Survey Methodology, June 1993
Vol. 19, No. 1, pp. 59–79
Statistics Canada

Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption

A.C. SINGH, H.J. MANTEL, M.D. KINACK and G. ROWE[1]

ABSTRACT

In the creation of micro-simulation databases which are frequently used by policy analysts and planners, several datafiles are combined by statistical matching techniques for enriching the host datafile. This process requires the conditional independence assumption (CIA) which could lead to serious bias in the resulting joint relationships among variables. Appropriate auxiliary information could be used to avoid the CIA. In this report, methods of statistical matching corresponding to three methods of imputation, namely, regression, hot deck, and log linear, with and without auxiliary information are considered. The log linear methods consist of adding categorical constraints to either the regression or hot deck methods. Based on an extensive simulation study with synthetic data, sensitivity analyses for departures from the CIA are performed and gains from using auxiliary information are discussed. Different scenarios for the underlying distribution and relationships, such as symmetric versus skewed data and proxy versus nonproxy auxiliary data, are created using synthetic data. Some recommendations on the use of statistical matching methods are also made. Specifically, it was confirmed that the CIA could be a serious limitation which could be overcome by the use of appropriate auxiliary information. Hot deck methods were found to be generally preferable to regression methods. Also, when auxiliary information is available, log linear categorical constraints can improve performance of hot deck methods. This study was motivated by concerns about the use of the CIA in the construction of the Social Policy Simulation Database at Statistics Canada.

KEY WORDS: Categorical constraints; Conditional correlation; Log normal contaminations; Shrinkage to the mean.

1. INTRODUCTION

Statistical matching can be viewed as a special case of imputation in which we have two distinct micro-data sources containing different information on different units. One data source serves as a host or recipient file to which new information is imputed for each record using data from the other source which is the donor file. Statistical matching, however, differs from the usual problem of imputation whenever the host file contains information about additional variables which are not present in the donor file. A typical use for the matched file is as input to micro-simulation models for which a complete file with all variables is required. Available micro-datafiles may correspond to samples from administrative files or survey data. Since the records from the different source files correspond to different units, the process of merging the information from the various files is unlike exact matching in which one would search through these other data sources for specific units. In fact, even if exact matching were possible, confidentiality concerns could prevent an exact matching of the files.

A general formulation is as follows. A host file A will contain information on variables (X, Y) and a donor file B will contain information on variables (X,Z). The common variable X can be used to identify similar units in the two files. The problem is to complete the records in file A by imputing live values for Z, using the information on the (X,Z) relationship in file B. In practice, the variables X, Y, and Z would generally be multivariate. An important advantage of imputing live values of Z is that relationships among components of multivariate Z are preserved. Throughout this paper, it will be assumed, for convenience, that X, Y and Z are univariate.

The Social Policy Simulation Database (SPSD; see Wolfson *et al.* 1987), a micro-simulation database created at Statistics Canada, provides an important application of statistical matching for use in economic policy analysis, *e.g.*, calculations of taxes and transfers for families on the database. The multistage construction process of the SPSD uses the technique of statistical matching at a number of points in order to enrich the host datafile, the Survey of Consumer Finance (SCF), with additional information from other data sources. Specifically, information from unemployment insurance claim histories, personal income tax returns, and the Family Expenditure Survey is added to the SCF records. If file A corresponds to the SCF and file B to the tax file, then X variables may represent

[1] A.C. Singh, H.J. Mantel and M.D. Kinack, Social Survey Methods Division; G. Rowe, Social and Economic Studies Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

demographic and economic variables, Y may denote transfer income, and Z may correspond to tax liability, investment income and deductions.

Statistical matching, as described above, suffers from a serious limitation in that information on the variable Y is completely ignored. This limitation amounts to the assumption of conditional independence of Y and Z given X ($Y \perp Z \mid X$), denoted CIA (conditional independence assumption). The importance of the CIA is obvious, since the purpose of the match is to analyze the joint relationships of X, Y and Z. If the true relationships of the variables are such that conditional independence does not hold, then the CIA would mask an important component of these relationships, and would bias some analyses involving the full set of variables. The potential seriousness of the CIA was noted by Sims (1978) and Rubin (1986), and, although statistical matches based on the CIA are not necessarily seriously flawed, Paass (1986) and Armstrong (1989) offer some empirical evidence that the problem is often real. The present study, in fact, is motivated from considerations of improving the content of the SPSD which assumes the CIA for the process of statistical matching; see also comments of Scheuren (1989) on the methodology used in the SPSD.

The literature on statistical matching extends over more than two decades. Early references are Budd and Radner (1969), Budd (1971) and Okner (1972). Sims (1972), in his comments on Okner's paper, was the first to point out the potential risk of statistical matching because of the implicit conditional independence assumption. Concerns were also expressed by Fellegi (1977) about the validity of joint distributions in the matched file and he suggested that thorough empirical testing of matching methods should be done. U.S. Department of Commerce (1980) provides a good review of statistical matching as well as exact matching methods; see also Kadane (1978) and Rodgers (1984). Barr and Turner (1990) describe a detailed empirical investigation of quality issues for file merging, and also present a good list of references. For a more recent review see Cohen (1991).

In this paper we consider the use of auxiliary information as an alternative to the CIA in statistical matching. Thus, it is assumed that there exists a third file C representing auxiliary information about the full set (X,Y,Z) or the reduced set (Y,Z). This information could be outdated, proxy (*i.e.* different but similar variables), or in the form of frequency tables and could come from small scale specially conducted surveys or from confidential datafiles. We wish to complete records in file A by adding Z from file B using information from files A, B, and C on the joint relationships of X, Y, and Z. A measure of success would be the extent to which the Z values on the completed file A could reasonably have come from the true underlying distribution conditional on X and Y. In the context of a simulation study we can compare the matched Z values to the suppressed true Z values by evaluation measures at the unit level or at the aggregate level. Some examples of unit level measures are mean absolute distance from the true Z values and the deviation of conditional covariance, $\text{Cov}(Y,Z \mid X)$, from the true value. Some examples of aggregate level evaluation measures are chi-square distance and P-values based on likelihood ratio tests for categorical distributions. It is often the case in practice that the completed file A is used to produce cross-classified tables of counts and, therefore, the aggregate level measures based on categorical distributions would generally be of main interest. Moreover, for any arbitrary distribution for (X,Y,Z), which could be quite complex in practice, the categorical transformation provides a simple unified approach for summarizing the joint distribution.

The statistical matching problem as mentioned above is clearly important from practical considerations. In practice, for a given problem the matching method should be appropriately chosen for the type of auxiliary information available. The methods proposed earlier in the literature are mainly due to Rubin (1986) and Paass (1986). Rubin proposed versions of parametric regression while Paass proposed versions of nonparametric regression. These are related respectively to the familiar regression (REG) and hot deck (HOD) methods of imputation.

Rubin's method (a version of which is denoted in this paper by REG*) basically consists of first finding an intermediate value, Z_{int}, from the regression predictor of Z on X and Y (obtained by using information about the unconditional correlation $\rho_{Y,Z}$ or the conditional correlation $\rho_{Y,Z|X}$ from file C) and then a live Z-value is determined from file B using hot deck with (X,Z) Euclidean distance; see Section 3 for details. If the form of the regression predictor function is known, then the REG* procedure for statistical matching could be easily implemented in practice. However, finding a suitable predictor for Z is in general not easy, especially when Z is multivariate. Moreover, if information in file C is in the form of a categorical distribution, which may be quite common in practice, the REG* method would not be applicable.

Paass's method (a version of which is denoted in this paper by HOD*), on the other hand, basically consists of first finding an intermediate value, Z_{int}, from file C by hot deck imputation (with Y- or (X,Y)-distance as the case may be) and then a live Z-value from file B is obtained using again hot deck with (X,Z)-distance. This is a simplified version of the original Paass's method which is iterative such that values of Z for file A, Y for file B, and X for file C (assuming C has only (Y,Z) information) are updated successively using files C, A and B respectively until some convergence criterion is satisfied; see Section 3 for details. To start the iteration, initial values of Z for A, Y for B and X for C are imputed suitably. In the evaluation study considered in this paper, we have considered only the simplified version of Paass's method due to the considerable computational effort required for the original method. As in the case of the REG* method, the HOD* method is not applicable if file C is in the form of a frequency table. Moreover, even if file C contains micro-data but its size is small (as in the case of a small scale specially conducted survey) or it is proxy or outdated, it may be better to extract some macro-level information such as the categorical distribution based on a fairly coarse partition.

It may be remarked that in the absence of auxiliary information, *i.e.* file C, both REG* and HOD* methods reduce simply to the usual methods of imputation, namely regression (REG) and hot deck (HOD). As part of the evaluation study, these methods are also included.

We propose modifications of Rubin's and Paass's methods, denoted by REG.LOGLIN* and HOD.LOGLIN* respectively, which are based on the log linear method of imputation as introduced by Singh (1988). The proposed modifications use auxiliary information to impose categorical constraints on the matched files obtained from REG* and HOD* methods. In this way, categorical association parameters (estimated via log linear modelling) which measure departure from conditional independence (in the categorical sense) are preserved in the matched file. These categorical constraints are expected to render joint distributions for the completed file A data robust to inferior quality or imperfect nature of the auxiliary data from file C. If auxiliary information is in the form of a categorical distribution and not at the micro-level, then CIA based matching methods can be modified by imposing categorical constraints; in this case the CIA is being used only within X, Y categories. For example, with the usual methods of imputation REG and HOD, which could be used to match by ignoring Y, we can get the corresponding modified versions as REG.LOGLIN and HOD.LOGLIN. These two methods are also considered in this paper.

Note that the categorically constrained matching methods are different from the usual constrained statistical matching methods where the constraints are in the form of a few characteristic measures from file B (such as mean and variance) that variables in the matched file must satisfy. Another key distinction is that the usual constrained matching methods focus on the marginal distribution of Z, whereas the focus here is on the conditional distribution, albeit categorical, which is more relevant for file A; thus there is a basic difference between the two approaches to constrained matching.

Following Rubin (1986) and Paass (1986), we investigate the performance of matching methods empirically. A Monte Carlo study was carried out to investigate the effect of the proposed modifications to the existing methods for the two cases, with and without auxiliary information. This would allow analysis of sensitivity to failure of the CIA and gains from using auxiliary data. The synthetic data for the simulation study was generated from multivariate normal distributions with some log normal contamination to induce asymmetry. An important advantage of using synthetic data is that relevant control parameters could be modified to yield different distributional scenarios for the matching problem. Eight methods (four existing ones, REG, REG*, HOD, HOD*, and four proposed ones, REG.LOGLIN, REG.LOGLIN*, HOD.LOGLIN, HOD.LOGLIN*) were compared by four evaluation measures (two at the unit level and two at the aggregate level) as mentioned earlier; see Section 6 for details. The main findings of the empirical study can be summarized as follows.

(i) Use of auxiliary information to avoid the CIA could considerably improve the quality of the matched file. However, if there is no auxiliary information, then among CIA based methods (*i.e.* REG and HOD), the HOD method has better overall performance. Furthermore, an interesting finding was that for small departures from conditional independence, use of auxiliary information may not improve performance of the HOD method with respect to aggregate level evaluation measures. This should have important practical implications in the absence of readily available auxiliary information.

(ii) The REG* method has very favourable performance with respect to unit level measures. By contrast, it has extremely unfavourable performance with respect to aggregate level measures. This is probably due to the shrinkage towards the mean phenomenon for regressions procedures.

(iii) The HOD* method does considerably better than REG* at the aggregate level but performs, in general, marginally worse than REG* at the unit level.

(iv) Categorical constraints, in general, improve performance of REG* and HOD* methods. Specifically, the REG.LOGLIN* method shows slight improvement at the aggregate level, but HOD.LOGLIN* shows considerable improvement at the aggregate level. Their performances at the unit level remain essentially unaffected.

(v) At the aggregate level, the HOD.LOGLIN method based only on categorical auxiliary information performs generally better than HOD.LOGLIN* based on micro-level auxiliary information. At the unit level, however, HOD.LOGLIN shows marginal deterioration in comparison to HOD.LOGLIN*. This finding may be important from practical considerations because HOD.LOGLIN is computationally much less demanding than HOD.LOGLIN* and does not require micro-level auxiliary information. The REG.LOGLIN method does not have such favourable performance, probably again due to the shrinkage to the mean effect.

(vi) If the auxiliary data is outdated or proxy, there may still be gain in using it. In this context, the HOD.LOGLIN method performs quite favourably and in fact, has fairly robust behaviour with respect to imperfect auxiliary information. Note that since this method uses only information about categorical associations from auxiliary data, it would seem reasonable for this to be affected only slightly by a limited degree of outdatedness or proxyness in file C. The REG.LOGLIN method, however, does not share this property.

It should be noted that there have been several empirical investigations in the past to evaluate statistical matching methods. Among those that do not consider the use of auxiliary information, some main references are Ruggles, Ruggles and Wolff (1977), Paass and Wauschkuhn (1980), Barr, Stewart and Turner (1981) and Rodgers and DeVol (1982). Paass (1986) provides an excellent review of these empirical tests on the quality of matching methods.

All of the studies cited above confirmed the seriousness of the CIA. This stresses the need for additional information to be incorporated in the matching process. There have been few empirical studies considering the use of auxiliary information and the impact of the CIA; Paass (1986) considered an evaluation with synthetic data only, whereas Armstrong (1989) considered simulations with both synthetic and real data. The present study could be considered as complementary to these studies in the sense that some new methods are included and the choice of underlying population distributions is reasonably broad.

The organization of this paper is as follows. Section 2 describes different types of auxiliary information. A brief review of alternative matching methods using auxiliary information is given in Section 3 and the proposed modifications using categorical constraints are described in Section 4. Different types of matching methods are illustrated in Section 5 by means of a simple numerical example. The description of the design of the empirical study on the proposed matching methods is given in Section 6 and the discussion of results in Section 7. Finally, Section 8 contains concluding remarks and some directions for further research.

2. TYPES OF AUXILIARY INFORMATION

Although a current and sufficiently large micro-datafile with information on the full set of variables is not available, it may be the case that an additional auxiliary source exists containing information on some of the joint relationships of either the full set of variables (X,Y,Z) or perhaps the reduced set (Y,Z). When this is the case it can be incorporated into the matching process to avoid the CIA and improve the quality of the completed file by reducing distortions in the joint relationships in the matched file.

Such auxiliary information may emanate from various possible sources and may reside in several different forms. Since the purpose of the auxiliary information is only to aid in avoiding the CIA, we limit its use in that information from the host or donor files is never overridden or modified by the auxiliary information. In other words, the objective is to borrow additional information from the auxiliary source not available in the source files. This is accomplished in such a way that confidentiality concerns associated with the auxiliary source would not be violated and implies that the auxiliary source could be a specially conducted small scale survey or a confidential datafile.

Another implication is that the auxiliary information need not be perfect. That is, it may be deficient in some sense. For instance, it may come from an outdated data source (perhaps a previous census or survey), but from which the required auxiliary information may still be valid, or at least represent an improvement over the otherwise default CIA. On the other hand, the auxiliary information may refer to a set of proxy variables expected to behave similarly to the variables of interest.

Auxiliary information could be at the macro-level or micro-level. At the macro-level, it could take the form of either correlations or categorical cell proportions or possibly some other parameters. If the auxiliary information in file C is on the conditional correlation of Y and Z given X, *i.e.* $\rho_{Y,Z|X}$, it can be used with the (X,Y) and (X,Z) correlations from files A and B to estimate the unconditional correlation of (Y,Z) using

$$\rho_{Y,Z} = \rho_{X,Y}\,\rho_{X,Z} + \rho_{Y,Z|X}\,(1 - \rho^2_{X,Y})^{1/2}\,(1 - \rho^2_{X,Z})^{1/2}. \qquad (2.1)$$

Now data from files A and B can be used to obtain a linear regression of Z on X and Y for the REG* method (see Section 3.1). If auxiliary information on only the unconditional correlation of Y and Z is available, then it can also be used in a similar manner.

The second type of macro-level auxiliary information from file C would be in the form of a categorical distribution for (X^*,Y^*,Z^*) where '*' denotes the categorical transformation of the original variable. If some variables were categorical to begin with, then it may not be necessary to change them. The frequency table required for categorically constrained matching methods can be obtained by raking the (X^*,Y^*,Z^*) table corresponding to file C such that its marginal tables (X^*,Y^*) and (X^*,Z^*) match respectively with the (X^*,Y^*) table from file A and (X^*,Z^*) table from file B. Note that the (X^*,Z^*) table from file B would have to be raked first to match its X^* marginal with that from file A. The method of raking preserves the (Y^*,Z^*) and (X^*,Y^*,Z^*) associations of the (X^*,Y^*,Z^*) table from file C in deriving the categorical constraints. The above adjustment of the (X^*,Y^*,Z^*) table from file C is reasonable on the grounds that information about the (X^*, Y^*) distribution from file A and about the (X^*,Z^*) distribution from B are believed to be more precise or appropriate than those from file C. If only the (Y^*,Z^*) distribution is available (or used) from file C, then the above raking procedure could be modified to obtain suitable categorical constraints. In this case, the (Y^*,Z^*) association from file C would be preserved and the three factor (X^*, Y^*,Z^*) association term would be assumed to be zero. To achieve this, first the (X^*,Z^*) table from B is raked as before to match the X^* margin from A and then the (Y^*,Z^*) table from C is raked to match the Y^* margin from A and the Z^* margin from B. Then, a three dimensional table of ones is raked to match the (X^*, Y^*) table from A, the adjusted (X^*,Z^*) table from B and the adjusted (Y^*,Z^*) table from C. The categorical counts obtained by these procedures need not be integer values. They are rounded randomly by redistributing fractional counts by sampling cells randomly without replacement with probabilities proportional to the fractions for each cell. This is done independently for each (X^*, Y^*) category.

The next section elaborates on the use of auxiliary information in statistical matching. It also describes the use of auxiliary micro-data. In most cases when micro-level auxiliary information is available, it is possible to roll it up to the macro-level and obtain reliable information on correlations and categorical cell proportions. The validity and reasonableness of this would depend in part on the size of the micro-level datafile.

3. REVIEW OF ALTERNATIVE STATISTICAL MATCHING METHODS

3.1 The Regression Method

We first describe a regression method which uses auxiliary information. This is a version of the method due to Rubin (1986). A parametric form of the regression of Z on X and Y is assumed and the corresponding parameters are then estimated from data in files A, B, and C. For example, in the case of a linear regression, we have the model

$$E(Z \mid X,Y) = \beta_0 + \beta_1 X + \beta_2 Y, \qquad V(Z \mid X,Y) = \sigma^2, \tag{3.1}$$

where β_0, β_1, and β_2 are estimated from equations similar to the usual least squares equations by combining information from files A, B, C suitably. Below we describe a procedure for doing this which is somewhat different from the one described in Rubin (1986). If file C has (X,Y,Z) information, then estimates can be obtained of the conditional correlation $\rho_{Y,Z|X}$ from C, the correlation $\rho_{X,Z}$, mean μ_Z, and standard deviation σ_Z from B and the correlation $\rho_{X,Y}$, means μ_X, μ_Y, and standard deviations σ_X, σ_Y from A. Thus file B will be used only if file A is deficient in information about the quantity of interest and file C will be used for some information only when A and B are deficient. Thus we assume a hierarchy of reliability or relevance of the files A, B, and C. Such a hierarchy was not assumed by Rubin. We can then get the required estimates from

$$\beta_2 = \rho_{Y,Z|X} \frac{\sigma_{Z|X}}{\sigma_{Y|X}}, \quad \beta_1 = \rho_{X,Z|Y} \frac{\sigma_{Z|Y}}{\sigma_{X|Y}},$$

$$\beta_0 = \mu_Z - \beta_1 \mu_X - \beta_2 \mu_Y, \tag{3.2}$$

where

$$\sigma_{Z|X} = (1 - \rho^2_{X,Z})^{1/2} \sigma_Z, \quad \sigma_{Y|X} = (1 - \rho^2_{X,Y})^{1/2} \sigma_Y,$$

$$\sigma_{Z|Y} = (1 - \rho^2_{Y,Z})^{1/2} \sigma_Z, \quad \sigma_{X|Y} = (1 - \rho^2_{X,Y})^{1/2} \sigma_X, \tag{3.3}$$

and $\rho_{X,Z|Y}$ is obtained from the standard formula after first calculating $\rho_{Y,Z}$ from (2.1), *i.e.*

$$\rho_{X,Z|Y} = (\rho_{X,Z} - \rho_{X,Y}\rho_{Y,Z})\ (1 - \rho^2_{X,Y})^{-1/2}. \tag{3.4}$$

It may be noted that under the normality assumption, departures from conditional independence are parametrized by $\rho_{Y,Z|X}$. Under conditional independence, $\rho_{Y,Z|X} = 0$ and the model (3.1) reduces to the simple linear regression of Z on X, *i.e.*

$$E(Z \mid X) = \beta_0 + \beta_1 X, \quad V(Z \mid X) = \sigma^2, \tag{3.5}$$

which can be specified by combining information from files A and B or from file B alone. The formulas (3.2) reduce to

$$\beta_2 = 0, \quad \beta_1 = \rho_{X,Z} \frac{\sigma_Z}{\sigma_X}, \quad \beta_0 = \mu_Z - \beta_1 \mu_X. \tag{3.6}$$

For the case when file C contains information about $\rho_{Y,Z}$ only, the parameters of (3.1) can be easily estimated in a similar manner by combining information from A, B and C.

After the regression model is determined, the REG* method can be applied in the following two steps. Step II is important because we want to have live values of Z so that relationships among components of multivariate Z are preserved.

REG* (Step I) For each (X,Y) in A, find an intermediate value Z_{int} from the regression model (3.1).

REG* (Step II) Replace each (X,Y,Z_{int}) obtained in Step I with (X,Y, Z_{match}) where Z_{match} denotes a live Z-value from B which is nearest under the Euclidean distance in (X,Z) where the components X and Z would be scaled by their respective standard deviations. In other words, the hot deck distance method is used to find the live value. This was termed "regression with predictive mean matching" by Rubin; see Little and Rubin (1987).

Another point of departure from the method described by Rubin (1986) is that in his method a predicted Y is found for records on file B using an equation analogous to (3.1) and then corresponding predicted Z values are found; then records on file A are matched to records on file B based on the difference in predicted Z-values.

If auxiliary information is not available then the matching method REG under CIA can be used. The two steps are

REG (Step I) For each (X,Y) in A, find Z_{int} from the simple regression model (3.5).

REG (Step II) Same as in REG*.

The method described by Rubin (1986) differs in that a predicted Z is also obtained for records on file B using (3.5), and then records on file A are matched to records on file B based on the difference in predicted Z-values. In the present example, where X is univariate, this is equivalent to matching on X.

3.2 The Hot Deck Method

We first describe a hot deck method using auxiliary data. This is a version of the method due to Paass (1986). Here, ideas of nonparametric regression are used. In parametric regression, the conditional distribution of Z given X and Y is specified in a wide sense by mean and variance functions in terms of a few parameters. In nonparametric regression the techniques of nonparametric density estimation are used to estimate the conditional distribution itself. For instance, in the case of the nearest neighbour method of density estimation, for each (X,Y), K nearest neighbours (with respect to a distance function such as the Euclidean distance in (X,Y) are determined and then the conditional distribution is represented by this sample (possibly weighted) of the K neighbours where K is an integer specified suitably. Thus, $P(Z\epsilon U \mid X,Y)$ can be specified as a conditional expectation,

$$E(I_U(Z) \mid X,Y) = \sum_{i=1}^{K} w_i(X,Y)\, I_U(Z_i), \qquad (3.7)$$

where w_i's denote weights which decrease with growing distance of (X_i,Y_i) from (X,Y) and I_U is the indicator function for the set U.

In Paass's method, first the conditional distribution of Z for each (X,Y) in A is determined by representing it with a set of K Z-values using nonparametric regression. In other words, K Z-values are added to each (X,Y). Then for each (X,Y) in A, a single live Z-value, Z_{match}, from B is obtained which is nearest under (X,Z)-distance. This gives the matched file with (X,Y,Z_{match}). The conditional distributions for file A are obtained by an iterative process in the case of file C with (Y,Z) information, as follows. Choose K initial values for nearest neighbours for Z in file A, for Y in file B, and for X in file C. This can be done by the usual hot deck method of imputation. Now each cycle consists of determining conditional distributions for elements (X,Y) in A from information in C, *i.e.* suitably updating K Z-values in A from Z-values in C using (X,Y) distance, and then updating K Y-values in file B from those of file A using (X,Z) distance, and finally updating K X-values in C from those of file B using (Y,Z) distance. This cycle is repeated until the maximal difference between some statistics for the three-dimensional distribution of (X,Y,Z) of successive iterations (*e.g.* covariance matrix) falls below a given threshold. At convergence, each file has K added values representing respective conditional distributions. In the other case in which file C has information about (X,Y,Z) the process becomes noniterative. We simply use file C to get K Z-values for A using (X,Y) distance and then get Z_{match} from B for each (X,Y) in A using (X,Z)-distance. This case was, however, not considered by Paass.

In the empirical study considered in this paper we did not use the above iterative version of Paass's method when file C had (Y,Z) data, because of its computationally intensive nature. Instead, we used a simplified noniterative version with $K = 1$. This method, denoted by HOD*, consists of the following two steps.

HOD* (Step I) For each (X,Y) in A, find an intermediate value Z_{int} from C using hot deck with Y-distance in the case of (Y,Z) auxiliary information and with (X,Y) Euclidean distance in the case of (X,Y,Z) auxiliary information.

HOD* (Step II) Replace each (X,Y,Z_{int}) obtained in Step I by (X,Y,Z_{match}) where Z_{match} is obtained from B using hot deck with (X,Z) Euclidean distance.

If file C were not available, then the matching method HOD under CIA can be used. The two steps for HOD are

HOD (Step I) Determine suitable X-categories as in usual hot deck imputation.

HOD (Step II) For each (X,Y) in A, impute a live Z-value from the corresponding X-category from B using hot deck with X-distance.

4. THE PROPOSED MODIFICATIONS BY CATEGORICALLY CONSTRAINED MATCHING

We propose modifications to REG, REG*, HOD and HOD* matching methods by imposing categorical constraints on the Z-values selected from B for completing A. The purpose of these constraints is to preserve categorical associations (as defined by log linear modelling) under a suitable partition of (X,Y,Z) for the matched file. These associations are obtained by combining information from A, B and C. The idea of categorically constrained matching is based on the method of log linear imputation (*cf.* Singh 1988, Singh *et al.* 1988). Here the constraints could be based on auxiliary information which could be used to estimate the categorical conditional distribution, or some aspects of it, but which would not be of sufficient quality to estimate the full conditional distribution.

We start with a suitable partition of X, Y and Z variables. Let X^*, Y^*, Z^* denote the corresponding categorically transformed variables. Now the distribution of cell proportions for the (X^*, Y^*, Z^*) table can be parametrized by a log linear model

$$\log p_{ijk} = u + u_{1i} + u_{2j} + u_{3k} + u_{12ij} + u_{13ik} + u_{23jk} + u_{123ijk}, \qquad (4.1)$$

where p_{ijk} denotes the proportion for (i,j,k)th cell and 1, 2, 3 denote respectively X^*, Y^*, and Z^*. It should be noted that the parametrization (4.1) holds for arbitrary underlying distributions of the original variables (X,Y,Z). The files A and B, of course, do not contain any information about the two-factor effects u_{23} and three-factor effects u_{123}. If these are set to zero, this amounts to assuming CIA in the categorical sense, *i.e.* $Y^* \perp Z^* \mid X^*$. However, with auxiliary information in file C, this assumption can be avoided because the parameters u_{23} and u_{123} could be estimable from C. Thus, regardless of the form of the joint distribution of (X,Y,Z), the above log linear modelling provides a unified approach for gauging departures from CIA at least in the categorical sense. In the linear regression approach, on the other hand, departures from CIA are parametrized by $\rho_{Y,Z|X}$ only in the case of normality.

As was explained in Section 2, the auxiliary information from file C (either on (Y,Z) or on (X,Y,Z)) is first used to construct categorical constraints in the form of a (X^*, Y^*, Z^*) distribution. This is done by means of raking such that u_{23} and u_{123} effects from file C are preserved. The categorically constrained version of REG*, denoted by REG.LOGLIN*, can now be defined by the following two steps.

REG.LOGLIN* (Step I) Same as in Step I of REG*.

REG.LOGLIN* (Step II) Same as in Step II of REG* except that categorical constraints are imposed, implying that match order is required when obtaining live Z-values from B. We first find the match with minimum distance in (X,Z). The (X^*, Y^*, Z^*) category of the completed record would be noted and if the resulting number of matched records in that (X^*, Y^*, Z^*) category does not exceed the count imposed by the categorical constraints that match is allowed. Otherwise, that match is rejected and the match with the second smallest distance is examined. The process continues until file A is completed, and then the distribution of (X^*, Y^*, Z^*) in the completed file must satisfy the categorical constraints.

Similarly, the categorically constrained version of HOD*, denoted by HOD.LOGLIN*, consists of the following two steps.

HOD.LOGLIN* (Step I) For each (X,Y) in A, find an intermediate value, Z_{int}, from C using hot deck with Y- or (X,Y)-distance as the case may be such that the categorical constraints are satisfied. This step is similar to Step II of REG.LOGLIN*.

HOD.LOGLIN* (Step II) For each (X,Y,Z_{int}), a live value, Z_{match}, from B is determined using hot deck with (X,Z)-distance while respecting the category of Z_{int}.

An alternative approach for HOD.LOGLIN* would have been to impute an intermediate Z_{int} without constraints and then to use categorically constrained distance matching to get a live value from file B, as in Step II of REG.LOGLIN*. This was also tried but did not work well so it was dropped from the study becuase of computational burden. One possible explanation for its poor performance is shrinkage to the mean for the Z_{int} values from file C due to file C being too small. That is, the Z_{int} values would tend to be near the centre of the distribution and when the categorical constraints are then imposed the final Z values would tend to be clumped at the inside boundaries of the outer Z categories.

Suppose file C has information only at the macro-level in the form of a categorical distribution, or the micro-level information in C is considered unreliable but the information in the categorical distribution under a somewhat coarse partition is considered reliable. We can then define categorically constrained versions of the REG and HOD methods, to be denoted by REG.LOGLIN and HOD. LOGLIN respectively. The two steps for REG.LOGLIN are

REG.LOGLIN (Step I) Same as in Step I of REG.

REG.LOGLIN (Step II) Same as in Step II of REG.LOGLIN*.

Similarly, HOD.LOGLIN consists of the following two steps.

HOD.LOGLIN (Step I) Same as in Step I of HOD.

HOD.LOGLIN (Step II) Same as in Step II of REG.LOGLIN* except that no intermediate values Z_{int} exist, so that matching is based on X-distance instead of (X,Z)-distance.

For both REG.LOGLIN and HOD.LOGLIN, which do not require micro-level information on file C, the CIA is being used only within X, Y categories. Thus a reduced form of conditional independence is being assumed and the consequences of this assumption should not be as severe as those of the full CIA.

5. AN ILLUSTRATIVE EXAMPLE

Before we investigate the empirical properties of the proposed modifications in relation to the previously proposed methods, it may be instructive to consider a simple numerical example to illustrate the types of computation involved with the eight methods. Suppose files A, B and C are as shown in Table 1 which are based on random samples drawn from a multivariate normal with mean 0 and covariance matrix specified by $\sigma_X = \sigma_Y = \sigma_Z = 1$, $\rho_{X,Y} = \rho_{X,Z} = .5$ and $\rho_{Y,Z} = .7$ (which implies that $\rho_{Y,Z|X} = .6$). Here, file C is assumed to have only (Y,Z) information. For file A, Z-values are suppressed in Table 1 but are shown in Table 3 for computing evaluation measures. Suppose we employ, for simplicity and in view of small file sizes, a rather coarse categorical transformation for X, Y, Z by considering only two categories, $(-\infty,0)$ and $[0,\infty)$. Then, the three two dimensional count tables corresponding to files A, B and C can be constructed as in Table 2(a). Table 2(b) shows the adjusted tables for B and C so that they match the appropriate marginals as described in Section 2. Table 2(c) gives the three-dimensional table obtained after raking and Table 2(d) gives the desired categorical constraints after random rounding of entries of Table 2(c) as explained earlier in Section 2.

The eight methods were applied to the data of Table 1 and the matching results are shown in Table 3 along with the true values of Z which were suppressed in Table 1.

The evaluation measures shown in Table 3 were briefly introduced earlier in the introduction and are fully explained in the next section. The categorical partition for the χ^2 measure was the same as the one used for deriving categorical constraints. Note that since the partitioning is not changed for evaluation, the χ^2 values for M3, M4, M7 and M8 would be identical. It should be pointed out that the evaluation measures are given only for the sake of illustrating the calculation and should not be construed as indicators for the relative performance of various methods because they are based on just one small sample realization.

The method M8 (HOD.LOGLIN*) happens to be the most computationally intensive, the details of which are shown in Table 4. From this, it would be relatively easy to visualize the computational steps required for other methods.

Table 1

Data for Files A, B, C

Record Identifier	File A		Record Identifier	File B		Record Identifier	File C	
	X	Y		X	Y		Y	Z
A1	−0.86	−0.32	B1	−0.95	−0.69	C1	−0.40	−0.60
A2	−0.77	−0.33	B2	−0.64	−0.83	C2	−2.33	−2.81
A3	−0.09	−0.26	B3	−1.58	−0.11	C3	−0.79	−0.47
A4	−0.42	0.62	B4	−0.42	0.36	C4	0.67	−0.29
A5	−0.81	0.56	B5	0.97	−0.42	C5	−0.65	1.19
A6	−0.56	0.00	B6	1.09	−1.16	C6	−1.32	0.05
A7	0.37	−0.04	B7	0.44	−0.49	C7	−0.55	0.70
A8	0.06	−1.29	B8	0.14	−0.38	C8	0.55	0.66
A9	0.95	−2.15	B9	1.33	1.24	C9	1.31	1.12
A10	1.90	−1.07	B10	0.80	0.85	C10	1.46	2.58
A11	1.32	0.61	B11	1.60	0.31			
A12	1.38	0.79	B12	1.42	0.99			
A13	1.63	1.03						
A14	0.50	1.24						
A15	0.90	1.19						

Table 2

Categorial distributions for files A, B, C under the given 2 × 2 × 2 partition

(a)

File A

	$Y < 0$	$Y \geq 0$
$X < 0$	3	3
$X \geq 0$	4	5

File B

	$Z < 0$	$Z \geq 0$
$X < 0$	3	1
$X \geq 0$	4	4

File C

	$Z < 0$	$Z \geq 0$
$Y < 0$	3	3
$Y \geq 0$	1	3

(b)

Unadjusted File A Table

	$Y < 0$	$Y \geq 0$
$X < 0$	3	3
$X \geq 0$	4	5

Adjusted File B Table

	$Z < 0$	$Z \geq 0$
$X < 0$	4.5	1.5
$X \geq 0$	4.5	4.5

Adjusted File C Table

	$Z < 0$	$Z \geq 0$
$Y < 0$	5.15	1.85
$Y \geq 0$	3.85	4.15

(c) Raked 2 × 2 × 2 table of ones to match the marginals in Table 2(b)

	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2.55	1.95	0.45	1.05
$X \geq 0$	2.60	1.90	1.40	3.10

(d) Categorical constraints by randomly rounding entries of Table 2(c)

	$Z < 0$		$Z \geq 0$	
	$Y < 0$	$Y \geq 0$	$Y < 0$	$Y \geq 0$
$X < 0$	2	2	1	1
$X \geq 0$	2	2	2	3

Table 3

Comparison of Eight Matching Methods for Completing File A

File A			Matched Z-Values							
			Versions of REG Method				Versions of HOD method			
X	Y	Z	M1	M2	M3	M4	M5	M6	M7	M8
−0.86	−0.32	−0.97	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69	−0.69
−0.77	−0.33	0.16	−0.69	−0.69	−0.69	−0.69	−0.83	−0.69	−0.83	−0.69
−0.09	−0.26	0.19	−0.38	−0.38	0.36	0.36	0.36	−0.38	0.36	0.36
−0.42	0.62	−0.44	−0.38	0.36	0.36	−0.38	0.36	−0.38	0.36	−0.38
−0.81	0.56	−0.76	−0.69	0.36	−0.69	0.36	−0.69	0.36	−0.69	0.36
−0.56	0.00	1.06	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83	−0.83
0.37	−0.04	−1.18	−0.38	−0.38	−0.38	0.36	−0.49	−0.49	−0.49	−0.49
0.06	−1.29	0.33	−0.38	−0.38	−0.36	−0.38	−0.38	−0.38	0.85	0.36
0.95	−2.15	−1.26	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16	−0.42	−1.16
1.90	−1.07	0.01	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31
1.32	0.61	2.08	0.31	0.99	0.31	0.99	1.24	−0.42	1.24	−0.42
1.38	0.79	0.32	0.31	0.99	0.31	0.99	0.99	−0.42	0.99	−0.42
1.63	1.03	1.53	0.31	0.99	0.31	0.99	0.31	0.99	0.31	0.99
0.50	1.24	1.34	−0.49	0.85	−0.49	−0.38	−0.49	0.85	−0.49	0.85
0.90	1.19	−1.01	−0.42	0.85	−0.42	−0.42	−0.42	0.85	−0.42	0.85
Evaluation Measures		MAD-Z	0.79	0.81	0.76	0.78	0.79	0.85	0.78	0.78
		x^2	13.07	13.34	1.75	1.75	2.70	10.78	1.75	1.75

Note: M1: REG M2: REG* M3: REG.LOGLIN M4: REG.LOGLIN* M5: HOD M6: HOD* M7: HOD.LOGLIN M8: HOD.LOGLIN*.

Table 4

Computational Steps Required for M8 (HOD.LOGLIN*)

(X,Y) cell	X	Y		Match Order	Y-dist	Z_{int}		(X,Z)-dist	Z_{match}	
$X < 0$	−0.86	−0.32	(A1)	2	.08	−0.60	(C1)	.12	−0.69	(B1)
$Y < 0$	−0.77	−0.33	(A2)	1	.07	−0.60	(C1)	.19	−0.69	(B1)
	−0.09	−0.26	(A3)	3	.29	0.70	(C7)	.47	0.36	(B4)
$X < 0$	−0.42	0.62	(A4)	2	.05	−0.29	(C4)	.56	−0.38	(B8)
$Y \geq 0$	−0.81	0.56	(A5)	1	.01	0.66	(C8)	.50	0.36	(B4)
	−0.56	0.00	(A6)	3	.40	−0.60	(C1)	.25	−0.83	(B2)
$X \geq 0$	0.37	−0.04	(A7)	4	.36	−0.60	(C1)	.14	−0.49	(B7)
$Y < 0$	0.06	−1.29	(A8)	1	.03	0.05	(C6)	.57	0.36	(B4)
	0.95	−2.15	(A9)	2	.18	−2.81	(C2)	1.66	−1.16	(B6)
	1.90	−1.07	(A10)	3	.25	0.05	(C6)	.40	0.31	(B11)
$X \geq 0$	1.32	0.61	(A11)	1	.06	−0.29	(C4)	.37	−0.42	(B5)
$Y \geq 0$	1.38	0.79	(A12)	3	.12	−0.29	(C4)	.43	−0.42	(B5)
	1.63	1.03	(A13)	5	.28	1.12	(C9)	.25	0.99	(B12)
	0.50	1.24	(A14)	2	.07	1.12	(C9)	.41	0.85	(B10)
	0.90	1.19	(A15)	4	.12	1.12	(C9)	.29	0.85	(B10)

Note: Match order between records in files A and C is within (X, Y) cell under the categorical constraints given by Table 2(d).

6. EMPIRICAL INVESTIGATION OF PROPOSED MATCHING METHODS

This section presents the details of an empirical evaluation through an extensive simulation study with synthetic data generated from symmetric as well as skewed multivariate distributions. Symmetry was introduced via normal distributions, while skewness was introduced via contaminations by multivariate log normal distributions. The reason for using synthetic data is to have control over all of the relevant parameters, including those specifying the joint relationships of the different variables. This permits evaluation of the various approaches to matching as the joint relationships are allowed to depart in a systematic manner from conditional independence. It also permits comparisons of the methods as the underlying distribution generating the data moves away from symmetry. Proxy auxiliary information was generated by changing parameters of the normal distribution generating file C or by inducing log normal contaminations. We thus have four types of matching problems; the two corresponding to symmetric and skewed distributions with nonproxy data for C and the two corresponding to symmetric distributions with two types of proxy data for C. Programming was done on micro-computers using the software GAUSS.

6.1 Design of the Monte Carlo Study

In order to simulate statistical matching three datafiles are needed: a host file A, a donor file B, and an auxiliary file C. These are generated synthetically from specified distributions, with each file containing the three variables X, Y and Z. In file A the variable Z is suppressed and in file B the variable Y is dropped. The suppressed Z-values in file A are used to evaluate the performance of the various methods of statistical matching. File C could have only (Y,Z) information (by suppressing X) or the full (X,Y,Z) information. The empirical results presented in this paper correspond to file C with only (Y,Z) variables although file C with (X,Y,Z) variables was also included in the study (see Singh *et al.* 1990).

Runs of 100 simulations apiece were performed for each combination of design parameters considered. Four evaluation measures were calculated for each simulation and then were combined over all 100 simulations.

Files A and B were always generated from the same underlying distribution, with each containing 500 independent and identically distributed observations. File C contained 250 observations, not necessarily from the same distribution as that for files A and B; that is, file C could contain either proxy or nonproxy auxiliary information.

The distribution of observations (X,Y,Z) was multivariate normal with some log normal contamination introduced by taking the exponentials of X, Y and Z for some of the observations. Individual observations were contaminations or not according to a Bernoulli process with probability fixed for any particular run of 100 simulations. Prior to contamination X, Y and Z were standard normal. The covariances of (X,Y) and (X,Z) prior to contamination were always .5, with the covariance of (Y,Z) varying from run to run. Consequently, the conditional correlation of Y and Z given X, $\rho_{Y,Z|X}$, was also varied from run to run.

For most runs the distribution of observations in the auxiliary file C was the same as that in files A and B. However, if in an application the source of auxiliary information is historical or via proxy variables this assumption may be unreasonable. Two series of runs were carried out with proxy auxiliary information. In the first series the auxiliary data had a different $\rho_{Y,Z|X}$. In the second series the auxiliary data had some log normal contamination.

For the proposed methods which use categorical constraints and for defining matching categories for the HOD method, it was necessary to choose a categorical partition. Two partitions were used. The first, called standard interval, divided the ranges of the X, Y and Z variables into the categories < -1, $[-1,0)$, $[0,1)$, ≥ 1; that is, the partition was centred on the mean of the marginal distribution before contamination, with break points at the centre and at plus or minus one standard deviation. The second partition, called equal probability, was similar but had break points at the quartiles of the pre-contamination marginal distributions; that is, the partition had the categories $< -.6745$, $[-.6745,0)$, $[0,.6745)$, $\geq .6745$. The partitions were defined in terms of the pre-contamination distributions; for simplicity the same partitions were used when there were log normal contaminations. It would, however, have been more realistic to let the partitions be data dependent.

6.2 The Matching Methods

The eight methods as defined earlier were considered. Except for REG and HOD, all others use auxiliary information. Thus, we have two variants for each depending on whether (Y,Z) or (X,Y,Z) information is available in file C. For the methods HOD and HOD.LOGLIN, three versions of hot deck (namely, rank, random, and X-distance) were considered for finding live Z-values from B although only results based on X-distance are reported here. For the other six methods, although we considered three types of hot deck (namely, Z-distance, (X,Z)-distance, and (X,Y,Z)-distance) for finding live Z-values from B, we show only results for (X,Z) distance here for simplicity. Section 7.3 does contain a brief description of results obtained with different distance measures. The report by Singh *et al.* (1990) contains other details not included here. It may be noted that for using hot deck with (X,Y,Z)-distance to get a live Z-value from B, intermediate Y-values would have to be first obtained for B from file C, analogous to Z_{int} for file A. Note also that the Euclidean distance was always employed whenever hot deck with distance metric was used. However, variables were not preadjusted by their standard deviations for convenience and because all the variables in the synthetic population had common variances.

6.3 The Evaluation Measures

Four evaluation measures were used to measure how well the different matching methods performed. All of the evaluations are based on comparisons of the matched file to the file with the suppressed true Z-values. Two of the measures are based on categorical comparisons, but the categories used for evaluations need not be the same as those used for categorical constraints by the LOGLIN procedures. The results reported here correspond to using the equal probability partition (see Section 6.1) for matching and the standard interval partition for evaluations. The first of the four evaluation measures is based on unit by unit comparison of the matched and suppressed Z-values. However, the objective of a statistical matching procedures cannot be to reproduce the suppressed Z-values exactly, but to produce Z-values that come from the same distribution given what is known, in this case given X and Y.The last three evaluation measures are based more on comparisons of the conditional distributional properties of Z.

(i) Average of Mean Absolute Differences of Z ($\overline{\text{MAD-}Z}$)

The simplest measure of performance is the mean absolute difference between the matched and suppressed Z-values for records in file A. Monte Carlo averages of these means as well as standard errors were obtained.

The formula for the MAD-Z statistic for any given simulation, is

$$\text{MAD-}Z = \sum_i | Z_{s,i} - Z_{m,i} | /500, \qquad (6.1)$$

where $Z_{s,i}$ is the suppressed Z-value for the ith record in file A, $Z_{m,i}$ is the matched Z-value, and the sum is over all 500 records of file A. $\overline{\text{MAD-}Z}$ denotes the average of the MAD-Z statistics over simulations.

(ii) Average of Absolute Difference of Covariances ($\overline{\text{AD-Cov}}$)

The second measure of performance is the absolute difference of the conditional covariances of Y and Z given X in the matched and suppressed files. Monte Carlo averages of these absolute differences as well as standard errors were obtained.

For a file with variables X, Y and Z we may define

$$\text{Cov}(Y,Z \mid X) = \text{Cov}(Y,Z) - \text{Cov}(X,Y)\text{Cov}(X,Z)/\text{Var}(X), \qquad (6.2)$$

where Cov and Var are the sample covariance and variance operators respectively. In the multivariate normal case this corresponds to the covariance of Y and Z given X. Otherwise it may be interpreted as the covariance of the residuals of a linear regression of Y on X with the residuals of a linear regression of Z on X. The AD-Cov statistic for any given simulation, would be the absolute difference between these quantities for the matched and suppressed files. $\overline{\text{AD-Cov}}$ denotes as usual the average over simulations.

(iii) Average of Chi-square Statistics ($\overline{\chi^2}$)

The third measure of performance, based on categorical comparisons, is a distance measure based on the Pearson chi-square statistic. What is reported is the average chi-square statistic over the 100 simulations, transformed to lie in the interval (0,1).

The formula for the chi-square statistic, is

$$\chi^2 = \sum_{i,j,k} (m_{ijk} - n_{ijk})^2/(m_{ijk} + .5), \qquad (6.3)$$

where m_{ijk} is the number of records in X^* category i, Y^* category j, and Z^* category k in the matched file, n_{ijk} is the same for the suppressed file, and the sum is over all (X^*, Y^*,Z^*) categories. A constant .5 is added to all of the denominators in this sum to avoid the problem of zeros.

Once the mean of the chi-square statistics from 100 simulations, say $\overline{\chi^2}$, is obtained, it is transformed to lie in the interval (0,1) using the transformation (see Bishop, Fienberg and Holland 1975, p 383; here 500 is the size of file A)

$$\text{Transformed } \overline{\chi^2} = \{\overline{\chi^2}/(\overline{\chi^2} + 500)\}^{1/2}. \qquad (6.4)$$

(iv) Likelihood Ratio Test (LRT)

The final measure of performance is also based on categorical comparisons. Within each (X^*, Y^*) category that has a minimum number of observations (in the present study, we set it at 20) a likelihood ratio test that the categorical Z-values from the matched and suppressed files come from the same multinominal distribution is performed. The tests for different (X^*, Y^*) categories are then combined to obtain an overall P-value. What is reported is the proportion of times, out of 100 simulations, that the overall P-value was less than .05. The larger this proportion, the greater the difference between the true and matched categorical distributions of Z^* given the (X^*, Y^*) categories.

The minimum sample size of 20 for (X^*, Y^*) categories in file A was required so that the chi-square approximation to the distribution of the test statistic might be reasonable. If the number of Z^* categories was increased, this minimum sample size might also need to be increased.

Using the same notation as in the previous measure, the formula for the likelihood ratio test statistic from the (i,j) (X^*, Y^*) category is

$$\begin{aligned} \text{LRT} = 2 \sum_k \{ & (n_{ijk} + .5) ln((n_{ijk} + .5)/ \\ & (n_{ijk} + m_{ijk} + 1)) + (m_{ijk} + .5) \\ & ln((m_{ijk} + .5)/(n_{ijk} + m_{ijk} + 1))\} \\ & + (4n_{ij} + 2K) ln 2, \qquad (6.5) \end{aligned}$$

where

$$\begin{aligned} n_{ij} = \sum_k n_{ijk} = \sum_k m_{ijk}, \quad & i = 1, \ldots, \text{I}, \\ & j = 1, \ldots, J, \quad k = 1, \ldots, K. \qquad (6.6) \end{aligned}$$

The asymptotic distribution of this statistic, when the m_{ijk}'s and n_{ijk}'s come from the same multinominal distribution, is chi-square with $(K - 1)$ degrees of freedom. An overall P-value is obtained by adding these statistics and their degrees of freedom for each (X^*, Y^*) category meeting the minimum sample size criterion, and finding the probability of a chi-square variable with the appropriate degrees of freedom being larger than the observed value.

7. RESULTS OF THE MONTE CARLO STUDY

In this section we describe the results of the simulation study. A more complete description is given in Singh *et al.* (1990). Tables of actual numbers underlying Figures 1 through 5 are available upon request.

We have not paid much attention to Monte Carlo standard errors of the evaluation measures in the presentation. This is because they were generally quite small, for example, coefficients of variation were generally less than two percent for the $\overline{\text{AD-Cov}}$ evaluation measure. Furthermore, the evaluations of different methods would be expected to be positively correlated so that the relative differences between matching methods would be even more precisely estimated than suggested by the standard errors. A further indication of the quality of the Monte Carlo evaluations of the various methods is the general smoothness of observed trends, for example, see Figures 2 to 5. In short, any discernible difference in the figures is likely to indicate a real difference.

7.1 Methods with no Auxiliary Information (REG and HOD)

Figures 2 through 5 show how departures from conditional independence affect performance of matching methods which use CIA. Apparently the use of such methods may result in serious bias in the joint relationship of (X,Y,Z) in the matched file. For example, Figure 2 shows a progressive deterioration as the true conditional correlation, $\rho_{Y,Z|X}$, moves away from zero with respect to all measures except $\overline{\text{MAD-}Z}$ which actually shows no deterioration at all. It may be due to the fact that $\overline{\text{MAD-}Z}$ is an unconditional measure which is based on unit by unit comparison of the matched and suppressed Z-values, while the other measures are based on comparisons of the conditional distributions of Z. It is interesting to note from Figure 2 that when the true value of $\rho_{Y,Z|X}$ is small, the performance of the HOD* method, which uses auxiliary information, can be worse with respect to the categorical or aggregate level evaluation measures than the performance of the HOD method which does not make use of auxiliary information. The point at which the use of auxiliary information would become advantageous would depend on the precision of the auxiliary information.

7.2 Methods with Auxiliary Information

Our empirical results do confirm, as expected, that the use of auxiliary information does protect against the failure of the CIA. The degree of protection would depend on the method and the type of auxiliary information used. A brief summary of performances of various methods was presented earlier in the introduction. Here, we will provide some details based on Figures 2 to 5.

In the regression family, the methods using auxiliary information on conditional correlations, namely REG* and REG.LOGLIN*, show very favourable performance with respect to the unit level measures (*i.e.* $\overline{\text{MAD-}Z}$ and $\overline{\text{AD-Cov}}$) for symmetric populations (see Figure 2). They continue to outperform hot deck methods for skewed populations (Figure 3) although the bias tends to increase as the degree of skewness grows. However, for proxy auxiliary information having different conditional correlation (Figure 4), the regression methods perform in a mixed fashion, *i.e.* they could be better or worse than hot deck methods at the unit level. In fact, for the second type of proxy auxiliary information (namely, with log normal contamination; see Figure 5), they tend to be slightly inferior to the HOD.LOGLIN method with respect to the $\overline{\text{AD-Cov}}$ measure. If we restrict ourselves to the regression family, then the REG* method can be recommended with regard to the unit level evaluation measures. However, with respect to the aggregate level, all regression methods show very unfavourable performance. This can probably be explained by the shrinkage to the mean effect as discussed in subsection 7.3.

Matching Methods for Figures 1 to 5

REG Z_{int} obtained from regression of Z on X, Z_{match} based on (X,Z) distance

REG* Z_{int} obtained from regression of Z on X and Y, Z_{match} based on (X,Z) distance

REG.LOGLIN Z_{int} obtained from regression of Z on X, Z_{match} based on (X,Z) distance using categorical constraints

REG.LOGLIN* Z_{int} obtained from regression of Z on X and Y, Z_{match} based on (X,Z) distance using categorical constraints

HOD Hot deck using X distance within X categories

HOD* Z_{int} obtained from file C using hot deck with Y distance, Z_{match} obtained from file B using hot deck with (X,Z) distance

HOD.LOGLIN Hot deck using X distance within X categories and using categorical constraints

HOD.LOGLIN* Z_{int} obtained using hot deck with Y distance and using categorical constraints, Z_{match} obtained using hot deck with (X,Z) distance within (X,Y,Z) categories

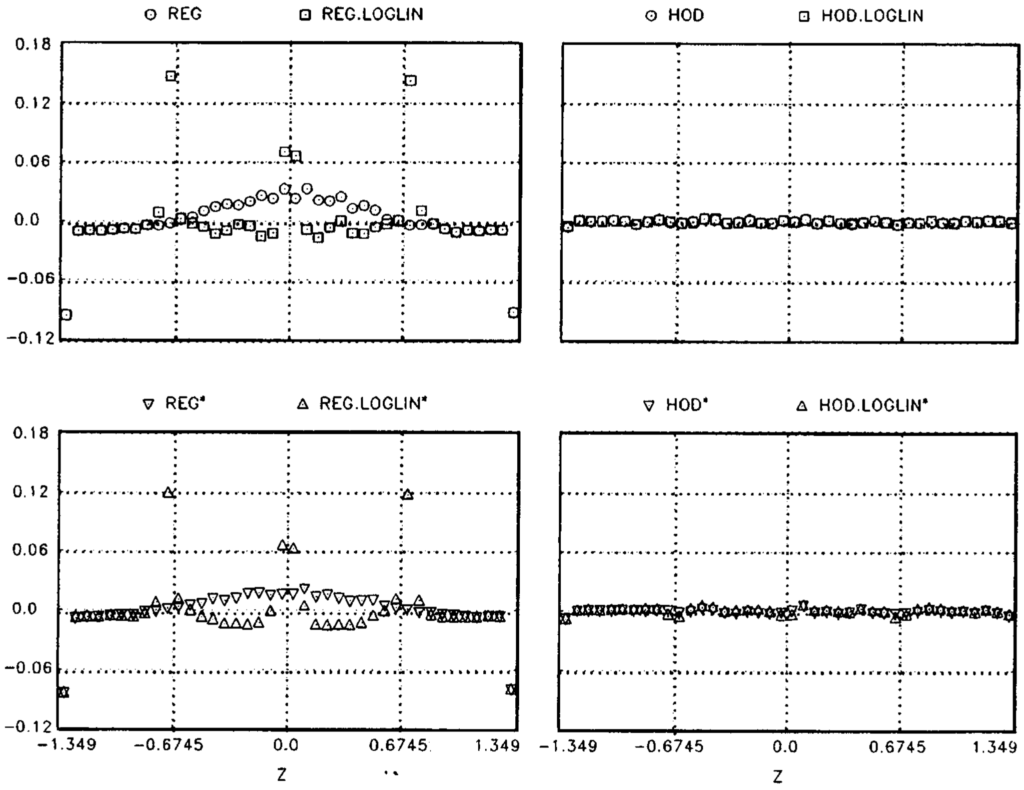

Figure 1. Difference of matched and suppressed marginal Z-histograms (symmetric data, $\rho_{Y,Z|X} = .4$)

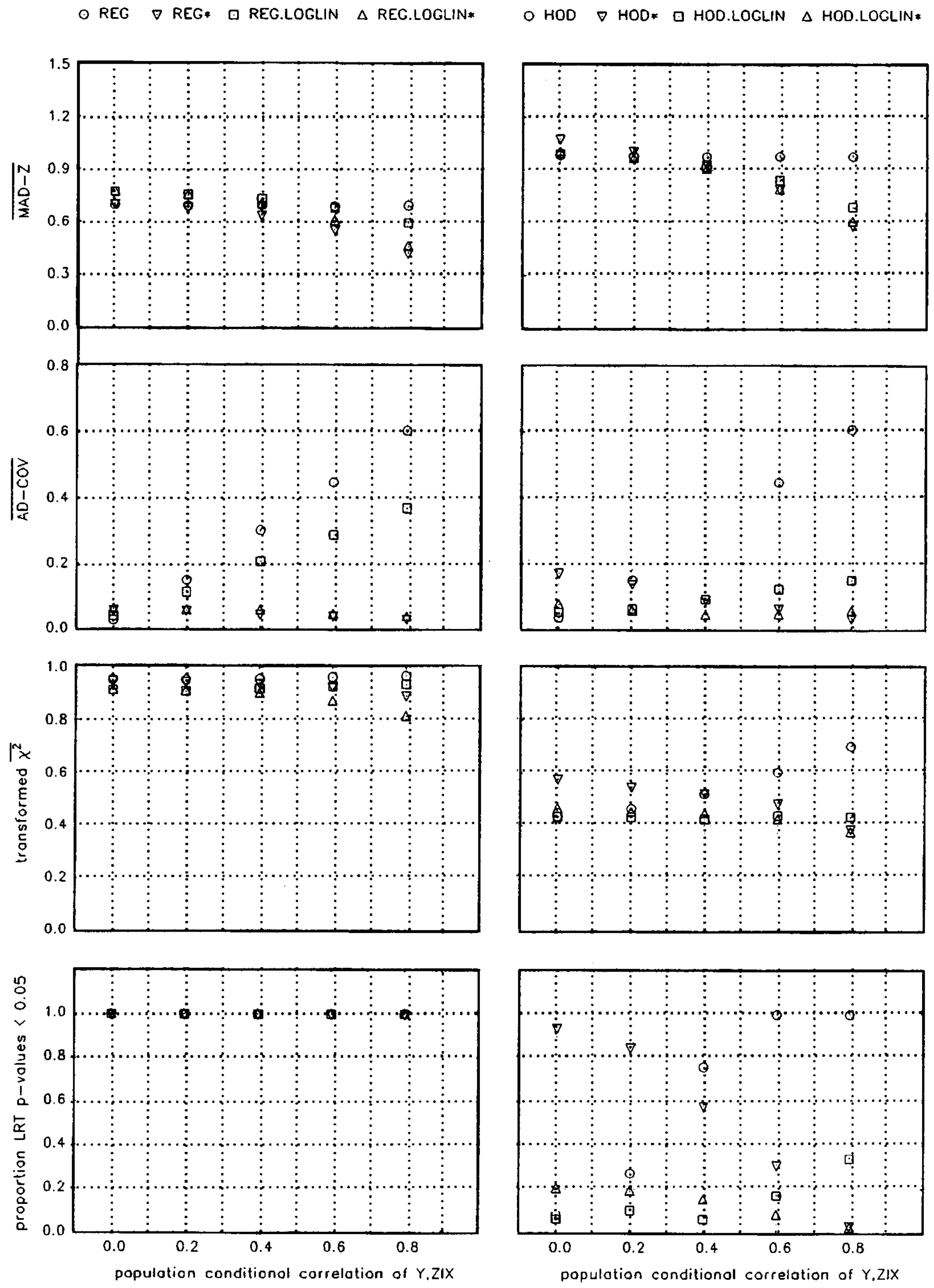

Figure 2. Comparison of statistical matching methods as $\rho_{Y,Z|X}$ varies for the symmetric population, non-proxy auxiliary information

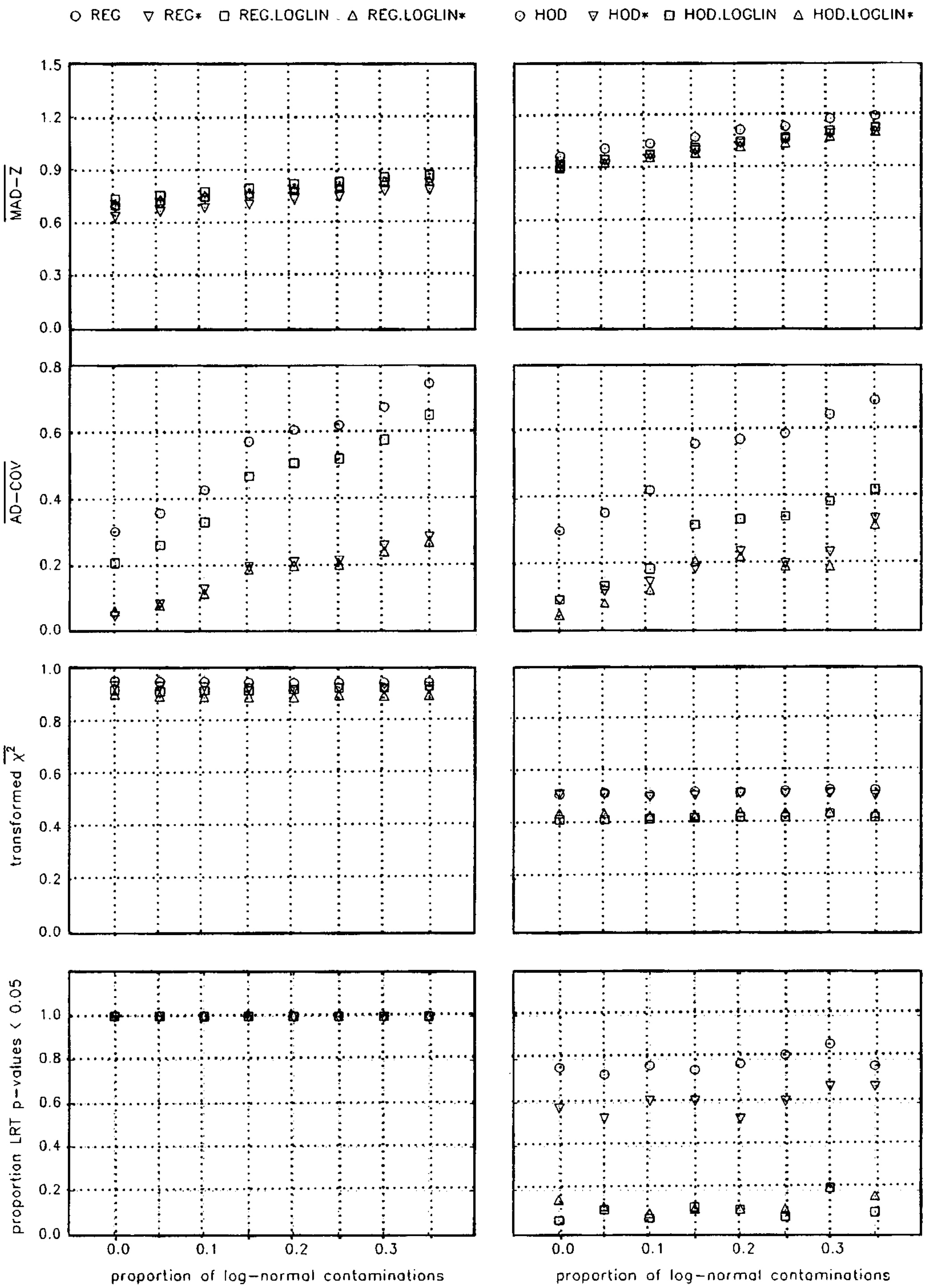

Figure 3. Comparison of statistical matching methods as the proportion of log-normal contamination varies ($\rho_{Y,Z|X}$ before contamination), non-proxy auxiliary information

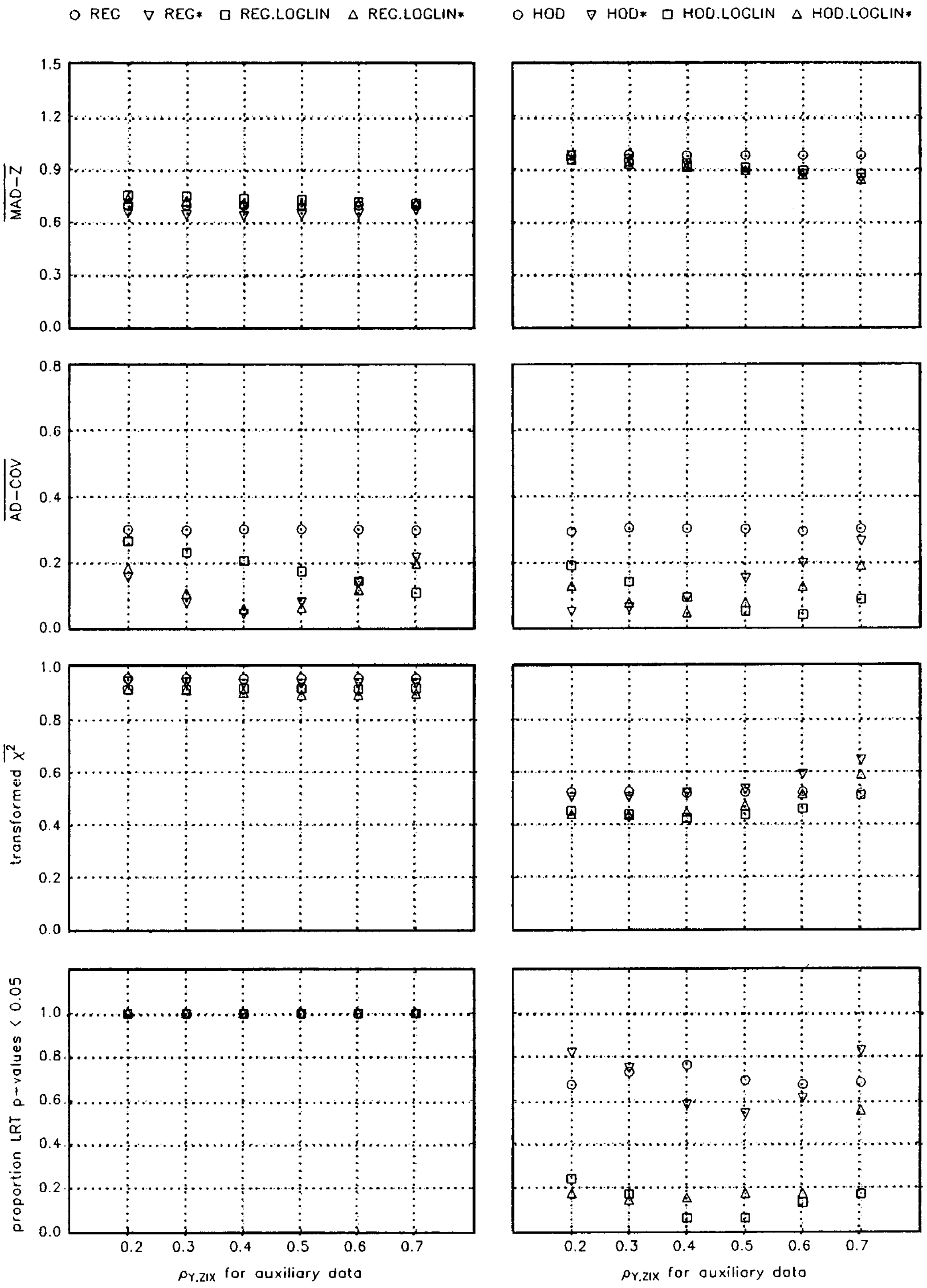

Figure 4. Comparison of statistical matching methods as $\rho_{Y,Z|X}$ varies for the auxiliary data file C ($\rho_{Y,Z|X}$ = .4 for files A and B)

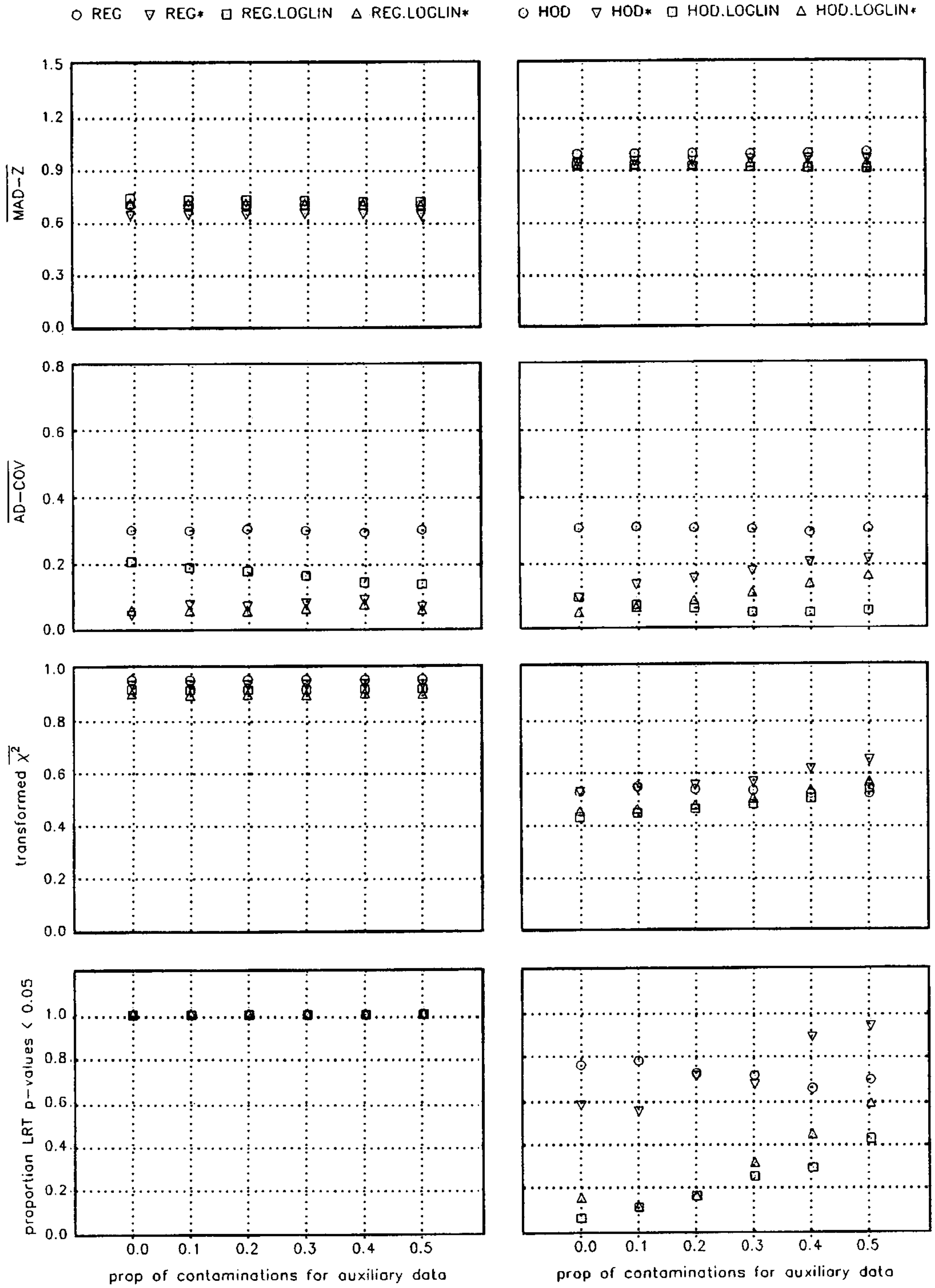

Figure 5. Comparison of statistical matching methods as the proportion of log normal contaminations varies for the auxiliary data file C ($\rho_{Y,Z|X}$ = .4 before contamination and files A and B have no log-normal contamination)

In the hot deck family of methods with auxiliary information, the two methods with categorical constraints, namely, HOD.LOGLIN and HOD.LOGLIN* show vary favourable performance at the aggregate level (*i.e.* with respect to transformed χ^2 and LRT) for all types of underlying populations; see Figures 2 to 5. In this case, the method HOD.LOGLIN generally outperforms HOD.LOGLIN*. Next we consider unit level measures. For symmetric and skewed populations (Figures 2 and 3), generally speaking, HOD* and HOD.LOGLIN* perform very similarly to each other and somewhat better than HOD.LOGLIN but slightly worse than REG*. However, for proxy auxiliary data (Figures 4 and 5) with respect to the AD-Cov measure, the method HOD.LOGLIN could be better or worse than the HOD.LOGLIN* and the REG* methods, and is more often better in the case of proxy data with log normal contaminations. Also HOD.LOGLIN in the case of proxy data tends to have fairly robust behaviour with respect to all four evaluation measures. Thus, in the hot deck family, based on overall performance, HOD.LOGLIN* can be recommended. However, in practice, HOD.LOGLIN may be preferable as a compromise because it performs moderately well at the unit level, extremely well at the aggregate level, is computationally much less demanding and shows robustness with respect to the proxy auxiliary data. Furthermore, HOD.LOGLIN does not require micro-level auxiliary information.

7.3 Miscellaneous Observations

In this subsection we describe separately some interesting findings, the corresponding empirical results for some of which are not included here, but are presented in Singh *et al.* (1990).

(i) Shrinkage to the Mean

An important and consistent finding was that matching methods in the regression family do not perform well with respect to the categorical measures. This can be explained by shrinkage towards the mean; that is, the matched Z-values are more tightly distributed about their mean than are the suppressed true Z-values. This is displayed in Figure 1 which shows the difference between the marginal histograms of matched and suppressed Z-values for various matching methods.

The positive differences for REG and REG* near the centre indicate that there are more Z-values in that region in the matched file than in the suppressed file. The very large negative observations at the extreme points of this plot are associated with open ended intervals, and it seems quite likely that had these intervals been broken down into several smaller intervals the plot would have shown several smaller negative numbers in the extreme tails, so that the interpretation of the plot should be that these methods are putting too many Z-values at the centre of the distribution at the expense of the extreme tails.

Figure 1 also shows shrinkage towards the mean for the REG.LOGLIN and REG.LOGLIN* methods. However, in this case the shrinkage is limited by the categorical constraints so that, while we still see that the tails of the Z-distribution of the matched file are too short, the displaced values are now not going to the centre of the distribution, but only to the partition boundary points which act like walls. The large positive values to either side of the central boundary point can be explained similarly if one bears in mind that what this plot is showing is actually an average of differences of histograms over 100 independent simulations. It seems reasonable that if we were to examine each of the 100 differences of histograms individually we would sometimes see a large positive value just to the left of the central boundary point, and sometimes just to the right, but never both at the same time.

Figure 1 also shows that shrinkage to the mean and boundary effects are not serious for methods in the hot deck family.

(ii) (Y,Z) *vs* (X,Y,Z) Auxiliary Information

Although only results based on (Y,Z) auxiliary information were presented in this paper, (X,Y,Z) auxiliary information was also considered as part of the simulation study as mentioned in Section 6. An interesting finding was that for the HOD.LOGLIN and HOD.LOGLIN* methods, the use of (Y,Z) auxiliary information leads, in general, to somewhat better performance at the aggregate level than the use of (X,Y,Z) information. This does not seem to be the case with the HOD* method. This phenomenon is probably due to instability in the estimation of (X^*,Y^*,Z^*) factor effects used in the categorical constraints on account of insufficient size of auxiliary data. An implication is that the true values were probably close to zero and so taking them as zero leads to better results. This suggests that the impact of different sample sizes on performance of matching methods should be considered, if possible, in future investigations. The above consideration also suggests an interesting new class of methods which would combine (X,Y,Z) micro-level auxiliary information for finding Z_{int} values along with the derived (Y^*,Z^*) categorical distribution only from file C for imposing constraints. These methods were, however, not included in the present study.

(iii) Comparison of Different Versions of Hot Deck Methods

In all the matching methods considered, except HOD and HOD.LOGLIN, the second step for finding Z_{match} consists of using hot deck imputation in which (X,Z)-distance is employed. For the remaining two, X-distance was considered. Some other options (for methods other than

HOD and HOD.LOGLIN), consist of using Z-distance or (X,Y,Z)-distance. For the latter, Y_{int} would have to be added first to file B. This was included in the original simulation study, although empirical results are not reported here. It was found that there is generally no difference though, for REG and REG* methods, (X,Z)-distance sometimes showed superior performance with respect to the $\overline{\text{AD-Cov}}$ measure. This is the reason for our choice of (X,Z)-distance in the methods considered here. However, in practice, it may be preferable to use Z-distance with hot deck matching methods because of computational convenience.

Further, it should be noted that for HOD and HOD.LOGLIN methods, there is the option of using random or rank in Step II instead of X-distance. In hot deck rank, records from files A and B are ranked separately according to the value of X, and then are matched based on ranks. This was proposed by G. Rowe for the SPSD application mentioned in the introduction. Clearly, this method is suitable for univariate X only. An advantage of ranking is that there will not be one record from file B acting as donor for many records from file A. The above three versions of hot deck were included in the Monte Carlo study although results for X-distance only are reported here. It was found that it generally does not make much difference which version is used. The choice of X-distance was made for HOD and HOD.LOGLIN because it was consistent with the hot deck distance version used for other methods. In practice the hot deck random version would be least demanding computationally; however, in a real application we would not know how much might be lost by using random matching instead of ranking or distance, and we would probably want to use as much information as would be feasible.

8. CONCLUDING REMARKS

In this paper, the problem of using auxiliary information in statistical matching was considered. The two main methods previously proposed are due to Rubin (1986) and Paass (1986), versions of which were denoted by REG* and HOD*. Some modifications of these methods, denoted by REG.LOGLIN* and HOD.LOGLIN*, were proposed by imposing categorical constraints derived from auxiliary information. These would reduce to REG.LOGLIN and HOD.LOGLIN if only categorical auxiliary information is available or useable. In the absence of auxiliary information, the usual methods of imputation, REG and HOD would be used. An empirical study was conducted to evaluate performance of the above eight methods with respect to four evaluation measures (two at the unit level, and two at the aggregate level). It was found that for the case of no auxiliary information, the HOD method is preferable. The case of auxiliary information is, however, more complex. If only unit level evaluation measures are deemed important, then the REG* method is recommended. If aggregate level measures are also considered important then if there is nonproxy auxiliary data HOD.LOGLIN* is recommended. As an alternative, a good compromise would be HOD.LOGLIN if computational burden is an important consideration or if proxy auxiliary data is believed to be present. If unit level measures are less important or are not of interest (this may often be the case because the matched data would generally be presented in tabular forms in practice), then HOD.LOGLIN would be recommended. With both HOD and HOD.LOGLIN methods, the similar performances of distance, random and rank versions might suggest the use of random versions in practice in view of its computational simplicity.

It may be remarked that we did not consider the fully iterative version of Paass's method. It would be interesting to find out in future investigations how this might perform. Another point that requires investigation is the implementation of categorical constraints with many variables. The application of the raking algorithm may be computationally prohibitive. In this connection, the results of Paass (1989) are expected to be useful.

In the present study we did not, due to limitations of computing, systematically vary the accuracy of the auxiliary data source; that is, we did not vary the size of the file C. We also did not vary the size of the files A or B. An interesting question that might have been addressed is how the performance of various methods might be affected by the size of these files.

Finally, it should be pointed out that although the results of this study are based on synthetic data (which was necessary to produce various scenarios mimicking real data), it is believed that the results would be relevant for real applications. Clearly, it would be interesting and useful to carry out a simulation study with real data to check whether the findings continue to hold and to see what sorts of substantive impact the biases in the joint distribution of the matched file have. A related question is how to account for such biases in inferences based on the matched file; that is, how to produce measures of uncertainty for parameter estimates from the matched file that reflect not only the variability within the matched file, but also the uncertainty inherent in the matching procedure itself. Although we are unable to answer this question, it is clear that matching procedures using auxiliary information would enhance the overall utility of the matched file. These and some other related questions will be investigated in the future.

ACKNOWLEDGEMENT

The authors would like to thank J. Armstrong, G. Gray, G. Hole, D. Royce and M. Wolfson for helpful comments. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University.

REFERENCES

ARMSTRONG, J. (1989). An evaluation of statistical matching methods. Methodology Branch Working Paper, BSMD, 90-003E. Statistics Canada.

BARR, R.S., and TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Surveys. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.

BARR, R.S., and TURNER, J.S. (1990). Quality issues and evidence in statistical file merging. In *Data Quality Control: Theory and Pragmatics* (Eds. G.E. Liepins and V.R.R. Uppuluri). New York: Marcel Dekker, 245-313.

BARR, R.S., STEWART, W.H., and TURNER, J.S. (1981). An empirical evaluation of statistical matching methodologies. Technical report, Edwin L. Cox School of Business, Southern Methodist University, Dallas, Texas.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, Mass: MIT Press.

BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.

BUDD, E.C., and RADNER, D.B. (1969). The OBE size distribution series: methods and tentative results for 1964. *American Economic Review, Papers and Proceedings*, LIX, 435-449.

COHEN, M.L. (1991). Statistical matching and microsimulation models. In *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*, Volume II, Technical Papers, (Eds. C.F. Citro and E.A. Hanushek). Washington, D.C.: National Academy Press, 62-85.

FELLEGI, I.P. (1977). Discussion paper. *Proceedings of the Section on Social Statistics, American Statistical Association*, 762-764.

FORD, B.L. (1983). An overview of hot-deck procedures. In *Incomplete Data in Sample Surveys*, (Vol. 2), (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 185-207.

KADANE, J.B. (1978). Some statistical problems in merging data files. In *1978 Compendium of Tax Research*. Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.

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

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social 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* (Eds. G.H. Orcutt, J. Merz and H. Quinke). Amsterdam: Elsevier Science.

PAASS, G. (1989). Stochastic generation of a synthetic sample from marginal information. *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 431-445.

PAASS, G., and WAUSCHKUHN, U. (1980). Experimentelle erprobung und vergleichende Bewertung statistischer Matchverfahren. Internal report, IPES.80.201, St. Augustin, *Gesellschaft für Mathematik und Datenverarbeitung*.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or Domains). *International Statistical Review*, 48,3-18.

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. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-132.

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

RUGGLES, N., RUGGLES, R., and WOLFF, E. (1977). Merging microdata: Rationale, practice and testing. *Annals of Economic and Social Measurement*, 6, 407-428.

SCHEUREN, F.J. (1989). Comment on Wolfson *et al.* (1989). *Survey of Current Business*, 69, 40-41.

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*, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.

SINGH, A.C. (1988). Log-linear imputation. Methodology Branch Working Paper, SSMD, 88-029E, Statistics Canada; also published in *Proceedings of the Fifth Annual Research Conference*, U.S. Bureau of the Census, 118-132.

SINGH, A.C., ARMSTRONG, J.B., and LEMAÎTRE, G.E. (1988). Statistical matching using log linear imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 672-677.

SINGH, A.C., MANTEL, H., KINACK, M., and ROWE, G. (1990). On methods of statistical matching with and without auxiliary information: Some modifications and an empirical evaluation. Methodology Branch Working Paper, SSMD, 90-016E. Statistics Canada.

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.

WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B., and ROWE, G. (1987). The social policy simulation database: an example of survey and administrative data integration. *Proceedings: Symposium on Statistical Uses of Administrative Data*, Statistics Canada, Ottawa, (Eds. J.W. Coombs and M.P. Singh), 201-229; another version published in *Survey of Current Business* (1989), 69, 36-40.