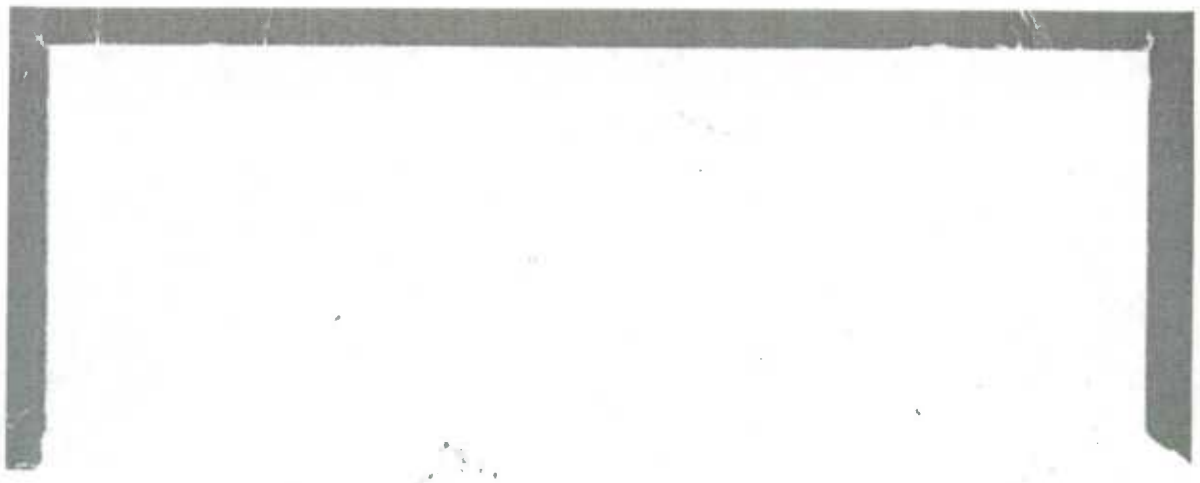
Statistics Canada Statistique Canada



M

Busines

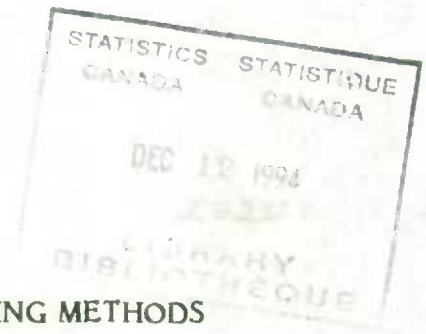
Cana

WORKING PAPER NO. BSMD-90-003E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-90-003E

DIRECTION DE LA MÉTHODOLOGIE



AN EVALUATION OF STATISTICAL MATCHING METHODS

by

J. Armstrong

December 1989

AN EVALUATION OF STATISTICAL MATCHING METHODS

J. Armstrong
Business Survey Methods Division, Statistics Canada, Ottawa
December 1989

Abstract

Statistical matching is a technique for combining information from two microdata sets that does not require the presence of unique identifiers. Validity of results obtained using statistical matching depends on assumptions about relationships between variables that are unique to each input file. Univariate and bivariate distributions on files created by statistical matching may be subject to additional distortions. Performances of a number of statistical matching methods are examined using synthetic data as well as a microeconomic data file created by exact matching of information from the Survey of Consumer Finance and Revenue Canada's tax file. Statistical matching results are compared to those obtained using imputation methods that require a small microdata set containing information about all variables.

Résumé

L'appariement statistique est une méthode utilisée afin de combiner de l'information provenant de deux bases de micro-données. La qualité des résultats de l'appariement statistique dépend des hypothèses concernant les rapports entre des variables provenant d'un seul fichier d'entrée. L'utilisation de l'appariement statistique peut causer d'autres déformations dans les distributions univariées et bivariées des variables. On examine la performance de plusieurs méthodes d'appariement statistique utilisant des données synthétiques ainsi qu'un fichier de micro-données économiques obtenu selon un appariement exact de l'information de l'Enquête sur les finances des consommateurs et des dossiers d'impôt de Revenu Canada. On compare les résultats de l'appariement statistique et ceux des méthodes d'imputation qui requièrent une petite base de micro-données renfermant de l'information pour toutes les variables.

1. Introduction

The Social Policy Simulation Database (SPSD) is a microdata base constructed using survey and administrative data from a variety of sources and intended for use in economic policy analysis. Construction of the database, described in Wolfson et. al. (1987), involves combination of information about individuals from a variety of sources. The basis of the SPSD is microdata from the Survey of Consumer Finance (SCF). During construction of the SPSD, information from the Family Expenditure Survey (FAMEX) and files containing unemployment insurance histories is added to the SCF microdata. In addition, certain income items are replaced, for high income earners, by information obtained from the Revenue Canada file of taxfiler data. The Revenue Canada file is also used to add various deductions, exemptions and tax credits needed to calculate income tax liability.

The operations used in SPSD construction to combine information from various sources are all examples of the problem of file merging. A general formulation of this problem involves three sets of variables (X, Y and Z) and two data files (A and B). File A contains values of X and Y for each individual in a subset of the population P and file B contains values of X and Z for each individual in another subset of P. The objective is to create a merged file C, containing values of X, Y and Z for each individual represented on the file A.

The file merging problem can be addressed using exact matching if the X variables contain sufficient information to positively identify individuals and information for each individual on file A is also available on file B. Exact matching was not used during SPSD construction for a number of reasons. Information sufficient to identify individuals was not always available and in most cases there was no information on file B about many individuals represented on file A. In addition, exact matching was considered undesirable for confidentiality reasons. In situations in which exact matching is undesirable or not feasible, the general problem of completing file A records with Z values can be addresed by selecting, for each record on file A, Z values from a file B record that is close to the file A record, based on values of the X variables. This approach is called statistical matching. All the processes in SPSD construction involving the addition of information from other sources to SCF microdata used statistical matching. A report by the U.S. Department of Commerce (1980) includes a review of both exact matching and statistical matching methods.

There is an obvious relationship between statistical matching and imputation. Consider the file created by combining records from A and B. This file contains (X,Y,*) records from file A and (X,*,Z) records from file B. (The symbol * denotes missing information.) The problem of imputing Z values to complete (X,Y,*) records in this combined file is the problem of statistical matching.

In this paper we will use the term imputation to refer to methods that can be used when an additional file, D, containing values of Y and Z for a subset of individuals in P, is available. Two general classes of imputation methods can be distinguished -- procedures that are appropriate when file D contains information about X, Y and Z; and those than can be used when file D contains information on Y and Z only. Thus far, information about joint distributions of Y and Z variables has not been available during the SPSD construction process. A file recently created by Alter (1988) by exact matching of information from the SCF and the Revenue Canada tax file could be employed as a file D in the context of SPSD construction operations involving the combination of Revenue Canada information with SCF microdata. In addition, the possibility of administering both the SCF and the FAMEX to a common sample, creating a file that could be used during the combination of FAMEX and SCF information, remains open.

In this paper, the results of an evaluation of a number of statistical matching and imputation strategies are presented. The purpose of this evaluation is to compare the statistical matching procedures currently used in construction of the SPSD to other approaches to statistical matching, as well as to provide some evidence concerning the advantages of imputation methods compared to statistical matching. Empirical results obtained using synthetic data, as well as results from use of data from the SCF/Revenue Canada exact match file, are reported here.

The comparisons of statistical matching and imputation techniques were designed to yield some evidence about the benefits of the availability of files containing information about the joint distribution of Y and Z. The comparisons reported in this paper could be used as input to decision processes concerning, for example, the creation of a more recent version of the SCF/Revenue Canada exact match file or the administration of both the SCF and the FAMEX to a common sample.

The plan of the paper is as follows. In section 2, literature related to a variety of statistical matching methods is reviewed. The third section contains a discussion of imputation methods that can be used when a file D is available. The fourth section includes results of an evaluation study involving simulations using synthetic data as well as information from the SCF/Revenue Canada exact match file. Conclusions are found in section 5.

2. Statistical Matching

2.1 Background

The first reference to statistical matching that appears in the literature is Okner (1972). The author describes an approach to the combination of information from two different sources involving the use of "equivalence classes". Classes are defined using ranges of similar X values and the (X,Z) record from file B needed to complete a particular (X,Y) record on file A is selected by minimizing the value of a "closeness" score defined using X values, respecting class boundaries.

Okner's approach was criticized by Sims (1972) on two grounds. First, it involves the implicit assumption that Y and Z are independent conditional on X. When the joint distribution of (X,Y,Z) is multivariate normal, this **conditional independence assumption** implies that the correlation between Y and Z, conditional on X, is zero. Second, if the conditional distribution of Z given X in file B depends on X within "equivalence classes" Okner's approach can produce distortions in the joint distribution of (X,Z) in file C (the file created by matching). Sims points out two types of distortions. First, the conditional variance of Z given X in file C will be biased upwards. Second, in cases in which the density of (X,Z) records in file B varies systematically with X, the mean of the Z distribution in file B will also tend to be biased.

In general, information on a file created by statistical matching may be subject to three types of distortion with potential impact on the analytical uses of the data. These are: (i) distortion in the marginal distributions of Z variables; (ii) distortion in the joint distribution of (X,Z) (equivalent to distortion in the conditional distribution $f(Z|X)$); and (iii) distortion in the joint distribution of (X,Y,Z) (equivalent to distortion in the conditional

distribution $f(Y,Z|X)$). In this context, distortion is measured relative to distributions involving the true (unknown) values of Z for file A records. When the records on files A and B are independent samples from a common distribution, distortions in marginal distributions of Z variables on file C can arise simply because particular records on file B may match with more than one file A record. Distortion in the conditional distribution $f(Z|X)$ on file C can be created, as Sims pointed out, when the conditional distribution $f(Z|X)$ depends on X within equivalence classes. Violations of the conditional independence assumption lead to the third type of distortion.

In the context of SPSD, distortions in the marginal distributions of Z variables on file C are of obvious importance. In analytical work involving the analysis of income and expenditure data, measurement of univariate dispersion and characterization of the extreme portions of univariate distributions is important. Distortions in the joint (X,Z) distribution on file C imply distortions in the marginal distributions of Z variables within categories defined using X. If one of the X variables indicates province, for example, distortions in joint (X,Z) distributions may imply distortions in the distributions of income variables within provinces. Distortions in the conditional distribution $f(Y,Z|X)$ are important when some Y variables are categorical and univariate analysis of Z distributions within Y categories is needed. For continuous variables, the importance of distortion in $f(Y,Z|X)$ depends on the analytical uses of the microdata. If regression equations involving Y and Z variables are estimated, for example, such distortion is important.

Use of the "equivalence class" statistical matching method employed by Okner is analogous, as Singh, Armstrong and Lemaître (1988) point out, to use of hot deck imputation in the file created by combining files A and B. Many statistical matching applications described in the literature have involved an "equivalence class" or "hot deck" approach with the use of distance functions (analogous to Okner's "closeness" score) defined using X variables. Hot deck statistical matching methods were, of course, included in the evaluation study.

Under the assumption that the records on files A and B represent independent samples from a common distribution, distortions in the marginal distributions of Z variables can be eliminated through the use of the constrained matching techniques that are discussed in section 2.2. Two suggestions have appeared in the literature designed, in the absence of

information concerning the distribution of (Y,Z), to militate against violation of the conditional independence assumption. These are the use of multiple imputation ideas (Rubin 1986) and the application to statistical matching of the log-linear imputation method proposed by Singh (1988), and are described in sections 2.3 and 2.4.

2.2 Constrained Matching

Use of constrained matching requires association of weights with records on file A and B. When the file contains data from two surveys of the same population, the survey weights can be employed. Suppose that files A and B include N_A and N_B records, respectively. Let w_i^A and w_j^B denote, respectively, weights of record i on file A and record j on file B. Denote by w_{ij}^C the weight of the file C record formed by combining Z information from record j on file B with file A record i. Finally, let d_{ij} denote distance. Following Rodgers (1984),constrained matching involves minimizing

$$\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} d_{ij} w_{ij}^C \quad ,$$

subject to

$$\sum_{j=1}^{N_B} w_{ij}^C = w_i^A \quad , \quad i=1,2,\ldots N_A$$

$$\sum_{i=1}^{N_A} w_{ij}^C = w_j^B \quad , \quad j=1,2,\ldots N_B$$

$$w_{ij}^C \geq 0 \quad , \qquad i=1,2,\ldots N_A \ , \quad j=1,2,\ldots N_B \ .$$

Since a file C record is created for all $w_{ij}^C > 0$, file C will generally contain more records than file A.

A discussion of constrained matching can be found in Barr and Turner (1980). The advantages of constrained matching are illustrated in simulation studies reported by Barr, Stewart and Turner (1982) and Rodgers and DeVol (1982). The disadvantage of constrained matching is its heavy computational requirement. In addition, the creation of a file C containing more records than file A may be undesirable in some practical applications.

The current SPSD construction process involves a statistical matching procedure incorporating an approximate constraint on the marginal distribution of Z. Categories are formed on file A and file B using all the X variables except one, denoted by X_1. File A and file B records are sorted within each category using X_1. Then the files are matched within each category according to sorted order -- the file A record with the largest value of X_1 is matched with the file B record with the largest value of X_1, etc. File B records are duplicated or skipped when numbers of file A and file B records within a particular category differ. Constrained matching methods were not included in the evaluation study reported in section 4 but the current SPSD approach (an "approximately constrained" method) was, of course, considered.

2.3 Use of Multiple Imputation Ideas

Rubin (1986) suggests the use of multiple imputations to incorporate uncertainty about the conditional correlation of Y and Z given X. The application of multiple imputation ideas is suggested in the context of a statistical matching method involving linear regression. In the absence of information about the conditional correlation of Y and Z given X, Z values are predicted for each file A record using a linear regression of Z on X, estimated using file B information. Let $\hat{Z}$ denote the predicted value of Z for a particular file A record. This record is matched to the file B record with observed Z closest to $\hat{Z}$. Little (1986) calls this method predictive mean matching. In order to incorporate uncertainty about the conditional correlation of Y and Z given X, predicted values of Z for file A records are calculated from linear regressions of Z on X and Y. Estimation of these regressions requires information about the correlation of Y and Z conditional on X. The Rubin scheme involves calculating a number of imputations corresponding to various values for this unknown conditional correlation. Calculations can be facilitated using the sweep operator described in Dempster (1969). The application of Rubin's scheme leads to a number of complete data sets, each corresponding to one set of values for the unknown conditional correlation.

In the context of SPSD, the use of the Rubin methodology poses three practical problems. First of all, its effectiveness depends on reasonable choices for the unknown conditional correlation. Such choices require good auxiliary information. For example, it might be appropriate to apply information about correlations between consumer income and expenditure items in another country in the Canadian context. The availability of this type of information is problematic. The second problem concerns the need to supply appropriate documentation and software to SPSD users. Suppose that the SCF-FAMEX statistical match, for example, was conducted using Rubin's scheme with k different sets of conditional correlations. In this case, it would probably be necessary for SPSD users to: (i) conduct most model experiments k times, using k different values for variables added to the database by statistical matching; and (ii) generate a linear combination of results to obtain a point estimate of the most probable outcome. While such a procedure does not involve any difficulties in principle, effort would be required to ensure that users understand its motivation and find the software sufficiently easy to use. The regression method with predictive mean matching and zero conditional correlation was included in the evaluation study reported in section 4. Variations involving multiple imputation were not considered.

It should be noted that the multiple imputation procedure proposed by Rubin (1978) for use in the context of item non-response in a sample survey can be applied to any non-deterministic statistical matching scheme. Application of multiple imputation would involve assigning two or more values of Z to each file A record by varying the stochastic component of the model used for statistical matching. If, for example, a statistical matching scheme involving random selection of records within categories defined using X variables was employed, application of multiple imputation would result in selection of two or more file B records for each missing value on file A.

Consequently, statistical matching could involve two levels of multiple imputation, one to incorporate uncertainty about the conditional correlation of Y and Z given X, and a second to reflect variation in the predicted values of Z for file A records, given a particular conditional correlation. (These predicted values would be calculated by adding random residuals to conditional expectations of Z obtained from fitted regressions.)

2.4 Application of Log-linear Imputation

The idea of using multivariate histograms during combination of information from two data files was suggested by Sims (1978). He proposed that the multivariate density function of (X,Y,Z) be estimated in accordance with the distribution of (X,Y) from file A and the distribution of (Y,Z) from file B. This estimation problem involves the definition of categorical variables (X^*,Y^*,Z^*). Sims intended that the results of this estimation could provide an alternative to matching — population values of quantities of interest involving both Y and Z could be calculated by integrating the multivariate density. The method was not intended to produce a complete data file.

An imputation method involving the use of categorical variables that can be readily applied to the statistical matching problem is suggested by Singh (1988). This method is one of the alternatives considered in the evaluation study described in section 4. Its theoretical development involves two major elements. First, some criteria for selection of a partition for use in the formation of the categorical variables (X^*,Y^*,Z^*) are introduced. The partition selection criteria are motivated by the idea that a suitable choice of partition is one for which the categorical variables Y^* and Z^* are conditionally independent given X^*. It is stated that appropriate application of these criteria will lead to the selection of an optimal partition. Second, it is suggested that the conditional distribution $f(Z^*|X^*)$ can be estimated using the empirical distribution from file B or a version of this distribution that has been smoothed using a log-linear model.

Given an estimate for the conditional distribution of Z^* given X^*, denoted by $\hat{f}(Z^*|X^*)$, the process of determining Z values to complete file A records involves two parts. First, the value of Z needed to complete a particular file A records is determined up to a Z^* category using this conditional distribution. The second step involves determining a value for Z within the Z^* category.

Depending on the details of the methodology used for each of these parts, variations of log-linear statistical matching can be distinguished. Suppose that these are k file A records in a particular X^* class, say X^*_j. The Z^* class associated with each of these records can be determined independently by prediction using the estimated conditional distribution. In this evaluation study, this method is called unconstrained prediction. Let k_i denote the number of file A records assigned class Z^*_i and let $[x]$ denote the integer part of x. Instead of using unconstrained prediction, the process of imputing a Z^* class can be constrained so that k_i/k

is close to $\hat{f}(Z_i^* | X_j = X_j^*)$. If $k \cdot \hat{f}(Z_i^* | X_j = X_j^*)$ is an integer for all values of i, exact constraints can be imposed. Otherwise, $[k \cdot \hat{f}(Z_i^* | X_j = X_j^*)]$ file A records can be assigned class Z_i^* for each value of i and an alternative method used to determine Z^* classes for the remaining file A records. We call this type of approach constrained prediction. In the context of constrained prediction, various methods can be used to determine Z^* classes for the remaining file A records. One alternative is unconstrained prediction. This method is used in the simulations reported in sections 4.2 and 4.3. Other possibilities include using weighted averages of file B records and creating file C records with fractional weights. More details are given in Singh (1988).

Once a Z^* category, say Z_i^*, has been associated with a file A record in class X_j^*, alternative methods of determing the Z value needed to complete the record can be considered. The use of a Z value taken from a file B record chosen at random from among those file B records in class (X_i^*, Z_j^*) is called random assignment. Alternatively, the file B record can be chosen using a distance measure defined as a function of X.

The Singh partition selection criteria could lead to the selection of partitions that produce files with relatively low levels of distortion in the conditional distribution $f(Y,Z|X)$. The criteria involve use of a non-parametric association measure and an upper bound on the chi-squared distance corresponding to the hypothesis of conditional independence. The results of some limited simulation experiments involving these quantitites (not reported here) suggest that the upper bound on the chi-squared distance is very weak.

Singh suggests a criterion for log-linear model selection based on a non-parametric measure of association. Computer software incorporating this criterion is not readily available. In addition, the problem of determining critical values for the criterion has not been addressed. In the evaluation study, log-linear models used to smooth the conditional distribution $f(Z^*|X^*)$ were chosen using the stepwise deletion procedure described in Benedetti and Brown (1978) and implemented in the BMDP package (Dixon *et al.* 1983). A modification of the Singh approach involving the use of Bayesian methods for estimation of log-linear models has been proposed by Stroud (1989).

3. Imputation

Procedures that can be used to complete (X,Y) records from file A when information about the joint distribution of (Y,Z) is available can be divided into two classes -- methods that require records with (X,Y,Z) information and those that need (Y,Z) information only. Given that the RCT-SCF exact match file provides a sample of records with (X,Y,Z) information, the procedures in the former group are the most important in the context of this paper. All of the methods involving information about the joint distribution of (Y,Z) included in the evaluation study reported in section 4 require (X,Y,Z) records. For completeness, methods that can be used with a sample of (Y,Z) records are also reviewed in this section.

3.1 Auxiliary Information about Distribution of (X,Y,Z)

When a file, say D, containing records with (X,Y,Z) information is available, any imputation method designed for the standard problem of item non-response in a single file can be used to complete file A records using information from file D. Kalton and Kasprzyk (1986) review a number of imputation methods. One can object to the use of imputation methods on the grounds that information on file B is not used. In most statistical matching applications, one would expect file A and file B samples to be much larger than the file D sample. Obviously, the implications of ignoring file B information will be greater in such situations. In addition, in cases in which the file A sample is much larger than the file D sample, each file D record may be used as a donor for a number of file A records and the distribution of Z values on file C may not be as smooth as is desirable.

Section four does not include an extensive evaluation of the use of conventional imputation methods. Attention is restricted to three alternatives:

(i) distance function matching using Euclidean distance;
(ii) distance function matching modified to reduce multiple use of donors;
(iii) regression imputation using stochastic residuals.

These methods are briefly described below.

The modified distance function matching method examined involves ideas from non-parametric regression similar to those incorporated in a method proposed by Paass (1986) for use when (Y,Z) information is available. An estimate of the conditional distribution of Z given X and Y is obtained from file D using the equation

$$E^{+}(g(Z)\,|\,X^{0},Y^{0}) = \sum_{i=1}^{k} b_{i}(X^{0},Y^{0})g(Z^{i}), \tag{1}$$

where $g(Z)$ is a measurable function. The weight $b_i(X^0,Y^0)$ usually decreases as the distance between (X^0,Y^0) and (X^i,Y^i) increases. Two approaches to obtaining weights -- the kernel method and the nearest neighbour method -- can be considered. The kernel method involves the assumption that the distribution of the population can be represented by a mixture of functions that integrate to unity and satisfy certain regularity conditions. The nearest neighbour method involves ranking observations according to their distance from (X^0,Y^0). The values of Z for the k points closest to (X^0,Y^0) are used (possibly after weighting) as a non-parametric estimate of the conditional distribution of Z.

The nearest neighbour method, using Euclidean distance and $k = .02N/\log(N)$, was included in the study reported in section 4. To construct a non-parametric Z value, each of the k points closest to (X^0,Y^0) is assigned a weight inversely proportional to its distance from (X^0,Y^0). Estimates of the conditional distribution obtained using this weighting scheme are asymptotically unbiased with respect to an integral norm (Stone 1977). Apart from the need to satisfy the asymptotic unbiasedness criterion, the weighting scheme was chosen arbitrarily.

The regressions imputation method involves estimation of the linear equation(s)

$$Z = A^{*}X + B^{*}Y + e \tag{2}$$

using file D information. Note that when there is more than one Z variable each regression can be estimated separately to produce maximum likelihood estimates for the system, even if the variance-covariance matrix for the disturbances, $V(e)$, is non-diagonal. This property holds because the same regressors appear in every equation (Zellner 1962). To determine imputed values, residuals that were random drawings from a multivariate normal

distribution with mean and variance-covariance matrix equal to their regression estimates were added to the regression estimates of the conditional expectation of Z given X and Y.

Imputation methods incorporating all available information (files A, B and D) that can be applied to the present problem are described in the literature. These methods require the assumption that the information on each of the three files represents a random sample from the same distribution. In addition it must be assumed that this distribution is completely specified except for certain unknown parameters, say $\underline{a}$. The general idea involves computing the maximum likelihood estimate of $\underline{a}$, denoted by $\underline{a}^*$. Given these estimates, the conditional distribution $f(Z|X,Y,\underline{a}^*)$ can be used for imputation. Two alternatives are possible -- each missing value of Z can be replaced by its conditional mean, or by a random draw from the conditional distribution.

Regardless of the form of the joint distribution of (X,Y,Z), the problem of maximum likelihood estimation of $\underline{a}$ can always be addressed using one of a number of algorithms for multivariate nonlinear optimization (see, for example, Goldfeld and Quandt 1972). Little (1982) mentions two ideas designed to simplify the computations involved -- factorization of the likelihood and use of the expectation-maximization (EM) algorithm described by Dempster, Laird and Rubin (1977). Factorization can be used only when the data pattern is monotone. A data pattern is monotone if all records that are missing a particular number of data items, say n, are missing the same n items, for any value of n. In the present problem, factorization techniques could be used if information on either file A or file B was ignored during estimation. The computations used in the evaluation reported in section 4 did not involve factorization techniques.

The EM algorithm can be used to compute maximum likelihood estimates for general data patterns. In the context of imputation, the EM algorithm has an additional computational appeal. Implementation of the algorithm involves repeated application of two steps -- an expectation step and a maximization step. Missing Z values are replaced by their conditional expectations (given X and Y) during the expectation step. Initially, the programming required for maximum likelihood estimation methods included in the evaluation study was done using the EM algorithm. However, the convergence of this algorithm for some preliminary test cases was very slow. As a result, direct maximization of the likelihood using a combination of the Newton-Raphson method and the method of

steepest ascent was employed. In the simulations using synthetic data, analytic first and second derivatives of the likelihood function corresponding to the assumption of a multivariate normal distribution for (X,Y,Z) were employed. Expressions for the derivatives are given in Srivastava and Carter (1986). Maximum likelihood methods were not evaluated in the simulations using real data.

It is interesting to note the relationship between the regression method proposed by Rubin (1986) and methods involving missing data maximum likelihood techniques. In particular, when file D is used to obtain an estimate of the conditional covariance of Y and Z given X, the predicted value of Z for a file A record produced using the Rubin procedure corresponds to the conditional expectation of Z computed using the distribution $f(Z|X,Y,a^{*})$. One other approach designed to incorporate a file D sample has been suggested in the literature. Singh (1988) proposed a modification of log-linear imputation to incorporate information available from file D. This modification was not evaluated.

3.2 Auxiliary Information about Distribution of (Y,Z)

In order to complete the review portion of this paper, methods available for use when the file D contains (Y,Z) information should be mentioned. In the SPSD context such methods might be considered if, for example, it was decided that more information about the conditional correlations of variables from the SCF with variables from the FAMEX was desirable and that joint administration of the SCF and the FAMEX would impose an unacceptably heavy burden on survey respondents. In this case, one option would involve splitting the SCF into two parts. Let X denote the set of variables collected by FAMEX. Half of the respondents to the FAMEX would be asked only those SCF questions corresponding to Y variables and the other half would be asked only SCF questions corresponding to Z variables.

Two alternative approaches to use of a file D containing (Y,Z) information are suggested in the literature. If one is willing to specify the distribution of (X,Y,Z) except for certain unknown parameters, imputation methods involving maximum likelihood estimation with missing data can be employed. Maximum likelihood estimates can be computed using a multivariate nonlinear optimization procedure or the EM algorithm. If it is not possible to specify the functional form of the (X,Y,Z) distribution, a non-parametric

procedure proposed by Paass (1986) can be employed. This method involves the use of non-parametric techniques to obtain estimates of conditional distributions. The idea is illustrated in section 3.1 in the context of imputation using a file with (X,Y,Z) information. Methods using a file D containing (Y,Z) information were not considered in the evaluation study reported in section 4.

4. Empirical Evaluation

An evaluation study involving the comparison of a number of statistical matching methods is described in this section. The results of three separate components of the study, conducted using Monte Carlo simulation techniques, are reported. In section 4.1, traditional statistical matching techniques and imputation methods requiring a file containing (X,Y,Z) information are examined using simulations with synthetic data. Maximum likelihood methods involving normality assumptions are included in this component of the evaluation study. A variety of evaluation measures, directed at measuring all types of distortions in the distributions of variables on files created by statistical matching, are reported.

Section 4.2 also contains the results of simulations using synthetic data. A variety of statistical matching methods involving the use of log-linear imputation ideas are considered. Maximum likelihood methods are not included. The main evaluation measure used in this component of the study is a measure of distortion in the (X,Z) distribution on the file created by matching. Section 4.3 includes the results of simulations using the same set of methods as in section 4.2 and data from a file created by exact matching of information from the Revenue Canada tax file and the Survey of Consumer Finance (Alter 1988). The evaluation measures reported are directed at distortion in the (X,Z) distribution. Eleven alternative statistical matching and imputation methods are considered in each of sections 4.1, 4.2 and 4.3.

Before providing detailed results of the simulations, some discussion of the overall objectives and limitations of the study is appropriate. The specificity of the results of Monte Carlo experiments — the difficulty involved in the application of Monte Carlo results to situations not directly considered in the original study — is an important limitation. The econometrics literature includes references to techniques that can be used to militate against the importance of specificity. For example, Hendry and Harrison (1974) consider a

situation in which the conditions determined by the experimentor for each set of simulations, involving alternative methods of estimation, can be summarized in terms of a reasonable number of known parameters. In this case, experimental design techniques can be used to choose appropriate parameter values as well as to model the behaviour of alternative methods at parameter points for which simulation results are not available. Such techniques have not been used in the statistical matching context, neither in evaluations of statistical matching that have appeared in the literature, nor in the current study. Statistical matching can be employed, from the operational point of view, to combine information from any pair of data files. Consequently, the parametric characterization of experimental conditions is very difficult. As a result, the generality of any conclusions that can be drawn from the results reported here is not clear.

4.1 Synthetic Data — Part I

The synthetic data simulations reported in this section are similar to other experiments mentioned in the literature and are intended to serve as a benchmark for the discussion in sections 4.2 and 4.3. Four variables, denoted by X_1, X_2, Y and Z are involved. At the beginning of each simulation three files — A, B and D — were available. Each file contained an independent sample of size 250 (files A and B) or 100 (file D) from the distribution $f(X_1, X_2, Y, Z)$. Values of Z on file A and values of Y on file B were suppressed.

Four sets of 25 simulations were conducted using two distributions for (X_1, X_2, Y, Z) and two schemes for definition of (X, Z) classes corresponding to categorical variables (X^*, Z^*). To generate data, the starting point was sets of independent observations (X_1, X_2, Y, Z') from a multivariate normal distribution with mean zero and variance-covariance matrix

$$V_1 = \begin{vmatrix} 1.0 & 0.4 & 0.3 & 0.3 \\ 0.4 & 1.0 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.0 & -0.35 \\ 0.3 & 0.3 & -0.35 & 1.0 \end{vmatrix}.$$

Z' was transformed to produce Z. For two sets of simulations, Z was set equal to Z' with probability 0.9 and equal to $\sqrt{10}Z'$ with probability 0.1. In these cases the marginal

distribution of Z is a mixture of normal distributions. For the other simulations, Z was set equal to Z' with probability 0.5 and equal to $e^{Z'}$ with probability 0.5. The marginal distribution of Z for these simulations is a mixture of a normal distribution and a lognormal distribution.

The conditional distribution of (Y, Z') given (X_1, X_2) corresponding to the multivariate normal joint distribution is bivariate normal with mean zero and covariance matrix

$$V_2 = \begin{vmatrix} .871 & -.479 \\ -.479 & .871 \end{vmatrix}$$

Consequently, the assumption that Y and Z are conditionally independent given (X_1, X_2) is violated in the data sets used in the simulations reported here.

It should be noted that the choice of variance-covariance matrix and the transformations used to generate data for these simulations were largely arbitrary. Paass (1986) reports results of simulations using data from a multivariate normal distribution with mean zero and variance-covariance matrix

$$V_3 = \begin{vmatrix} 1.0 & 0.8 & 0.3 & 0.3 \\ 0.8 & 1.0 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.0 & -0.7 \\ 0.3 & 0.3 & -0.7 & 1.0 \end{vmatrix}.$$

Some experiments were conducted using this variance-covariance matrix. V_1 was used for more extensive simulations because, with files A and B of size 250 and file D of size 100, implementation of matching methods involving classes was difficult due to relatively large numbers of classes with zero counts.

In the four sets of simulations conducted, each data generating procedure was used with two different schemes for defining classes. In the first scheme, three classes were used for X_1, X_2, Y and Z (in the context of those matching methods requiring classes for these variables). Class boundaries for X_1 were determined to make the number of file A records in

each class as equal as possible. The same method was used to choose class boundaries for X_2 and Y. Class boundaries for Z were determined analogously using file B information.

Eleven methods of determining Z values needed to complete file A records were considered. The methods are described below. Methods (i) - (vi) are imputation methods that use file D information. Methods (vii) - (xi) are matching methods.

(i) Hot deck imputation using distance (HDI-D). The Z value needed to complete a file A record is taken from the file D record that is "closest" according to the Euclidean distance measure defined using values of X_1, X_2 and Y. Estimated standard deviations are computed using file A information.

(ii) Hot deck imputation using non-parametric smoothing (HDI-NPS). For each file A record, a set of file D records that are nearest neighbours according to Euclidean distance defined using values of X_1, X_2 and Y is determined. The record from file D that is matched to the file A record is chosen from the set of nearest neighbours using selection probabilities inversely proportional to distance. More details are given in section 3.1.

(iii) Regression imputation with stochastic residuals (RI-SR). The regression of Z on (X_1, X_2, Y) is estimated using file D information. The Z value needed for a particular file A record is computed as the sum of the estimated conditional mean of Z given X_1, X_2 and Y plus a residual drawn from the normal distribution with mean zero and variance equal to the estimated residual variance from the regression.

(iv) Maximum likelihood assuming multivariate normality with zero residuals (ML-ZR). Assuming multivariate normality, maximum likelihood estimates of the mean vector and variance-covariance matrix of the joint distribution of (X_1, X_2, Y, Z) are computed using information from files A, B and D. The Z value used to complete each file A record is the conditional mean of Z given X_1, X_2 and Y, based on the estimated joint distribution.

(v) Maximum likelihood assuming multivariate normality with stochastic residuals (ML-SR). This method is similar to (v) except that a residual drawn from the normal

distribution with mean zero and variance equal to the estimated conditional variance of Z given (X_1, X_2, Y) is added to the estimated conditional mean.

(vi) Maximum likelihood using predictive mean matching (ML-PM). This method is similar to (v). After the estimated conditional mean of Z given X_1, X_2 and Y is computed for a file A record, nearest neighbour matching is used to determine the value from file B that should be used to complete the record.

(vii) Hot deck matching using distance (HDM-D). Let x^A_{1i} and x^A_{2i} denote values of X_1 and X_2 for a particular file A record. Similarly, let x^B_{1j} and x^B_{2j} denote values of X_1 and X_2 for file B record j. Let $\hat{s}_1$ and $\hat{s}_2$ denote estimated standard deviations of, respectively, X_1 and X_2, computed using file A information. The Z value used to complete record i on file A is taken from the file B record for which

$$D_{ij} = ((x^A_{1i} - x^B_{1j})/\hat{s}_1)^2 + ((x^A_{2i} - x^B_{2j})/\hat{s}_2)^2 \qquad (3)$$

is minimized.

(viii) Hot deck matching using random assignment within classes (HDM-RWC). Each file A record is matched with a file B record chosen at random from among those records with the same value for (X^*_1, X^*_2).

(ix) Hot deck matching using distance within classes (HDM-DWC). This method is similar to (i) except that only records with the same values for (X^*_1, X^*_2) are matched.

(x) Hot deck matching using ranks within classes (HDM-RANKWC). This method is similar to the statistical matching method currently used during construction of the SPSD that is described in section 2.2. Within each X_2 category, file A and file B records are ranked according to X_1. Records are matched in rank order, duplicating or skipping file B records as required.

(xi) Regression using predictive mean matching (R-PM). The regression of Z on (X_1, X_2) is estimated using file B information. For each file A record, the estimated

conditional mean of Z given (X_1, X_2) is calculated. The Z value from file B that is closest to this estimate is used to complete the file A record. More details are given in section 2.3.

To evaluate the performance of the matching methods, various measures of distortions in the distribution of Z and distortions in bivariate distributions involving Z on the file created by matching were calculated.

Let $\hat{R}_{ij}$ and $\hat{r}_{ij}$ denote, respectively, empirical rank correlation and empirical Pearson correlation coefficients between variables i and j (using index 1 to refer to X_1, 2 for X_2 and 4 for Z), calculated using information on the file created by statistical matching. Similarly let $\hat{R}^A_{ij}$ and $\hat{r}^A_{ij}$ denote empirical correlation coefficients calculated using file A information (before the suppression of Z values). For each set of simulations and each matching or imputation method examined, Monte Carlo means and standard errors of the quantities

$$B_{i4} = \hat{R}_{i4} - \hat{R}^A_{i4}, \quad i = 1, 2, 3,$$

$$b_{i4} = \hat{r}_{i4} - \hat{r}^A_{i4}, \quad i = 1, 2, 3, 4,$$

were calculated. These quantities give information about the distortions in the univariate distribution of Z on files created by matching for i=4 and information about distortions in bivariate distributions involving Z for other values of i.

The two-sided Kolmogorov-Smirnov test statistic was calculated, using the original file A values of Z and the values of Z on the file created by matching, for each simulation and all methods examined. Since the file A sample size is 250, we will subsequently denote empirical values of this statistic by $\hat{D}_{250}$. Let $\bar{z}^A$ and $\bar{z}$ denote the empirical means of Z values on the original file A and the file created by matching respectively. Monte Carlo means and standard deviations of

$$M = \bar{z}^A - \bar{z}$$

were computed.

Runs tests were also used to assess distortions on files created by matching. The well-known univariate runs test was used to assess distortion in the univariate distribution of Z. For bivariate distributions involving Z, the multivariate runs test proposed by Friedman and Rafsky was employed. To calculate the multivariate runs test statistic for (X_1, Z), for example, the 250 values of (X_1, Z) on the file created by matching was combined with an independent sample of 250 (X_1, Z) values from the distribution used to generate data on the original file A. A minimal spanning tree was constructed for the two-dimensional graphic representation of the set of 500 (X_1, Z) values using the algorithm described by Whitney (1972). The multivariate runs test statistic was computed as one plus the number of nodes in the spanning tree connecting points corresponding to different samples. Subsequently, univariate runs test statistics will be denoted by $\hat{U}_4$ and bivariate runs test statistics by $\hat{U}_{i4}$, $i=1, 2, 3$. Since file A is of size 250, these statistics have expected value 250 under the hypothesis of zero distortion in the corresponding distribution.

The final evaluation measure was suggested by Stroud (1989). The distributions of variables involved in applications like SPSD are often highly skewed and quantile estimates are particularly important in analytical work using income and expenditure data. Let $\hat{Z}_{(r),j}$ denote the empirical estimate of quantile r of the distribution of Z within decile j of the one of the matching variables, say X_1, obtained from the file created by matching. Let $\hat{Z}^A_{(r),j}$ denote the empirical estimate of the same quantity using original file A values. Monte Carlo means and standard errors of

$$Q_{(0.9)} = \sum_{j=1}^{10} \left| \hat{Z}^A_{(0.9),j} - \hat{Z}_{(0.9),j} \right|$$

were computed for both X variables, each set of simulations and each method examined.

References to statistically significant results in subsequent discussion in this section and in sections 4.2 and 4.3 are based on the assumption that Monte Carlo means of evaluation statistics have normal distributions with standard deviations given by the corresponding Monte Carlo standard errors.

Monte Carlo means for differences in the values of Spearman correlation coefficients before and after matching are given in Table 1. These values were computed using 25 simulations. The marginal distribution of Z was a mixture of normals. The numbers of

classes used by methods involving classes are given in parentheses. From an implementation point of view, a set of simulations including HDM-RWC(3), HDM-DWC(3), HDM-RANKWC(3) and the eight methods not involving classes was conducted. Then a second set of simulations involving HDM-RWC(4), HDM-DWC(4), HDM-RANKWC(4) and the eight methods that do not use classes was calculated using the same data. Monte Carlo standard errors for pairwise comparisons of the Monte Carlo means reported in Table 1 (as well as means of other evaluation statistics) are available except for comparisons involving a hot deck method with three categories and a hot deck method with four categories.

Table 1: Normal Mixture

Monte Carlo Means of B_{i4} and their Monte Carlo Standard Errors

Method	i = 1	i = 2	i = 3
HDI-D	-.014 (.026)	.012 (.022)	.050 (.025)
HDI-NPS	-.053 (.027)	-.012 (.019)	.075 (.026)
RI-SR	-.041 (.025)	.009 (.022)	.032 (.024)
ML-ZR	.159 (.044)	.228 (.054)	-.164 (.044)
ML-SR	-.061 (.024)	-.027 (.019)	.069 (.026)
ML-PM	.158 (.044)	.227 (.054)	-.164 (.044)
HDM-D	-.015 (.015)	-.002 (.023)	.454 (.093)
HDM-RWC(3)	-.033 (.016)	-.015 (.022)	.419 (.086)
HDM-RWC(4)	-.023 (.017)	-.011 (.019)	.461 (.094)
HDM-DWC(3)	-.011 (.015)	-.010 (.026)	.458 (.093)
HDM-DWC(4)	-.013 (.014)	-.005 (.026)	.466 (.095)
HDM-RANKWC(3)	-.005 (.013)	-.022 (.019)	.452 (.092)
HDM-RANKWC(4)	-.013 (.013)	.001 (.017)	.460 (.094)
R-PM	.553 (.113)	.513 (.106)	.678 (.136)

Methods using estimation based on normality assumptions that do not involve adding a residual to the estimated conditional mean of Z (ML-ZR; ML-PM; R-PM) produced relatively large values for the Monte Carlo means of B_{14} and B_{24}. The values for B_{14} and B_{24} obtained for these three methods are larger (at the 1% level of statistical significance) than those produced by the other eleven methods. Note that the Table 1 results suggest that maximum likelihood imputation with zero residuals performs very much like maximum likelihood using predictive mean matching. The Monte Carlo standard error for the comparison of Monte Carlo means of B_{14} for these two methods is less than .001. Monte Carlo standard errors for comparisons involving methods other than ML-ZR, ML-PM and R-

PM are between .01 and .03. Consequently, except for the poor performance of ML-ZR, ML-PM and R-PM, the Monte Carlo means of B_{14} and B_{24} do not provide much evidence about the relative merits of the methods considered.

As one would expect, the Monte Carlo means of B_{34} for imputation methods that incorporate information about the (Y,Z) distribution are closer to zero in absolute magnitude than the Monte Carlo means for matching methods that do not use (Y,Z) information. Observed differences between Monte Carlo means for pairs of methods including one matching method and one imputation procedure are all statistically significant at the 5% level. Analogous to the results for B_{14} and B_{24}, ML-ZR and ML-PM performed poorly compared to other methods using (Y,Z) information and the Monte Carlo mean of B_{34} for R-PM does not compare well with the means obtained for other matching methods (not incorporating (Y,Z) information).

The Monte Carlo means of b_{14}, b_{24} and b_{34} reported in Table 2 are qualitatively consistent with the Monte Carlo means of B_{14}, B_{24} and B_{34}. The poor performances of ML-ZR, ML-PM and R-PM are evident. The Monte Carlo means of b_{44} indicate that these methods also lead to more distortion than other alternatives in the marginal distribution of Z on files created by matching.

Table 2: Normal Mixture

Monte Carlo Means of b_{i4} and their Monte Carlo Standard Errors

Method	i			
	1	2	3	4
HDI-D	-.009 (.022)	.002 (.022)	.051 (.026)	-.062 (.14)
HDI-NPS	-.052 (.024)	-.010 (.021)	.087 (.027)	-.226 (.154)
RI-SR	-.018 (.021)	.027 (.022)	-.007 (.023)	-.020 (.135)
ML-ZR	.184 (.048)	.258 (.058)	-.205 (.050)	-1.45 (.301)
ML-SR	-.038 (.022)	-.007 (.018)	.037 (.023)	-.035 (.109)
ML-PM	.184 (.048)	.258 (.058)	-.206 (.05)	-1.45 (.301)
HDM-D	-.015 (.017)	.007 (.022)	.426 (.087)	-.116 (.13)
HDM-RWC(3)	-.052 (.017)	-.033 (.02)	.395 (.081)	.005 (.118)
HDM-RWC(4)	-.017 (.016)	-.000 (.015)	.447 (.091)	-.058 (.104)
HDM-DWC(3)	-.009 (.017)	-.011 (.025)	.425 (.087)	-.126 (.133)
HDM-DWC(4)	-.016 (.016)	.005 (.023)	.437 (.089)	-.101 (.130)
HDM-RANKWC(3)	-.004 (.012)	-.041 (.021)	.419 (.086)	-.087 (.097)
HDM-RANKWC(4)	-.012 (.013)	-.010 (.016)	.429 (.088)	-.089 (.098)
R-PM	.579 (.117)	.537 (.111)	.668 (.134)	-1.76 (.359)

Monte Carlo means of M and their Monte Carlo standard errors are reported in Table 3. These numbers provide evidence of bias in the mean of Z on files created by matching that is statistically significant at the 5% level for six methods — HDI-D, HDI-NPS, RI-SR, ML-ZR, ML-SR and ML-PM — when the marginal distribution of Z is a mixture of normals. The fact that the simulations involving a symmetric Z distribution provide more evidence of bias in the mean of Z on files created by matching than the simulation with a skewed Z distribution is not easy to interpret.

More evidence about distortions in the marginal distribution of Z on files created by matching is provided in Table 4. Using the Smirnov approximation to the mean of the Kolmogorov-Smirnov two-sample test statistic for large samples discussed in Kim (1969), the mean of $\hat{D}_{250}$ under the null hypothesis of no distortion is .0777. Consequently, the Table 4 results provide evidence of distortions in the marginal distribution of Z on files created by matching for all methods except the hot deck matching methods using ranks.

Table 3: Monte Carlo Means of M and their Monte Carlo Standard Errors

Method	Marginal Distribution of Z			
	Normal Mixture		Normal-Lognormal	
HDI-D	.141	(.046)	.003	(.039)
HDI-NPS	.115	(.041)	.020	(.039)
RI-SR	.120	(.042)	.002	(.045)
ML-ZR	.063	(.030)	-.041	(.042)
ML-SR	.053	(.029)	-.062	(.050)
ML-PM	.063	(.030)	-.046	(.046)
HDM-D	.008	(.036)	-.025	(.051)
HDM-RWC(3)	.046	(.038)	-.042	(.045)
HDM-RWC(4)	.026	(.037)	-.043	(.046)
HDM-DWC(3)	.015	(.035)	-.032	(.053)
HDM-DWC(4)	-.005	(.039)	-.005	(.047)
HDM-RANKWC(3)	.035	(.032)	-.044	(.041)
HDM-RANKWC(4)	.031	(.033)	-.035	(.043)
R-PM	.028	(.031)	-.035	(.038)

The poor performances of the methods using modelling based on normality assumptions (RI-SR, ML-ZR, ML-SR, R-PM, ML-PM), particularly when the marginal distribution of Z is a mixture of a normal distribution and a lognormal, are clearly indicated by the Table 4 results and are expected.

Monte Carlo means of runs test statistics are reported in Table 5. Under the null hypothesis of no distortion in the corresponding univariate or bivariate distribution, the expected values of these statistics are 250. The Monte Carlo means provide evidence of distortion in almost all cases.

Some of the Table 5 results may appear to contradict information reported earlier. Note, for example, the small values of the Monte Carlo means of $\hat{U}_4$ (indicating large distortion in the univariate Z distribution) for the hot deck imputation methods (HDI-D and HDI-NPS). In addition, these methods produce lower Monte Carlo means of $\hat{U}_{34}$ than hot deck matching methods. One would expect, of course, that imputation methods involving

Table 4: Monte Carlo Means of $\hat{U}_{250}$ and their Monte Carlo Standard Errors

Method	Marginal Distribution of Z			
	Normal Mixture		Normal-Lognormal	
HDI-D	.114	(.006)	.117	(.007)
HDI-NPS	.114	(.006)	.112	(.007)
RI-SR	.124	(.009)	.177	(.013)
ML-ZR	.158	(.011)	.181	(.014)
ML-SR	.110	(.005)	.191	(.011)
ML-PM	.159	(.011)	.182	(.014)
HDM-D	.099	(.006)	.100	(.005)
HDM-RWC(3)	.096	(.004)	.090	(.006)
HDM-RWC(4)	.096	(.006)	.095	(.006)
HDM-DWC(3)	.097	(.007)	.097	(.005)
HDM-DWC(4)	.100	(.006)	.100	(.005)
HDM-RANKWC(3)	.082	(.004)	.080	(.005)
HDM-RANKWC(4)	.084	(.004)	.081	(.005)
R-PM	.254	(.009)	.256	(.015)

use of information about the (Y,Z) distribution would produce less distortion in the (Y,Z) distribution on file C than matching methods that do not involve such information. The relatively low values of Monte Carlo means of runs test statistics for HDI-D and HDI-NPS can be attributed to the fact that the file D used as a source of Z values included only 100 observations, compared to 250 observations on the file A. Repetition of Z values on file C obviously leads to smaller values for runs test statistics. Note that the best results for (Y,Z) distortion are produced by two imputation methods (RI-SR and ML-SR). Among

matching methods the hot deck method using ranks within categories gives the best results with respect to (Y,Z) distortion as well as distortion in other univariate and bivariate distributions involving Z.

Monte Carlo means of $Q_{(r)}$ statistics, calculated for simulations involving Z data from a mixture of normal distributions, are reported in Table 6. Mean values for these statistics under the null hypothesis of no distortion in the (X,Z) distribution are not known. These means are, of course, non-zero since the number of observations in file A is finite. (Estimates could be obtained by simulation although, since the distribution of $Q_{(r)}$ depends on the (X,Z) distribution, a separate estimate would be needed for each set of Monte Carlo conditions.) Based on previous results it is plausible to assume that the Monte Carlo means reported in Table 6 are larger than the mean of $Q_{(r)}$ under the null hypothesis of no distortion. Standard errors for pairwise comparisons of methods are smaller than the standard errors shown in the table, ranging in general between 0.1 and 0.3. Consequently, the Table 6 results suggest that the hot deck method using ranks produces less (X,Z) distortion than alternative methods and many of the observed differences are statistically significant.

Table 5: Normal Mixture

Monte Carlo Means of Runs Test Statistics and their Monte Carlo Standard Errors

Method	Statistic			
	$\hat{U}_{14}$	$\hat{U}_{24}$	$\hat{U}_{34}$	$\hat{U}_{4}$
HDI-D	205.2 (2.58)	203.7 (2.05)	204.4 (2.82)	128.9 (1.09)
HDI-NPS	219.0 (1.81)	219.8 (2.25)	221.0 (2.53)	131.7 (1.33)
RI-SR	242.6 (2.12)	242.9 (3.29)	243.6 (2.3)	241.0 (3.1)
ML-ZR	223.3 (5.75)	214.8 (5.77)	216.0 (5.19)	224.6 (5.35)
ML-SR	241.4 (2.8)	245.0 (1.69)	243.3 (2.38)	243.3 (2.36)
ML-PM	224.0 (5.9)	214.7 (5.54)	216.0 (5.01)	176.3 (4.49)
HDM-D	205.7 (2.23)	204.8 (2.56)	229.3 (2.27)	183.8 (1.43)
HDM-RWC(3)	242.0 (2.18)	235.8 (1.96)	234.3 (2.0)	191.2 (2.31)
HDM-RWC(4)	235.0 (2.69)	228.3 (2.1)	231.7 (2.0)	191.0 (1.89)
HDM-DWC(3)	210.0 (1.79)	206.4 (2.00)	228.6 (2.32)	180.6 (1.88)
HDM-DWC(4)	206.7 (2.12)	202.3 (2.03)	230.6 (2.26)	182.1 (1.41)
HDM-RANKWC(3)	248.0 (2.26)	247.2 (1.72)	234.1 (1.9)	243.8 (1.99)
HDM-RANKWC(4)	245.7 (2.48)	248.0 (2.43)	237.8 (2.37)	240.3 (2.21)
R-PM	134.4 (7.63)	146.6 (4.39)	176.0 (3.68)	142.2 (3.75)

Table 6: Normal Mixture
Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

Method	r = 0.9			
	X_1		X_2	
HDI-D	6.25	(0.65)	6.47	(0.59)
HDI-NPS	5.59	(0.41)	6.21	(0.45)
RI-SR	5.60	(0.42)	5.72	(0.36)
ML-ZR	5.42	(0.40)	5.61	(0.45)
ML-SR	5.24	(0.22)	5.18	(0.32)
ML-PM	5.42	(0.41)	5.62	(0.45)
HDM-D	5.12	(0.38)	5.77	(0.37)
HDM-RWC(3)	5.56	(0.36)	4.91	(0.24)
HDM-RWC(4)	5.38	(0.27)	5.33	(0.37)
HDM-DWC(3)	6.40	(0.46)	5.98	(0.36)
HDM-DWC(4)	5.48	(0.38)	6.35	(0.58)
HDM-RANKWC(3)	4.55	(0.18)	4.43	(0.28)
HDM-RANKWC(4)	4.92	(0.30)	4.77	(0.30)
R-PM	9.2	(0.32)	9.9	(0.37)

Monte Carlo means of B_{34} and b_{34}, computed for simulations in which the marginal distribution of Z is a normal-lognormal mixture, are given in Table 7. These results are qualitatively similar to the values reported in Tables 1 and 2 for the normal mixture case. Imputation methods involving (Y,Z) information produce less distortion than matching methods. Methods involving modelling based on normality assumptions without the use of stochastic residuals perform relatively poorly. Otherwise, most differences in Monte Carlo means reported in Table 7 are not statistically significant.

Table 7: Normal-Lognormal Mixture

Monte Carlo Means of B_{34} and b_{34} and their Monte Carlo Standard Errors

Method	Statistic			
	B_{34}		b_{34}	
HDI-D	.007	(.033)	.032	(.027)
HDI-NPS	.071	(.033)	.071	(.03)
RI-SR	.053	(.03)	.038	(.027)
ML-ZR	-.224	(.06)	-.253	(.063)
ML-SR	.042	(.036)	.026	(.033)
ML-PM	-.223	(.06)	-.251	(.063)
HDM-D	.384	(.079)	.371	(.076)
HDM-RWC(3)	.366	(.076)	.340	(.069)
HDM-RWC(4)	.363	(.076)	.346	(.072)
HDM-DWC(3)	.376	(.078)	.366	(.075)
HDM-DWC(4)	.378	(.079)	.363	(.075)
HDM-RANKWC(3)	.385	(.079)	.371	(.075)
HDM-RANKWC(4)	.385	(.079)	.364	(.075)
R-PM	.601	(.121)	.608	(.122)

Monte Carlo means of runs test statistics for the normal-lognormal mixture case are reported in Table 8. Comparing this table to the normal mixture results in Table 5, one notes the relatively poor performance of methods involving modelling using normality assumptions when the marginal distribution of Z is asymmetric. According to the runs tests results, the hot deck method using ranks produces the lowest distortion for all univariate and bivariate distributions involving Z. It is interesting to note that the runs test results suggest that the hot deck matching methods using ranks outperforms imputation methods with respect to (Y,Z) distortion. Some imputation methods involve distributional assumptions not consistent with the Z data — others rely on the use of a relatively small number of Z values from file D to complete all file A records.

Monte Carlo means of $Q_{(r)}$ statistics reported in Table 9 provide more evidence of the poor performances of modelling methods compared to hot deck methods for the simulations involving Z data from a normal-lognormal mixture.

Table 8: Normal-Lognormal Mixture

Monte Carlo Means of Runs Test Statistics and their Monte Carlo Standard Errors

Method	Statistic			
	$\hat{U}_{14}$	$\hat{U}_{24}$	$\hat{U}_{34}$	$\hat{U}_{4}$
HDI-D	202.8 (1.88)	199.8 (2.74)	199.0 (2.56)	123.1 (1.88)
HDI-NPS	219.3 (1.84)	220.3 (2.15)	219.1 (1.92)	125.1 (1.98)
RI-SR	221.6 (3.27)	220.9 (3.79)	224.9 (4.68)	220.2 (4.14)
ML-ZR	216.6 (4.46)	214.1 (4.33)	216.7 (4.8)	218.8 (3.95)
ML-SR	220.3 (4.06)	218.7 (3.87)	220.4 (4.33)	218.9 (4.09)
ML-PM	216.1 (4.4)	214.1 (4.42)	215.4 (4.73)	169.4 (3.92)
HDM-D	206.8 (1.87)	207.4 (2.22)	235.2 (2.21)	185.0 (1.51)
HDM-RWC(3)	236.1 (1.76)	237.9 (1.95)	239.9 (2.08)	194.6 (1.79)
HDM-RWC(4)	232.4 (1.77)	233.0 (2.3)	237.9 (2.09)	192.7 (2.08)
HDM-DWC(3)	206.2 (1.88)	206.8 (1.97)	236.8 (2.29)	185.3 (1.57)
HDM-DWC(4)	204.8 (2.26)	206.2 (1.92)	236.0 (2.47)	182.0 (1.74)
HDM-RANKWC(3)	246.7 (2.1)	248.2 (2.58)	246.8 (1.87)	244.7 (2.22)
HDM-RANKWC(4)	244.1 (1.99)	249.7 (2.07)	243.3 (2.17)	243.0 (1.83)
R-PM	139.3 (9.67)	154.7 (7.22)	188.8 (4.8)	146.8 (5.23)

Table 9: Normal-Lognormal Mixture

Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

Method	r=0.1		r = 0.9	
	X_1	X_2	X_1	X_2
HDI-D	4.92 (0.30)	4.67 (0.26)	8.56 (0.36)	8.15 (0.41)
HDI-NPS	4.71 (0.27)	4.66 (0.27)	8.50 (0.46)	8.36 (0.31)
RI-SR	8.60 (1.59)	9.01 (1.94)	13.0 (1.89)	13.1 (2.25)
ML-ZR	7.69 (0.54)	7.47 (0.54)	10.6 (0.60)	10.2 (0.58)
ML-SR	9.29 (1.28)	9.49 (1.45)	13.7 (1.48)	14.0 (1.64)
ML-PM	7.57 (0.52)	7.33 (0.50)	10.4 (0.55)	10.1 (0.52)
HDM-D	5.37 (0.32)	5.22 (0.23)	9.43 (0.50)	8.76 (0.41)
HDM-RWC(3)	4.41 (0.22)	4.30 (0.23)	8.60 (0.46)	7.91 (0.35)
HDM-RWC(4)	4.40 (0.23)	4.30 (0.23)	8.91 (0.53)	8.42 (0.40)
HDM-DWC(3)	5.23 (0.31)	5.20 (0.25)	9.32 (0.50)	8.74 (0.47)
HDM-DWC(4)	5.09 (0.33)	5.41 (0.25)	8.91 (0.46)	8.87 (0.36)
HDM-RANKWC(3)	4.34 (0.24)	4.31 (0.18)	8.64 (0.43)	8.51 (0.36)
HDM-RANKWC(4)	4.76 (0.27)	4.24 (0.24)	8.72 (0.36)	8.52 (0.30)
R-PM	11.3 (0.52)	11.4 (0.50)	13.5 (0.53)	13.6 (0.51)

4.2 Synthetic Data — Part II

In this section, the results of the second component of the evaluation study involving the use of synthetic data are described. Eleven methods were considered. Traditional matching methods as well as various log-linear statistical matching methods were examined. Regression methods and maximum likelihood methods involving normality assumptions were excluded. The methods included are listed below. Some of the methods were also included in the first component of the evaluation study. For those that were not included in the first component, a brief description is given. Methods (i) and (ii) are imputation methods that use file D information. The others are matching methods.

(i) Hot deck imputation using distance (HDI-D).

(ii) Hot deck imputation using non-parametric smoothing (HDI-NPS).

(iii) Hot deck matching using distance (HDM-D).

(iv) Hot deck matching using random assignment within classes (HDM-RWC).

(v) Hot deck matching using distance within classes (HDM-DWC).

(vi) Hot deck matching using ranks within classes (HDM-RANKWC).

(vii) Log-linear matching, no smoothing, constrained prediction, distance assignment (LLM-CPD). The empirical distribution from file B is used as an estimate of the conditional distribution $f(Z^* | X^*)$. Suppose that there are J classes for Z, equivalent to J levels for the categorical variable Z^*. If there are k file A records in class X_i^*, $[k \cdot f(Z_j^* | X^* = X_i^*)]$ records are assigned class Z_j^*, for j=1, 2, ..., J, using the following procedure. Determine the pair of records (a,b) that are closest according to Euclidean distance where a is a file A record that has not yet been assigned a Z^* class, b is a file B record in class Z_j^* and the number of file A records that have already been assigned class Z_j^* is less than $[k.f(Z_j^* | X^* = X_i^*)]$. Record a is assigned class Z_j^*. . This process is repeated until there are no more record pairs (a,b) that are eligible (satisfy the conditions mentioned above). At this point, if any file A records have not been assigned a Z^* class, they are assigned class Z_j^* with probability proportional to $k \cdot f(Z_j^* | X^* = X_i^*) - [k \cdot f(Z_j^* | X^* = X_i^*)]$.

The Z value needed to complete a particular file A record that has been assigned class Z_j^* is taken from the file B record in class Z_j^* that is closest according to Euclidean distance.

(viii) Log-linear matching, no smoothing, constrained prediction, random assignment (LLM-CPR). This method is similar to (vii). During assignment of Z^* classes, the "closest" record pair is chosen randomly from among eligible pairs. The Z value needed to complete a file A record in class X_i^* that has been assigned class Z_j^* is taken from a randomly chosen file B record in class (X_i^*, Z_j^*).

(ix) Log-linear matching, smoothing, constrained prediction, distance assignment (LLMS-CPD). This method is identical to (vii) except that an estimate of the conditional distribution $f(Z^*|X^*)$ is obtained by smoothing the empirical joint distribution of (X^*, Z^*) from file B using a log-linear model.

(x) Log-linear matching, smoothing, constrained prediction, random assignment (LLMS-CPR). This method is identical to (viii) except that an estimate of the conditional distribution $f(Z^*|X^*)$ is obtained by smoothing the empirical joint distribution of (X^*, Z^*) from file B using a log-linear model.

(xi) Log-linear matching, smoothing, unconstrained prediction, random assignment (LLMS-UPR). As in methods (ix) and (x), an estimate of the conditional distribution $f(Z^*|X^*)$ is obtained by smoothing using a log-linear model. The Z^* class associated with each file A record in class X_i^* is determined by independent prediction using $f(Z^*|X^*=X_i^*)$. The Z value needed to complete a file A record in class X_i^* that has been assigned class Z_j^* is taken from a randomly chosen file B record in class (X_i^*, Z_j^*).

Six sets of 50 simulations were conducted. Four of these sets involved the same conditions (distribution of (X_1, X_2, Y, Z) and criteria for determining numbers of classes and their boundaries) as the simulations reported in section 4.1. The starting values used for the random number generators and, consequently, the actual data sets were different from those used earlier.

In the fifth set of simulations, data from the multivariate normal distribution with mean zero and variance V_1 was employed. Four classes for X_1, X_2, Y and Z were used (when classes were required). Class boundaries were determined using the method described in section 4.1. The hot deck matching method using ranks requires the assumption that files A and B contain random samples from a common distribution. The other methods considered here involve a weaker assumption, namely that the conditional distribution $f(Z|X)$ or, for some methods, $f(Z^*|X^*)$, is the same for both files. The sixth set of simulations was intended to provide some evidence about the robustness of the hot deck matching method using ranks to departures from the assumption of random samples from a common

distribution. Data from the multivariate normal distribution was used but file B was restricted to observations with $|X_2| < 1.28$. Four classes for X_1, X_2, Y and Z, with boundaries determined using the method of section 4.1, were employed when classes were needed.

With the exception of the runs test statistics, the evaluation measures used in the simulations reported in Section 4.1 were calculated. Monte Carlo means and standard errors of $Q_{(r)}$ statistics were computed for both X_1 and X_2 and $r = 0.1, 0.5, 0.9$. Most of the detailed results that are reported here involve $Q_{(r)}$ statistics. Before $Q_{(r)}$ results are discussed, a summary of results for other evaluation measures will be provided. Results for B_{i4} and b_{i4} $(i=1, 2, 3)$ are generally similar to those obtained in the first component of the evaluation study. As expected, the imputation methods (HDI-D and HDI-NPS) produced less distortion in B_{34} and b_{34} than matching methods that do not involve use of information about the (Y,Z) distribution. Otherwise, most differences between alternative methods are not statistically significant. The results for B_{i4} obtained from the set of simulations involving a mixture of a normal distribution and a lognormal for the marginal distribution of Z (using three categories when categories were necessary), given in Table 10, are typical. Standard errors for most pairwise comparisons of means in this table are between 0.1 and 0.2.

One interesting result not found in the first component of the evaluation study is evident in Table 10. Monte Carlo means of B_{34} are smaller for the log-linear matching methods involving smoothing (LLMS-UPR, LLMS-CPR, LLMS-CPD) than for other matching methods. In addition these three methods produce Monte Carlo means for B_{14} and B_{24} that are larger (in absolute terms) than the corresponding means for other matching methods. Many pairwise comparisons are statistically significant. This result appeared in all six sets of simulations included in the second component of the evaluation study and was also present for Pearson correlations. It has an interpretation in the context of log-linear modelling. Some of the correlation between the categorical variables Y^* and Z^* may be captured in the coefficients of the X^* variables in the equations used to predict proportions of observations in various Z^* categories. Better understanding of this result is needed before one can recommend use of log-linear matching methods with smoothing in practice.

Results for this component of the evaluation study related to Monte Carlo means of $\hat{D}_{250}$ are similar to those reported in section 4.1. Monte Carlo means of $\hat{D}_{250}$ for simulations involving a normal-lognormal mixture for the marginal distribution of Z are

given in Table 11. The Monte Carlo mean for the hot deck method using ranks is closest to 0.777, the mean of the distribution of the Kolmogorov-Smirnov test statistic according to the Smirnov approximation. There is statistically significant evidence (at the 0.05% level) of distortion in the marginal distribution of Z for most of the other methods.

Monte Carlo means of $Q_{(r)}$ statistics and their Monte Carlo standard errors obtained from simulations involving a mixture of normal distributions for the marginal distribution of Z are given in Table 12. Note that the medians of the Z distribution within deciles of X_1 are reproduced more closely than quantile 0.9 for all methods. Results for X_2 were computed by the simulation program but are not included in the table since the distribution used to generate the data for these simulations is symmetric with respect to X_1 and X_2. Monte Carlo standard errors for pairwise comparisons are comparable to the standard errors of the Monte Carlo means shown in the table — approximately 0.1 for $Q_{(0.5)}$ and between 0.2 and 0.3 for $Q_{(0.9)}$.

Most of the comments that can be made about alternative methods based on results involving $Q_{(r)}$ statistics obtained from this component of the evaluation study are apparent from Table 12. The hot deck method using ranks appears to be the best method with respect to distortions in the marginal distribution of Z within X deciles. Hot deck distance imputation produces less distortion than the hot deck matching method involving distance. This result is expected given that the imputation method involves the use of Y, as well as the X variables used by the matching method, in the calculation of distances. There is statistically significant evidence that non-parametric smoothing reduces the distortion involved in hot deck distance imputation.

The log-linear methods using a saturated model (LLM-CPR and LLM-CPD) produce less (X,Z) distortion, according to $Q_{(r)}$, than the corresponding hot deck matching methods (HDM-RWC and HDM-DWC). Intuitively, this result is plausible since the log-linear methods impose constraints on the (X,Z) distribution that are not involved in hot deck matching. The results in Table 12 do not clearly suggest that smoothing leads to a reduction of (X,Z) distortion in the context of log-linear methods. For example, values of $Q_{(.05)}$ for the log-linear matching methods involving smoothing are all higher than those for log-linear methods without smoothing, although none of the differences are statistically significant.

Results in Table 13 are similar to those in Table 12. Since the simulation for methods involving four categories were conducted separately from the simulations with three categories, using different data sets, standard errors for pairwise comparisons involving, for example, a $Q_{(0.5)}$ statistic from Table 12 and a $Q_{(0.5)}$ statistic from Table 13 can be computed using the standard errors in the tables (covariance terms are zero). There is no clear evidence to prefer use of four categories rather than three. This situation is not unexpected given the somewhat arbitrary way in which categories were determined for both sets of simulations.

Tables 14 and 15 provide Monte Carlo means of $Q_{(r)}$ statistics for percentiles of Z within deciles of X_1, computed for simulations with a mixture of a normal distribution and a log-normal distribution for the marginal distribution of Z. Note that Monte Carlo means of $Q_{(0.9)}$ are much larger than those for $Q_{(0.1)}$ due to the skewness of the marginal distribution of Z. Some Monte Carlo means for $Q_{(r)}$ statistics for simulations using data from a multivariate normal distribution for (X_1, X_2, Y, Z) are reported in Table 16. The results in these three tables are similar to those summarized above. Note, in particular, the good relative performance of the hot deck matching methods using ranks.

The hot deck matching method using ranks depends on the assumption that information on files used for statistical matching represent independent samples from a common distribution. Alternative methods involve weaker assumptions, namely that the conditional distribution of Z given X (and the conditional distribution of Z^* given X^* for log-linear methods) does not differ between the two files. The final set of simulations conducted as part of this component of the evaluation study was motivated by an interest in the robustness of the hot deck method using ranks to departures from this independent sample assumption. File B was restricted to observations with $|X_2|<1.28$. This corresponds to truncation of 10% in both tails of the X_2 distribution. Monte Carlo means of $Q_{(r)}$ statistics for these simulations are reported in Table 17. Note that the hot deck method using ranks appears to give the best overall performance despite the violation of the independent samples assumption. The conditions for this set of simulations were chosen arbitrarily. The specificity of Monte Carlo results must be emphasized.

Table 10: Normal-Lognormal Mixture, Part II, Three Categories
Monte Carlo Means of B_{i4} and their Monte Carlo Standard Errors

Method	i		
	1	2	3
HDI-D	-.038 (.018)	-.037 (.016)	.021 (.019)
HDI-NPS	-.058 (.018)	-.056 (.016)	.068 (.019)
HDM-D	.001 (.011)	.001 (.016)	.381 (.056)
HDM-RWC	-.015 (.013)	-.025 (.015)	.354 (.052)
HDM-DWC	-.006 (.012)	-.005 (.017)	.377 (.055)
HDM-RANKWC	.013 (.011)	-.025 (.013)	.375 (.055)
LLM-CPD	-.006 (.010)	-.014 (.014)	.364 (.053)
LLM-CPR	-.018 (.010)	-.029 (.015)	.354 (.051)
LLMS-CPD	-.087 (.020)	-.084 (.019)	.341 (.050)
LLMS-CPR	-.091 (.019)	-.098 (.020)	.325 (.047)
LLMS-UPR	-.111 (.023)	-.088 (.022)	.322 (.047)

Table 11: Normal-Lognormal Mixture, Part II, Three Categories
Monte Carlo Means of $\hat{D}_{250}$ and their Monte Carlo Standard Errors

Method	Mean	(Standard Error)
HDI-D	.115	(.005)
HDI-NPS	.109	(.004)
HDM-D	.095	(.004)
HDM-RWC	.092	(.004)
HDM-DWC	.094	(.004)
HDM-RANKWC	.078	(.003)
LLM-CPD	.088	(.004)
LLM-CPR	.082	(.003)
LLMS-CPD	.088	(.004)
LLMS-CPR	.081	(.003)
LLMS-UPR	.091	(.003)

Table 12: Normal Mixture, Part II, Three Categories

Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

Method	X_1			
	r = 0.5		r = 0.9	
HDI-D	2.94	(0.11)	5.70	(0.30)
HDI-NPS	2.54	(0.11)	5.27	(0.27)
HDM-D	3.21	(0.13)	5.98	(0.27)
HDM-RWC	3.10	(0.11)	5.14	(0.22)
HDM-DWC	3.20	(0.13)	5.96	(0.28)
HDM-RANKWC	2.86	(0.09)	4.66	(0.14)
LLM-CPD	2.84	(0.10)	5.51	(0.22)
LLM-CPR	2.91	(0.10)	5.20	(0.24)
LLMS-CPD	3.00	(0.10)	5.39	(0.21)
LLMS-CPR	3.01	(0.10)	5.21	(0.25)
LLMS-UPR	3.08	(0.11)	4.92	(0.19)

Table 13: Normal Mixture, Part II, Four Categories

Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

Method	X_1			
	r = 0.5		r = 0.9	
HDI-D	2.98	(0.11)	5.82	(0.27)
HDI-NPS	2.96	(0.12)	5.08	(0.23)
HDM-D	3.26	(0.14)	6.35	(0.30)
HDM-RWC	3.06	(0.10)	5.37	(0.23)
HDM-DWC	3.21	(0.13)	6.60	(0.32)
HDM-RANKWC	2.86	(0.09)	4.79	(0.23)
LLM-CPD	2.95	(0.11)	5.77	(0.26)
LLM-CPR	2.85	(0.09)	4.88	(0.18)
LLMS-CPD	2.98	(0.12)	5.61	(0.26)
LLMS-CPR	2.75	(0.09)	5.08	(0.18)
LLMS-UPR	2.94	(0.11)	5.54	(0.29)

Table 14: Normal-lognormal Mixture, Part II, Three Categories
Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

Method	X_1		
	r = 0.1	r = 0.5	r = 0.9
HDI-D	4.81 (0.17)	2.71 (0.13)	11.0 (0.53)
HDI-NPS	4.77 (0.16)	2.55 (0.10)	10.0 (0.52)
HDM-D	5.12 (0.18)	2.76 (0.10)	10.9 (0.53)
HDM-RWC	4.81 (0.17)	2.38 (0.08)	9.89 (0.44)
HDM-DWC	5.15 (0.18)	2.62 (0.10)	10.8 (0.46)
HDM-RANKWC	4.34 (0.14)	2.2 (0.06)	8.36 (0.32)
LLM-CPD	4.81 (0.17)	2.37 (0.10)	10.3 (0.44)
LLM-CPR	4.6 (0.15)	2.15 (0.08)	8.99 (0.47)
LLMS-CPD	5.0 (0.21)	2.35 (0.07)	10.8 (0.51)
LLMS-CPR	4.65 (0.17)	2.28 (0.10)	10.3 (0.99)
LLMS-UPR	4.67 (0.19)	2.36 (0.08)	9.75 (0.40)

Table 15: Normal-lognormal Mixture, Part II, Four Categories

Method	X_1		
	r = 0.1	r = 0.5	r = 0.9
HDI-D	4.69 (0.17)	2.78 (0.12)	9.09 (0.41)
HDI-NPS	4.50 (0.15)	2.37 (0.09)	9.31 (0.49)
HDM-D	5.08 (0.18)	2.81 (0.11)	10.4 (0.47)
HDM-RWC	4.53 (0.17)	2.24 (0.07)	9.80 (0.42)
HDM-DWC	5.13 (0.19)	2.81 (0.10)	11.2 (0.47)
HDM-RANKWC	4.44 (0.19)	2.20 (0.08)	8.63 (0.33)
LLM-CPD	4.70 (0.16)	2.33 (0.07)	10.1 (0.65)
LLM-CPR	4.32 (0.16)	2.20 (0.08)	9.06 (0.50)
LLMS-CPD	4.78 (0.15)	2.47 (0.08)	9.72 (0.42)
LLMS-CPR	4.27 (0.17)	2.28 (0.08)	9.67 (0.41)
LLMS-UPR	4.48 (0.16)	2.41 (0.07)	10.7 (0.49)

Table 16: Normal, Four Categories

Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

	X_1			
Method	r = 0.5		r = 0.9	
HDI-D	2.64	(0.16)	3.71	(0.17)
HDI-NPS	2.70	(0.15)	3.52	(0.14)
HDM-D	2.87	(0.12)	3.82	(0.14)
HDM-RWC	2.56	(0.12)	3.54	(0.12)
HDM-DWC	3.05	(0.12)	3.74	(0.14)
HDM-RANKWC	2.52	(0.10)	3.33	(0.13)
LLM-CPD	2.74	(0.11)	3.59	(0.12)
LLM-CPR	2.55	(0.12)	3.34	(0.14)
LLMS-CPD	2.76	(0.12)	3.69	(0.14)
LLMS-CPR	2.84	(0.13)	3.54	(0.13)
LLMS-UPR	2.93	(0.14)	3.67	(0.15)

Table 17: Normal with X_2 Truncation, Four Categories

Monte Carlo Means of $Q_{(r)}$ and their Monte Carlo Standard Errors

	X_1			
Method	r = 0.5		r = 0.9	
HDI-D	2.83	(0.10)	3.73	(0.13)
HDI-NPS	2.57	(0.09)	3.54	(0.12)
HDM-D	3.04	(0.10)	3.77	(0.12)
HDM-RWC	2.74	(0.11)	3.51	(0.14)
HDM-DWC	2.96	(0.11)	3.80	(0.14)
HDM-RANKWC	2.51	(0.09)	3.47	(0.13)
LLM-CPD	2.70	(0.09)	3.61	(0.12)
LLM-CPR	2.54	(0.09)	3.43	(0.15)
LLMS-CPD	2.72	(0.09)	3.64	(0.16)
LLMS-CPR	2.60	(0.09)	3.49	(0.14)
LLMS-UPR	2.91	(0.09)	3.44	(0.15)
	X_2			
Method	r = 0.5		r = 0.9	
HDI-D	2.71	(0.11)	3.71	(0.15)
HDI-NPS	2.59	(0.11)	3.42	(0.15)
HDM-D	3.08	(0.11)	4.12	(0.15)
HDM-RWC	2.53	(0.10)	3.43	(0.13)
HDM-DWC	3.01	(0.11)	3.88	(0.15)
HDM-RANKWC	2.47	(0.08)	3.00	(0.11)
LLM-CPD	2.60	(0.08)	3.53	(0.13)
LLM-CPR	2.64	(0.07)	3.51	(0.12)
LLMS-CPD	2.83	(0.10)	3.60	(0.16)
LLMS-CPR	2.58	(0.09)	3.41	(0.14)
LLMS-UPR	2.92	(0.09)	3.33	(0.11)

4.3 Real Data

In this section the results of simulations conducted in order to evaluate the eleven methods considered in section 4.2 using real data are reported. Five sets of 50 simulations were performed. The data was obtained from a file created by exact matching (using record linkage techniques) of information from the 1984 Survey of Consumer Finance (reference year 1983) and the Revenue Canada file of taxfiler information for 1983. The procedures used to create this file are described in Alter (1988). Revenue Canada tax data and information from the Survey of Consumer Finance is used during construction of the Social Policy Simulation Database. Consequently, the distributions of variables used in these simulations should be similar to distributions of variables involved in statistical matching needed to build the SPSD.

The variables used were:

DIVS - dividend income;
EMP - earnings from employment;
AGE - age in years;
PENS - pension income (excluding CPP, QPP);
DUES - union dues;
CHAR - charitable donations;
TOTH - reported on "other deductions" line of personal tax return.

Not all variables were involved in all sets of simulations. When DIVS, EMP and PENS were used, there were considered X variables. DUES, CHAR and TOTH, part of the Revenue Canada component of exact match file records, were Z variables. PENS, part of the SCF component, was the Y variable in all simulations.

The general strategy for data generation involved determining which exact match file records satisfied an eligibility criterion based on earnings from employment. The set of records satisfying the criterion provided a finite population of (X,Y,Z) observations. Data for each Monte Carlo trial was generated by selecting random samples with replacement from this finite population. There were 250 observations in Files A and B and 100 File C observations in all simulations.

The evaluation statistics computed for the real data simulations are measures of distortion in the (X,Z) distribution. Monte Carlo means and standard errors of $Q_{(0.9)}$ were calculated for each matching variable. An analogous statistic, intended to assess the performance of the alternative methods in the prediction of zero and non-zero values was also used. Let $N_{0,j}$ denote the number of zero values of Z within decile j of one of the matching variables, say X_1, on the file C created by matching. Let $N^A_{0,j}$ denote the number of zero Z values within decile j of X_1 for the original file A data. Monte Carlo means and standard errors of

$$N_0 = \sum_{j=1}^{10} \left| N^A_{0,j} - N_{0,j} \right|$$

were computed for each matching variable. Runs test statistics and correlations were not computed for the real data simulations based on the thinking that, due to the large proportion of zero values in the data, they would provide relatively little discrimination.

Four sets of simulations involved the use of exact match file records with positive values for earnings from employment. There were 14,521 such records. Univariate description statistics for the variables involved, computed based on the population of 14,521 observations, are given in Table 18. One notes that, except for AGE, the variables all have highly skewed distributions. In addition, PENS, DUES, CHAR and TOTH are often zero. Spearman and Pearson correlations between (X,Y) and Z are given in Table 19. Real data correlations are generally lower than those for the synthetic data used earlier in this evaluation study. Simulations performed using records with EMP>0 each involved use of two X variables, one Y variable and one Z variable.

The final set of simulations with real data involved use of all exact match file observations with EMP>20000. DIVS, EMP and AGE were X variables and DUES was the Z variable. The finite population used to generate data for these simulations involved 5,035 observations. Descriptive statistics related to this finite population are shown in Tables 20 and 21. Note that correlations between DUES and the X and Y variables are weak. The use of DIVS as an X variable was motivated by the fact that it is used as a matching variable (with EMP and AGE) in the statistical matching of information for high income earners from the Survey of Consumer Finance and Revenue Canada's tax file conducted during construction of the Social Policy Simulation Database. It was necessary to restrict

observations included in the finite population used in data generation to records with EMP>20000 in order to obtain a reasonable proportion of cases with positive values for DIVS.

For most variables two categories, corresponding to zero and non-zero values, were used when categories were needed. Subsequent references to methods involving use of more than two categories refer to the number of categories used for EMP and AGE. For these variables the method described earlier for synthetic data was used to determine category boundaries.

Table 18: Univariate Descriptive Statistics for Exact Match File Variables (based on 14,521 records with positive EMP)

Variable	Std. Dev.	Skewness	Kurtosis	Proportion Non-Zero
EMP	17,130	12.4	434	1.0
AGE	13.2	0.52	-0.67	1.0
PENS	1,551	12.4	200	0.03
DUES	146	2.3	9.9	0.39
CHAR	927	75.7	7,593	0.13
TOTH	1,665	17	472	0.15

Table 19: Correlation Coefficients for Exact Match File Variables (based on 14,521 records with positive EMP)

		DUES	CHAR	TOTH
EMP	Pearson	.339	.097	.032
	Spearman	.447	.232	.050
AGE	Pearson	.116	.091	.081
	Spearman	.145	.242	.103
PENS	Pearson	-.033	.041	.135
	Spearman	-.023	.108	.140

Table 20: Univariate Descriptive Statistics for Exact Match File Variables (based on 5,035 records with EMP>20,000)

Variable	Std. Dev.	Skewness	Kurtosis	Proportion Non-Zero
DIVS	12,736	47	2,562	0.11
EMP	20,436	18.6	559	1.0
AGE	11.1	0.35	-0.83	1.0
PENS	1,285	17.6	435	0.02
DUES	191	1.34	5.37	0.59

Table 21: Correlation Coefficients for Exact Match File Variables (based on 5,035 records with EMP>20,000)

		DUES
DIVS	Pearson	-.038
	Spearman	-.012
EMP	Pearson	-.005
	Spearman	.10
AGE	Pearson	-.001
	Spearman	.006
PENS	Pearson	-.034
	Spearman	-.040

Table 22: Charitable Donations, Three Categories

Monte Carlo Means of $Q_{(0.9)}$, N_0

Method	X_1(EMP)		X_2(AGE)	
	$Q_{(0.9)}$	N_0	$Q_{(0.9)}$	N_0
HDI-D	2,536	19.2	5,134	20.2
HDI-NPS	2,375	17.7	2,602	17.4
HDM-D	2,422	16.6	2,709	17.1
HDM-RWC	2,374	17.1	2,165	15.9
HDM-DWC	2,375	16.7	2,727	17.4
HDM-RANKWC	1,943	15.0	2,287	14.8
LLM-CPD	2,323	15.0	2,636	15.9
LLM-CPR	2,162	14.9	2,327	14.7
LLMS-CPD	2,357	16.1	2,786	16.6
LLMS-CPR	1,909	15.0	2,296	15.2
LLMS-UPR	2,601	17.5	2,527	17.6

Table 23: Other Deductions, Three Categories
Monte Carlo Means of $Q_{(0.9)}$, N_0

Method	X_1(EMP)		X_2(AGE)	
	$Q_{(0.9)}$	N_0	$Q_{(0.9)}$	N_0
HDI-D	10,171	23.2	12,281	22.7
HDI-NPS	8,880	21.3	9,017	20.3
HDM-D	9,936	21.0	8,365	20.8
HDM-RWC	7,695	17.2	8,446	18.7
HDM-DWC	9,705	21.4	8,466	20.9
HDM-RANKWC	6,738	17.2	7,588	17.0
LLM-CPD	8,403	18.9	8,168	18.6
LLM-CPR	6,622	18.3	7,432	18.7
LLMS-CPD	7,934	18.4	8,274	17.9
LLMS-CPR	6,569	18.6	6,789	19.1
LLMS-UPR	7,617	17.6	7,688	18.3

Table 24: Union Dues, Two Matching Variables
Monte Carlo Means of $Q_{(r)}$, N_0 (for Deciles of EMP)

Method	Three Categories		Four Categories	
	$Q_{(0.9)}$	N_0	$Q_{(0.9)}$	N_0
HDI-D	622	32.3	588	32.5
HDI-NPS	589	29.3	558	28.6
HDM-D	553	28.7	509	28.9
HDM-RWC	629	26.4	544	24.9
HDM-DWC	559	29.0	511	28.6
HDM-RANKWC	422	24.0	470	23.4
LLM-CPD	533	26.6	505	25.6
LLM-CPR	639	24.4	486	25.4
LLMS-CPD	537	26.9	503	26.5
LLMS-CPR	579	25.2	485	24.8
LLMS-UPR	645	24.2	557	25.8

Table 25: Union Dues: Three Matching Variables

Monte Carlo Means of $Q_{(0.9)}$, N_0

Method	X_2(EMP)		X_3(AGE)	
	$Q_{(0.9)}$	N_0	$Q_{(0.9)}$	N_0
HDI-D	781	31.9	866	37.0
HDI-NPS	642	29.5	778	33.1
HDM-RWC	744	26.2	662	27.3
HDM-DWC	645	27.2	730	29.1
HDM-RANKWC	525	25.7	704	25.5
LLM-CPD	604	25.0	728	26.8
LLM-CPR	657	25.4	745	25.9
LLMS-CPD	614	26.1	734	27.6
LLMS-CPR	614	25.2	728	25.1
LLMS-UPR	773	27.3	769	27.4

Results from simulations with real data are reported in Tables 22-25. Earnings from employment was used as the ranking variable by the hot deck matching method using ranks for all simulations. In deference to the reader's fatigue, not to mention that of the author, these results will be discussed very briefly. For Table 22, the Z variable is charitable donations. Monte Carlo standard errors for pairwise comparisons (detailed results available from the author) are generally between 150 and 200 for $Q_{(0.9)}$ and between 0.5 and 1.0 for N_0 when these measures are calculated using deciles of EMP. They are slightly lower for distortion measures calculated with respect to AGE.

The large value of $Q_{(0.9)}$ with respect to AGE for hot deck distance imputation may be related to the fact that Euclidean distance is not a good metric for the multivariate distribution of AGE, EMP and PENS. It is particularly bad for PENS, which is used to calculate distances in hot deck distance imputation but is not employed in the analogous matching method. HDI-D often produces large distortion measures in the real data simulations. Otherwise, the important elements of the simulation results for real data are identical to those for the synthetic data simulations. Non-parametric smoothing reduces distortion in hot deck distance imputation. The hot deck matching method using ranks performs well. Log-linear matching methods without smoothing produce less distortion than the analogous hot deck methods but the benefits of smoothing are not clear. Results for both distortion measures, N_0 and $Q_{(0.9)}$, support these conclusions.

The Z variable for Table 23 is other deductions. Monte Carlo standard errors for pairwise comparisons are generally between 600 and 1100 for $Q_{(0.9)}$ with respect to EMP and are slightly lower for the same measure computed with respect to AGE. Standard errors for pairwise comparisons of N_0 values are generally between 0.5 and 1.0.

In Table 24, results from two set of simulations with dues as Z variable are reported. The first two columns contain results obtained when three categories were used in methods requiring categories. The other two columns give results based on use of four categories. Monte Carlo standard errors for pairwise comparisons are usually between 20 and 40 for $Q_{(0.9)}$ and 0.8 and 1.4 for N_0.

Dues is the Z variable for simulations summarized in Table 25. The X variables used were DIVS, EMP and AGE. The large proportion of zero values for DIVS prevented calculation of a distortion measure involving deciles of this matching variable. Monte Carlo standard errors for pairwise comparisons are similar in magnitude to those for Table 24.

5. Conclusions

Distributions of variables on data files created by statistical matching may be subject to various types of distortions. Many statistical matching methods involve the use of categories for the variables (X) common to both input data files. Notwithstanding the specificity of the results of Monte Carlo simulations, the evaluation study reported here provides strong evidence that the use of categories for variables found on only one input data file (Y and Z) as well as for the common variables leads to reduced distortion in the joint distribution of (X,Z) on files created by matching. The study also suggests, as one would expect, that auxiliary information about the distribution of (Y,Z) (obtained, for example, from a sample of (X,Y,Z) observations) is necessary to reduce distortion in the conditional distribution of (Y,Z) given X.

The use of categories for all variables is an idea that is incorporated in log-linear statistical matching, a method that involves application of ideas from the log-linear imputation method proposed by Singh (1988). Various log-linear statistical matching methods involving smoothing estimates of the distribution of the categorical variables (X^*,Z^*) were evaluated here. The evidence concerning the benefits of smoothing in terms of reduced distortion in the joint distribution of (X,Z) is mixed.

The hot deck matching method involving the use of ranks that is currently employed during construction of the Social Policy Simulation Database (SPSD) performed well in the evaluation study. In most simulation experiments this method lead to the lowest distortion in the distribution of (X,Z) on the file created by matching. The performance of this method probably depends on the strength of correlations between the variable used to determine ranks and variables unique to one input file. During statistical matching applications involved in the construction of the SPSD, a total income variable is usually used to determine ranks. Consequently the real data simulations, in which the hot deck matching method using ranks determined by earnings from employent performed well, are of particular relevance for the SPSD applications of statistical matching.

The hot deck matching method using ranks relies on the assumption that the input files used for statistical matching contain random samples from a common distribution. Alternative methods involve weaker assumptions, namely that the conditional distribution of Z given X (and the conditional distribution Z^* given X^* for methods involving Z^* categories) does not differ between input files. The limited simulation experiment reported here suggests that the hot deck method using ranks is relatively robust to departures from the independent sample assumption.

Although the result is somewhat tangential to the main objectives of the evaluation study, it is interesting to note the strong evidence of the benefits of non-parametric smoothing in the context of hot deck imputation. This type of smoothing is apparently not widely used in practice. For example, it is not mentioned in the review of imputation methods by Kalton and Kasprzyk (1986). Non-parametric smoothing should also reduce (X,Z) distortion for hot deck statistical matching methods.

It is appropriate to conclude with some discussion of issues that were not addressed in this study and could serve as a focus for further work. First, the criteria for category definition and log-linear model choice proposed by Singh (1988) as part of the log-linear imputation method were not extensively examined here. In addition, only a few choices for correlations of (X,Y,Z) were considered. Consequently, the possibility that the performance of the log-linear statistical matching methods (particularly those involving smoothing) could be improved by different choices of categories and/or models remains open. The log-linear statistical matching methods might perform better, in relative terms, for correlation structures that differ from those considered here.

Finally, there is evidence that log-linear statistical matching methods involving assignment of Z values using a distance measure or random matching with X classes produce less (X,Z) distortion than the corresponding hot deck methods. The hot deck matching method using ranks produced less (X,Z) distortion than random or distance hot deck matching methods in the evaluation study. Consequently, the use of a log-linear matching method involving assignment based on ranks should be investigated.

Acknowledgements

The author would like to thank John Kovar, Geoff Rowe, Avinash Singh, Michael Wolfson and Patricia Whitridge for useful comments on a draft version of this paper, as well as Judy Clarke for efficient text processing. Most of the work described here was done when the author was employed in the Social Survey Methods Division of Statistics Canada.

References

Alter, H. (1988). Linked records as a foundation for analysis. Staff report, Labour and Household Surveys Analysis Division. Statistics Canada.

Barr, R.S., Stewart, W.H. and Turner, J.S. (1982). An empirical evaluation of statistical matching methodologies. Unpublished manuscript, Southern Methodist University, Dallas.

Barr, R.S. and Turner, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey. Report prepared for the Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.

Benedetti, J.K. and Brown, M.B. (1978). Strategies for the selection of log-linear models. *Biometrics*, 34, 680-686.

Dempster, A.P. (1969). *Elements of Continuous Multivariate Analysis*, Don Mills: Addison-Wesley.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Dixon, W.J., Brown, M.B., Engelman, L., Frane, J.W., Hill, M.A., Jennrich, R.I. and Toporek, J.D. (1983). *BMDP Statistical Software: 1983 Printing with Additions.* Berkeley: University of California Press.

Friedman, J.H. and Rafsky, L.C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Annals of Statistics*, 7, 697-717.

Goldfeld, S.M., and Quandt, R.E. (1972). *Nonlinear Methods in Econometrics*, Amsterdam: North-Holland.

Hendry, D.F. and Harrison, R.W. (1974). Monte Carlo methodology and the small sample behaviour of ordinary and two-stage least squares. *Journal of Econometrics*, 2, 151-174.

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kim, P.J. (1969). On the exact and approximate sampling distribution of the two-sample Kolmogorov-Smirnov criterion D_{mn}, $m \leq n$. Journal of the American Statistical Association, 64, 1625-1637.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A. (1986). Missing data in Census Bureau surveys. *Proceedings of the Second Annual Research Conference*, Washington, D.C.: U.S. Bureau of the Census, 442-454.

Okner, B.A. (1972). Constructing a new data base from existing microdata sets: the 1966 merge file. Annals of Social and Economic Measurement, 1, 325-342.

Paass, G. (1986). Statistical match: evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*, (Orcutt, G.H., Merz, J. and Quinke, H., eds.), Amsterdam: Elsevier Science.

Rodgers, W.L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2, 91-102.

Rodgers, W.L. and DeVol, E. (1982). An evaluation of statistical matching. American Statistical Association, Proceedings of the Section of Survey Research Methods, 128-132.

Rubin, D.B. (1978). Multiple imputations in sample survey - a phenomenological Bayesian approach to non-response. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 20-34.

Rubin, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4, 87-94.

Sims, C.A. (1972). Comment (on Okner 1972). *Annals of Economic and Social Measurement*, 1, 343-345.

Sims, C.A. (1978). Comment (on Kadane 1978). In *1978 Compendium of Tax Research*, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.

Singh, A. (1988). Log-linear imputation. Working Paper 88-029E, Methodology Branch, Statistics Canada. (See also *Proceedings of the Fifth Annual Research Conference* (1989), Washington, D.C.: U.S. Bureau of the Census.)

Singh, A., Armstrong, J. and Lemaitre, G. (1988). Statistical matching using log-linear imputation. *American Statistical Association, Proceedings of the Section on Survey Research Methods* (to appear).

Srivastava, M.S. and Carter, E.M. (1986). The maximum likelihood method for non-response in sample surveys. *Survey Methodology*, 12, 61-72.

Stone, C.J. (1977). Consistent non-parametric regression. *Annals of Statistics*, 5, 595-620.

Stroud, T.W.F. (1989). A Bayesian approach to the problem of statistical matching. Unpublished manuscript, Queen's University.

U.S. Department of Commerce (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, Washington, D.C.: Federal Committee on Statistical Methodology.

Whitney, V.K.M. (1972). Algorithm 422, minimal spanning tree. *Communications of the ACM*, 15, 273-274.

Wolfson, M., Gribble, C., Bordt, M., Murphy, B. and Rowe, G. (1987). The social policy simulation database: an example of survey and administrative data integration. *Proceedings, Statistical uses of Administrative Data, an International Symposium*, 201-229.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.