z|X)$ 15
1.9.3 Distance functions 15
1.10 Experimentation Overview 15

2. STATISTICAL FRAMEWORK AND EXPERIMENTAL DESIGN 17

2.1 Notation and overview of the study 18
2.2 Statistical matching issues to be addressed 19
2.2.1 Statistical inference with matched files 19
2.2.2 Constrained versus unconstrained procedures 20
2.2.3 Cov $(Y,Z|X)$ 21
2.3 Distance functions and matching methodologies 22
2.3.1 Weighted absolute difference measures 23
2.3.2 Mahalanobis distance metrics 24
2.4 Experimental design 24
2.5 Phase I: Randomly - generated sample files 25
2.5.1 Simulated populations 26
2.5.2 Computational procedure 27
2.6 Phase II: Testing of national microdata files 27
2.7 Statistical Evaluation of the merged files 29
2.7.1 Bias for matched file estimatiors 29
2.7.2 Matched samples versus valid random samples 30

3. PHASE I: ANALYSIS OF SIMULATED POPULATIONS 32

3.1 Procedures 33
3.1.1 Populations Studied 33
3.1.2 Merging Methodologies and Experiment Steps 35
3.1.3 Questions addressed 36

3.2 Estimation of Cov(Y,Z) 37
3.3 Estimation of the Joint Y-Z Distribution 41
3.4 Estimation of Var(Z) 43
3.5 Estimation of Cov(X1,Z) and Cov(X2,Z) 44
3.6 Summary and Conclusions 45

4. PHASE II: ANALYSIS BASED ON NATIONAL DATA SET 57

4.1 Population, Samples, and Data Item Descriptions 59
4.2 Matched Files Generated Using the Transportation Model 68
4.2.1 Weighted Absolute Differences Model 68
4.2.2 Mahalanobis Distance Model 72
4.2.3 Other Constrained Models 72
4.2.4 Matched Files Created 73
4.2.5 Tests Used to Compare Matched File Distributions with the Population Distributions 74
4.3 Comparison of Absolute Difference and Mahalanobis Distance Functions 76
4.4 Comparison of Matched Files Generated with an Absolute Value Distance Function Using a Range of Common Variables 82
4.5 Matching Under Conditions of Noise and Bias 86
4.6 Results of Matching a Sample with Itself 89
4.7 Analysis of Unconstrained Procedures 91
4.8 Summary and Results of the Real Data Empirical Investigation 92

5. SUMMARY AND CONCLUSIONS 95

5.1 The Viability of Merging 96
5.2 Choice of Merging Model 98
5.3 Effects of Data Perturbations 99
5.4 Improvements Needed in Existing Methods 99
5.5 Further Research Directions 100

Bibliography and References

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

LIST OF TABLES

Table(s)		Page(s)
1.1	SOI Item Distributions	13
1.2	SOI Variance-Covariance Differences	14
3.1	Populations Used in the Simulation	34
3.2A-F	Estimation of Cov(Y, Z) using Methods 1-6	38-40
3.3A-F	Frequency counts for Chi-Square Goodness of Fit Statistics for the Matched Y-Z Samples: Methods 1-6	47-52
3.4A-D	Estimation of Var(Z) = 1 using Methods 1-6	53-54
3.5A-D	Estimation of Cov(X1, Z) and Cov(X2, Z) Using Methods 1 and 2	55-56
4.1	Files Created from Perturbations of SIE5	60
4.2A	Full SIE Population Y-Z Correlation Matrix	62
4.2B-F	Difference Between SIE1-SIE5 Correlation Matrices and Population Correlation Matrix	62-63
4.3A-I	Population Item Cross-Tabulations	64-67
4.4	C_2 = Interest Income Difference Index	69
4.5	C_4 = Penalty Index for Difference in Ages	70
4.6	Matched Files Created Using the Transportation Matching Algorithm	77
4.7	Matched File 1 Total Income and Dividend Joint Distribution	78
4.8	Contingency Table χ^2 Values Based on Population Percentages as Expected Values	80
4.9A-B	Average Correlation Matrix for Matched files 1-10 (11-20) Minus the Population Correlation Matrix (Methods 1 and 2)	81-82
4.10	Contingency Table χ^2 Values Using Population Percentages as Base	84
4.11-12	Average Correlation Matrix for Matched Files 40-43 (44-47) Minus Population Matrix, Method 1 (4)	85
4.13	χ^2 Statistics For Bias and Noise Tests	88
4.14-15	Correlation Difference Matrix for Matched File 21 (22) Minus SIE5	90
4.16A	Dividend Income Statistics for Sample Files	93
4.16B	SIE2 Dividend Income Statistics After Unconstrained Merging	93
4.16C	Unconstrained Merges: Noise and Bias Tests	93
A.1	Item Tabulations for Example Files (Weighted)	A-2

LIST OF FIGURES

Figure		Page
1.1	Statistical File Merging	6
1.2	Constrained Merge Model	9
3.1	Y-W Plane	42
A.1	Example Files A and B	A-2
A.2	Scatter Diagram of Selected Records	A-3

LIST OF TABLEAUX

Tableau		Page
0.	Sample Tableau for Example Merge Problem	A-4
1	Unconstrained File B	A-7
2	Unconstrained File A	A-8
3	Constrained Match	A-9
3a-b	Constrained Match, Improved Solutions	A-11, A-12
4	Optimal Constrained Match	A-14

PREFACE

The objective of this research study is to empirically analyze microdata files which have been matched using statistical merging techniques. To achieve this objective, two major research efforts were undertaken: one based on randomly-generated data from theoretically ideal distributions, and another which used sample survey data taken from a national data set. Although interrelated, these two sets of experiments and their respective conclusions are presented separately in this report in order that a reader might focus on different sections according to his or her interests.

The report has the following organization. Section 1 gives an overview of statistical file matching, highlighting the techniques in current use and the major research issues. Section 2 describes the statistical framework and experimental design for each phase of this study and introduces terminology used throughout the remainder of the paper.

In Section 3, the randomly-generated data experiments and their associated results are described in detail. Twelve distinct statistical populations were selected for test file generation, and six merge methods implemented. A total of 7,200 matched files were created and analyzed to test all combinations of merge scheme and population with 100 replications of each for statistical confidence.

The findings from the second study phase, based on sample files drawn from the Survey of Income and Education, are presented in Section 4. Fifty constrained-optimal matched files were formed to evaluate the effects on merge file quality, if any, of distance function and data condition (bias and noise). Contrasts are made with unconstrained procedures employed under the same circumstances.

Finally, an overall summary and set of conclusions for the study are given in Section 5.

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 behavioral model DYNASIM, and the tax policy simulations at the U.S. Department of the Treasury and at Brookings Institute.

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 modelling technique is very flexible and can accommodate as much detail as desired.

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 the 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 to "correct" the data, for underreporting for example, through editing procedures. 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, as with 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.

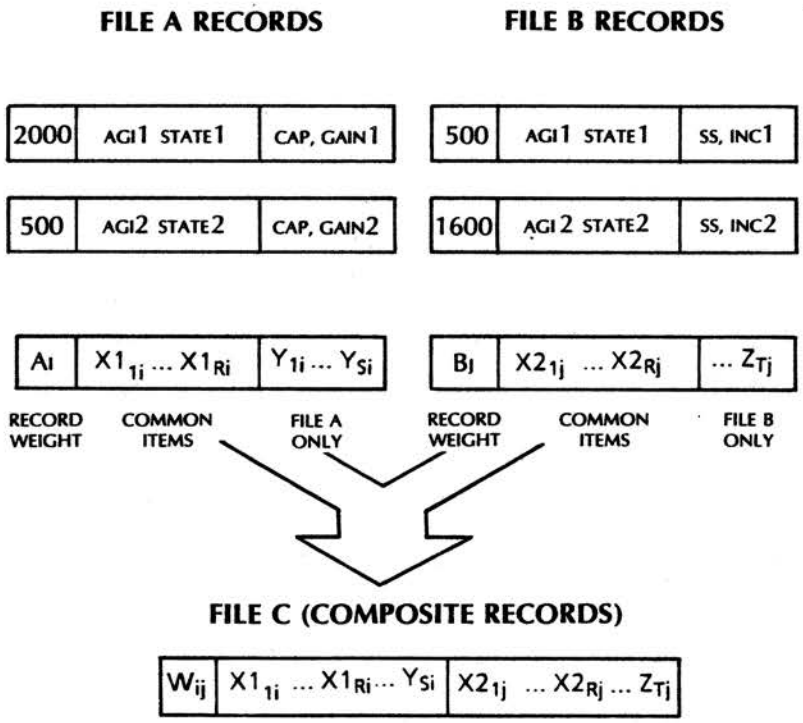
In choosing a methodology, exact matching is obviously preferable. But where such a match is not possible, statistical merging is often employed.

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. File C, the merged file, 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

Figure 1.1.

Statistical File Merging



INTERRECORD DISSIMILARITY MEASURE (DISTANCE FUNCTION):

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

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

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j \qquad (1.1)$$

$$\sum_{i=1}^{n} w_{ij} = a_i, \; i=1,\ldots,m, \qquad (1.2)$$

$$\sum_{i=1}^{m} w_{ij} = b_j, \; j=1,\ldots,n, \qquad (1.3)$$

$$w_{ij} >= 0, \text{ for all } i \text{ and } j. \qquad (1.4)$$

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 Mathematica 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).

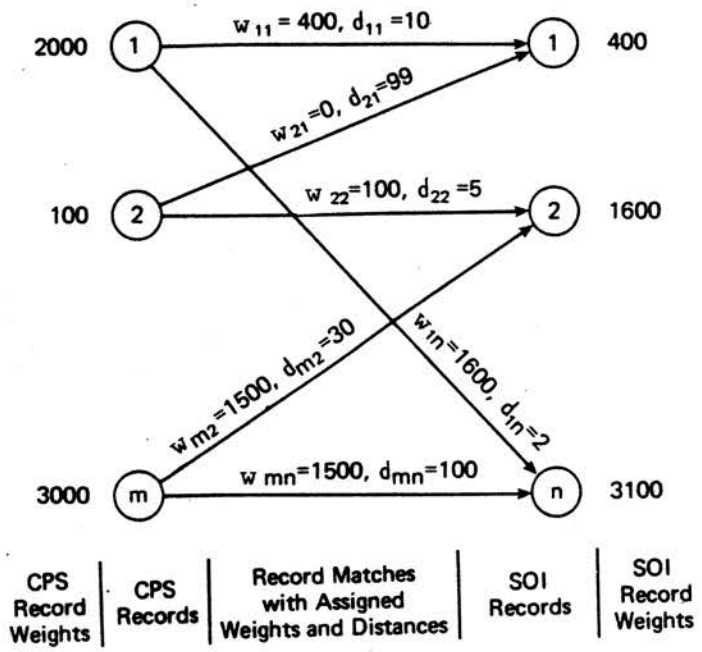
This merge technique can be further refined by requiring the procedure to

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

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



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, 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 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 frequences 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 points out potential data problems measured by these types of 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.

Table 1.1

SOI Item Distributions

Income Class ($000's)	Total SOI Wages ($ Millions) Original	Unconstrained	Total SOI Business Income ($ Millions) 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	601
> 200	244	0	111	0
Total	115,784	121,291	5,055	8,187
Mean:	$9,108	$9,542	$398	$644

Table 1.2

SOI Variance-Covariance Differences

Common Variable	Median Variance-Covariance Error Relative to Original	
	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.

1.9.1. 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.

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 questions 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.

There were two phases to this work. The first phase makes use of randomly-generated data in a highly-structured experimental design which tests for statistical biases introduced by merge methodology/distance

function employed. By knowing the statistical distribution of the population from which these hypothetical microdata files are drawn, strong statistical tests are available for hypothesis testing in an "ideal" environment.

A similar design is employed in the second study phase based on a public-use national data set. The 1975 SIE file was designated to be a test population from which a selected series of samples were drawn. Each record item is 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 characteristic 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 (X1,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 nonmatching 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

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

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 $\text{cov}(Y,Z|X=x)=0$. 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 $\text{cov}(Y,Z|X=x)\neq 0$.

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:

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

In unconstrained and constrained optimal 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 ℓ of Y, an additional component of the distance function could be $s_{\ell k} \cdot |y1_{i\ell} - x2_{jk}|$. 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

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

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 $(X1,Y,\hat{Z})$ and $(X2,\hat{Y},Z)$, 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_{YZ} 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 function, and measurement error in the data, a set of experiments have been designed. The experimentation is organized as two phases, each using the same general methodology, but applied to different sources of data. Initially, randomly-generated data files are used to form "best case" conclusions. A second phase builds on these results by applying similar statistical tests to samples drawn from national data files.

2.5. Phase I: Randomly-Generated Sample Files

In the face of the difficult theoretical problems involved with statistical matching, Monte Carlo simulations of the matching methods are a practical initial approach to understanding how these mathods perform. In this phase of the study, a known population is defined from which two independent samples (X1,Y) and (X2,Z) can be generated and a statistically-merged file (X12,Y,Z) formed. A known population is required so that statistics from the matched file can be compared with known population parameters for evaluation of the performance of the matching method.

It is not enough to match a single pair of samples. Since it is the validity of the matching process itself we wish to study, it is necessary to repeat the simulation enough times to adequately study the sampling distribution of statistics based on the matching method. Such repetitions allow us to approach two questions concerning statistical matching methods: first, are estimators obtained from matched files

biased and, second, is the sampling variability of matched files approximately the same as the variability in valid random samples?

In choosing populations from which to generate the data, those having variables with normal distributions are the best for the initial phase of this study for several reasons. First, strong statistical tests are available which will give insight into the behavior of the merge methodologies to more complicated sample survey populations on statistically well-defined data. The transferability of the conclusions reached will be studied in the second testing phase. Also, normal populations are central to much of the existing theory of statistical matching [25] and many populations of interest can be roughly approximated with normal distributions. Therefore, studying the methodologies under this controlled environment will yield "best-case" conclusions against which "actual case" results can be benchmarked.

A balance must be struck between simulating a realistic problem and using a problem which is simple enough to be computationally feasible. Populations with four multivariate items (x_1,x_2,y,z) will be used to generate files with $(x1_1,x1_2,y)$ and $(x2_1,x2_2,z)$ observations. This is simple, but still allows nontrivial matching problems by having bivariate X.

The size of the generated files, while much smaller than true sample survey files, must be large enough to allow some investigation of the variability within a matched file. Hence, each simulated file A and B will have 200 records.

Finally, the number of replications must be large enough to allow the large sample statistical analysis described in the following section to apply while also staying within existing computational limits. This

number is also influenced by the size of the biases we wish to have the power to detect. In this study 100 replications were used.

2.5.1. Simulated Populations

It is important to choose populations which address the controversial areas of statistical matching. The following is a list of four pairs of population properties which were expected to be critical to the performance of the various merging methodologies:

1. $Cov(Y,Z|X=x) = 0$ or $Cov(Y,Z|X=x) \neq 0$,
2. $R^2_Z(X1,X2)$ high or $R^2_Z(X1,X2)$ low,
3. $R^2_Y(X1,X2)$ high or $R^2_Y(X1,X2)$ low,
4. $Cov(X_1,X_2)$ high or $Cov(X_1,X_2)$ low.

Properties 1 and 2 are deemed the most crucial as these may have a general bearing on the performance of all of the matching schemes. Properties 3 and 4 are more related to the choice of which matching methodology to employ. Cases from all 16 possible combinations of these properties are studied, although more attention is paid to properties 1 and 2.

2.5.2. Computational Procedure

For each population considered, it was necessary to generate k pairs of independent $(X1_1,X1_2,Y)$ and $(X2_1,X2_2,Z)$ files. Then, for each matching methodology of interest with that population, k matched files were created by merging these file pairs. Sample statistics were kept for each matched file to statistically compare the matched results with the actual population and the results known for valid random samples. Also by using each methodology on the same pairs of files, a randomized block design is created for comparing the competing methodologies.

2.6. Phase II: Testing of National Microdata Files

In contrast with artifically-generated data, the impact of matching methodology on the merging of actual microdata files is studied. This provides a more practical setting to compare the weighted absolute differences distance functions, which are not specifically designed for normal data, with the Mahalanobis distances which are.

The 1975 Survey of Income and Education is treated as a population and, from this file, five randomly-drawn samples of 1000 records each were drawn. The records of each file are 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 be merged using the 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.

2.7. Statistical Evaluation of the Merged Files

2.7.1. 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) also 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 possibly 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

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

where

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

and

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

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.

2.7.2. 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<= d	d < z <= e	e < z <= g	g < z
y<= a	f_{11}	f_{12}	f_{13}	f_{14}
a < y<= b	f_{21}	f_{22}	f_{23}	f_{24}
b < y<= 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 known 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 the 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 could be pooled for a given methodology into a single table. The pooled distribution could then be 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 could be used.

PART 3

PHASE I: ANALYSIS OF SIMULATED POPULATIONS

3.1 Procedures

3.1.1 Populations Studied

Table 3.1 shows the 12 normal populations used in the Monte Carlo generation of random (X1,X2,Y) and (X1,X2,Z) samples. In each of these 12 populations the means and variances of X1, X2, Y, Z were set equal to zero and one respectively, while the covariances were chosen to give a variety of interesting and comparable cases. The populations 1A, 2A, 3A, 4A, 5A, 6A were all selected so that $cov(Y,Z|X) = 0$. The populations 1B, 2B, 3B, 4B, 5B, 6B were then constructed so that each of these contained the same covariances as the corresponding A population, except cov(Y,Z) was chosen to ensure that $cov(Y,Z|X)$ was far from 0. For example, populations 1A and 1B have the same covariance structure except that cov(Y,Z) of 1A forces $cov(Y,Z|X) = 0$, whereas 1B has a different cov(Y,Z) with $cov(Y,Z|X) \neq 0$.

This distinction of $cov(Y,Z|X) = 0$ vs. $cov(Y,Z|X) \neq 0$ proved to be extremely important. It was clear from the outset that matching methods would have difficulty distinguishing a B population from the corresponding A population, since the marginal distributions of (X1,X2,Y) and (X1,X2,Z) would be the same for both A and B populations. Thus, the sample data would contain no information about cov(Y,Z) which would allow a B population to be distinguished from the corresponding A population. The only real hope to make this distinction must necessarily come from some outside information not derived from the sample, which could then be incorporated in the matching methodology. Only Kadane's full (X,Y,Z) information method (see Section 2.3.2) proported to do this, so we felt it very important to include this method in the simulation.

Table 3.1

Populations Used in the Simulation
[All are Multivariate Normal (X1,X2,Y,Z) with E(X1) = E(XZ) = E(Y) = E(Z)=0
and Var(X1) = Var(X2) = Var(Y) = Var(Z) = 1.]

POP	COVX1X2	COVX1Y	COVX2Y	COVX1Z	COVX2Z	COVYZ
1A	0.8	0.9	0.9	0.9	0.9	0.900
1B	0.8	0.9	0.9	0.9	0.9	0.810
2A	0.8	0.1	0.1	0.9	0.9	0.100
2B	0.8	0.1	0.1	0.9	0.9	0.400
3A	0.1	0.7	0.7	0.7	0.7	0.891
3B	0.1	0.7	0.7	0.7	0.7	0.980
4A	0.1	0.2	0.2	0.7	0.7	0.255
4B	0.1	0.2	0.2	0.7	0.7	0.500
5A	0.8	0.3	0.3	0.3	0.3	0.100
5B	0.8	0.3	0.3	0.3	0.3	-0.700
6A	0.2	0.3	0.3	0.3	0.3	0.150
6B	0.2	0.3	0.3	0.3	0.3	-0.400

NOTE: For A-populations, $cov(Y,Z|X)=0$ and for B-populations, $|cov(Y,Z|X)|>>0$.

3.1.2 Merging Methodologies and Experiment Steps

The simulation procedure was as follows. For each of the 12 populations, 100 pairs of independent (X1,X2,Y) and (X1,X2,Z) samples, each of size 200, were generated. For each pair of generated samples, six matched files were created, one for each matching method under consideration. These methods are listed below.

Method 1: absolute difference distance function on X-data, constrained.

Method 2: Mahalanobis distance function on X-data, constrained.

Method 3: Kadane's full (X,Y,Z) information distance function, constrained.

Method 4: absolute difference distance function on X-data, unconstrained.

Method 5: Mahalanobis distance function on X-data, unconstrained.

Method 6: Kadane's full (X,Y,Z) information distance function, unconstrained.

It was decided to use the true population values of the (X1,X2,Y,Z) covariance matrix in the distance functions of Methods 2, 3, 5, and 6. In actual practice one would employ estimates from the (X1,X2,Y) and (X1,X2,Z) samples in the distance functions. However, real problems will typically involve much greater sample sizes than the n = 200 used here, so that sampling error in the estimation of the covariance matrix would be almost negligible. We felt that we would get a better estimate of the realistic performance of these methods by using the true covariance matrix, rather than letting the sampling error in our simulation influence the results.

For each combination of matching method and population we then generated 100 matched samples in this manner. We obtained estimates of the known population parameters from these matched samples. The quality

of these estimators could then be judged by considering the variation of these estimates over the 100 matched samples for each case.

3.1.3 Questions Addressed

This data enabled us to address the following important questions.

1. How well is cov(Y,Z) estimated by these matching methods?
2. Do the matched Y-Z samples behave as valid random samples from the Y-Z population?
3. What is the relationship of the cov(Y,Z|X) to these methods?
4. How well is var(Z) estimated by these methods?
5. How well are cov(X1,Z) and cov(X2,Z) estimated by these methods?

One of the main motivations for matching is to obtain information on the Y-Z distribution. Questions 1 and 2 address this directly. By including the A and B populations in this study we were in a position to answer question 3. Since these methods used the (X1,X2,Y) sample as base file, statistics involving only these variables would necessarily behave as valid statistics based on random sampling. However, since the matched samples would be appending Z-values to this base file through the matching procedures, it was important to see how statistics involving Z would behave. Constrained matching would necessarily retain the original Z sample, so any estimators based only on the Z-values would be done validly. However, unconstrained matching could possibly distort the marginal distribution of Z. Thus, question 4 was designed primarily for evaluating unconstrained matching.

There was the possibility in all methods that cov(X1,Z) and cov(X2,Z) might not be well estimated from the matched samples, even though the

original (X1,X2,Z) samples could provide valid estimates of these parameters. Question 5 deals with this point.

3.2 Estimation of Cov(Y,Z)

Since the (X1,X2,Y,Z) populations are multivariate normal, a complete description of a population is given by the means, variances, and covariances of its variables. From the (X1,X2,Y) and (X1,X2,Z) samples everything can be validly estimated except the cov(Y,Z). Thus, the key question for the matching methods is: How well do the matched samples do for estimating cov(Y,Z)?

Tables 3.2A-F summarize the performances of the six matching methods for estimating cov(Y,Z). The true cov(Y,Z) can be compared with the mean estimate obtained over the 100 replications using a particular matching method. The sample standard deviation of the estimates then gives some indication as to the variability of the estimates from sample to sample. To test for a bias in estimating cov(Y,Z) with a particular method one can look at the statistic T. If there is no bias caused by the matching procedure, then the sampling distribution of T should be approximately normal with mean 0 and variance 1. One can then do a test of significance for biased estimation. Values of T greater than 1.96 in absolute value are evidence for biased estimation at the .05 significance level.

It is clear from Tables 3.2A-F that all the matching procedures perform extremely poorly on the B populations where $\text{cov}(Y,Z|X) \neq 0$. However, with some of the procedures on some of the A populations, where $\text{cov}(Y,Z|X) = 0$, the estimation of cov(Y,Z) is done quite well. Note that for each pair of A and B populations, which differ only in cov(Y,Z), the matched

Table 3.2A

Estimation of Cov(Y,Z) using Method 1
Mean is the Mean Estimate over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovYZ)/SD.

POP	COVYZ	MEAN	SD	T
1A	0.900	0.887	0.064	-1.98
1B	0.810	0.886	0.062	12.22
2A	0.100	0.102	0.064	0.30
2B	0.400	0.102	0.062	-48.15
3A	0.890	0.859	0.066	-4.79
3B	0.980	0.861	0.067	-17.88
4A	0.255	0.250	0.064	-0.72
4B	0.500	0.252	0.063	-39.15
5A	0.100	0.096	0.075	-0.48
5B	-0.700	0.100	0.066	120.51
6A	0.150	0.136	0.065	-2.09
6B	-0.400	0.134	0.065	82.56

Table 3.2B

Estimation of Cov(Y,Z) using Method 2
Mean is the Mean Estimate over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovYZ)/SD.

POP	COVYZ	MEAN	SD	T
1A	0.900	0.891	0.063	-0.137
1B	0.810	0.890	0.062	12.810
2A	0.100	0.102	0.065	0.365
2B	0.400	0.100	0.070	-42.780
3A	0.890	0.867	0.065	-3.720
3B	0.980	0.868	0.065	-17.060
4A	0.255	0.252	0.058	-0.400
4B	0.500	0.255	0.061	40.520
5A	0.100	0.099	0.068	-0.170
5B	-0.700	0.104	0.075	107.520
6A	0.150	0.147	0.061	-0.520
6B	-0.400	0.145	0.062	103.800

Table 3.2C

Estimation of Cov(Y,Z) Using Method 3
Mean is the Mean Estimate over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovYZ)/SD.

POP	COVYZ	MEAN	SD	T
1A	0.900	0.876	0.063	-3.82
1B	0.810	0.874	0.060	10.66
2A	0.100	0.097	0.061	-0.55
2B	0.400	0.099	0.063	-47.60
3A	0.890	0.864	0.065	-4.13
3B	0.980	0.865	0.066	-17.36
4A	0.255	0.250	0.061	-0.75
4B	0.500	0.252	0.066	-39.25
5A	0.100	0.089	0.066	-1.70
5B	-0.700	0.092	0.071	110.80
6A	0.150	0.139	0.064	-1.66
6B	-0.400	0.140	0.067	81.12

Table 3.2D

Estimation of Cov(Y,Z) Using Method 4
Mean is the Mean Estimate over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovYZ)/SD.

POP	COVYZ	MEAN	SD	T
1A	0.900	0.882	0.077	-2.34
1B	0.810	0.886	0.078	9.74
2A	0.100	0.102	0.061	0.33
2B	0.400	0.102	0.059	-50.51
3A	0.890	0.846	0.078	-5.64
3B	0.980	0.845	0.079	-17.09
4A	0.255	0.248	0.065	-1.08
4B	0.500	0.248	0.065	-38.77
5A	0.100	0.101	0.070	0.14
5B	-0.700	0.100	0.079	101.27
6A	0.150	0.145	0.075	-0.67
6B	-0.400	0.146	0.080	68.25

Table 3.2E

Estimation of Cov(Y,Z) Using Method 5
Mean is the Mean Estimate over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovYZ)/SD.

POP	COVYZ	MEAN	SD	T
1A	0.900	0.882	0.079	-2.28
1B	0.810	0.885	0.079	9.49
2A	0.100	0.102	0.063	0.32
2B	0.400	0.103	0.062	-47.90
3A	0.890	0.850	0.080	-5.00
3B	0.980	0.847	0.079	-16.84
4A	0.255	0.249	0.066	-0.91
4B	0.500	0.248	0.066	-38.18
5A	0.100	0.099	0.072	-0.14
5B	-0.700	0.097	0.078	102.18
6A	0.150	0.149	0.074	-0.14
6B	-0.400	0.149	0.078	70.38

Table 3.2F

Estimation of Cov(Y,Z) Using Method 6
Mean is the Mean Estimate over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovYZ)/SD.

POP	COVYZ	MEAN	SD	T
1A	0.900	0.817	0.069	-12.03
1B	0.810	0.817	0.069	1.01
2A	0.100	0.093	0.057	-1.23
2B	0.400	0.094	0.059	-51.86
3A	0.890	0.803	0.075	-11.60
3B	0.980	0.805	0.071	-24.65
4A	0.255	0.234	0.061	-3.44
4B	0.500	0.235	0.057	-46.49
5A	0.100	0.090	0.023	-4.35
5B	-0.700	0.093	0.023	344.78
6A	0.150	0.135	0.030	-5.00
6B	-0.400	0.136	0.028	191.43

samples' mean and standard deviation are nearly the same. This points out that for matching purposes the observed data gives no information for distinguishing the A and B populations. It is surprising that this happens even for Methods 3 and 6 which used Kadane's "full information" distance function. However, on close examination, we see that Kadane's procedure does not actually use cov(Y,Z) in the covariance matrix for the distance function. It enters only as part of a transformation of the original (X1,X2,Y) and (X1,X2,Z) data, with the matching only being accomplished through consideration of the original samples.

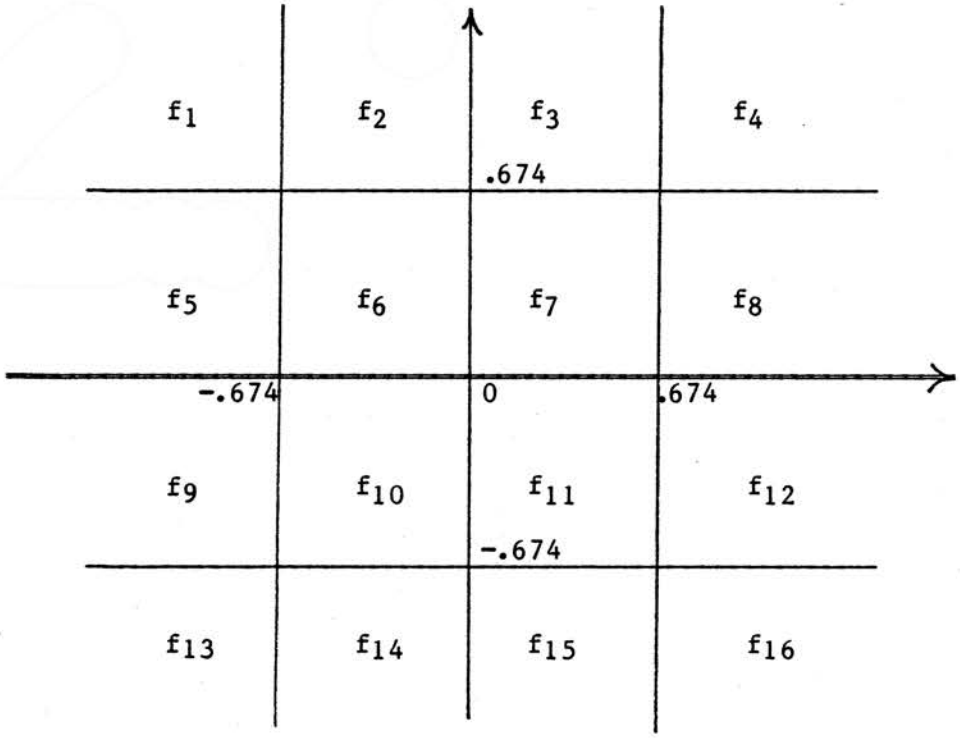
In comparing the six methods, none of them do well for B populations, but for the A populations Method 2 did best. It only showed a bias for estimating cov(Y,Z) on population 3A with T = -3.72. The other five methods were all about the same in their performance for estimating cov(Y,Z) on the A populations. Generally they did not perform well on populations 1A and 3A. The results are quite striking in the enormous failure of all these methods on the B populations where cov $(Y,Z|X) \neq 0$.

3.3 Estimation of the Joint Y-Z Distribution

Rather than focus on the cov(Y,Z) alone, it is also interesting to consider the overall estimation of the Y-Z distribution. To do this the Y-Z plane was divided into 16 equally probable regions for each population. This was easily done by transforming Z to $W = (Z-rY)/(1-r^2)^{.5}$, where r = cov(Y,Z). (Here Y and W are independent standard normal random variables.) The Y-W plane was then divided by the grid given in Figure 3.1, and the f_i, the number of matched sample (Y,W) pairs falling in region i, were recorded for each sample.

Figure 3.1

Y-W Plane



A Chi-Square goodness of fit statistic

$$\chi^2 = \sum_{i=1}^{16} \frac{(f_i - 12.5)^2}{12.5}$$

was then calculated for each sample. This χ^2 measures how well the sampled data corresponds to or "fits" the population distribution. Small values of χ^2 indicate a good fit while large values of χ^2 indicate a poor fit. If the matched samples behave like valid random samples, then the χ^2 values should have an approximate Chi-Square distribution with 15 degrees of freedom.

Tables 3.3A-F (at end of section) give the empirical sampling distribution of the χ^2 values calculated for the 100 reps on each population using Methods 1 through 6. These empirical distributions may be judged against the Chi-Square distribution with 15 degrees of freedom whose exected values are shown in the same tables. We may declare a sample to be "unacceptable" if its χ^2 value is greater than 25. For valid sampling this will occur on 5% of the samples. So if a procedure exhibits much more than 5% of its 100 χ^2 values greater than 25, then we have evidence that the procedure performs poorly.

Methods 1 and 2 did better than the other four procedures on this χ^2 approach. On examining the tables we again see that for all B populations, even methods 1 and 2 perform terribly. Methods 1 and 2 did well on populations 2A, 4A, 5A, and 6A, but not on populations 1A and 3A, although the results on 1A and 3A are not as disastrous as on the B populations. We have very strong evidence that the condition $\text{cov}(Y,Z|X) = 0$ is absolutely necessary for these procedures to do well, but in some instances there still may be difficulties even when $\text{cov}(Y,Z|X) = 0$.

3.4 Estimation of Var(Z)

The estimation of var(Z) provides the most striking contrast between constrained and unconstrained matching. Since the original Z sample is wholly retained in constrained matching, the var(Z) is validly estimated by a constrained matched sample. This is again empirically demonstrated in Table 3.4A at the end of this section. However, for unconstrained matching some of the original Z-values may not be included in the matched sample, while other Z-values are included more than once. This seems to bias the estimates of var(Z) quite dramatically. Tables 3.4A-D show that

unconstrained matching will consistently underestimate var(Z). Apparently, extreme values of Z are linked to X values which are hard to match, so consequently they get removed from the matched sample. This provides evidence that constrained matching is generally preferable to unconstrained matching.

3.5 Estimation of Cov(X1, Z) and Cov(X2, Z)

In Tables 3.5A-D we consider the estimation of cov(X1, Z) and cov(X2, Z) with Methods 1 and 2. Generally there is an indication of underestimation of these parameters when their values are high or, more accurately, when the multiple correlation $R^2(X, Z)$ is high. It is easy to see why this underestimation will occur in the extreme case where X1 and Z have correlation 1. In this case all the (X1, Z) values from the original (X1, X2, Z) file will lie on a straight line. But when the matched sample is created, the new (X1, Z) values in the matched sample will not all be on a straight line, due to the impossibility of getting the X-values matched to values which are exactly equal. Thus, the correlation between X1 and Z in the matched sample would be less than 1. This slight underestimation may be related to the small sample size used here, and may not prove to be so great a difficulty for the large samples of most real matching problems.

This underestimation of cov(X1, Z) and cov(X2, Z) seems to be the reason why cov(Y, Z) was not well estimated for populations 1A and 3A. The combination of high cov(X1, Z) and cov(X2, Z) with high cov(Y, Z) can cause a difficulty even when $\text{cov}(Y, Z|X) = 0$. It is interesting to note that in these cases the empirical estimates of $\text{cov}(Y, Z|X)$ were near zero, despite the direct estimates of cov(X1, Z), cov(X2, Z), and cov(Y, Z) being biased downward. For example, with Method 2 on population 3A, using the mean estimates, we have

$$\mathrm{cov}(Y,Z|X) = \mathrm{cov}(Y,Z) - \begin{bmatrix}\mathrm{cov}(X1,Z)\\ \mathrm{cov}(X2,Z)\end{bmatrix}^t \begin{bmatrix}\mathrm{var}(X1) & \mathrm{cov}(X,X2)\\ \mathrm{cov}(X1,X2) & \mathrm{var}(X2)\end{bmatrix}^{-1} \begin{bmatrix}\mathrm{cov}(X1,Z)\\ \mathrm{cov}(X2,Z)\end{bmatrix},$$

so the estimate of $\mathrm{cov}(Y,Z|X)$ is

$$.024 = .867 - \begin{bmatrix}.081\\ .678\end{bmatrix}^t \begin{bmatrix}1.012 & .091\\ .091 & .997\end{bmatrix}^{-1} \begin{bmatrix}.681\\ .678\end{bmatrix}.$$

This suggests that these matching methods will effectively be forcing the matched sample to exhibit $\mathrm{cov}(Y,Z|X) = 0$.

3.6 Summary of Conclusions

The following conclusions are strongly suggested by the simulation data.

1. For the matching procedures considered here, the population must have $\mathrm{cov}(Y,Z|X) = 0$ if the procedures are to perform well.
2. For unconstrained matching, var(Z) is consistently underestimated. Thus constrained matching is generally preferable as it completely avoids this difficulty.
3. Cov(X,Z) can be underestimated when X and Z are highly correlated, and this can lead to bias in the estimation of cov(Y,Z).
4. Of the six methods considered here, Methods 1 and 2 seem best with Method 2 being slightly better than the remaining approaches.

As a final remark, since the condition of $\mathrm{cov}(Y,Z|X) = 0$ is an extremely strong one, which may not be realistically expected in many cases, there is concern as to the accuracy of these procedures in practice. The strategy of matching on the closeness of the observed X-values alone should

perhaps be modified to employ alternative procedures which can be devised where in good matches are directly defined in terms of their preservation of the distributional properties of the original samples, while imputing some outside information about the Y-Z distribution into the matched samples.

While it is tempting to generalize these strong conclusions to the merging of actual microdata, there are three aspects of the simulation experiments that tend to mitigate these findings in a broader setting. First, the number of X-variables was very small (two). This is shown in the following section to have a strong effect on the quality of merge files. Second, the data was normally distributed, a characteristic which is not typical for econometric and social science data, which tends to be either discrete or to follow unique distributions. Third, with file sizes of only 100 records, the number of file B match possibilities for a given file A record was limited; larger sizes would produce a greater variety of B-records in a given file and increase the likelihood of producing "attractive" matches. It should be noted that most of these limitations were brought about by machine time and capacity considerations.

Table 3.3A

Frequency Counts for Chi-Square Goodness of Fit Statistics for the Matched Y-Z Samples: Method 1 (absolute value distance function, constrained)

χ^2 Value

Population	0-11.04	11.04-14.34	14.34-18.25	18.25-25.00	25.00-30.58	30.58-
(Expected)	(25)	(25)	(25)	(20)	(4)	(1)
1A	16	13	21	14	12	24
1B	0	3	5	22	12	58
2A	31	20	23	21	4	1
2B	0	3	5	21	24	47
3A	5	8	19	30	18	20
3B	0	0	0	0	0	100
4A	23	22	24	24	7	0
4B	0	2	7	38	15	38
5A	27	26	21	19	7	0
5B	0	0	0	0	0	100
6A	28	21	24	21	4	2
6B	0	0	0	1	1	98

Table 3.3B

Frequency Counts for Chi-Square Goodness of Fit Statistics for the Matched Y - Z Samples: Method 2 (Mahalanobis distance function, constrained)

χ^2 Value

Population	0-11.04	11.04-14.34	14.34-18.25	18.25-25.00	25.00-30.58	30.58
(Expected)	(25)	(25)	(25)	(20)	(4)	(1)
1A	8	13	24	17	11	27
1B	0	1	2	15	19	63
2A	26	27	27	13	3	0
2B	0	2	8	17	25	48
3A	8	7	25	21	16	23
3B	0	0	0	0	0	100
4A	22	22	26	21	8	1
4B	0	3	11	24	26	36
5A	34	22	22	15	4	3
5B	0	0	0	0	0	100
6A	28	22	27	21	2	0
6B	0	0	0	0	3	97

Table 3.3C

Frequency Counts for Chi-Square Goodness of Fit Statistics of Matched Y-Z Samples: Method 3 (Kadane distance function, constrained)

χ^2 Value

Population	0-11.04	11.04-14.34	14.34-18.25	18.25-25.00	25.00-30.58	30.58-
(Expected)	(25)	(25)	(25)	(20)	(4)	(1)
1A	7	15	18	19	13	28
1B	1	7	8	25	18	41
2A	27	21	23	24	4	1
2B	0	2	2	27	24	47
3A	11	7	17	29	12	24
3B	0	0	0	0	0	100
4A	21	20	25	26	6	2
4B	1	4	9	27	27	32
5A	23	31	24	21	1	0
5B	0	0	0	0	0	100
6A	21	25	31	20	3	0
6B	0	0	0	0	2	98

Table 3.3D

Frequency Counts for Chi-Square Goodness of Fit Statistics of Matched Y - Z Samples: Method 4 (absolute value distance function, unconstrained)

χ^2 Value

Population	0-11.04	11.04-14.34	14.34-18.25	18.25-25.00	25.00-30.58	30.58-
(Expected)	(25)	(25)	(25)	(20)	(4)	(1)
1A	19	17	17	38	5	4
1B	0	2	2	15	24	57
2A	20	26	16	26	9	3
2B	1	2	4	19	22	52
3A	11	21	24	27	10	7
3B	0	0	0	0	0	100
4A	18	22	24	24	11	1
4B	0	2	10	34	14	40
5A	16	12	26	33	8	5
5B	0	0	0	0	2	100
6A	11	20	21	28	14	6
6B	0	0	0	0	2	98

Table 3.3E

Frequency Counts for Chi-Square Goodness of Fit Statistics of Matched Y -Z Samples: Method 5 (Mahalanobis distance function, unconstrained)

χ^2 Value

Population	0-11.04	11.04-14.34	14.34-18.25	18.25-25.00	25.00-30.58	30.58-
(Expected)	(25)	(25)	(25)	(20)	(4)	(1)
1A	17	15	26	26	12	4
1B	2	0	3	17	19	59
2A	21	21	20	27	7	4
2B	0	3	3	21	24	49
3A	14	20	19	28	14	5
3B	0	0	0	0	0	100
4A	18	25	22	26	8	1
4B	0	5	10	27	22	36
5A	18	17	20	30	10	5
5B	0	0	0	0	0	100
6A	12	16	23	32	11	6
6B	0	0	0	2	1	97

Table 3.3F

Frequency Counts for Chi-Square Goodness of Fit Statistics of Matched Y - Z Samples: Method 6 (Kadane's distance function unconstrained)

χ^2 Value

Population	0-11.04	11.04-14.34	14.34-18.25	18.25-25.00	25.00-30.58	30.58-
(Expected)	(25)	(25)	(25)	(20)	(4)	(1)
1A	0	1	10	26	23	40
1B	0	0	0	0	0	100
2A	9	15	27	23	18	8
2B	0	1	0	21	23	55
3A	0	1	3	29	16	51
3B	0	0	0	0	0	100
4A	9	21	20	28	12	10
4B	0	1	6	25	27	41
5A	0	0	0	0	0	100
5B	0	0	0	0	0	100
6A	0	0	0	0	0	100
6B	0	0	0	0	0	100

Table 3.4A

Estimation of Var(Z) = 1 Using Methods 1,2,3
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - 1)/SD

POP	MEAN	SD	T
1A	0.9998	0.0946	-0.02
2A	0.9998	0.0946	-0.02
3A	0.9976	0.1045	-0.23
4A	0.9976	0.1045	-0.23
5A	0.9902	0.1051	-0.93
6A	0.9903	0.1040	-0.93

Table 3.4B

Estimation of Var(Z) = 1 Using Method 4
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - 1)/SD

POP	MEAN	SD	T
1A	0.9633	0.0901	-4.07
2A	0.9633	0.0901	-4.07
3A	0.9283	0.0877	-8.18
4A	0.9283	0.0877	-8.18
5A	0.9797	0.1365	-1.49
6A	0.9718	0.1436	-1.96

Table 3.4C

Estimation of Var(Z) = 1 Using Method 5
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - 1)/SD

OBS	POP	MEAN	SD	T
1	1A	0.9630	0.0927	-3.99
2	2A	0.9630	0.0927	-3.99
3	3A	0.9336	0.0943	-7.04
4	4A	0.9336	0.0943	-7.04
5	5A	0.9777	0.1388	-1.61
6	6A	0.9703	0.1436	-2.07

Table 3.4D

Estimation of Var(Z) = 1 Using Method 6
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - 1)/SD

POP	MEAN	SD	T
1A	0.7931	0.0748	-27.66
2A	0.7931	0.0748	-27.66
3A	0.7793	0.0775	-28.48
4A	0.7793	0.0775	-28.48
5A	0.1113	0.0182	-487.36
6A	0.1528	0.0236	-358.98

Table 3.5A

Estimation of Cov(X1,Z) Using Method 1
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovX1Z)/SD

POP	COVX1Z	MEAN	SD	T
1A	0.9	0.890	0.062	-1.61
2A	0.9	0.890	0.062	-1.61
3A	0.7	0.676	0.057	-4.21
4A	0.7	0.676	0.057	-4.21
5A	0.3	0.296	0.067	-0.60
6A	0.3	0.291	0.066	-.136

Table 3.5B

Estimation of Cov(X2,Z) Using Method 1
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovX2Z)/SD

POP	COVX2Z	MEAN	SD	T
1A	0.9	0.883	0.066	-2.58
2A	0.9	0.883	0.066	-2.58
3A	0.7	0.671	0.064	-4.53
4A	0.7	0.671	0.064	-4.53
5A	0.3	0.294	0.073	-0.82
6A	0.3	0.291	0.070	-1.29

Table 3.5C

Estimation of Cov(X1,Z) Using Method 2
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovX1Z)/SD

POP	COVX1Z	MEAN	SD	T
1A	0.9	0.893	0.063	-1.11
2A	0.9	0.893	0.063	-1.11
3A	0.7	0.681	0.060	-3.17
4A	0.7	0.681	0.060	-3.17
5A	0.3	0.298	0.066	-0.30
6A	0.3	0.293	0.067	-1.04

Table 3.5D

Estimation of Cov(X2,Z) Using Method 2
Mean is the Mean Estimate Over the 100 Replications.
SD is the Estimated Standard Deviation of the Estimates.
T = 10(Mean - CovX2Z)/SD

POP	COVX2Z	MEAN	SD	T
1A	0.9	0.887	0.065	-2.00
2A	0.9	0.887	0.065	-2.00
3A	0.7	0.678	0.061	-3.61
4A	0.7	0.678	0.061	-3.61
5A	0.3	0.297	0.073	-0.41
6A	0.3	0.295	0.071	-0.70

PART 4

PHASE II: ANALYSIS BASED ON 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. This same function was studied for sensitivity to the number of common variables used and the Mahalanobis distance function was tested on non-simulated data. 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 using 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.

4.1 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 (see Appendix B for codebook descriptions).

Five subsamples of approximately 1000 records were randomly selected for matching purposes. In addition, one of the subsamples (SIE5) was used to form "other subsamples" 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 on each file summed to approximately 1.4 million. The subsamples created from SIE5 and their characteristcs are described in Table 4.1.

Table 4.1

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 4.2A-F. In addition, the population percentage frequency counts for all Y-Z item pairs are given in Tables 4.3A-I.

Table 4.2A

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	- .05	- .02	- .04	.05	1.00

Table 4.2B

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	0

Table 4.2C

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 4.2D

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 4.2E

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 4.2F

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 4.3A

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 4.3B

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 4.3C

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 4.3D

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 4.3E

Population

Social Security and Family Size Joint Distribution (Percentage Counts)

	Family Size			
Social Security	1	2	3	4 Plus
$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 4.3F

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 4.3G

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 4.3H

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 4.3I

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

4.2. 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 subsamples 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.

4.2.1 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

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

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,

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

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 4.4. 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 4.5. The age penalty is based upon the age of the first person in the tax record.

Table 4.4

C_2 = Interest Income Difference Index

File A Interest Income	File B Interest Income				
	$0	$1-1000	$1001-10000	$10001-100000	$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 4.5

C_4 = Penalty Index for Difference in Ages

File A Age	File B Age				
	≤ 17	$18 \leq 22$	$23 \leq 61$	$62 \leq 65$	65 and Over
≤ 17	0	12	32	80	80
$18 \leq 22$	12	0	24	80	80
$23 \leq 61$	32	24	0	64	80
$62 \leq 65$	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:

$$C_5 = 16 \cdot \text{INT}\left(\left|\text{Grade of A} - \text{Grade of B}\right| \div 3\right)$$

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 difference 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 summation of variables C_k.

$$\text{Distance Function Value for a given pair of A and B records} = \sum_{k=1}^{6} C_k$$

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.

d_{ij} = 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.

4.2.2 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:

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

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

Σ_{XX} is the covariance matrix of the X variable from the population file.

4.2.3 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,

highest grade attended, and sex components of Model 1, i.e., C_4, C_5, and C_6 of Model 1.

4.2.4 Matched Files Created

The specifications of the 50 generated matched files using the transportation algorithm are given in Table 4.17. 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 samples 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 file 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).

4.2.5 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 with the corresponding distributions in the population file. One method is to calculate a simple χ^2 statistic for each Y-Z pair, based upon cross-tabulated, percentage for the categories specified in Tables 4.3-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,

$$\chi_k^2 = (fm - fp)^2/fp$$

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

fp = weighted percentage of population records in the k-th cell taken from Table 4.3A through 4.3I

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

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 by itself

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 4.3A, 4.3D, and 4.3G 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 4.3D-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.

4.3 Comparison of Absolute Difference and Mahalanobis Distance Functions

Matched files 1-20 identified in Table 4.6 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 4.7 summarizes this χ^2 statistic for each of the nine frequency count tables representing all Y-Z distributions. The rows in Table 4.7 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.

Table 4.6

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[M1]	#24[M1]	#25[M1]	#26[M1]	#27[M1]	#28[M1]	#29[M1]	#30[M1]

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.

The following table is given to illustrate one of the χ^2 calculations in Table 4.8. Table 4.7 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 distribution in Table 4.7 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, and the marginal distribution of the Z variable dividend, in this table is identical to the marginal Z distribution in sample SIE2.

Table 4.7

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 4.7 are given in Table 4.3A 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 4.7 is, for all practical purposes, the same as the distribution in Table 4.3A, 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 4.3A and 4.7, 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 4.8. 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 4.8 using the χ^2 test that the obsolute 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 4.9A gives the result of subtracting the population correlation matrix given in Table 4.1 from the average correlation matrix obtained from matched files 1-10.

The blocked-in portion of Table 4.8 represents the Y-Z distribution where for the ideal match all entries should be zero or close to zero. It is observed from Table 4.9A that these are significant differences from zero where the big differences are for the correlations between total income and

Table 4.8

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 Family Size Omitted.

Table 4.9A

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

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 was also very evident from Table 4.8. However, the difference between the population and matched file distribution 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 4.9A is that six of the differences are negtive 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 4.9B gives the difference between the average correlation matrix for matched files 11-20 and the population correlation matrix given in Table 4.1.

From Table 4.9B it is observed that the Mahalamobis distance function produces larger deviations from the population than the absolute difference distance function represented in Table 4.9A. The blocked-in portion of Table 4.9B represents the Y-Z distributions and it is observed that correlation between family size and number of adults, between dividends and total income,

Table 4.9B

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

and between family size and total income are very different from the population correlations. As observed in Table 4.9A, 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 4.9A and 4.9B that the absolute value distance function is better than the Mahalanobis distance function.

4.4 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 and 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 4.10 displays the 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 Table 4.10 it is observed from the row average column that Model 4 yields the largest average χ^2 statistic in three of the four groupings. It is also observed from Table 4.10 that Model 4 has the largest column average in seven 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 not too bad given the fact 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 4.8. The

Table 4.10

Contingency Table χ^2 Values Using 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, and 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, and 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.

difference between the average correlation matrix using Model 3 and the population correlation matrix is given in Table 4.11. The difference between the average correlation matrix using Model 4 and the population correlation matrix is given in Table 4.12.

Table 4.11

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 4.12

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 4.11 and 4.12 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 4.8, 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 4 correlation is .18.

It should be noted that the entries in Tables 4.8, 4.11, and 4.12 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 the 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.

4.5 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 4.7. 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 4.6. In all cases, matching Model 1 is used.

Table 4.13 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 29%, 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 a 15% downward bias and noise factor in assets in 25% of the records in SIE5, and matched file 38 has a 15% downward bias and noise factor in assets for all records in SIE5.

In Table 4.13 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 for matched files 31-38 when compared with the row entries for matched file 10 that, on the average, the 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 4.13 is that moderate amounts of noise, bias, and combined noise and bias do not greatly affect the Y-Z distributions in matched files.

Table 4.13

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 from Row Average.

4.6 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 4.6 as numbers 21 (Models 5 and 1), 2 (Models 5 and 2), 39 (Models 5 and 3), 48 (Models 5 and 4), 49 (Models 5 and 5), 50 (Models 5 and 6) and 23-30 (noise and bias tests).

Matched files 39, 48, 49, and 50 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 this file 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 4.14 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 matched 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 4.15.

Table 4.14

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 4.15 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 4.15

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.

4.7 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 4.2.1. 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 4.16A for all five sample files. Table 4.16B 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 $\pm$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 4.16C. Files SIE6A through SIE8B are identical to file SIE5 but with the X2 variables perturbed with noise or bias as described in Table 4.1. Table 4.16C 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

Table 4.16A

Dividend Income Statistics for Sample Files

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

Table 4.16B

SIE2 Dividend Income Statistics
After Unconstrained Merging

File A	File B	Dividend Income from SIE2 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 4.16C

Unconstrained Merges: Noise and Bias Tests

File A	File B	Dividends on File B Mean	Dividends on File B 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

weights and hence preserve all of the X2 and Z data items and their interrelationships.

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.

4.8 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 4.2.4 and the unconstrained matches discussed in Section 4.7. 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 Mahalanobis distance function.
2. Noise and bias have a nominal effect on 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 8 of the 9 Y-Z distributions specified, using percentage distribution functions and the categories specified in Tables 4.3A-I.

It is extremely interesting to note that the absolute difference distance function used with the transportation model generated the unacceptable Y-Z disbtributions 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 collections. In this situation the population Y-Z characteristics are used to match X1-Y and X2-Z file such that the matched file Y-Z distribution conform to expected patterns.

PART 5

SUMMARY AND CONCLUSIONS

Herein 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 7,000 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 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.

5.1 The Viability of Merging

There are several instances suggested by the studies 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 these Y-Z relationships be preserved in "modestly broad" categories can obtain

generally good results from a merge file created by the transportation model and distance function described in (5.3) below. 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 items 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, [cov(Y,Z|X)=0] this condition is replicated well by merge techniques. Of course this situation is not as useful as where there is such a relationship [cov(Y,Z|X)≠0], a case current merging techniques have difficulty replicating. While the simulation study suggested that all nonzero relationships were 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 to reflect Y-Z relationships to a high degree of accuracy.

5.2 Choice of Merging Model

In summary, both studies indicated that the best results can be obtained by applying an optimal-constrained merge model with an absolute difference distance function.

With the simulation study, the constrained models yielded much better results than the unconstrained models and there was very little difference between constrained absolute difference and Mahalanobis distance functions. These conclusions were based, however, on experiments with small, normally distributed data files having few common variables, but with the advantage of a large number of replications to enhance the generality of the results.

The study based on "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 used.

5.3 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.

5.4 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 $cov(Y,Z|X)\neq 0$ 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 at much larger 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 $cov(Y,Z|X)$ significantly different from zero.

5.5 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. Much data was generated, but the time available for analysis has been extremely limited.

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.

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 Applications, 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 Microanalytic 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 Labelling 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 appears 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: 353-372.

[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). "Comment." 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

APPENDIX A

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

It can be shown algebrically and understood intuitively that the statistical merging technique chosen directly affects the statistical structure of the resultant composite 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

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, avg.	n.a.	2.65
Transfer income (000s)	n.a.	5,200

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 defined as

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

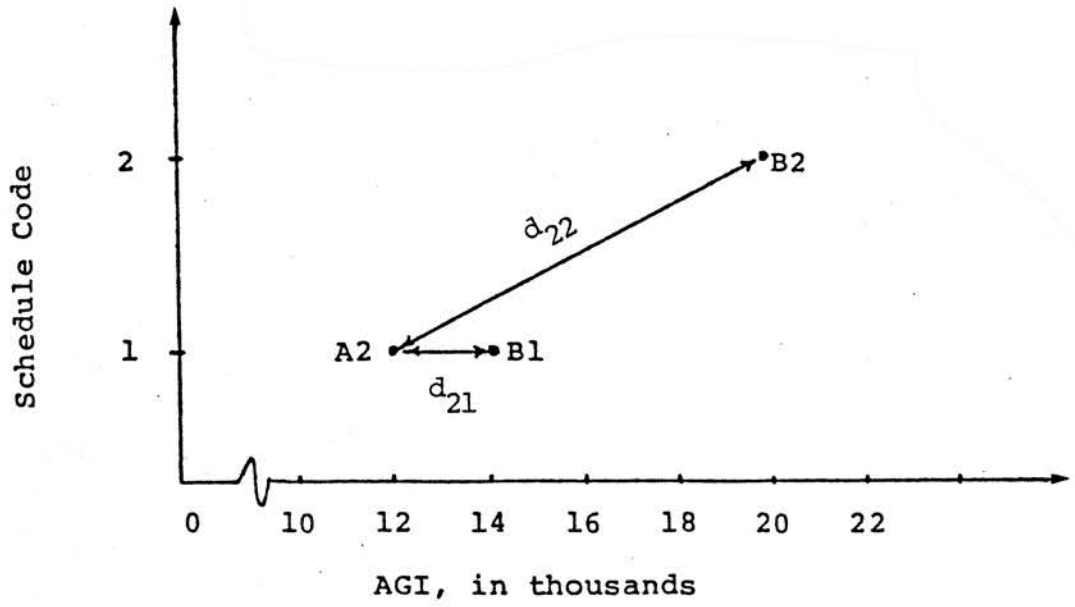


Figure A.2. Scatter Diagram of Selected Records

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.

Tableau 0. Sample Tableau for Example Merge Problem

FILE A RECORD	FILE B 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	

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:

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

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

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

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

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

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:

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

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).

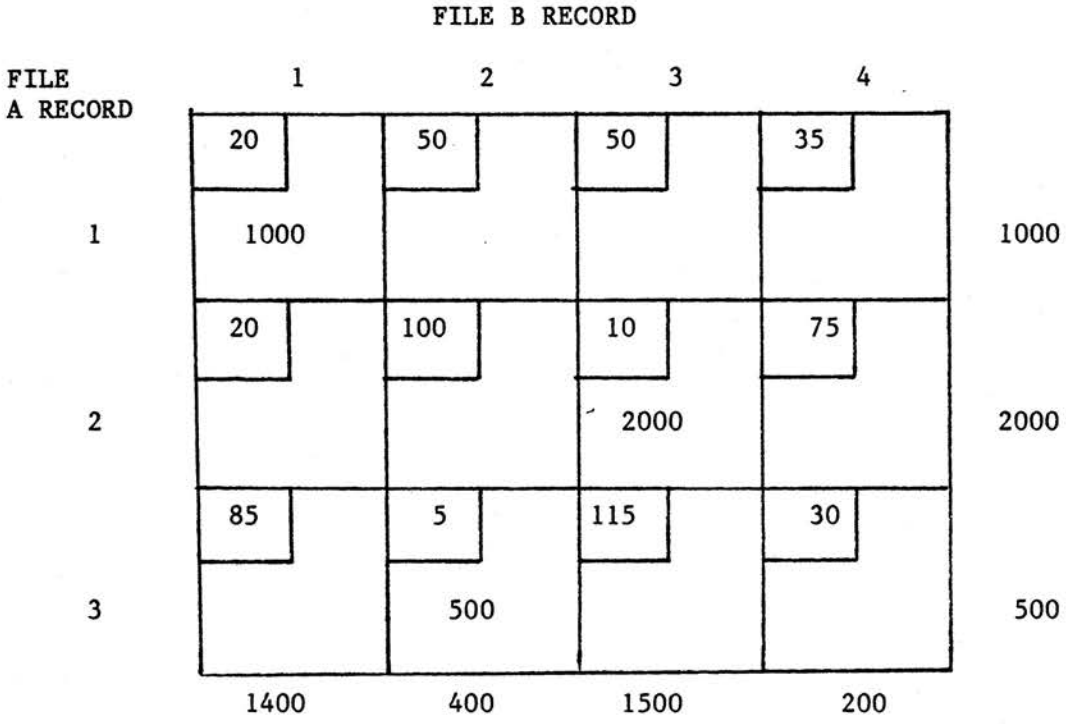
The first, unconstrained matching uses one file as a base and matches each record with the minimum-distance record in the other file. The merged records use the weights from the base file records. This problem can be described mathematically by the set of expressions (A.1),(A.3),(A.4),(A.5) or (A.1),(A.2),(A.4),(A.5).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



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 AGI Difference	1,214	n.a.

Tableau 2. Unconstrained File A (Ignore A Weights)

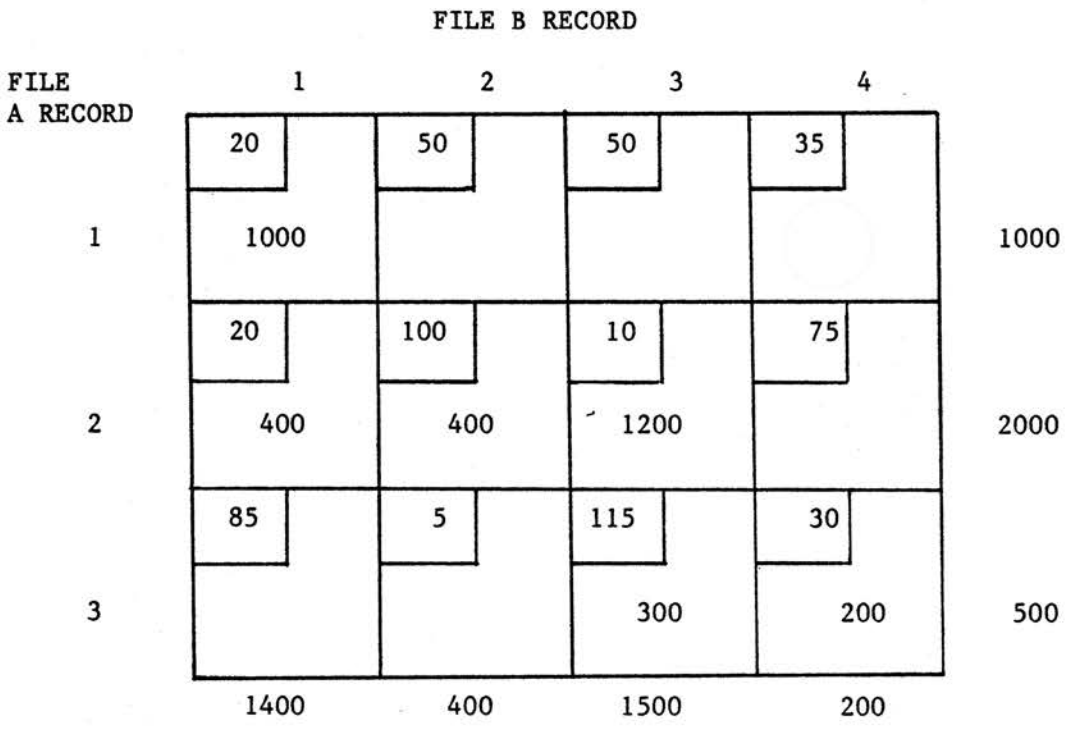
FILE A RECORD	FILE B RECORD 1	2	3	4	
1	20 1400	50	50	35	1000
2	20	100	10 1500	75	2000
3	85	5 400	115	30 200	500
	1400	400	1500	200	

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 AGI Difference	1,457	n.a.

Tableau 3. Constrained Match



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 AGI 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.

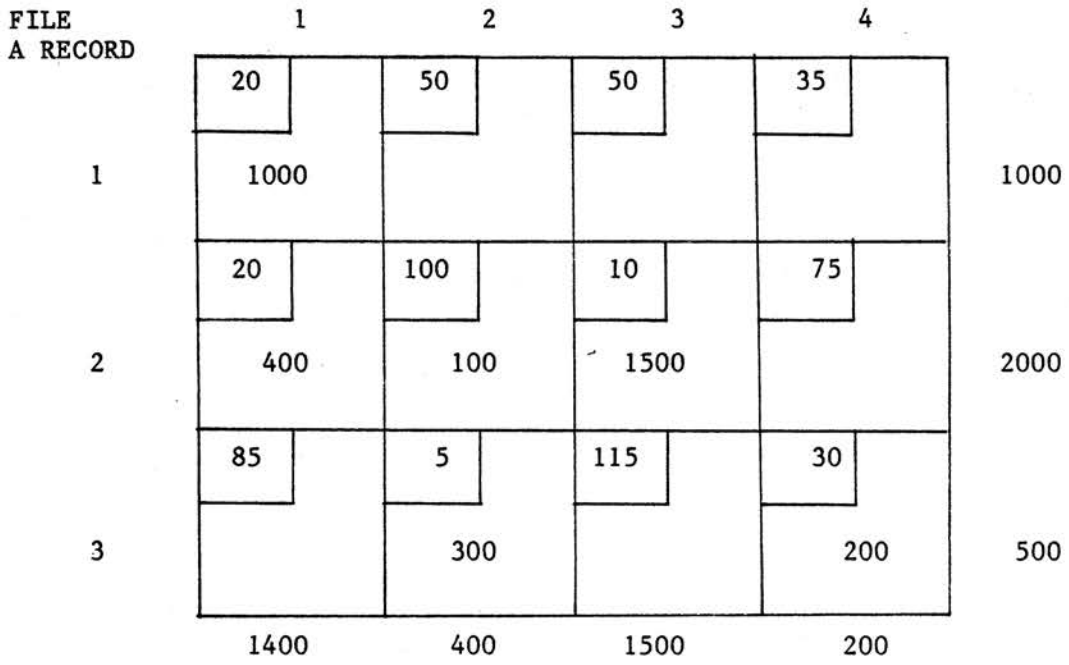
Tableau 3 depicts a constrained match. Note that all match weights in a row sum to the row total (original file A record weight) and column sums are similarly kept. By so doing, the original data structures are maintained and all tabulations are the same as those for the original files.

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 unconstained matches.

To improve this solution to the greatest extent possible, expression (A.5) can be included and a constained-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.

Calculations from Tableau 3 indicate that if the third record of A were matched with the second record of B ($w_{32}>0$) and w_{33} set to zero, and

Tableau 3a. Constrained Match, Improved Solution 1

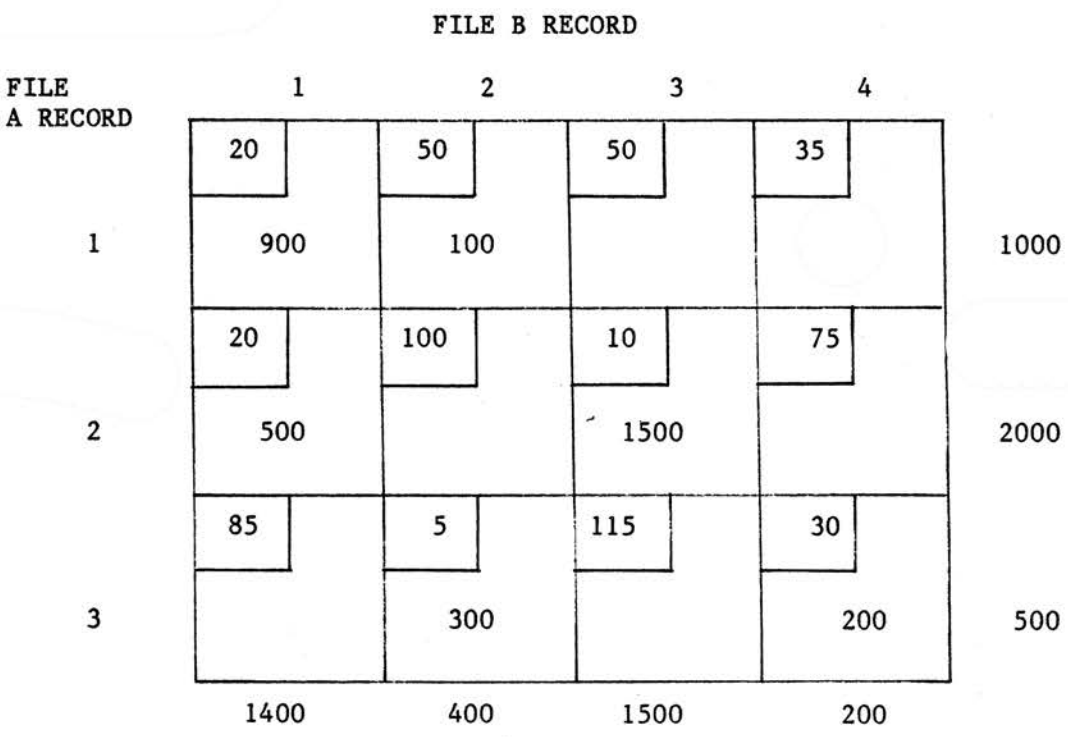


Total solution distance = 60,500

****
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 AGI Difference	1,657	n.a.

Tableau 3b. Constrained Match, Improved Solution 2



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 AGI Difference	1,542	n.a.

adjustments made to the other positive weights to maintain the row and column totals, then the improved solution of Tableau 3a could be obtained. Note that with this match, total distance drops by one-half and there is a 97 percent schedule code agreement, improvement of $1,285 in average AGI discrepancy and the row and column totals are maintained. Further, by increasing w_{12} and dropping w_{22}, the next improved Tableau 3b can be obtained with a total distance lowered by 1000 and the average AGI discrepancy is proved by $115 over Tableau 3a.

Finally, by increasing w_{14} and deleting w_{12}, the constrained-optimal solution of Tableau 4 can be obtained. This match not only has a total solution value and schedule code agreement which is comparable to the unconstrained matches, but has an average AGI discrepancy that is <u>better</u> than that of the unconstrained file A match of Tableau 2.

In summary, this small example illustrates that while unconstrained file matching can yield a closer overall match, serious data distortions can be introduced. These distortions can be overcome by constrained matching which requires that the match file record weights sum to the original individual record weights. And by devising a constrained-optimal match, not only are such data distortions eliminated but extremely close matches at the record level can be obtained.

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 AGI Difference	1,400	n.a.

The following papers are currently available in the Edwin L. Cox School of Business Working Paper Series.

79-100 "Microdata File Merging Through Large-Scale Network Technology," by Richard S. Barr and J. Scott Turner

79-101 "Perceived Environmental Uncertainty: An Individual or Environmental Attribute," by Peter Lorenzi, Henry P. Sims, Jr., and John W. Slocum, Jr.

79-103 "A Typology for Integrating Technology, Organization and Job Design," by John W. Slocum, Jr., and Henry P. Sims, Jr.

80-100 "Implementing the Portfolio (SBU) Concept," by Richard A. Bettis and William K. Hall

80-101 "Assessing Organizational Change Approaches: Towards a Comparative Typology," by Don Hellriegel and John W. Slocum, Jr.

80-102 "Constructing a Theory of Accounting--An Axiomatic Approach," by Marvin L. Carlson and James W. Lamb

80-103 "Mentors & Managers," by Michael E. McGill

80-104 "Budgeting Capital for R&D: An Application of Option Pricing," by John W. Kensinger

80-200 "Financial Terms of Sale and Control of Marketing Channel Conflict," by Michael Levy and Dwight Grant

80-300 "Toward An Optimal Customer Service Package," by Michael Levy

80-301 "Controlling the Performance of People in Organizations," by Steven Kerr and John W. Slocum, Jr.

80-400 "The Effects of Racial Composition on Neighborhood Succession," by Kerry D. Vandell

80-500 "Strategies of Growth: Forms, Characteristics and Returns," by Richard D. Miller

80-600 "Organization Roles, Cognitive Roles, and Problem-Solving Styles," by Richard Lee Steckroth, John W. Slocum, Jr., and Henry P. Sims, Jr.

80-601 "New Efficient Equations to Compute the Present Value of Mortgage Interest Payments and Accelerated Depreciation Tax Benefits," by Elbert B. Greynolds, Jr.

80-800 "Mortgage Quality and the Two-Earner Family: Issues and Estimates," by Kerry D. Vandell

80-801 "Comparison of the EEOCC Four-Fifths Rule and A One, Two or Three σ Binomial Criterion," by Marion Gross Sobol and Paul Ellard

80-900 "Bank Portfolio Management: The Role of Financial Futures," by Dwight M. Grant and George Hempel

80-902 "Hedging Uncertain Foreign Exchange Positions," by Mark R. Eaker and Dwight M. Grant

80-110 "Strategic Portfolio Management in the Multibusiness Firm: An Implementation Status Report," by Richard A. Bettis and William K. Hall

80-111 "Sources of Performance Differences in Related and Unrelated Diversified Firms," by Richard A. Bettis

80-112 "The Information Needs of Business With Special Application to Managerial Decision Making," by Paul Gray

80-113 "Diversification Strategy, Accounting Determined Risk, and Accounting Determined Return," by Richard A. Bettis and William K. Hall

80-114 "Toward Analytically Precise Definitions of Market Value and Highest and Best Use," by Kerry D. Vandell

80-115 "Person-Situation Interaction: An Exploration of Competing Models of Fit," by William F. Joyce, John W. Slocum, Jr., and Mary Ann Von Glinow

80-116 "Correlates of Climate Discrepancy," by William F. Joyce and John Slocum

80-117 "Alternative Perspectives on Neighborhood Decline," by Arthur P. Solomon and Kerry D. Vandell

80-121 "Project Abandonment as a Put Option: Dealing with the Capital Investment Decision and Operating Risk Using Option Pricing Theory," by John W. Kensinger

80-122 "The Interrelationships Between Banking Returns and Risks," by George H. Hempel

80-123 "The Environment For Funds Management Decisions In Coming Years," by George H. Hempel

81-100 "A Test of Gouldner's Norm of Reciprocity In A Commercial Marketing Research Setting," by Roger Kerin, Thomas Barry, and Alan Dubinsky

81-200 "Solution Strategies and Algorithm Behavior in Large-Scale Network Codes," by Richard S. Barr

81-201 "The SMU Decision Room Project," by Paul Gray, Julius Aronofsky, Nancy W. Berry, Olaf Helmer, Gerald R. Kane, and Thomas E. Perkins

81-300 "Cash Discounts To Retail Customers: An Alternative To Credit Card Performance," by Michael Levy and Charles Ingene

81-400 "Merchandising Decisions: A New View of Planning and Measuring Performance," by Michael Levy and Charles A. Ingene

81-500 "A Methodology For The Formulation and Evaluation of Energy Goals And Policy Alternatives For Israel," by Julius Aronofsky, Reuven Karni, and Harry Tankin

81-501 "Job Redesign: Improving The Quality of Working Life," by John W. Slocum, Jr.

81-600 "Managerial Uncertainty and Performance," by H. Kirk Downey and John W. Slocum, Jr.

81-601 "Compensating Balance, Rationality, and Optimality," by Chun H. Lam and Kenneth J. Boudreaux

81-700 "Federal Income Taxes, Inflation and Holding Periods For Income-Producing Property," by William B. Brueggeman, Jeffrey D. Fisher, and Jerrold J. Stern

81-800 "The Chinese-U.S. Symposium On Systems Analysis," by Paul Gray and Burton V. Dean

81-801 "The Sensitivity of Policy Elasticities to the Time Period Examined in the St. Louis Equation and Other Tests," by Frank J. Bonello and William R. Reichenstein

81-900 "Forecasting Industrial Bond Rating Changes: A Multivariate Model," by John W. Peavy, III

81-110 "Improving Gap Management As A Technique For Reducing Interest Rate Risk," by Donald G. Simonson and George H. Hempel

81-111 "The Visible and Invisible Hand: Source Allocation in the Industrial Sector," by Richard A. Bettis and C. K. Prahalad

81-112 "The Significance of Price-Earnings Ratios on Portfolio Returns," by John W. Peavy, III and David A. Goodman

81-113 "Further Evaluation of Financing Costs for Multinational Subsidiaries," by Catherine J. Bruno and Mark R. Eaker

81-114 "Seven Key Rules For Successful Stock Market Speculation," by David Goodman

81-115 "The Price-Earnings Relative As An Indicator of Investment Returns," by David Goodman

81-116 "Strategic Management for Wholesalers: An Environmental Management Perspective," by William L. Cron and Valarie A. Zeithaml

81-117 "Sequential Information Dissemination and Relative Market Efficiency," by Christopher B. Barry and Robert H. Jennings

81-118 "Modeling Earnings Behavior," by Michael F. van Breda

81-119 "The Dimensions of Self-Management," by David Goodman and Leland M. Wooton

81-120 "The Price-Earnings Relatives - A New Twist To The Low-Multiple Strategy," by David A. Goodman and John W. Peavy, III.

82-100 "Risk Considerations in Modeling Corporate Strategy," by Richard A. Bettis

82-101 "Modern Financial Theory, Corporate Strategy, and Public Policy: Three Conundrums," by Richard A. Bettis

82-102 "Children's Advertising: The Differential Impact of Appeal Strategy," by Thomas E. Barry and Richard F. Gunst

82-103 "A Typology of Small Businesses: Hypothesis and Preliminary Study," by Neil C. Churchill and Virginia L. Lewis

82-104 "Imperfect Information, Uncertainty, and Credit Rationing: A Comment and Extension," by Kerry D. Vandell

82-200 "Equilibrium in a Futures Market," by Jerome Baesel and Dwight Grant

82-201 "A Market Index Futures Contract and Portfolio Selection," by Dwight Grant

82-202 "Selecting Optimal Portfolios with a Futures Market in a Stock Index," by Dwight Grant

82-203 "Market Index Futures Contracts: Some Thoughts on Delivery Dates," by Dwight Grant

82-204 "Optimal Sequential Futures Trading," by Jerome Baesel and Dwight Grant

82-300 "The Hypothesized Effects of Ability in the Turnover Process," by Ellen F. Jackofsky and Lawrence H. Peters

82-301 "Teaching A Financial Planning Language As The Principal Computer Language for MBA's," by Thomas E. Perkins and Paul Gray

82-302 "Put Budgeting Back Into Capital Budgeting," by Michael F. van Breda

82-400 "Information Dissemination and Portfolio Choice," by Robert H. Jennings and Christopher B. Barry

82-401 "Reality Shock: The Link Between Socialization and Organizational Commitment," by Roger A. Dean

82-402 "Reporting on the Annual Report," by Gail E. Farrelly and Gail B. Wright

82-403 "A Linguistic Analysis of Accounting," by Gail E. Farrelly

82-600 "The Relationship Between Computerization and Performance: A Strategy For Maximizing The Economic Benefits of Computerization," by William L. Cron and Marion G. Sobol

82-601 "Optimal Land Use Planning," by Richard B. Peiser

82-602 "Variances and Indices," by Michael F. van Breda

82-603 "The Pricing of Small Business Loans," by Jonathan A. Scott

82-604 "Collateral Requirements and Small Business Loans," by Jonathan A. Scott

82-605 "Validation Strategies For Multiple Regression Analysis: A Tutorial," by Marion G. Sobol

82-700 "Credit Rationing and the Small Business Community," by Jonathan A. Scott

82-701 "Bank Structure and Small Business," by William C. Dunkelberg and Jonathan A. Scott

82-800 "Transportation Evaluation in Community Design: An Extension with Equilibrium Route Assignment," by Richard B. Peiser

82-801 "An Expanded Commercial Paper Rating Scale: Classification of Industrial Issuers," by John W. Péavy, III and S. Michael Edgar

82-802 "Inflation, Risk, and Corporate Profitability: Effects on Common Stock Returns," by David A. Goodman and John W. Peavy, III

82-803 "Turnover and Job Performance: An Integrated Process Model," by Ellen F. Jackofsky

82-804 "An Empirical Evaluation of Statistical Matching Methodologies," by Richard A. Barr, William H. Stewart, and J. Scott Turner