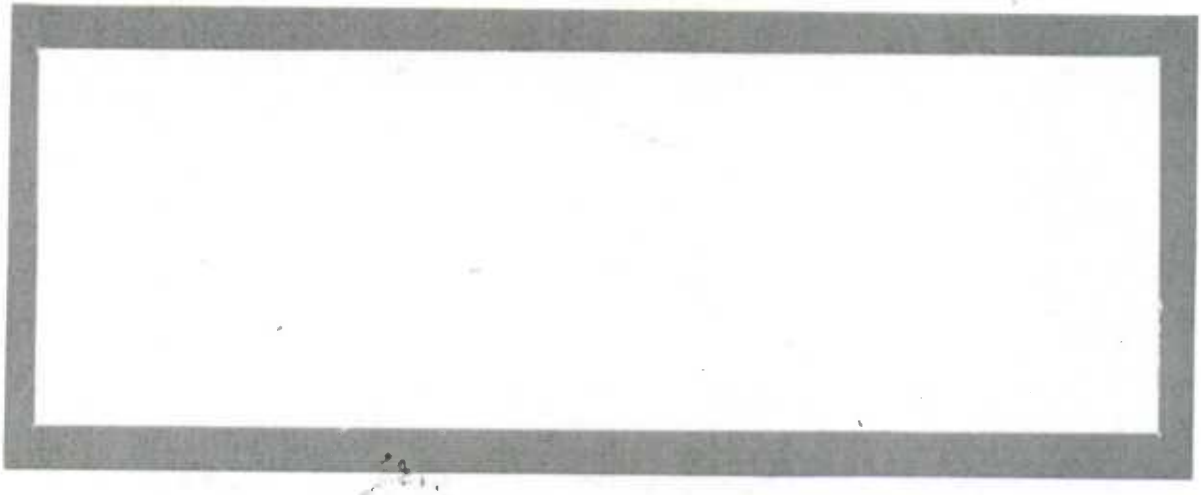
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER NO. SSMD-90-016E

METHODOLOGY BRANCH

ON METHODS OF STATISTICAL MATCHING WITH AND WITHOUT AUXILIARY INFORMATION

Some Modifications and an Empirical Evaluation

SSMD-90-016E

A.C. Singh[1], H. Mantel[1], M. Kinack[1] and G. Rowe[2]

December 1990

[1] Social Survey Methods Division, Statistics Canada

[2] Social and Economic Studies Division, Statistics Canada

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 seriously bias the resulting joint relationships between variables. The use of appropriate auxiliary information could alleviate this problem to a great extent. In this report, methods of statistical matching corresponding to three methods of imputation, namely, hot deck, linear regression and log linear, with and without auxiliary information are considered. The log linear methods consist of adding categorical constraints to either the hot deck or linear regression methods. Based on an extensive simulation study with synthetic data, sensitivity analyses for departures from CIA are performed and gains from using auxiliary information are discussed. Different scenarios for the underlying distribution and relationships are created using synthetic data such as normal *versus* nonnormal data and proxy *versus* nonproxy auxiliary data. Some recommendations on the use of statistical matching methods are also made. Specifically, it was confirmed that CIA could be a serious limitation which could be overcome by the use of 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 CIA used in the construction of the Social Policy Simulation Database at Statistics Canada.

Key Words: Macro/Micro and Proxy/Nonproxy Auxiliary Information, Categorical Constraints, CIA, Conditional Correlation, Log-normal Contaminations, Regression to the Mean, Unit *vs* Aggregate Evaluation Measures

RESUME

Lors de la création des bases de données de micro-simulation, lesquelles sont fréquemment utilisées par les concepteurs et analystes de politiques, certains fichiers de données sont combinés par des techniques statistiques d'appariement afin d'enricher le fichier de données principal. Ce procédé requiert l'hypothèse d'indépendance conditionnelle (HIC) qui pourrait sérieusement biaiser les relations entre les variables résultant de l'appariement. L'utilisation de l'information auxiliaire appropriée pourrait simplifier grandement ce problème. Dans ce rapport, des méthodes statistiques d'appariement correspondant à trois méthodes d'imputation sont considérées, "hot deck", régression linéaire et log linéaire, avec ou sans information auxiliaire. La méthode log linéaire consiste en l'addition de contraintes de catégories à la méthode "hot deck" ou à la méthode de régression linéaire. Basées sur une vaste étude de simulation avec des données artificielles, des analyses de sensibilité aux écarts à la HIC sont effectuées, et les gains obtenus par l'utilisation de l'information auxiliaire sont discutés. Différents scénarios sont créés pour les relations et la distribution sous-jacente en utilisant les données artificielles telles que les données normales *versus* non-normales et les données auxiliaires "proxy" *versus* "non-proxy". Quelques recommandations sur l'utilisation des méthodes statistiques d'appariement sont aussi faites. Plus spécifiquement, il a été confirmé que la HIC pourrait être une sérieuse limitation qui peut être contrée par l'utilisation d'informations auxiliaires appropriées. Les méthodes "hot deck" se sont montrées généralement préférables aux méthodes de régression. Aussi, lorsque l'information auxiliaire est disponible, les contraintes de catégories de la méthode log linéaire peuvent améliorer les performances des méthodes "hot deck".

Cette étude a été motivée par l'intérêt porté à la HIC utilisée lors de la construction de la base de données de simulation de politique sociale à Statistique Canada.

TABLE OF CONTENTS

1. INTRODUCTION

The literature on statistical matching is spread over the last two decades, starting probably with the work of 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 strong conditional independence assumption (CIA). 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.

The present paper considers several methods based on auxiliary information to avoid the CIA which are developed from the original ideas of Rubin (1986) and Paass (1986). Rubin proposed versions of parametric regression while Paass proposed versions of nonparametric regression which are related to the familiar hot deck method of imputation. A simplified version of Paass's method is considered in this paper due to the considerable computational effort required for the original method. Another class of methods using auxiliary information based on log linear imputation (Singh, 1988) along with some modifications of the above methods is also considered. An extensive simulation study with synthetic data, which also included matching methods that assume CIA, was conducted in order to analyze sensitivity to failure of the CIA and gains from using auxiliary information.

There have been several empirical investigations in the past on evaluating 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. A recent reference is Barr and Turner (1990), which describes a detailed empirical investigation of quality issues for file merging, and also contains a good list of references.

All 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 it's impact on 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 quite broad.

This research was motivated from considerations of improving the content of the Social Policy Simulation Database (SPSD), a microsimulation database created at Statistics Canada by merging various files for use in economic policy analysis; see Wolfson, Gribble, Bordt, Murphy and Rowe (1987). Some preliminary results from this study were presented to the Statistical Society of Canada, cf. Singh, Mantel, Kinack and Rowe (1990).

The organization of this paper is shown in the table of contents. More specifically, section 2 introduces the problem of statistical matching as it arises in the context of the SPSD. Included is a brief overview of the problem of statistical matching in general. A description of some methods of statistical matching is presented, all of which are limited by the CIA. The section concludes with a short discussion on the method of statistical matching used in SPSD construction. Section 3 contains some proposals for modifications to the methods of statistical matching described in section 2. These modified methods are aimed at avoiding the CIA and require the existence of auxiliary information on the joint relationships of variables in the different source datafiles. Section 4 provides the details of an extensive simulation study with synthetic data designed to evaluate and compare various strategies of statistical matching. Included is a description of the different versions of the methods examined and the evaluation measures used. Section 5 contains the results from the Monte Carlo trials. Limitations of the Monte Carlo study are also mentioned. Section 6 presents a summary of the report, highlighting the major findings. Some practical recommendations and directions for future work are included.

2. THE SPSD AND STATISTICAL MATCHING

2.1 SPSD Background

In the words of Wolfson, Gribble, Bordt, Murphy and Rowe (1987, p202 top) , "The Social Policy Simulation Database with its related Social Policy Simulation Model software (SPSD/M) has as its general goal to provide a comprehensive, publicly available, microsimulation-based, integrated individual tax/transfer policy analysis capability." The typical uses to which SPSD are put include calculations of taxes and transfers for families on the database, tabulations and cross tabulations.

The multi-stage 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 Finances (SCF), with additional information from other data sources. Specifically, information from unemployment insurance claim histories (UI), personal income tax returns (T1) and the Family Expenditure Survey (FAMEX) is added to the SCF records.

Since the UI and T1 data come from specially drawn samples from the complete administrative files, and the FAMEX data come from a survey, it is not necessarily the case that an individual will appear in more than one of the data sources. Hence, the process of adding this extra information to the SCF records is unlike exact matching in which one would search through these other data sources for specific individuals. In fact, even if the three additional data sources were complete, confidentiality concerns would prevent an exact matching of the files. Thus, the alternative of statistical matching of records from the various files is used.

2.2 Description of Statistical Matching

Statistical matching can be viewed as a special case of imputation in which we have two or more distinct 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 similar records from the other donor file(s). A typical use for the matched file is as input to micro-simulation models for which a complete file with all variables is required. In the case of the SPSD, the SCF serves as the host file with the UI, T1 and FAMEX data corresponding to donor files. Since the statistical matching of these three files to the SCF is carried out sequentially, for our purpose here it is sufficient to restrict

the discussion that follows to the general case of matching two files, a host file A and a donor file B.

Both files A and B will normally contain information on vectors of variables. We assume the existence of some set of common variables $\underline{X}$ in the two files that can be used to identify similar units. In the case of the SPSD these are usually demographic or income variables. The remaining variables unique to file A are designated as $\underline{Y}$, while those unique to file B are designated as $\underline{Z}$. The problem is to complete the records in file A by imputing values for $\underline{Z}$, using the information on the $(\underline{X}',\underline{Z}')'$ relationships in file B.

A more detailed discussion on statistical matching can be found in U.S. Department of Commerce (1980); see also Kadane (1978), Rodgers (1984), Rubin (1986) and Paass (1986).

Note: Although this setting describes the most general framework for the problem of statistical matching, we will restrict our attention here to the case of univariate values for $\underline{X}$, $\underline{Y}$ and $\underline{Z}$; that is, one X, one Y and one Z variable. This is partly for reasons of simplicity, but also because of computational limitations involved with the simulation study undertaken. A restriction to continuous variables only is also imposed at the present time. It is important to note that in addition to computational concerns, some concepts (such as ranking of records) do not have natural analogues in the multivariate and/or discrete setting.

2.3 Common Methods of Statistical Matching

Two commonly used classes of methods of statistical matching are analogous to linear regression imputation and hot deck imputation (see Kalton and Kasprzyk, 1986). The basic idea of all of these methods is to use information on the common variable X to find a similar record in file B for each record in file A.

2.3.1 Linear Regression Method

The linear regression method is a two-stage procedure. File B is used to fit a linear regression of Z on X, with the estimated regression coefficients used to obtain an intermediate value for Z based on X for each record in file A. A live value for Z is then obtained from file B by selecting the closest record according to a distance measure, such as the distance in Z or the Euclidean distance in (X,Z).

This last step is analogous to the addition of stochastic residuals to the predicted values in regression imputation. The version using the distance in Z is also known as "regression with predictive mean matching", as proposed by Rubin; see Little and Rubin (1987).

2.3.2 Hot Deck Methods

The term hot deck comes from a class of methods of imputation for item non-response in which "the value assigned for a missing response is taken from a respondent to the current survey" (Kalton and Kasprzyk, 1986), though it should be noted that Ford (1983) says "there is no general agreement on the exact definition of a hot-deck procedure".

The term hot deck is used here to describe methods of matching which obtain live values based on comparison of files at the unit level. In this sense, the second stage of the regression procedures described above, in which a live value is obtained from file B based on some distance measure is also hot deck. However, by hot deck matching methods we mean methods that do not make use of any synthetic intermediate values. Hot deck methods may be considered as non-parametric analogues of regression methods.

Three types of hot deck methods are considered: random, distance and rank. As well, each of these methods can be applied globally to all records at one time, or for computational or other reasons, independently to records within categories of the X variable, denoted by X^*.

Hot deck random corresponds to matching each record in file A with a randomly selected record from file B.

Hot deck distance requires the calculation of a distance measure between records in files A and B. Each record in file A is matched to the closest record in file B.

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 these ranks. This was proposed by G. Rowe for the SPSD (see Wolfson, Gribble, Bordt, Murphy and Rowe, 1987). A complication arises with hot deck rank when the number of records in the two files differs. One way to circumvent this problem is by matching percentage points of the empirical cumulative distribution functions of X for host and donor records.

2.4 Log Linear Method

A modification of the linear regression and hot deck methods is to impose categorical constraints on the records selected from file B in the process of completing file A, as suggested by Singh (1988). This new type of constrained statistical matching is a special case of log linear imputation. The constraints imposed are in the form of a categorical distribution of (X^*,Y^*,Z^*) that the completed file A must satisfy. Here, X^*, Y^* and Z^* denote respectively the categorical transformations of X, Y and Z. Note that this represents a difference from other types of constrained statistical matching in which constraints are in the form of a few characteristic measures (such as mean and variance) that variables in the completed file must satisfy. The categorical constraints for log linear imputation can be obtained in a number of ways through log linear modelling; see Singh, Armstrong and Lemaitre (1988).

The basic idea here is, within (X^*,Y^*) categories in file A, to distribute counts to Z^* categories according to the categorical (X^*,Z^*) distribution in file B. There are two approaches here which do not require fitting log linear models. Both require that the (X^*,Z^*) categorical distribution from file B is first raked to the X^* margin from file A. A simple example demonstrating the difference between the two approaches can be found in appendix 3.

The first procedure, which we call rakeyz, is to rake a 2-dimensional table of 1's to the categorical Y^* and Z^* margins from files A and B to obtain a categorical (Y^*,Z^*) marginal distribution. This amounts to assuming Y^* and Z^* are unconditionally independent. A 3-dimensional table of 1's is then raked to the categorical (X^*,Y^*), (X^*,Z^*) and (Y^*,Z^*) distributions. This amounts to setting the (Y^*,Z^*) interaction terms in a log linear model for the categorical (Y^*,Z^*) distribution to zero and the (X^*,Y^*,Z^*) interaction terms in a log linear model for the categorical (X^*,Y^*,Z^*) distribution to zero (see Purcell and Kish, 1980).

The second procedure, which we call rakexyz, is to rake a 3-dimensional table of 1's to the categorical (X^*,Y^*) and (X^*,Z^*) distributions from files A and B. This amounts to assuming that, conditional on the X^* category, Y^* and Z^* are categorically independent, which does not imply Y^* and Z^* are unconditionally independent. In terms of log linear models, it is equivalent to setting the (Y^*,Z^*) and (X^*,Y^*,Z^*) interaction terms in a log linear model for the categorical (X^*,Y^*,Z^*) distribution to zero.

Which procedure is most appropriate would depend on which assumptions are nearest to being correct.

Under the categorical constraints either the linear regression method or a hot deck method can be used to impute live Z-values from file B onto records in file A. For example, the hot deck distance method would be modified in the following manner. Within an (X^*,Y^*) category in file A one would compute the distances to all records in the same X^* category in file B. The first records to be matched would correspond to the pair with minimum distance in X. The (X^*,Y^*,Z^*) category of the completed record would be noted and a running count of the number of matched records in that (X^*,Y^*,Z^*) category incremented. If the resulting count 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 so that the categorical distribution of (X^*,Y^*,Z^*) satisfies the categorical constraints.

2.5 An Important Limitation with these Methods

The methods of statistical matching described above all suffer from a similar limitation in that information on the variable Y is completely ignored in the matching process. This limitation amounts to the assumption of conditional independence of Y and Z given X ($Y \perp Z | X$), denoted CIA (conditional independence assumption). Note that when the categorical constraints of log linear imputation are imposed this assumption is made for the categorical distribution ($Y^* \perp Z^* | X^*$), and within (X^*,Y^*,Z^*) categories Y and Z are assumed to be independent given X. 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 seriousness of the CIA is well-documented in Sims (1978), Rubin (1986), Paass (1986) and Armstrong (1989). In section 3 we will see how the existence of additional auxiliary information can be used to help avoid this serious assumption.

2.6 The Method of Statistical Matching in SPSD Construction

One important restriction placed on the statistical matching of records in the SPSD is the requirement that all records from file B be used, since a primary purpose of the match is to exploit as much information as possible that exists in file B. This differs somewhat from the general objective of statistical matching which is focused primarily on completing the records in file A for the missing variables, and does not necessarily mean that all records from file B have to be used.

With this requirement in mind, the method of statistical matching applied in SPSD construction is a slightly modified version of hot deck rank. Records from both files A and B are first classified into specified "bins" ($\underline{X}^*$ categories) and then ranked separately on one of the common continuous $\underline{X}$ variables (usually total income). Records are selectively duplicated to overcome the problem of different numbers within bins so that corresponding bins will have the same number of records from each of the two files. Within each bin, records are matched one-to-one across the two files.

Currently, categorical constraints are not imposed in the version of hot deck rank used in SPSD construction.

Note: One other important consideration with the SPSD is not addressed at the present time. All records on the SPSD have record weights associated with them, since individuals on the contributing datafiles come from surveys or samples of administrative files with different probabilities of selection. The existence of these weights can complicate the matching process. The methods considered in this simulation study would need suitable modifications in order to handle the case of record weights.

3. STATISTICAL MATCHING WITH AUXILIARY INFORMATION

3.1 Existence of Auxiliary Information

The serious CIA described in section 2.5 can severely affect the meaningfulness of analyses conducted with regards to joint relationships of variables that come from different source files. When additional auxiliary information is available on these joint relationships it can be incorporated into the matching process to avoid the CIA and improve the quality of the completed file by reducing distortions in these joint relationships.

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, it's use is limited to the extent that information that exists in the source files is never overridden 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 exact match of confidential datafiles.

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.

3.2 Types of Auxiliary Information

Since the information missing from the source files pertains to the joint relationships of (Y,Z) and (X,Y,Z), it is natural to consider two general classes of auxiliary information: one with information solely on the (Y,Z) joint relationships, and one with information on the full set of (X,Y,Z) joint relationships.

Within each of these two classes of auxiliary information, one can group the forms the auxiliary

information can take into two levels: macro-level auxiliary information in the form of summary statistics or measures, and micro-level auxiliary information in the form of individual unit records.

3.2.1 Macro-level Auxiliary Information

Auxiliary information at the macro-level could be in the form of either correlations or categorical cell proportions. Given that the auxiliary information could be either on the complete set (X,Y,Z) or simply on (Y,Z), four kinds of macro-level auxiliary information are possible. These are:

(X,Y,Z) Auxiliary Information

i) $\rho_{Y,Z|X}$ conditional correlation of Y,Z given X

ii) π_{ijk} 3-dimensional categorical cell proportions

(Y,Z) Auxiliary Information

iii) $\rho_{Y,Z}$ unconditional correlation of Y,Z

iv) π_{jk} 2-dimensional categorical cell proportions

One might also have macro-level auxiliary information of more than one kind, for example both i) and ii), or both iii) and iv).

3.2.2 Micro-level Auxiliary Information

Auxiliary information at the micro-level would be in the form of a third datafile C with individual unit records containing information on either the complete set of variables (X,Y,Z) or on the reduced set (Y,Z). However, this file could be outdated, proxy or confidential so that it could not be used directly.

3.3 Using Auxiliary Information

Clearly, the method of incorporating auxiliary information with statistical matching will depend on the level it is at, and the specific form it has.

3.3.1 Macro-level Auxiliary Information

The existence of information on correlations means that the linear regression method can be modified so that a regression of Z on X and Y can be used, instead of simply a regression of Z on X, in obtaining the intermediate imputed value for Z.

Information on the cell proportions can be incorporated into the procedure of constructing the categorical constraints of log linear imputation. In the case of (X,Y,Z) auxiliary information, the set of categorical proportions can be used as the starting table in place of the table of 1's in the procedures described in section 2.4. Thus there are two possibilities.

The first possibility, which corresponds to rakeyz in section 2.4, is to collapse the (X^*,Y^*,Z^*) categorical distribution to a (Y^*,Z^*) categorical distribution and rake it to categorical Y^* and Z^* margins as in the previous case, and then the (X^*,Y^*,Z^*) categorical distribution would be raked to the categorical (X^*,Y^*), (X^*,Z^*), and (Y^*,Z^*) margins. This amounts to borrowing the (Y^*,Z^*) interaction terms of a saturated log linear model for the categorical (Y^*,Z^*) distribution from the auxiliary categorical (Y^*,Z^*) distribution and the (X^*,Y^*,Z^*) interaction terms of a saturated log linear model for the categorical (X^*,Y^*,Z^*) distribution from the auxiliary categorical (X^*,Y^*,Z^*) distribution.

The second possibility, which corresponds to rakexyz in section 2.4, is to rake the (X^*,Y^*,Z^*) categorical distribution to the (X^*,Y^*) and (X^*,Z^*) categorical margins. This amounts to borrowing both the (Y^*,Z^*) and (X^*,Y^*,Z^*) interaction terms of a saturated log linear model for the categorical (X^*,Y^*,Z^*) distribution from the auxiliary categorical (X^*,Y^*,Z^*) distribution. Intuitively, the second procedure would be preferred if the information about the (X^*,Y^*,Z^*) categorical distribution was very good; the first procedure would be expected to perform better if the auxiliary information was less precise because in that case the information about the (Y^*,Z^*) categorical distribution would be more precise.

Information on the 2-dimensional cell proportions for the (Y^*,Z^*) margin can be used in place of the 2-dimensional table of 1's in the first raking procedure described in section 2.4. In terms of a saturated log linear model for the categorical (Y^*,Z^*) distribution, this amounts to taking the (Y^*,Z^*) interaction terms from the auxiliary data while leaving the marginal terms obtained from files A and B intact. Next, a 3-dimensional table of 1's is raked to the categorical (X^*,Y^*), (X^*,Z^*), and (Y^*,Z^*) margins which, in a saturated log linear model, amounts to setting the second order interaction terms to zero.

In the case of macro-level auxiliary information on both correlations and categorical cell proportions, it would be possible to use a modified linear regression method and also apply modified categorical constraints.

3.3.2 Micro-level Auxiliary Information

Similarly to the linear regression method, the approach with micro-level auxiliary information is to use the data in file C to impute an intermediate value for Z, and then use this intermediate value to help find a live record from file B.

For example, based on methods developed by Paass (1986), a modification of hot deck distance is obtained as a two-step procedure. In the first step, hot deck distance matching is applied to files A and C to add intermediate imputed values for Z onto file A. In the second step, hot deck distance matching is applied to the new file created in the first step and file B to obtain final values for Z, and hence the completed file A. The difference between this approach and usual hot deck distance is that the variable Z can be used in the distance function when searching for live values from file B.

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. In such cases, the options of using the modified linear regression method and of applying modified categorical constraints would also both be present. The term modified refers to the versions of the methods that use macro-level auxiliary information.

3.4 Comparing Methods of Statistical Matching

Given the large number of considerations that may be involved in selecting a method of statistical matching, several questions arise as to how the various possible methods compare and how one should select a method appropriate for a particular application. Listed below are some questions which come to mind in considering the different factors involved:

- for methods that use no auxiliary information
  - a) how serious is the CIA ?

b) does imposing categorical constraints improve the situation ?

c) is there a best method ?

- to what extent does macro-level auxiliary information on correlations improve the linear regression method ?

- to what extent does micro-level auxiliary information improve the hot deck methods ?

- when macro-level auxiliary information on categorical cell proportions is available
  a) does imposing categorical constraints improve the performance of the methods ?
  b) how does the choice of partitioning points and associated degree of fineness affect the performance ?

- how do the methods using auxiliary information perform when the auxiliary information is from an outdated or proxy source, or from an insufficiently large auxiliary source ?

- how robust are the linear regression methods to non-normality ?

- how do computational requirements balance against performance for the various methods ?

- for given auxiliary information, is there a best method ?

An empirical evaluation through an extensive simulation study with synthetic data generated from multivariate normal distributions was undertaken to examine these questions. Non-normality was introduced via contaminations from 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 vary in a systematic departure from conditional independence. It also permits comparisons of the methods as the underlying distribution generating the data moves away from normality. The design of this study and subsequent results constitute the content of the following two sections.

4. A MONTE CARLO STUDY OF STATISTICAL MATCHING METHODS

This section presents the details of the Monte Carlo study undertaken. Programming was done on micro-computers using the software GAUSS. Execution time for one simulation varied from 20 minutes to one hour, depending on the machine used. Documentation of the programs written can be found in Mantel and Kinack (1991).

4.1 Design of the Monte Carlo Study

4.1.1 Strategy

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 suppressed. The suppressed Z-values in file A are used to evaluate the performance of various methods of statistical matching that use the information in files B and C to impute live Z-values from file B onto the records in file A.

Each method of statistical matching investigated is applied to the same data. That is, for a given file A and a given file B each method of statistical matching is performed. When a particular method requires auxiliary information it is always taken from the same file C. This could be the micro-level file itself, or summary measures or statistics calculated from the data in file C, such as correlations or categorical cell proportions.

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.

4.1.2 The Data

For this study both files A and B are always generated from the same underlying distribution,

with each containing 500 independent and identically distributed observations. File C contains 250 independent and identically distributed observations from an underlying distribution that may or may not be the same as that for files A and B. A different underlying distribution for file C would be used to represent proxy auxiliary information.

The basic distribution of observations (X,Y,Z) is multivariate normal with the marginal distributions of X, Y and Z being standard normal. The covariances of (X,Y) and (X,Z) in the basic distribution are always .5, while the covariance of (Y,Z) varies from one run to another. Correspondingly, the conditional correlation of Y and Z given X, with the formula given by $\rho_{Y,Z|X} = (\rho_{Y,Z} - \rho_{X,Y}\rho_{X,Z}) / (1 - \rho_{X,Y}^2)^{1/2} (1 - \rho_{X,Z}^2)^{1/2}$, also varies. In some runs the basic distribution is contaminated by taking the exponentials of X, Y and Z instead of X, Y and Z for some of the observations. The probability of any particular observation coming from this log-normal contamination distribution is fixed for any particular run of 100 simulations and individual observations are chosen independently to be contaminations.

4.1.3 Proxy Auxiliary Information

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; that is, auxiliary information that comes from a distribution different from the distribution generating observations in files A and B. In the first series the auxiliary data has a different $\rho_{Y,Z|X}$. In the second series the auxiliary data has some log-normal contamination.

4.1.4 Categorical Partition

For the methods applying categorical constraints (to be fully described in section 4.2), it is necessary to choose a categorical partition. For this study 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 centered on the mean of the basic marginal distribution with break points at the center 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 basic marginal distributions; that is, the partition had the categories $<-.6745$, $[-.6745,0)$, $[0,.6745)$, $\geq .6745$. Note that these

partitions are defined in terms of the basic distributions, for simplicity the same partitions are used when there are log-normal contaminations. It would have been most realistic to let the partitions be data dependent, but we used fixed partitions for simplicity.

Only one of these partitions may be used in any particular run of 100 simulations for matching, but either or both may be used for the categorical evaluation measures to be described in section 4.3, namely, chi-square statistics and conditional likelihood ratio tests.

4.1.5 Summary of Control Parameters

In summary, the parameters that will be varied for different runs are:

1) the conditional correlation of Y and Z given X ($\rho_{Y,Z|X}$); recall that $\rho_{X,Y}$ and $\rho_{X,Z}$ are fixed at .5

2) the proportion of log-normal contaminations for generating the underlying non-normal populations

3) the auxiliary data may come from a distribution different from that of the data in files A and B through:
   a) a different conditional correlation of Y and Z given X ($\rho_{Y,Z|X}$)
   b) log-normal contaminations

4) the categorical partitions for matching methods with categorical constraints and for categorical evaluation measures may vary

4.2 The Matching Methods

We first describe in detail the specific matching methods included in the study and a naming convention that will be used to refer to them.

There are four main types of methods included in the study. These are regression methods and hot deck methods, both with and without categorical constraints.

4.2.1 Regression Methods

Regression methods, denoted by REG, use an estimated regression relationship to impute an intermediate Z-value to records in file A, and perhaps also an intermediate Y-value to records in file B. A live value from file B is then obtained on the basis of z-distance, xz-distance, or xyz-distance. File B may be used to estimate a linear regression of Z on X. If there is an auxiliary file C with all three variables available then it may be used to estimate the conditional correlation of Y and Z given X and this, together with (X,Y) and (X,Z) correlations from files A and B, may be used to estimate a linear regression of Z on X and Y. If file C contains only variables Y and Z we would take the unconditional (Y,Z) correlation from it. Note that we are only taking certain correlations from file C; the full micro data is not needed by these methods.

4.2.2 Hot Deck Methods

In the absence of any auxiliary datafile hot deck methods, denoted by HOD, are based on comparisons of the X-values in files A and B. Within X^* categories the matching may be done randomly or be based on the ranked X-values or be based on x-distance. Matching based on x-distance is also done without X^* categories.

When there is an auxiliary micro-datafile C available we would use hot deck distance to obtain intermediate Z-values for file A and Y-values for file B from file C and then use hot deck z, xz, or xyz-distance to obtain live Z-values from file B for records in file A. If the auxiliary micro-datafile C contains variables X, Y and Z then xy-distance would be used to obtain the intermediate Z-values; if it contained only variables Y and Z then y-distance would be used. This method may be considered as a simplified version of the method proposed by Paass (1986). The original method of Paass (1986) was not included in this study because it was thought to be too computationally intensive.

4.2.3 Log Linear Categorical Constraints

The categorical constraints mentioned above are obtained by raking a categorical distribution to the categorical (X^*,Y^*) and (X^*,Z^*) margins from files A and B respectively. There are four different ways in which this is done in the present study. The two raking procedures that do not use auxiliary information are described in section 2.4. We also use the raking procedure described in section 3.3.1

for (Y^*,Z^*) auxiliary information and the first of the two raking procedures described there for (X^*,Y^*,Z^*) auxiliary information. We use the first of the procedures for (X^*,Y^*,Z^*) auxiliary information because in this simulation study the auxiliary categorical (X^*,Y^*,Z^*) distribution may not be sufficiently accurate to be acceptable.

The categorical counts obtained by these procedures need not be integer values. We force them to be integer values 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.

Because of the close connections to log linear models we use the word LOGLIN to refer to these procedures of finding categorical constraints. After the constraints are obtained it is still necessary to do the matching respecting the constraints. We then obtain REG.LOGLIN and HOD.LOGLIN as the last two major types of methods included in the study.

4.2.4 Regression with Categorical Constraints

For REG.LOGLIN intermediate Z-values for file A and Y-values for file B are obtained in the same way as for the REG procedures. However, when live values from file B are obtained on the basis of z, xz, or xyz-distance, the categorical constraints are respected. This is done by finding for each remaining record of file A the closest record from file B with a Z-value in one of the Z^* categories that is not yet filled, then choosing the match that is the closest, and then repeating the procedure until all of the records from file A have a live Z-value from file B.

4.2.5 Hot Deck with Categorical Constraints

For HOD.LOGLIN with an auxiliary micro-datafile C the intermediate Z-values for file A and Y-values for file B are obtained subject to categorical constraints. The constraints for intermediate Y-values for file B need not be the same as those obtained for the Z-values for file A; they are obtained in the same way, but with files A and B playing reversed roles. The procedure for finding the intermediate Z-values from file C subject to the constraints is the same as the procedure described above for finding live Z-values from file B in the REG.LOGLIN procedures, but using xy or y-distance depending on whether file C contains variables (X,Y,Z) or (Y,Z). Once the intermediate values are

obtained, live values are found from file B on the basis of z, xz, or xyz-distance within Z^* categories.

HOD.LOGLIN with no auxiliary micro-datafile and x-distance is very similar to the final stage of the REG.LOGLIN procedures with an auxiliary micro-datafile C. If random matching is used it is done within (X^*,Z^*) categories from file B to satisfy the constraints. Matching based on ranking satisfying categorical constraints is not as straightforward. For each (X^*,Y^*) category, records from file B are randomly selected respecting the X^* category to satisfy the constraints and then the records in file A from that (X^*,Y^*) category are matched to these randomly selected records from file B based on X ranking.

4.2.6 A Naming Convention

A complete list of naming conventions is given in appendix 1. The major indicators of matching type are REG, HOD, REG.LOGLIN, and HOD.LOGLIN. More detailed descriptions of the various methods within these major types would be indicated by modifiers within parentheses. Thus, for example, REG(auxcorrxyz, xz-dist) indicates a regression procedure with auxiliary information about the conditional correlation of Y and Z given X to obtain the regression relationship used to impute intermediate Z-values to file A, and with xz-distance being used to obtain live Z-values from file B.

4.3 The Evaluation Measures

A variety of 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. Categorical evaluation measures, which look at the overall distribution, are relevant to the typical uses of the SPSD. 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 procedure 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 distributional properties of Z.

The various evaluation measures were calculated for each of the 100 simulations run for each

set of parameter values considered. What is reported are a variety of summary statistics for the evaluation measures.

4.3.1 Mean Absolute Differences of Z (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 means of these means as well as standard errors were obtained.

4.3.2 Difference of Covariances (MAD-Cov)

The second evaluation measure is the absolute difference of the conditional covariances of Y and Z given X in the matched and suppressed files. These conditional covariances are calculated from the observed covariance matrices as if the data were multivariate normally distributed. Monte Carlo means of these absolute differences as well as standard errors were obtained.

4.3.3 Chi-square Statistics (χ^2)

The third measure of performance, based on categorical comparisons, is simply a Pearson chi-square statistic to test that the suppressed categorical Z-values come from a multinomial distribution with probabilities equal to the categorical proportions from the matched file. What is reported are the mean chi-square statistics over the 100 simulations, transformed to lie in the interval (0,1) (see appendix 2).

4.3.4 Conditional Likelihood Ratio Test (CLRT)

The final measure of performance is also based on categorical comparisons. Within each (X^*, Y^*) category that has a minimum number of observations, in this case 20, a likelihood ratio test that the categorical Z-values from the matched and suppressed files come from the same multinomial distribution is performed. The tests for different (X^*, Y^*) categories are then combined to obtain an overall p-value (see appendix 2). What is reported is the proportion of times, out of 100 simulations

at each set of parameter values, 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.

Note that the minimum sample size of 20 for (X^*,Y^*) categories in file A is 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.

5. RESULTS OF THE MONTE CARLO STUDY

In this section we describe the results of the simulation study. The plots referred to throughout the section can be found at the end of the paper.

We first consider in section 5.1 methods that depend on CIA, that is methods that do not use auxiliary information. The effects of violation of CIA as measured by the conditional correlation of Y and Z given X ($\rho_{Y,Z|X}$) are investigated. We also investigate the effect of non-normality of the data via log-normal contamination.

We next look in section 5.2 at REG methods that attempt to avoid CIA by using auxiliary information on correlations. Their performance under non-zero conditional correlations ($\rho_{Y,Z|X}$), under non-normality of the data and under proxy auxiliary information will be investigated. In section 5.3 we consider LOGLIN methods using categorical auxiliary information to impose categorical constraints on HOD and REG methods that use no auxiliary information. In section 5.4 we look at combining categorical auxiliary information with auxiliary information about the correlations. Finally, in section 5.5 we consider the use of an auxiliary micro-datafile which allows for HOD methods with auxiliary information, possibly with categorical constraints derived from that auxiliary information.

We have not paid much attention to Monte Carlo standard errors in the presentation. This is because they were generally quite small, for example, less than two percent for the mean absolute differences of covariances (MAD-Cov). Furthermore, the evaluations of different methods for each run 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 in the plots. In short, any discernible difference in the plots is likely to indicate a real difference.

In studying the evaluation measures via plots we have tended to look at rank versions for hot deck methods (HOD) with no auxiliary micro-data. This is because ranking is an approach that is currently very much used in SPSD and, as demonstrated by the series C plots, it generally doesn't make much difference in this study whether rank, random or x-distance is used. Similarly, we have generally used xz-distance for HOD methods with auxiliary micro-data and for regression (REG) methods. There is generally no difference, though for REG methods xz-distance sometimes showed superior performance with respect to the MAD-Cov evaluation measure, see plot D.2.

5.1 Methods that do not use Auxiliary Information

To investigate the performances of various matching methods that do not use auxiliary information we looked at a series of runs in which the true conditional correlation of Y and Z given X varied from 0 to .8. In a second series $\rho_{Y,Z|X}$ was held fixed at .4 while the proportion of log-normal contaminations was varied to investigate the effect of non-normality. The partition used in the LOGLIN procedures was the equal probability partition, but both equal probability and standard interval partitions were used for categorical evaluation measures.

We consider each performance measure individually. The results of this exercise will be summarized at the end.

5.1.1 MAD-Z under no Auxiliary Information

The MAD of Z-values for methods using no auxiliary information are shown in plots A.1 and B.1.

It can be seen immediately that the methods fall into two groups, REG and HOD methods. REG methods do somewhat better with respect to this measure. There is little or no dependence on the true conditional correlation of Y and Z given X ($\rho_{Y,Z|X}$), showing that with respect to this measure CIA is not serious. All methods show slight degradation of performance as the distribution becomes non-normal, but the basic relationships among methods remain the same.

5.1.2 MAD-Cov under no Auxiliary Information

The MAD of conditional covariances for methods using no auxiliary information are displayed in plots A.2 and B.2.

With respect to this measure the failure of CIA seems to have serious consequences. The effect of non-normality seems to be somewhat stronger too.

The performances of REG, REG.LOGLIN(rakexyz), HOD(xcat), and HOD.LOGLIN(rakexyz) are very close to each other. An interesting finding was that the two LOGLIN methods using the other type

of raking, namely rakeyz, do more poorly than the rest. This relatively poor performance of methods using rakeyz could perhaps be explained by the particular parameter values chosen. For rakeyz it may be argued that it is departures of the unconditional correlation of Y and Z from zero which are important. In the present case, for example, when $\rho_{Y,Z|X} = .4$, $\rho_{Y,Z} = .55$.

5.1.3 Transformed χ^2 under no Auxiliary Information

See plots A.3, A.4, B.3, and B.4.

With respect to this measure the REG method does not perform well. The transformed χ^2 mean is constant around a high of 0.95 even for low values of $\rho_{Y,Z|X}$. Its poor performance can be explained by shrinkage towards the mean, which means that the matched Z-values are more tightly distributed about their mean than are the suppressed Z-values. This is displayed in plot I.1 which shows the difference between the marginal histograms of suppressed and matched Z-values for REG(xz-dist) and REG.LOGLIN(xz-dist,rakexyz). The positive differences near the centre of the plot indicate that there are more Z-values in that region on 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 the REG method is putting too many Z-values at the centre of the distribution at the expense of the extreme tails. We are not aware of any previously published findings noting this regression to the mean effect in the context of statistical matching.

HOD methods do rather better. They display some dependence on $\rho_{Y,Z|X}$, but seem to be robust against non-normality. Although we did not show it in plot I.1, the HOD method also displays some shrinkage towards the mean, but it is much less serious than it is for the REG method, as suggested by the relatively good performance of HOD with respect to the categorical evaluation measures.

REG.LOGLIN does much better than REG and even slightly better than HOD when the same partition is used for constraints and for testing, but the improvement practically disappears when the testing partition is different. This can again be explained by shrinkage towards the mean as displayed in plot I.1. However, in this case the shrinkage towards the mean 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.

HOD.LOGLIN(rakexyz) is slightly better than HOD when the same partitions are used for matching and testing but no better when the testing partition is different. As with MAD-Cov, HOD.LOGLIN(rakeyz) does not perform as well as HOD or HOD.LOGLIN(rakexyz).

5.1.4 Conditional Likelihood Ratio Test under no Auxiliary Information

See plots A.5, A.6, B.5 and B.6.

Again, the REG method does not do very well with respect to this measure.

REG.LOGLIN(rakexyz) does well when the same partition is used for matching and testing but otherwise does no better.

HOD methods do rather better, though they display a strong dependence on $\rho_{Y,Z|X}$. They seem to do well until $\rho_{Y,Z|X}$ reaches about .3 at which point they deteriorate very quickly. However, they seem to show some robustness against non-normality, see plots B.5 and B.6. Although it is not shown in the plots, this is the only measure that shows a distinction among hot deck random, rank and x-distance, with rank doing slightly better than random and x-distance. The series C plots display how hot deck random, rank and x-distance generally perform similarly.

HOD.LOGLIN(rakexyz) shows no consistent improvement or degradation over HOD. HOD.LOGLIN(rakeyz) shows some degradation as compared to HOD. The slight advantage of rank over random and x-distance for HOD methods seems to disappear for HOD.LOGLIN.

5.1.5 Summary of Results for Methods with no Auxiliary Information

The best methods overall seem to be the HOD methods. Though REG methods do best with respect to the mean absolute differences of Z-values, their relatively poor performance with respect to the categorical evaluation measures, χ^2 and conditional likelihood ratio tests, should be kept in mind. Within the class of HOD methods there is little to choose between random, rank and x-distance, leaving other concerns such as computational requirements as important factors in deciding between methods for a particular application.

Based on this limited study, we may speculate that for HOD.LOGLIN procedures the choice of raking procedure is important and the best one to use depends on the underlying distributions of the data. In any case, HOD.LOGLIN procedures in the case of no auxiliary information never gave dramatic improvement over simple HOD procedures.

5.2 Methods that use Auxiliary Information on Correlations

5.2.1 REG Methods with Auxiliary Information on $\rho_{Y,Z|X}$

5.2.1.1 MAD-Z for REG with Auxiliary Information on $\rho_{Y,Z|X}$

See plots E.1, F.1, G.1 and H.1.

REG methods generally do best with respect to this measure, and REG(auxcorrxyz) does better than REG with no auxiliary information. As the true conditional correlation of Y and Z given X increases this measure decreases which can be attributed to the greater explanatory power of (X,Y) for Z.

This measure displays slight deterioration as the data move away from normal, but seems to be robust against the use of proxy auxiliary information.

5.2.1.2 MAD-Cov for REG with Auxiliary Information on $\rho_{Y,Z|X}$

See plots E.2, F.2, G.2 and H.2.

REG methods also tend to do well with respect to this measure, and the use of auxcorrxyz

offers good protection against the breakdown of CIA (compare plots A.2 and E.2). However, this is the one measure that showed a distinction among z-dist, xz-dist and xyz-dist with xz-dist doing well but the other two doing only slightly worse, see plot D.2.

There is some degradation of performance as the distributions move away from normal. The method displays sensitivity to the use of proxy auxiliary information with a different $\rho_{Y,Z|X}$, which is naturally to be expected as it is precisely this quantity which is borrowed from the auxiliary data.

5.2.1.3 Transformed χ^2 for REG with Auxiliary Information on $\rho_{Y,Z|X}$

See plots E.3, E.4, F.3, F.4, G.3, G.4, H.3 and H.4.

The performance of REG methods with respect to the χ^2 statistics is not very good. The reason for this seems to be regression towards the mean, see plot I.2.

5.2.1.4 Conditional Likelihood Ratio Test for REG with Auxiliary Information on $\rho_{Y,Z|X}$

See plots E.5, E.6, F.5, F.6, G.5, G.6, H.5 and H.6.

With respect to this measure also, REG methods do not perform well, see plot I.2 to see the regression towards the mean which explains this.

5.2.2 REG Methods with Auxiliary Information on $\rho_{Y,Z}$

With respect to all measures the performance of REG methods with auxiliary information on $\rho_{Y,Z}$ is practically identical to the REG methods with auxiliary information on $\rho_{Y,Z|X}$.

5.3 Methods that use Auxiliary Information on Categorical Distributions

5.3.1 LOGLIN Methods with Auxiliary Information on (X^*,Y^*,Z^*) Distribution

5.3.1.1 MAD-Z for LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution

See plots E.1, F.1, G.1 and H.1.

REG.LOGLIN(auxcatxyz,xz-dist) does slightly worse with respect to this measure than REG(auxcorrxyz,xz-dist), but otherwise there are no important differences. HOD.LOGLIN(auxcatxyz,x-dist) on the other hand is slightly worse than REG.LOGLIN. However, both methods show some improvement over corresponding methods with no auxiliary information.

5.3.1.2 MAD-Cov for LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution

See plots E.2, F.2, G.2 and H.2.

With respect to this measure HOD.LOGLIN performs generally much better than REG.LOGLIN. For normal and non-normal data, HOD.LOGLIN does somewhat worse than the REG method, but generally does very similarly or sometimes better when the auxiliary information is proxy. There seems to be substantial improvement over methods that do not use any auxiliary information.

5.3.1.3 Transformed χ^2 for LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution

See plots E.3, E.4, F.3, F.4, G.3, G.4, H.3 and H.4.

With respect to this measure HOD.LOGLIN(auxcatxyz,x-dist) is better than any of the other methods with (X,Y,Z) auxiliary information; this includes the methods to be introduced in sections 5.4.1 and 5.5.1. It shows little or no dependence on $\rho_{Y,Z|X}$ and does not seem to be affected by non-normality of the data. It is slightly sensitive to the use of proxy auxiliary information, but perhaps less so than other methods. However, this measure also demonstrates that there may be some cost associated with the use of auxiliary information when $\rho_{Y,Z|X}$ is fairly small. In comparing plots A.3 and E.3, for example, we see that HOD.LOGLIN(rank,rakexyz) actually outperforms HOD.LOGLIN(auxcatxyz,x-dist) with respect to this measure for $\rho_{Y,Z|X} < .6$. This is likely due to ineffective estimation of the second order interaction terms of the saturated log linear model for the

categorical proportions and more precise auxiliary information would be expected to improve performance.

REG.LOGLIN(auxcatxyz,xz-dist) does poorly when different partitions are used for matching and testing. The reason for this is regression towards the mean, see plot I.2.

5.3.1.4 Conditional Likelihood Ratio Test for LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution

See plots E.5, E.6, F.5, F.6, G.5, G.6, H.5 and H.6.

Also with respect to this measure HOD.LOGLIN(auxcatxyz,x-dist) performs very well. Looking through the plots one notices that HOD.LOGLIN(auxcatxyz,x-dist) is sensitive to the use of log-normal contaminated proxy auxiliary information. When the same partition is used for constraints and testing, it is slightly sensitive to departures in $\rho_{Y,Z|X}$ in the proxy auxiliary data, but plot G.6 suggests that it is more robust against this than other methods are. It remains the best according to this measure among all methods using (X,Y,Z) auxiliary information, as with the χ^2 measure. There is also some apparent improvement as $\rho_{Y,Z|X}$ increases. This may be explained by the greater explanatory power of (X,Y) for Z, which is likely to translate to the categorical variables. As for the previous measure, HOD.LOGLIN(auxcatxyz,x-dist) is outperformed by HOD.LOGLIN(rank,rakexyz) for $\rho_{Y,Z|X} < .4$, though this is probably related to the precision of the auxiliary information.

REG.LOGLIN(auxcatxyz,xz-dist) again does not do well due to regression towards the mean, see plot I.2.

5.3.2 LOGLIN Methods with Auxiliary Information on (Y^*,Z^*) Distribution

The evaluations for the two methods REG.LOGLIN and HOD.LOGLIN with (Y^*,Z^*) auxiliary information are generally very similar to the evaluations for the two methods with (X^*,Y^*,Z^*) auxiliary information. However, there is one surprising difference and that is that for the categorical measures, χ^2 statistics and conditional likelihood ratio test measures, methods using (Y^*,Z^*) auxiliary information do better. Moreover, HOD.LOGLIN(auxcatyz,x-dist) is never outperformed by HOD.LOGLIN(rank,rakeyz), though HOD.LOGLIN(rank,rakexyz) is superior with respect to χ^2 when $\rho_{Y,Z|X} < .2$. Perhaps this relatively poor performance of (X^*,Y^*,Z^*) auxiliary information could be

explained by poor estimation from the auxiliary data of the second order interaction terms of the saturated log linear model for the categorical proportions. If the auxiliary information about these terms were more reliable we would expect some improvement in the methods that use (X^*,Y^*,Z^*) auxiliary information. However, the issue of sample size for the auxiliary data was not investigated empirically.

5.4 Methods that use Auxiliary Information on Correlations and Categorical Distributions

5.4.1 REG.LOGLIN Methods with Auxiliary Information on (X^*,Y^*,Z^*) Distribution and $\rho_{Y,Z|X}$

5.4.1.1 MAD-Z for REG.LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution and $\rho_{Y,Z|X}$

See plots E.1, F.1, G.1 and H.1.

With respect to this measure REG.LOGLIN(auxcorrxyz,auxcatxyz) lies between REG(auxcorrxyz) and REG(auxcatxyz).

5.4.1.2 MAD-Cov for REG.LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution and $\rho_{Y,Z|X}$

See plots E.2, F.2, G.2, and H.2.

With respect to this measure REG.LOGLIN(auxcorrxyz,auxcatxyz) is practically indistinguishable from REG(auxcorrxyz).

5.4.1.3 Transformed χ^2 for REG.LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution and $\rho_{Y,Z|X}$

See plots E.3, E.4, F.3, F.4, G.3, G.4, H.3, and H.4.

With respect to this measure REG.LOGLIN(auxcorrxyz,auxcatxyz) is very close to REG.LOGLIN(auxcatxyz), that is, it does not do well when the partitions for constraints and for testing are not the same, due to regression towards the mean.

5.4.1.4 Conditional Likelihood Ratio Test for REG.LOGLIN with Auxiliary Information on (X^*,Y^*,Z^*) Distribution and $\rho_{Y,Z|X}$

See plots E.5, E.6, F.5, F.6, G.5, G.6, H.5, and H.6.

With respect to this measure also REG.LOGLIN(auxcorrxyz,auxcatxyz) is very close to REG.LOGLIN(auxcatxyz), that is, it does not perform well, due to regression towards the mean, when the partitions for constraints and for testing are not the same.

5.4.2 REG.LOGLIN Methods with Auxiliary Information on (Y^*,Z^*) Distribution and $\rho_{Y,Z}$

There does not seem to be much difference between the two types of auxiliary information for this method; that is, REG.LOGLIN(auxcorrxyz,auxcatxyz) and REG.LOGLIN(auxcorryz,auxcatyz) perform similarly.

5.5 Methods using an Auxiliary Micro-datafile

5.5.1 (X,Y,Z) Auxiliary Micro-data

With this type of auxiliary information we add two methods, namely HOD(auxmicxyz) and HOD.LOGLIN(auxmicxyz,auxcatxyz), to the four methods considered previously.

5.5.1.1 MAD-Z with (X,Y,Z) Auxiliary Micro-data

See plots E.1, F.1, G.1, and H.1.

With respect to this measure HOD(auxmicxyz) and HOD.LOGLIN(auxmicxyz,auxcatxyz) both perform very similarly and are slightly better than HOD.LOGLIN(auxcatxyz). The REG method does the best and the HOD methods seem to be moderately worse than REG.

5.5.1.2 MAD-Cov with (X,Y,Z) Auxiliary Micro-data

See plots E.2, F.2, G.2, and H.2.

Both of these methods perform very similarly with respect to this measure and seem to do about as well as any other methods using (X,Y,Z) auxiliary information.

5.5.1.3 Transformed χ^2 with (X,Y,Z) Auxiliary Micro-data

See plots E.3, E.4, F.3, F.4, G.3, G.4, H.3, and H.4.

For this measure also the two methods perform very similarly, but HOD.LOGLIN(auxmicxyz,auxcatxyz) is slightly superior, especially when the same partition is used for constraints and for testing. In fact, for this measure HOD.LOGLIN(auxmicxyz,auxcatxyz) is almost the best of the methods using (X,Y,Z) auxiliary information, doing only slightly worse than HOD.LOGLIN(auxcatxyz). It is of interest to note the superiority of HOD methods using no auxiliary information when $\rho_{Y,Z|X}$ is small. This, as before, is probably due to insufficient auxiliary information.

5.5.1.4 Conditional Likelihood Ratio Test with (X,Y,Z) Auxiliary Micro-data

See plots E.5, E.6, F.5, F.6, G.5, G.6, H.5, and H.6.

This measure demonstrates the superiority of HOD.LOGLIN(auxmicxyz,auxcatxyz) over HOD(auxmicxyz). The gain due to the categorical constraints would likely be diminished if the auxiliary micro-datafile were larger. HOD.LOGLIN(auxcatxyz) seems to show considerable further gains with respect to this measure. Again, note the superior performance at small values of $\rho_{Y,Z|X}$ of HOD methods that do not use auxiliary information.

5.5.2 (Y,Z) Auxiliary Micro-data

The evaluations of these methods with (Y,Z) auxiliary micro-data are generally similar to those with (X,Y,Z) auxiliary micro-data except that the performance of HOD.LOGLIN(auxmicyz,auxcatyz) improves and is generally much closer to HOD.LOGLIN(auxcatyz) for the categorical evaluation measures. Also, for small $\rho_{Y,Z|X}$, the superior performance of methods using no auxiliary information almost disappears when compared to HOD methods using (Y,Z) auxiliary micro-data.

In comparing (Y,Z) to (X,Y,Z) auxiliary information it is striking that HOD.LOGLIN based on (Y,Z) auxiliary micro-data does much better than HOD.LOGLIN based on (X,Y,Z) auxiliary micro-data

for the categorical measures while performing similarly for the other measures, MAD of Z-values and MAD of conditional covariances. Again, a reasonable explanation for this would be poor estimation from the auxiliary data of the second order interaction terms of the saturated log linear model for the categorical proportions.

5.6 Summary of Results for the use of Auxiliary Information

One important and consistent finding was that REG and REG.LOGLIN methods do not perform well with respect to categorical measures because of regression towards the mean. This unfavourable performance tends to outweigh their favourable performances with respect to the other two evaluation measures.

A second important finding is that the use of auxiliary information does protect against the failure of CIA, and even the use of proxy auxiliary information is very helpful. It is also interesting to note that when the true value of $\rho_{Y,Z|X}$ is small, the performance of methods that use auxiliary information can be worse with respect to some categorical evaluation measures than the performance of those that do 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. However, for other measures there is considerable gain from the use of auxiliary information even at small values of $\rho_{Y,Z|X}$. Methods that use auxiliary information are still sensitive, with respect to the MAD-Cov measure, to non-normality of the underlying distribution produced via log-normal contamination. This could reasonably be explained by non-linearity of the relationship between X, Y, and Z in the log-normal contaminated distribution.

Especially in the case of (Y,Z) auxiliary micro-data, but also with (X,Y,Z) auxiliary micro-data, we have seen that the use of categorical constraints improves performance with respect to the categorical measures with only a marginal deterioration in performance with respect to the other measures. The difference between the effects of (Y,Z) and (X,Y,Z) auxiliary micro-data is likely to disappear for larger auxiliary micro-datafiles.

An interesting finding is that for HOD.LOGLIN methods the use of (Y,Z) auxiliary information leads to better performance than (X,Y,Z) auxiliary information with respect to the categorical measures, while not affecting the other measures. Since HOD methods with (Y,Z) auxiliary information perform no better than HOD methods with (X,Y,Z) auxiliary information, it seems that a reasonable explanation

of this phenomenon is that for the set-up of our study the estimation of the second order interaction terms of the saturated log linear model for the categorical proportions is of too poor a quality to be useful, so that simply assuming these terms to be zero leads to better results. This implies that the true values are in some sense not very far from zero to begin with. In any case, we may suppose that if the auxiliary information about the categorical proportions were of better quality, (X,Y,Z) auxiliary information would perform better; however, it is difficult to judge where the break point would be. This finding also suggests that we should consider using (X,Y,Z) auxiliary micro-data together with derived (Y^*,Z^*) categorical auxiliary information, something that we did not include in our study.

As we had noted in section 5.3, hot deck methods that use only categorical auxiliary information may be more robust against the use of proxy auxiliary information than other methods that use auxiliary information. See, for example, plots G.8, G.10, and G.12.

A complete summary of the results of the Monte Carlo study along with some recommendations is given in section 6.

5.7 Limitations of the Monte Carlo Study

A simulation study of this sort is always limited in the sense that there are situations or parameter settings which might have been studied but were not. Along with the general limitations affecting all simulation studies it is worthwhile to list some more specific limitations affecting this particular study.

Perhaps the most important limitation of this study was that it was done entirely with synthetic data. Despite this, it is believed that our results are relevant to real applications. In a simulation study with real data, it would be of particular interest to investigate whether hot deck methods continue to be preferable to regression methods.

A second limitation is that we did not investigate in this study the effect of the size of the files A, B, and C. We have, at various points in sections 5.1 through 5.6, speculated on how the picture would change if one or more of these files had been larger, but limitations of time and especially of computing resources have prevented empirical studies with different file sizes.

Another limitation is that we did not fully investigate the effect of choice of partition for the

LOGLIN procedures. Again, this was largely a matter of time and computing resources. Nevertheless, we would generally expect that the finer the partition the better the performance of methods using categorical constraints, at least to the point that the categorical proportions could be well estimated. If the partition were too fine then the noise in the estimation of the categorical proportions would probably nullify the advantage of using categorical constraints.

There are other matching methods which might have been included in the study but were not, partly because the number of methods that could reasonably be included was limited, and partly because we could not have prospectively imagined all of the possibilities (nor could we retrospectively imagine all of them, but some additional methods were suggested).

We have investigated only a limited number of scenarios as far as data sources were concerned. The effects of non-normality were investigated via log-normal contamination of the distributions but it would be very easy to suggest many other ways to generate non-normal data. In considering proxy auxiliary information we considered only two ways in which the distribution of the auxiliary data could be different.

Other possible scenarios, such as having a proxy file B, with possibly a small nonproxy auxiliary datafile C, were not considered. Such a scenario might conceivably arise if files A and B represented data from different time points and the auxiliary file C was obtained by a small scale special survey.

Despite these limitations a lot was learned about the various approaches to statistical matching and some general conclusions and guidelines are given in the next section.

6. SUMMARY WITH DISCUSSION

In this report the problem of statistical matching of two files, A (the host file) and B (the donor file) was considered as it arises in the creation of micro-simulation databases, e.g. SPSD (Wolfson, Gribble, Bordt, Murphy and Rowe, 1987). This problem can be viewed as one of imputation by regarding files A and B respectively as the item nonrespondent and complete respondent data sets. Statistical matching, however, differs from the usual problem of imputation whenever file A contains information about certain variables which are not included in file B.

A Monte Carlo evaluation of four methods of statistical matching was performed for synthetic datafiles A and B. The first two methods correspond to the commonly used imputation methods, namely, hot deck (HOD) and linear regression (REG). The last two methods correspond to log linear imputation (LOGLIN) in which either HOD or REG is used under categorical constraints. The purpose of categorical constraints is to preserve log linear associations in the categorical distribution of the completed data set under a suitable partition. These associations would be obtained from files A and B and possibly from auxiliary data.

With synthetic data from multivariate normal and those under log-normal contaminations, a number of scenarios for Monte Carlo simulation were studied. First, the impact of the commonly made assumption of conditional independence (CIA) is considered which, in fact, distinguishes statistical matching from the usual imputation problem. Denoting file A variables by (X,Y) and file B variables by (X,Z), under CIA one can ignore Y in matching Z variable information to file A. The CIA is generally expected to introduce serious bias (or distortion) in the Y-Z relationship. A sensitivity analysis is performed as the underlying distribution moves away from the CIA. It is possible to avoid CIA if additional information (file C) about (Y,Z) or (X,Y,Z) is available. The question of the extent of gain by using auxiliary information was considered next. Ways of using auxiliary information for statistical matching by HOD, REG and LOGLIN methods were based on procedures proposed by Paass (1986), Rubin (1986) and Singh (1988) respectively. The impact of having proxy auxiliary data (i.e. from a different or outdated universe) was also examined. The particular choice of partition for LOGLIN methods was somewhat coarse, and for partition points either boundaries of equal intervals on the standardized scale or equal probability intervals were used. Various choices of the underlying distribution were made in order to get a range of conditional correlations ($\rho_{Y,Z|X}$) and a range of proportions of log-normal contaminations for a given $\rho_{Y,Z|X}$. Four evaluation measures, two at the unit level (denoted by MAD-Z and MAD-Cov, see section 4.3), and two at the aggregate level (denoted by transformed χ^2 and CLRT, see section 4.3) were considered.

The main findings of the empirical study are listed below.

(i) The CIA based methods may cause serious bias in the joint relationship of (X,Y,Z). In other words, none of the methods considered is robust in general. However, the HOD methods generally perform best. Moreover, there seems to be no gain in imposing categorical constraints on HOD methods (i.e. with HOD.LOGLIN methods). This is because there is no natural way to define categorical constraints in the absence of auxiliary information. However, an interesting finding was that, with respect to the categorical evaluation measures, auxiliary information does not improve performance of HOD methods at smaller values of $\rho_{Y,Z|X}$. This seems quite reasonable when the auxiliary data set is not large.

(ii) If a nonproxy micro-level auxiliary datafile C is available, then both HOD and HOD.LOGLIN methods work well; that is, they tend to produce a substantial reduction in distributional distortion as compared to CIA based methods. For a small datafile C, HOD.LOGLIN with (Y,Z) categorical constraints is expected to do well. However, for a sufficiently large datafile C, the two methods HOD and HOD.LOGLIN should perform very similarly. It may be noted that the HOD.LOGLIN method based on only categorical auxiliary information performs moderately well as compared to the HOD.LOGLIN method based on micro-level auxiliary information. Moreover, even a fairly coarse categorical partition for LOGLIN methods can lead to very favourable performances.

(iii) If the auxiliary data were outdated or proxy, there may still be gain in using it. The HOD.LOGLIN method with (Y,Z) categorical auxiliary information seems to perform quite favourably with proxy auxiliary data. If file C were large, then the difference between HOD.LOGLIN methods based on (Y,Z) and (X,Y,Z) categorical auxiliary information would be expected to disappear. Since LOGLIN methods only require log linear associations from categorical auxiliary information, it would seem reasonable for these to be affected only marginally by a limited amount of outdatedness or proxy values in the auxiliary data.

Some other interesting findings are as follows. Methods based on different distance measures (z, xz and xyz) perform very similarly, although methods based on xz-distance have a slight superiority for the MAD-Cov evaluation measure. There is also hardly any difference between random, distance, and rank versions of hot deck with no auxiliary micro-data. Regression methods do well with respect to the two unit level evaluation measures (MAD-Z and MAD-Cov). However, this seems to be outweighed by their very unfavourable performance with respect to the categorical measures. This

striking phenomenon can be explained by the regression to the mean effect.

Based on the above discussion, it would be reasonable to recommend HOD or HOD.LOGLIN methods for the purpose of statistical matching. However, an important factor in practical application is the computational burden associated with each method. For example, the similar performances of hot deck random, rank and distance with no auxiliary micro-data might suggest the use of hot deck random since this is the least demanding computationally. In addition, for all types of matching, methods based on xyz-distance are significantly more computationally intensive than those based on z or xz-distance which require similar levels of resources. HOD.LOGLIN using auxiliary micro-data together with derived categorical constraints is the most computer intensive procedure that we have considered. As files B and C become larger this could become a serious problem and the advantage over HOD using only auxiliary micro-data would also be expected to disappear. If file C is smaller then categorical constraints can lead to substantial improvement in the performance of HOD methods with auxiliary micro-data, and the computational burden is not excessive. If file C is thought to be proxy it is recommended to use only the derived categorical constraints.

Although this study with synthetic data has demonstrated the possible gains from the use of auxiliary data, a further study using real data is desirable and is currently being planned to measure the potential gains and to confirm what was learned here.

ACKNOWLEDGEMENTS

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 as an adjunct research professor. The authors would like to thank J. Armstrong, G. Gray, G. Hole, D. Royce and M. Wolfson for helpful comments, as well as J. Clarke for her efficient assistance with manuscript processing.

REFERENCES

Armstrong, J. (1989). An Evaluation of Statistical Matching Methods. Methodology Branch Working Paper, Statistics Canada, BSMD 90-003E.

Barr, R.S., Stewart, W.H. and Turner, J.S. (1981). An Empirical Evaluation of Statistical Matching Methodologies, Dallas, Texas, Edwin L. Cox School of Business, Southern Methodist University.

Barr, R.S. and Turner, J.S. (1990). Quality Issues and Evidence in Statistical File Merging. In Data Quality Control: Theory and Pragmatics (G.E. Liepins and V.R.R. Uppuluri, eds.), Marcel Dekker, 245-313.

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, volume 2 (W.G. Madow, I. Olkin and D.B. Rubin, eds.), Academic, New York, 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.

Mantel, H. and Kinack, M. (1991). Program Documentation for a Monte Carlo Evaluation of Statistical Matching Methods (forthcoming).

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 (G.H.Orcutt, J. Merz and H. Quinke, eds.), Amsterdam: Elsevier Science.

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.

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, Statistics Canada, SSMD 88-029E; also published in Proceedings of the Fifth Annual Research Conference (1989), Washington, D.C.: U.S. Bureau of the Census, 118-132.

Singh, A.C., Armstrong, J.B. and Lemaitre, 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). Use of Auxiliary Information for Statistical Matching. Abstracts of the Annual Meeting of the Statistical Society of Canada, St. John's, Newfoundland, p. 23.

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 of the Symposium on Statistical Uses of Administrative Data, Statistics Canada, Ottawa, November 23-25, 1987 (J.W. Coombs and M.P. Singh, eds.), 201-229.

A.1 APPENDIX 1 - NAMING CONVENTION FOR STATISTICAL MATCHING METHODS

Major Types:

REG	- regression methods
HOD	- hot deck methods
REG.LOGLIN	- regression methods with categorical constraints
HOD.LOGLIN	- hot deck methods with categorical constraints

Auxiliary Information Types:

auxcatxyz	- (X^*,Y^*,Z^*) categorical auxiliary information for LOGLIN procedures
auxcatyz	- (Y^*,Z^*) categorical auxiliary information for LOGLIN procedures
auxcorrxyz	- conditional correlation of Y and Z given X for regression procedures
auxcorryz	- unconditional correlation of Y and Z for regression procedures
auxmicxyz	- (X,Y,Z) auxiliary micro-datafile for hot deck procedures
auxmicyz	- (Y,Z) auxiliary micro-datafile for hot deck procedures

Miscellaneous:

rakeyz	- raking procedure for LOGLIN that assumes (Y^*,Z^*) are unconditionally categorically independent
rakexyz	- raking procedure for LOGLIN that assumes (Y^*,Z^*) are categorically independent conditional on X^* category
xcat	- for HOD procedures without LOGLIN, indicates that matching is done within X^* categories
rand	- for HOD procedures with no auxiliary micro-datafile, indicates that final matching step is random within categories
rank	- for HOD procedures with no auxiliary micro-datafile, indicates that final matching step is based on ranks within categories
x-dist	- for HOD procedures with no auxiliary micro-datafile, indicates that final matching step is based on x-distance, perhaps within X^* categories
z-dist	- indicates that final matching step is based on z-distance
xz-dist	- indicates that final matching step is based on xz-distance
xyz-dist	- indicates that final matching step is based on xyz-distance

A.2 APPENDIX 2 - DETAILS OF EVALUATION MEASURES

A.2.1 Mean Absolute Differences of Z

The formula for the MAD-Z statistic for one simulation, described in section 4.3.1, is

$$\sum_i |z_{s,i} - z_{m,i}|/500$$

where $z_{s,i}$ is the suppressed Z-value for the i^{th} record in file A, $z_{m,i}$ is the matched Z-value, and the sum is over all 500 records of file A.

A.2.2 Differences of Covariances

For a file with variables X, Y and Z the conditional covariance of Y and Z given X is defined as

$$cov(Y,Z) - cov(X,Y)cov(X,Z)/var(X)$$

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 in the distribution conditional on X; in general it may be thought of as a measure of the strength of the relationship between Y and Z given X. The MAD-Cov statistic for one simulation, described in section 4.3.2, would be the absolute difference between these quantities for the matched and suppressed files.

A.2.3 Chi-square Statistics

The precise formula for the chi-square statistics, described in section 4.3.3, is

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

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. Because the (X^*,Y^*) margins are fixed, the degrees of freedom of this statistic are IJ(K-1), where I is the number

of X^* categories, J is the number of Y^* categories and K is the number of 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 X, is obtained, it is transformed to lie in the interval (0,1) using the transformation

$$\{X/(X+500)\}^{\frac{1}{2}}$$

The number 500 here is not arbitrary. It is the size of file A; that is, the number of observations that the chi-square statistic is based on.

A.2.4 Conditional Likelihood Ratio Test

Using the same notation as in section A.2.3 the precise formula for the conditional likelihood ratio test statistic from one (X^*,Y^*) category, described in section 4.3.4, is

$$2\sum_k \{(n_{ijk}+.5)\,ln((n_{ijk}+.5)/(n_{ijk}+m_{ijk}+1)) + (m_{ijk}+.5)\,ln((m_{ijk}+.5)/(n_{ijk}+m_{ijk}+1))\} + (4n_{ij}+2K)ln2$$

where

$$n_{ij} = \sum_k n_{ijk} = \sum_k m_{ijk}$$

The asymptotic distribution of this statistic, when the m_{ijk}'s and n_{ijk}'s come from the same multinomial 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.

A.3 APPENDIX 3 - EXAMPLE OF TWO RAKING PROCEDURES

In order to illustrate the difference between the two raking procedures described in section 2.4 consider the following example in which a 2x2x2 table of 1's is raked to two 2x2 tables corresponding to X^*-Y^* and X^*-Z^* marginal tables, using first rakexyz and then rakeyz. Note that the X^*-Z^* marginal table has been raked so that the X^* margin corresponds to the X^* margin from the X^*-Y^* marginal table:

X^*-Y^* marginal table

	Y^*	
X^*	50	40
	27	35

X^*-Z^* marginal table

	Z^*	
X^*	41	49
	55	7

When the starting table is a table of 1's the rakexyz procedure will always converge in one iteration. In this example the 2x2x2 table of 1's can first be raked to the X^*-Z^* marginal table giving

		Z^*		
	Y^*		Y^*	
X^*	20.5	20.5	24.5	24.5
	27.5	27.5	3.5	3.5

and then raked to the X^*-Y^* marginal table giving a table which satisfies both marginal tables:

		Z^*		
	Y^*		Y^*	
X^*	22.8	18.2	27.2	21.8
	24.0	31.0	3.0	4.0

The alternative raking procedure rakeyz differs from rakexyz in that a third marginal table on Y^*-Z^* is also used. This additional marginal table is constructed from the X^*-Y^* and X^*-Z^* marginal tables by raking a 2x2 table of 1's to the Y^* margin from the X^*-Y^* marginal table and the Z^* margin from the X^*-Z^* marginal table. This process will converge in one iteration, in this case yielding the Y^*-Z^* marginal table:

	Z*	
Y*	48.6	28.4
	47.4	27.6

The rakeyz procedure now calls for a 2x2x2 table of 1's to be raked to the three marginal tables for X*-Y*, X*-Z* and Y*-Z*. This process will not necessarily converge in one iteration. In the present example the 2x2x2 table of 1's can first be raked to the X*-Z* marginal table giving

	Z*			
	Y*		Y*	
X*	20.5	20.5	24.5	24.5
	27.5	27.5	3.5	3.5

and then raked to the X*-Y* marginal table giving

	Z*			
	Y*		Y*	
X*	22.8	18.2	27.2	21.8
	24.0	31.0	3.0	4.0

and then raked to the Y*-Z* marginal table giving

	Z*			
	Y*		Y*	
X*	23.7	17.5	25.5	23.4
	24.9	29.8	2.9	4.2

After several iterations the rakeyz process converges to the table below, which satisfies all three marginal tables and is clearly different from the final table obtained from the rakexyz procedure.

	Z*			
	Y*		Y*	
X*	24.3	16.7	25.7	23.3
	24.4	30.6	2.6	4.4

A: Methods with no auxiliary information

× HOD(xcat,rank)

□ HOD.LOGLIN(rank,rakexyz)

△ HOD.LOGLIN(rank,rakeyz)

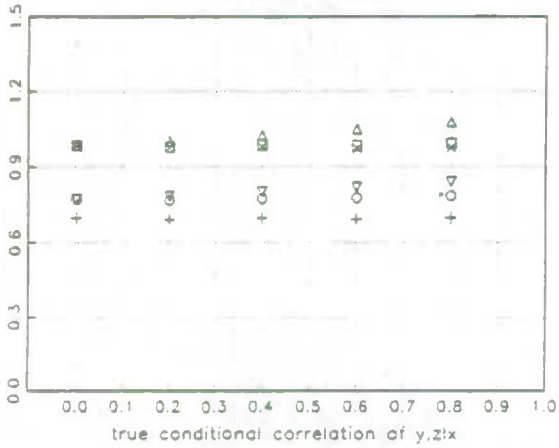
\+ REG(xz-dist)

○ REG.LOGLIN(xz-dist,rakexyz)

▽ REG.LOGLIN(xz-dist,rakeyz)

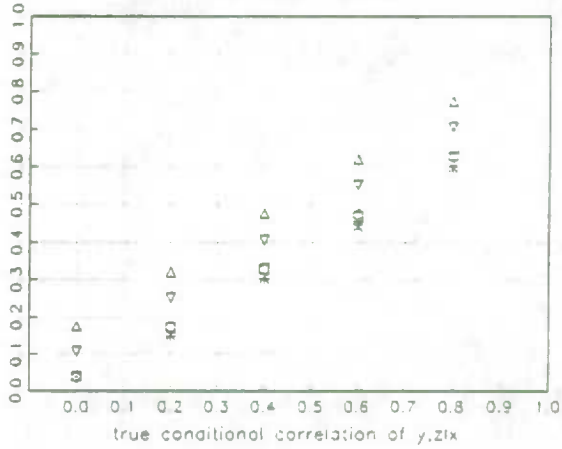
Plot A.1: MAD of z-values

Partition is equal probability



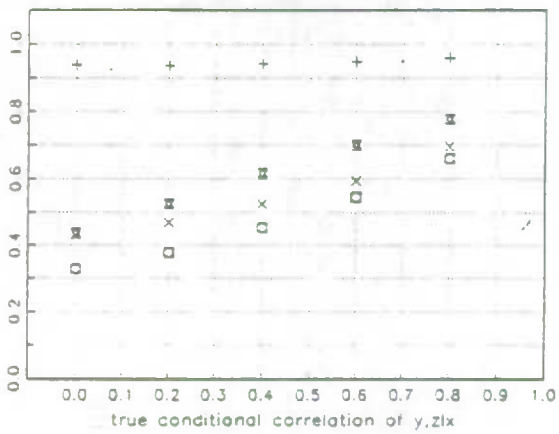
Plot A.2: MAD of covariances of y,z|x

Partition is equal probability



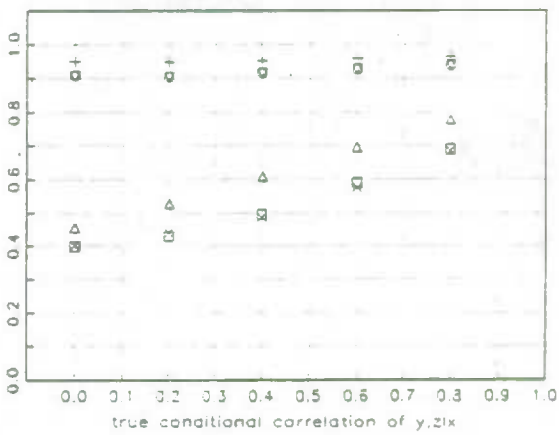
Plot 5.1.3: Transformed mean chi-square statistics

Imputation partition is equal probability
Testing partition is equal probability



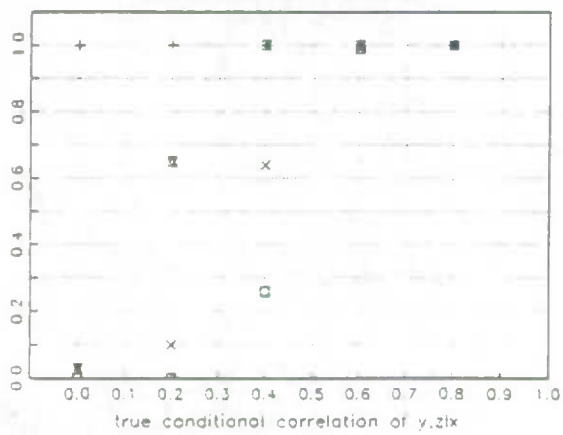
Plot A.4: Transformed mean chi-square statistics

Imputation partition is equal probability
Testing partition is standard interval



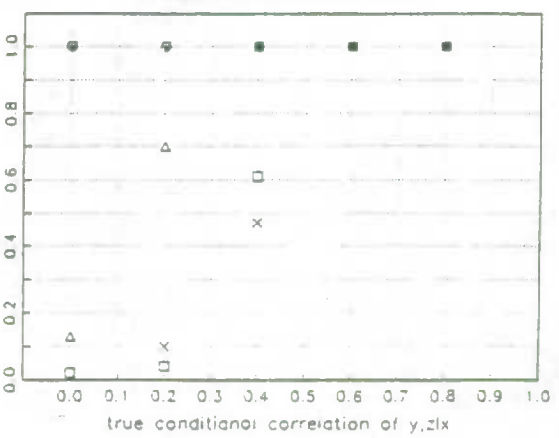
Plot A.5: Proportion of CLRT p-values < 0.05

Imputation partition is equal probability
Testing partition is equal probability



Plot A.6: Proportion of CLRT p-values < 0.05

Imputation partition is equal probability
Testing partition is standard interval

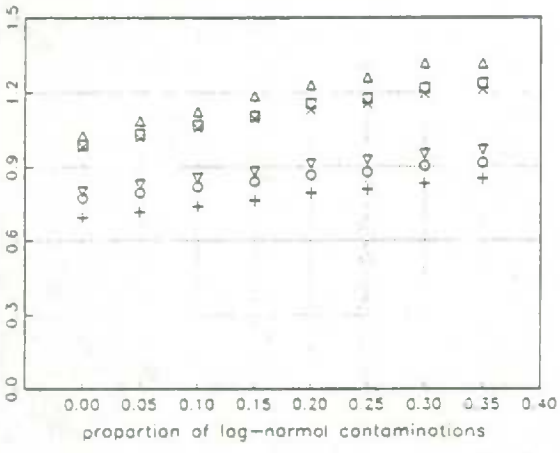


B: Methods with no auxiliary information

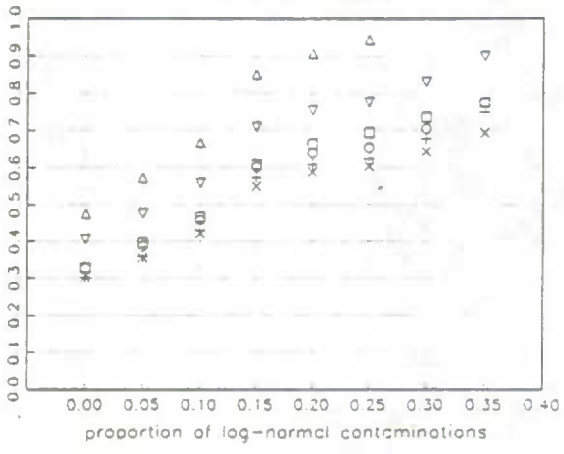
× HOD(xcat,rank)
□ HOD.LOGLIN(rank,rakexyz)
△ HOD.LOGLIN(rank,rakeyz)
\+ REG(xz-dist)
○ REG.LOGLIN(xz-dist,rakexyz)
▽ REG.LOGLIN(xz-dist,rakeyz)

data with log-normal contaminations, rho(y,z|x)=.4

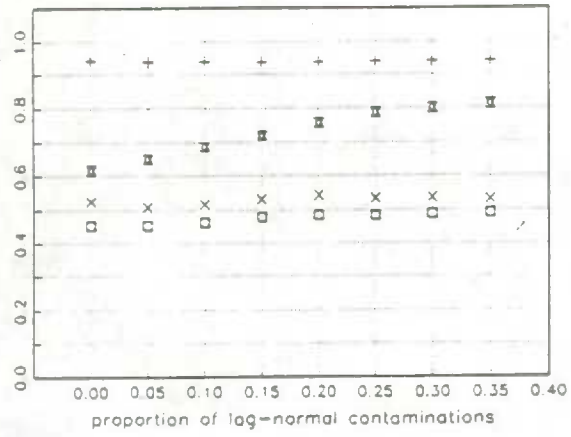
Plot B.1: MAD of z-values
Partition is equal probability



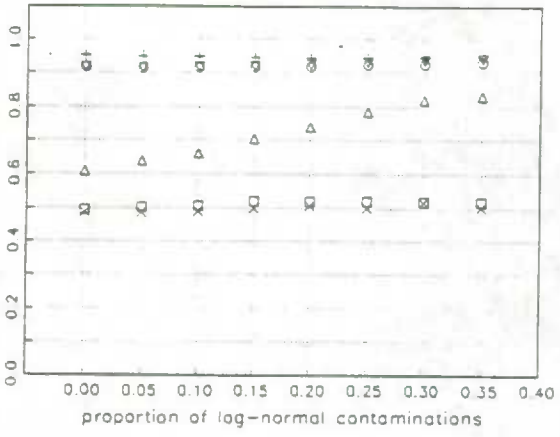
Plot B.2: MAD of covariances of y,z|x
Partition is equal probability



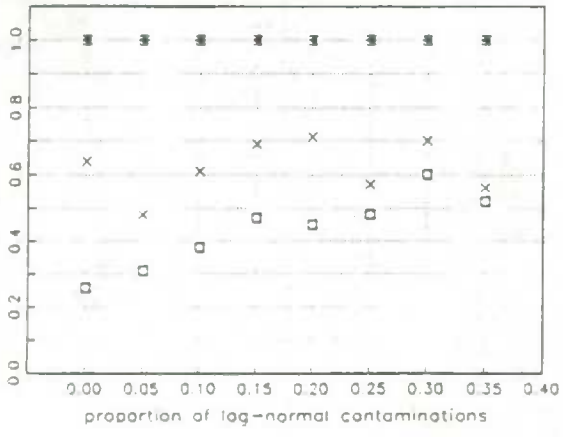
Plot B.3: Transformed mean chi-square statistics
Imputation partition is equal probability
Testing partition is equal probability



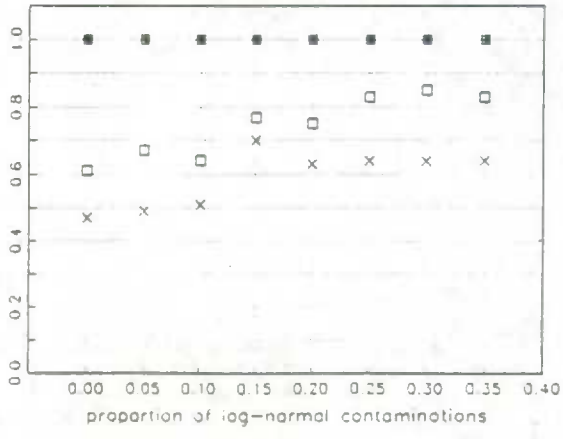
Plot B.4: Transformed mean chi-square statistics
Imputation partition is equal probability
Testing partition is standard interval



Plot B.5: Proportion of CLRT p-values < 0.05
Imputation partition is equal probability
Testing partition is equal probability



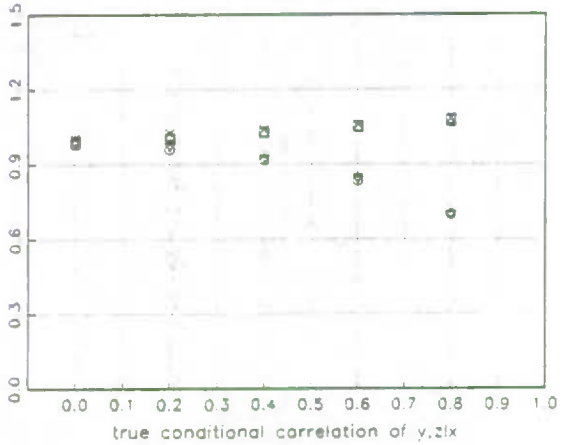
Plot B.6: Proportion of CLRT p-values < 0.05
Imputation partition is equal probability
Testing partition is standard interval



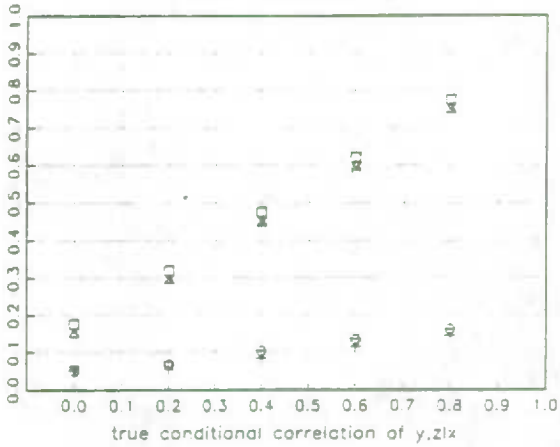
C: Hot-Deck Methods with Log-Linear Constraints

× HOD.LOGLIN(rand,rakeyz)
□ HOD.LOGLIN(rank,rakeyz)
△ HOD.LOGLIN(x-dist,rakeyz)
+ HOD.LOGLIN(auxcatyz,rand)
○ HOD.LOGLIN(auxcatyz,rank)
▽ HOD.LOGLIN(auxcatyz,x-dist)

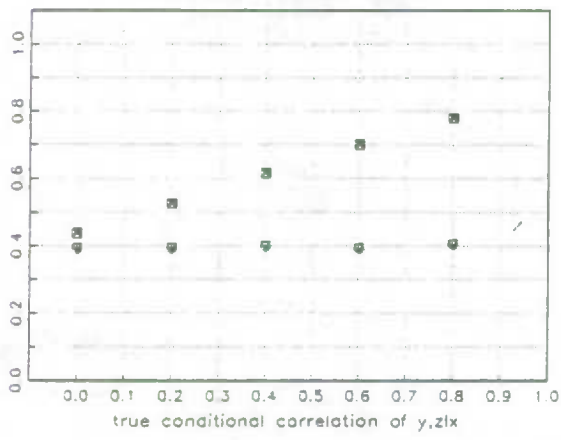
Plot C.1: MAD of z-values
Partition is equal probability



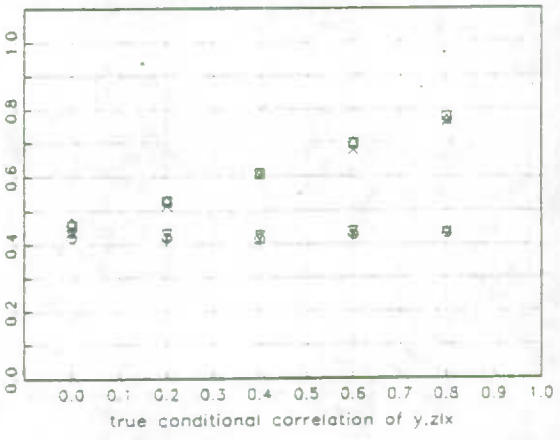
Plot C.2: MAD of covariances of y,z|x
Partition is equal probability



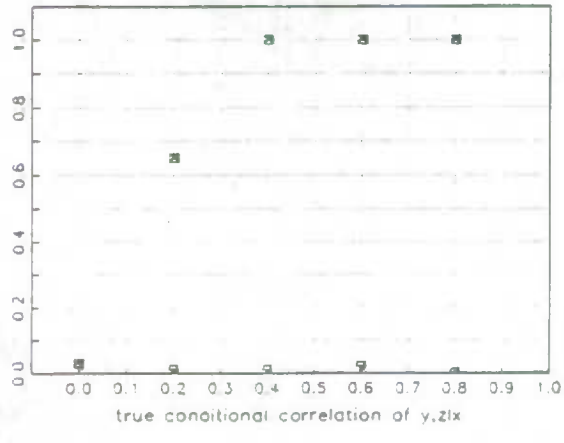
Plot C.3: Transformed mean chi-square statistics
Imputation partition is equal probability
Testing partition is equal probability



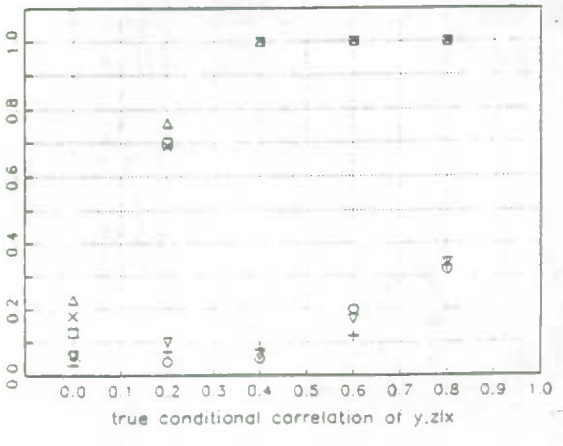
Plot C.4: Transformed mean chi-square statistics
Imputation partition is equal probability
Testing partition is standard interval



Plot C.5: Proportion of CLRT p-values < 0.05
Imputation partition is equal probability
Testing partition is equal probability



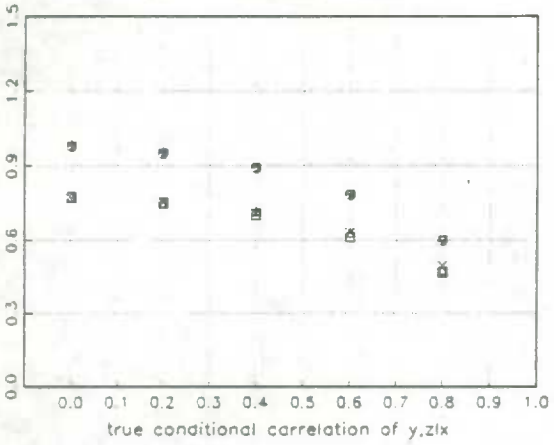
Plot C.6: Proportion of CLRT p-values < 0.05
Imputation partition is equal probability
Testing partition is standard interval



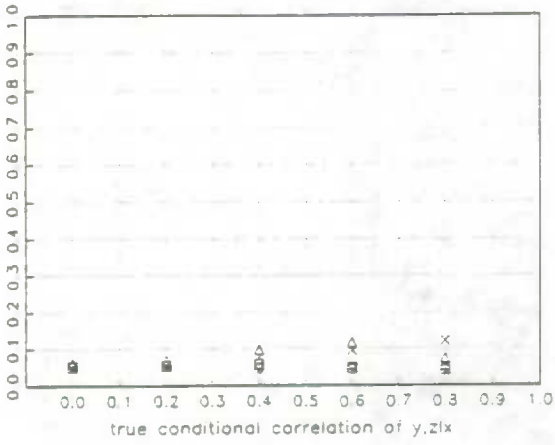
D: Different Distance Measures

× REG.LOGLIN(auxcorrxyz,auxcatxyz,z−dist)
□ REG.LOGLIN(auxcorrxyz,auxcatxyz,xz−dist)
△ REG.LOGLIN(auxcorrxyz,auxcatxyz,xyz−dist)
+ HOD.LOGLIN(auxmicxyz,auxcatxyz,z−dist)
◇ HOD.LOGLIN(auxmicxyz,auxcatxyz,xz−dist)
▽ HOD.LOGLIN(auxmicxyz,auxcatxyz,xyz−dist)

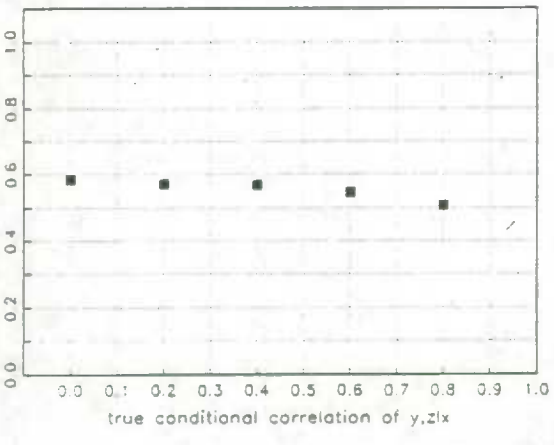
Plot D.1: MAD of z−values



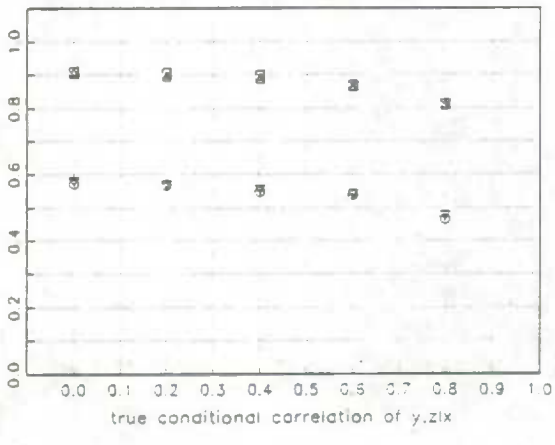
Plot D.2: MAD of covariances of y,z|x



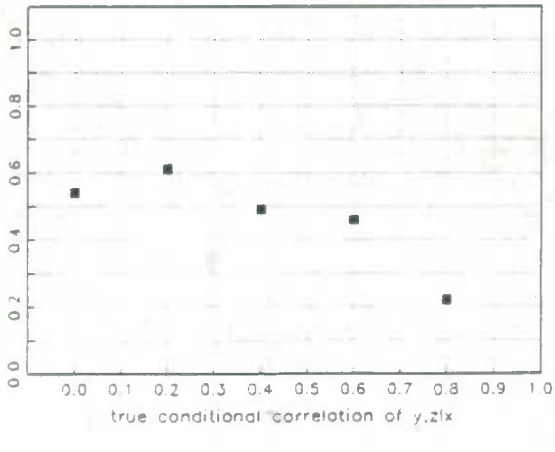
Plot D.3: Transformed mean chi−square statistics
Testing partition is equal probability



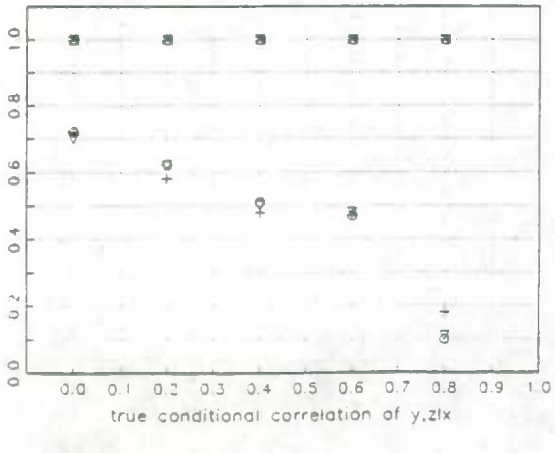
Plot D.4: Transformed mean chi−square statistics
Testing partition is standard interval



Plot D.5: Proportion of CLRT p−values < 0.05
Testing partition is equal probability



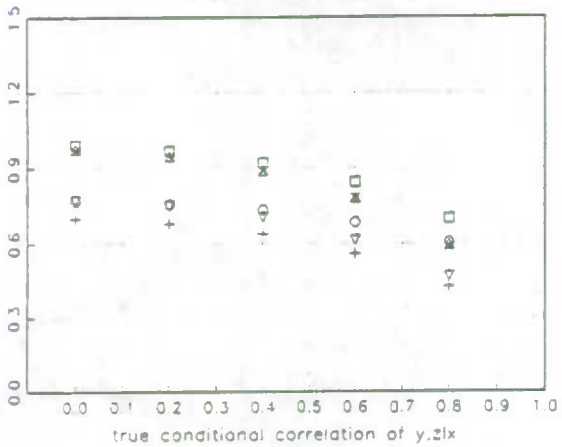
Plot D.6: Proportion of CLRT p−values < 0.05
Testing partition is standard interval



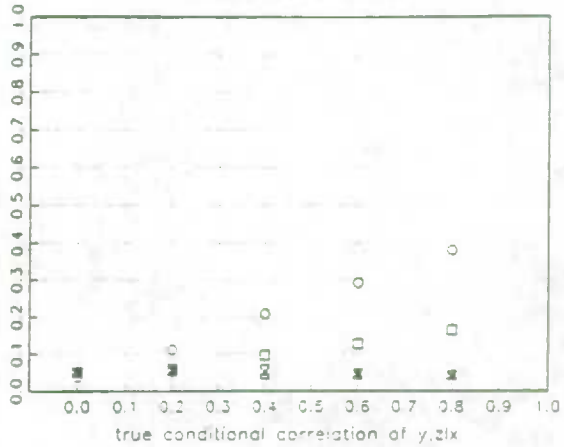
E: Methods using xyz auxiliary information

× HOD(auxmicxyz,xz−dist)
□ HOD.LOGLIN(auxcatxyz,x−dist)
△ HOD.LOGLIN(auxmicxyz,auxcatxyz,xz−dist)
+ REG(auxcorrxyz,xz−dist)
◇ REG.LOGLIN(auxcatxyz,xz−dist)
▽ REG.LOGLIN(auxcorrxyz,auxcatxyz,xz−dist)

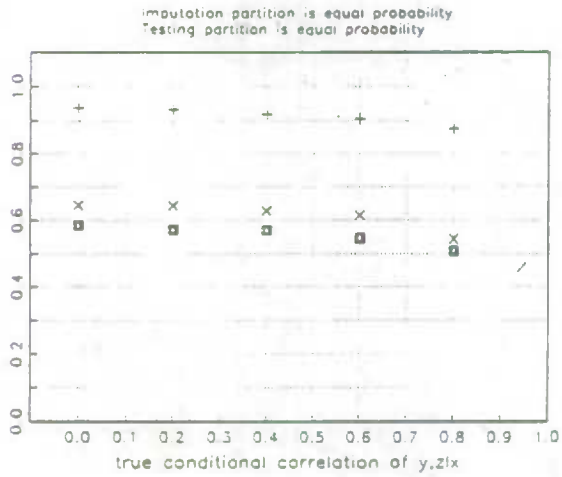
Plot E.1: MAD of z−values



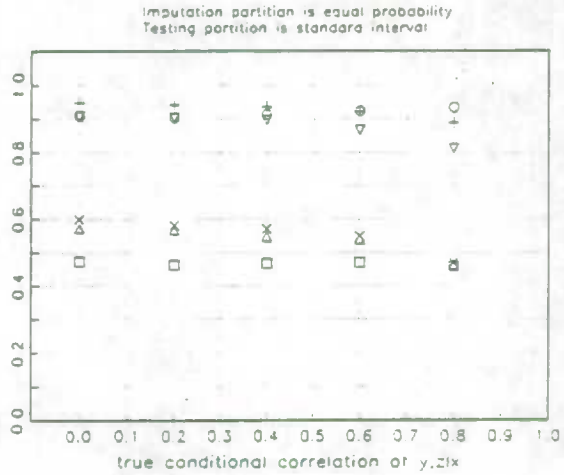
Plot E.2: MAD of covariances of y,z|x



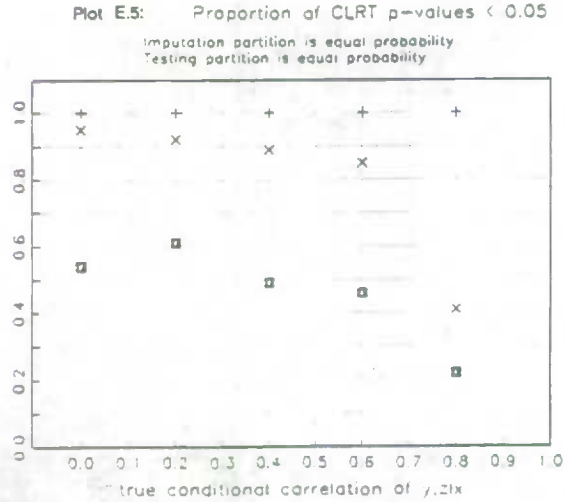
Plot E.3: Transformed mean chi−square statistics



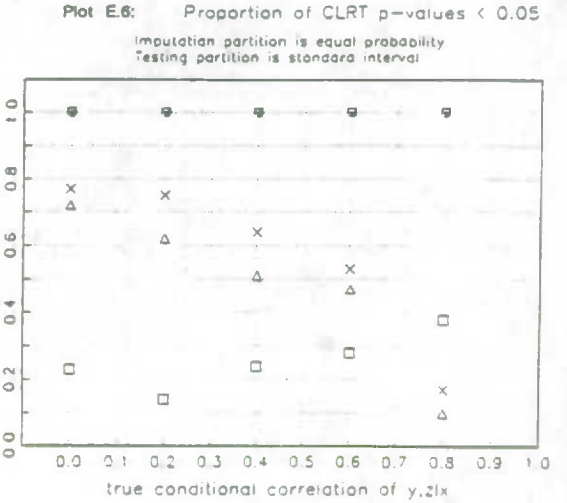
Plot E.4: Transformed mean chi−square statistics



Plot E.5: Proportion of CLRT p−values < 0.05



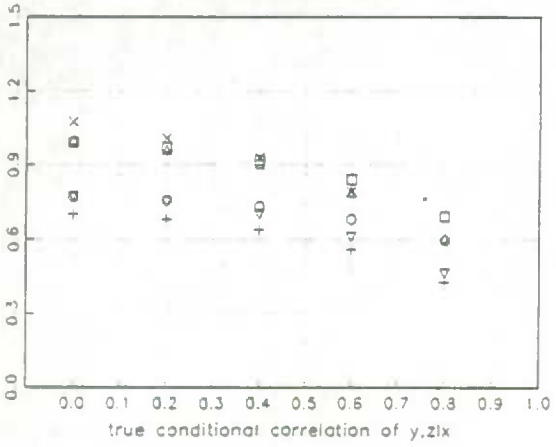
Plot E.6: Proportion of CLRT p−values < 0.05



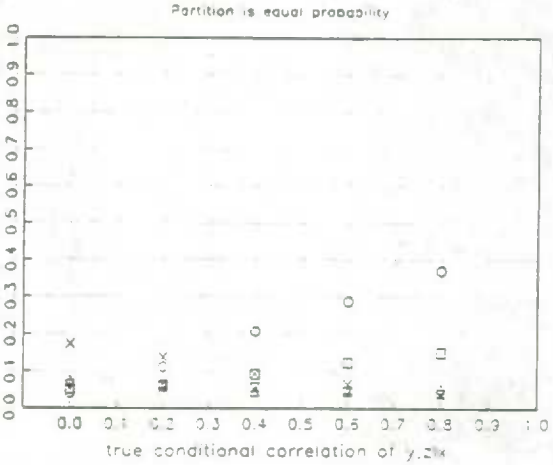
E: Methods using yz auxiliary information

× HOD(auxmicyz,xz−dist)
□ HOD.LOGLIN(auxcatyz,x−dist)
△ HOD.LOGLIN(auxmicyz,auxcatyz,xz−dist)
+ REG(auxcorryz,xz−dist)
○ REG.LOGLIN(auxcatyz,xz−dist)
▽ REG.LOGLIN(auxcorryz,auxcatyz,xz−dist)

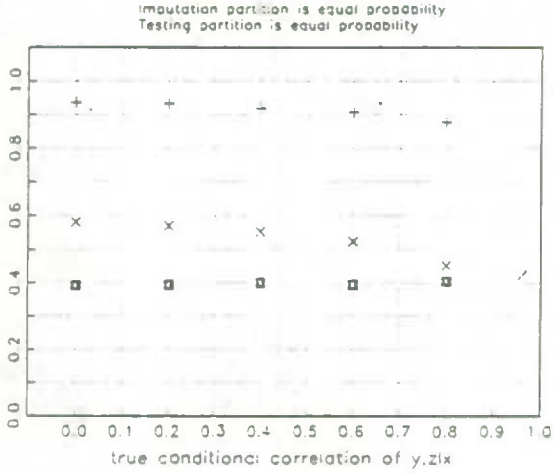
Plot E.7: MAD of z−values



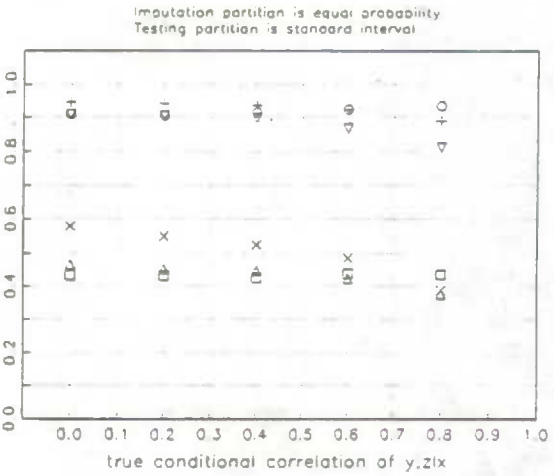
Plot E.8: MAD of covariances of y,zlx



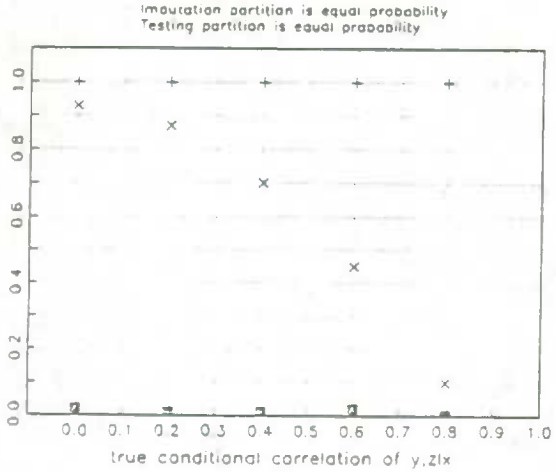
Plot E.9: Transformed mean chi−square statistics



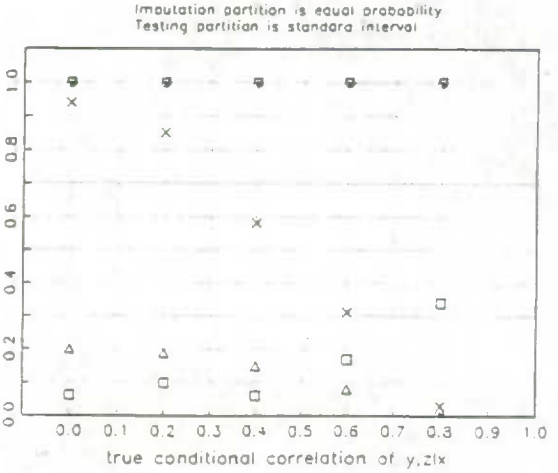
Plot E.10: Transformed mean chi−square statistics



Plot E.11: Proportion of CLRT p−values < 0.05



Plot E.12: Proportion of CLRT p−values < 0.05

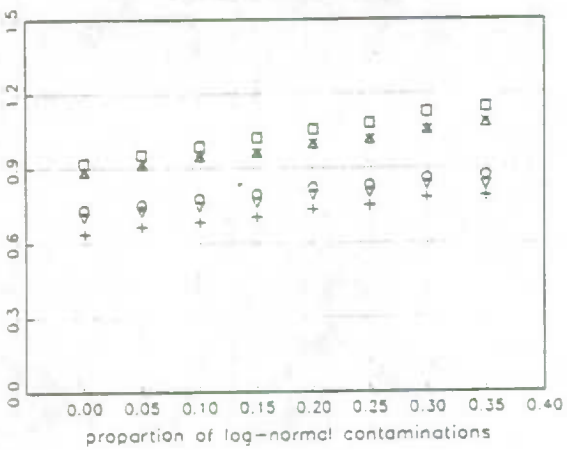


F: Methods using xyz auxiliary information

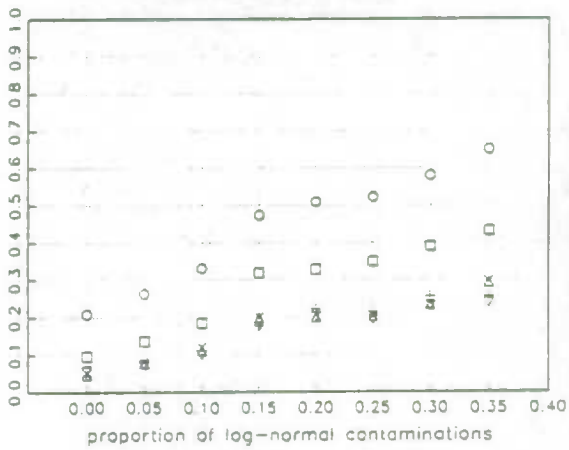
× HOD(auxmicxyz,xz−dist)
□ HOD.LOGLIN(auxcatxyz,x−dist)
△ HOD.LOGLIN(auxmicxyz,auxcatxyz,xz−dist)
+ REG(auxcorrxyz,xz−dist)
○ REG.LOGLIN(auxcatxyz,xz−dist)
▽ REG.LOGLIN(auxcorrxyz,auxcatxyz,xz−dist)

data with log−normal contaminations, rho(y,z|x)=.4

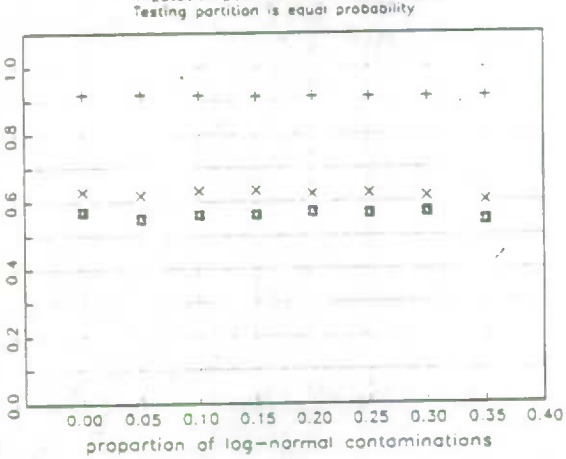
Plot F.1: MAD of z−values
Partition is equal probability



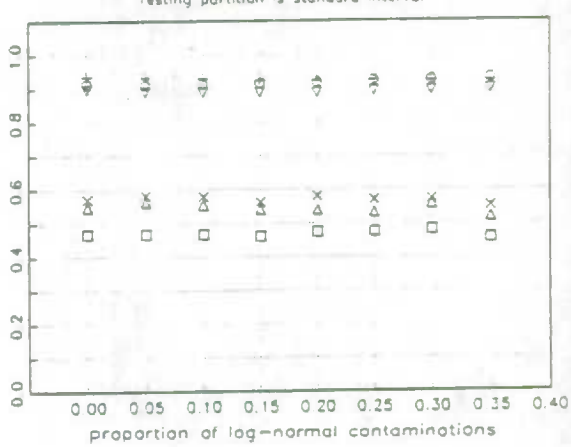
Plot F.2: MAD of covariances of y,z|x
Partition is equal probability



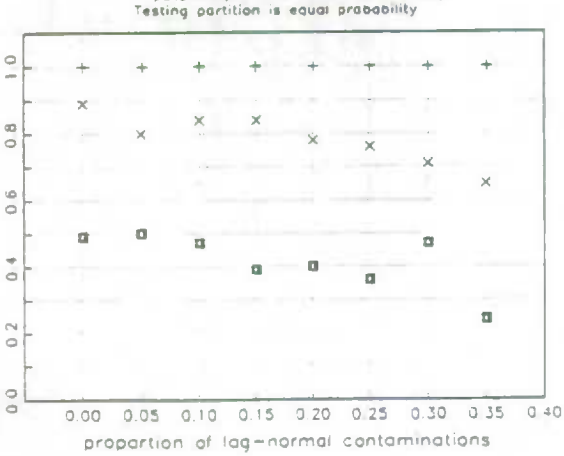
Plot F.3: Transformed mean chi−square statistics
Imputation partition is equal probability
Testing partition is equal probability



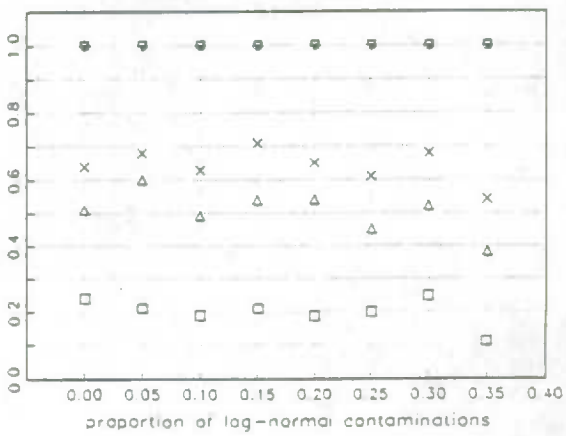
Plot F.4: Transformed mean chi−square statistics
Imputation partition is equal probability
Testing partition is standard interval



Plot F.5: Proportion of CLRT p−values < 0.05
Imputation partition is equal probability
Testing partition is equal probability



Plot F.6: Proportion of CLRT p−values < 0.05
Imputation partition is equal probability
Testing partition is standard interval

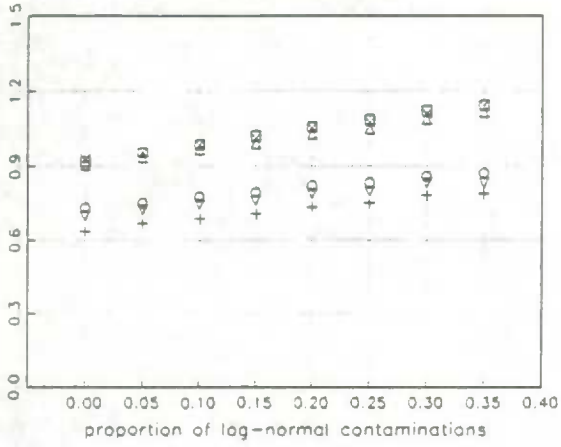


F: Methods using yz auxiliary information

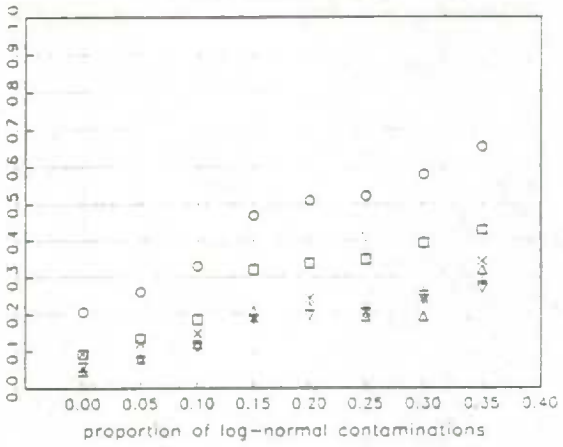
× HOD(auxmicyz,xz−dist)
□ HOD.LOGLIN(auxcatyz,x−dist)
△ HOD.LOGLIN(auxmicyz,auxcatyz,xz−dist)
+ REG(auxcorryz,xz−dist)
○ REG.LOGLIN(auxcatyz,xz−dist)
▽ REG.LOGLIN(auxcorryz,auxcatyz,xz−dist)

data with log−normal contaminations, rho(y,z|x)=.4

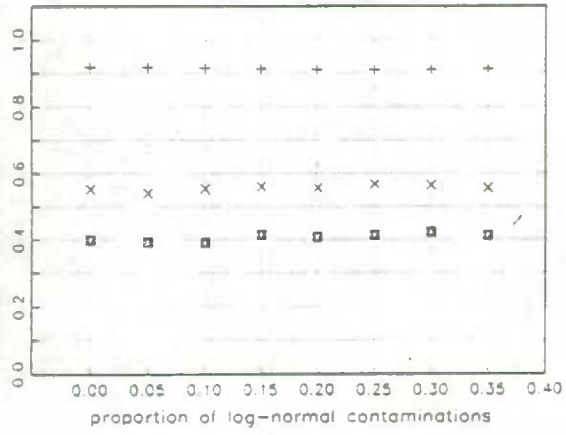
Plot F.7: MAD of z−values
Partition is equal probability



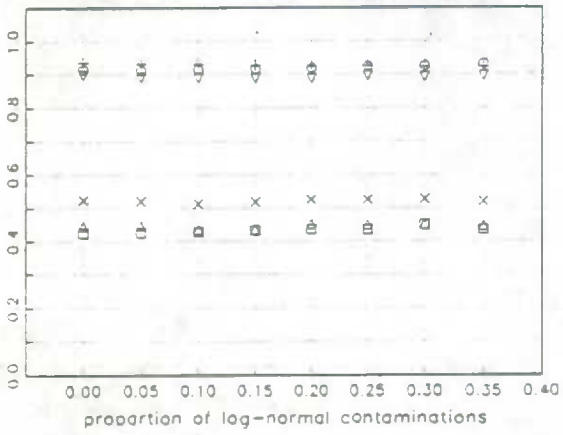
Plot F.8: MAD of covariances of y,z|x
Partition is equal probability



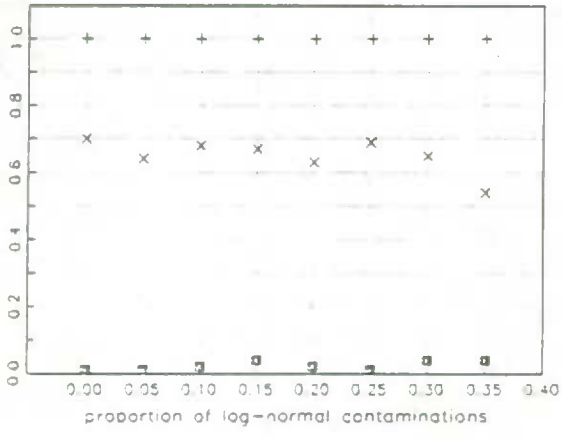
Plot F.9: Transformed mean chi−square statistics
Imputation partition is equal probability
Testing partition is equal probability



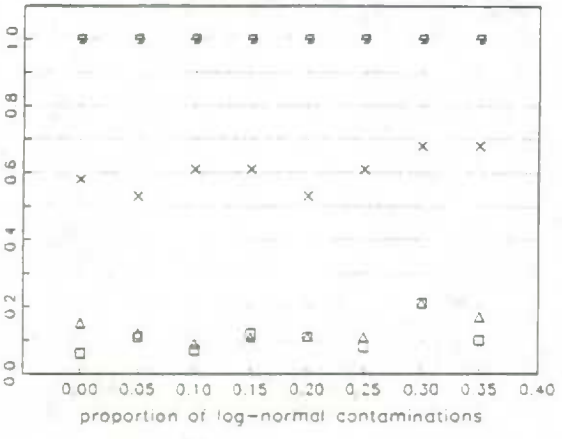
Plot F.10: Transformed mean chi−square statistics
Imputation partition is equal probability
Testing partition is standard interval



Plot F.11: Proportion of CLRT p−values < 0.05
Imputation partition is equal probability
Testing partition is equal probability



Plot F.12: Proportion of CLRT p−values < 0.05
Imputation partition is equal probability
Testing partition is standard interval



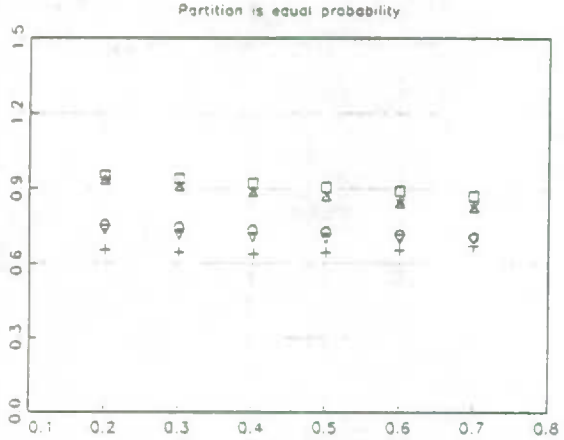
G: Methods using xyz auxiliary information

× HOD(auxmicxyz,xz−dist)
□ HOD.LOGLIN(auxcatxyz,x−dist)
△ HOD.LOGLIN(auxmicxyz,auxcatxyz,xz−dist)
+ REG(cuxcorrxyz,xz−dist)
○ REG.LOGLIN(auxcatxyz,xz−dist)
▽ REG.LOGLIN(auxcorrxyz,auxcatxyz,xz−dist)

Proxy auxiliary information, true conditional correlation is .4

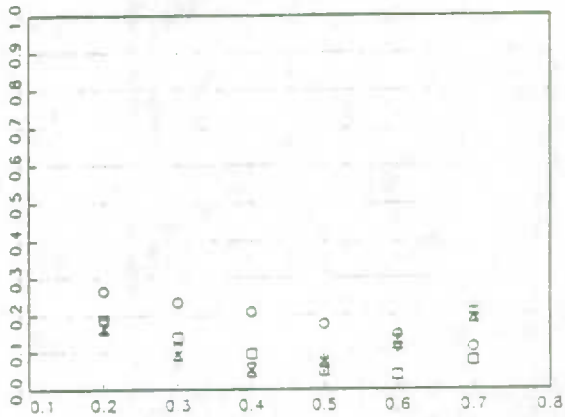
Plot G.1: MAD of z−values

Partition is equal probability



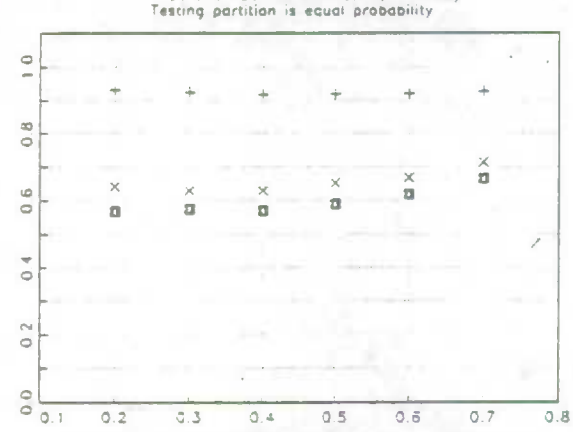
Plot G.2: MAD of covariances of y,zlx

Partition is equal probability



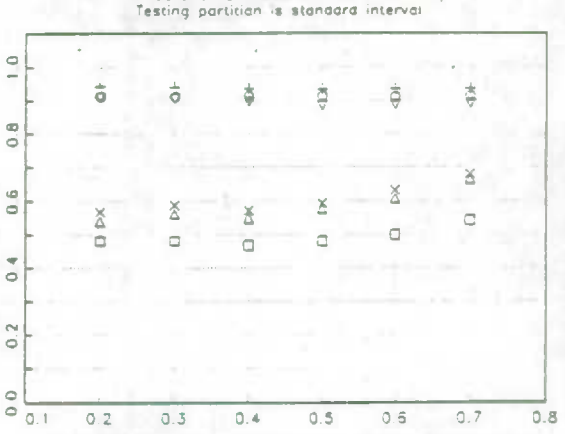
Plot G.3: Transformed mean chi−square statistics

Imputation partition is equal probability
Testing partition is equal probability



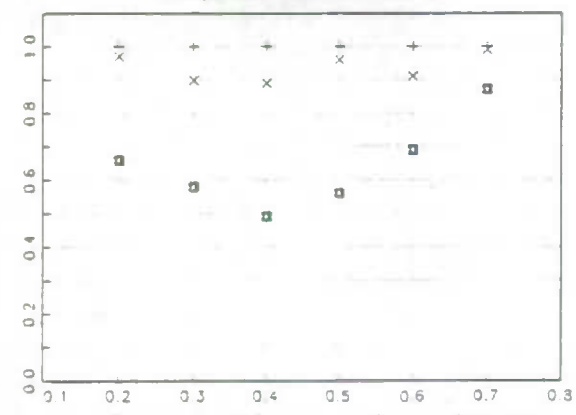
Plot G.4: Transformed mean chi−square statistics

Imputation partition is equal probability
Testing partition is standard interval



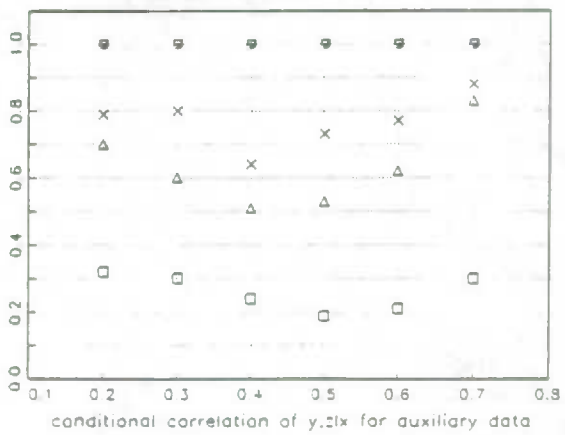
Plot G.5: Proportion of CLRT p−values < 0.05

Imputation partition is equal probability
Testing partition is equal probability



Plot G.6: Proportion of CLRT p−values < 0.05

Imputation partition is equal probability
Testing partition is standard interval

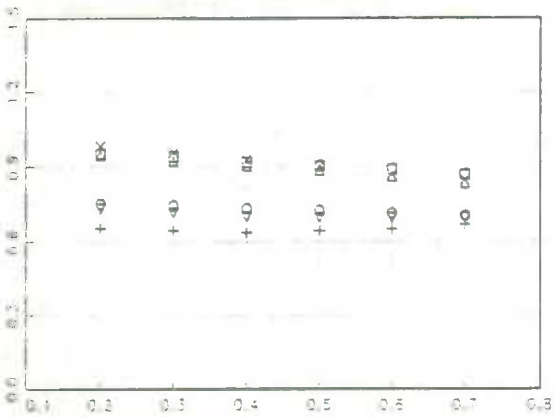


G: Methods using yz auxiliary information

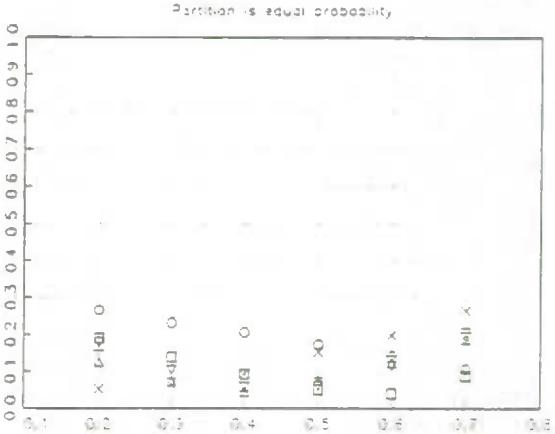
× HOD(auxmicyz,xz−dist)
□ HOD.LOGLIN(auxcatyz,x−dist)
△ HOD.LOGLIN(auxmicyz,auxcatyz,xz−dist)
\+ REG(auxcorryz,xz−dist)
○ REG.LOGLIN(auxcatyz,xz−dist)
▽ REG.LOGLIN(auxcarryz,auxcatyz,xz−dist)

Proxy auxiliary information, true conditional correlation is .4

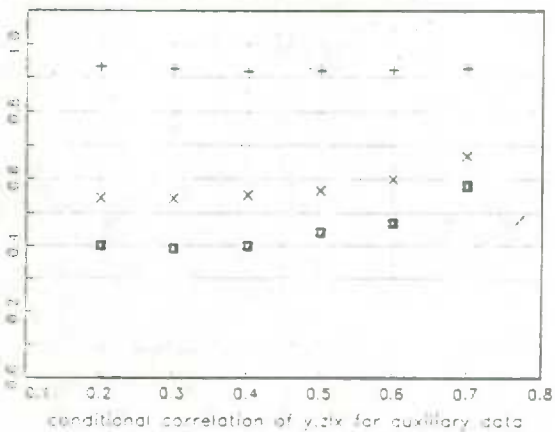
Plot G.7: MAD of p−values
Partition is equal probability



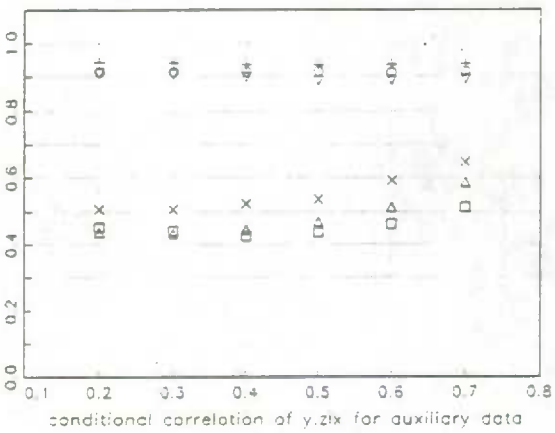
Plot G.8: MAD of covariances of y,z|x
Partition is equal probability



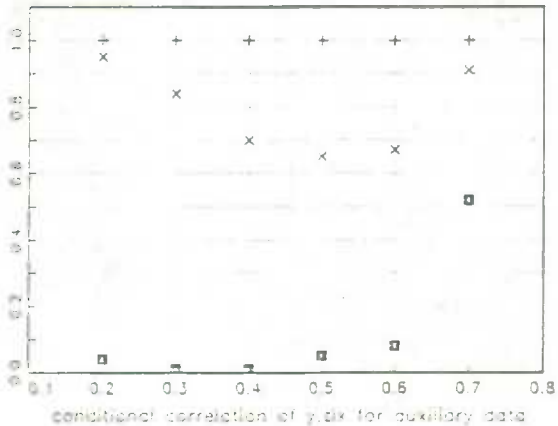
Plot G.9: Transformed mean chi−square statistics
Imputation partition is equal probability
Testing partition is equal probability



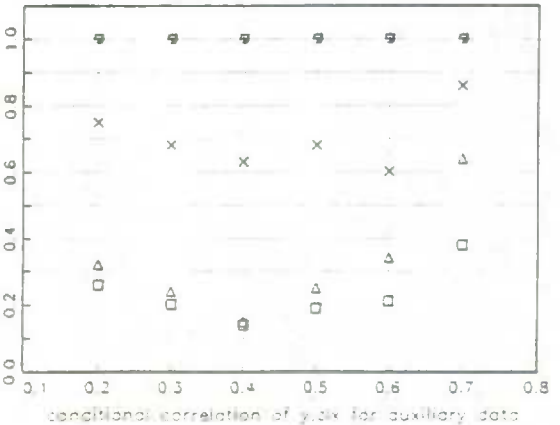
Plot G.10: Transformed mean chi−square statistics
Imputation partition is equal probability
Testing partition is standard interval



Plot G.11: Proportion of CLRT p−values < 0.05
Imputation partition is equal probability
Testing partition is equal probability



Plot G.12: Proportion of CLRT p−values < 0.05
Imputation partition is equal probability
Testing partition is standard interval

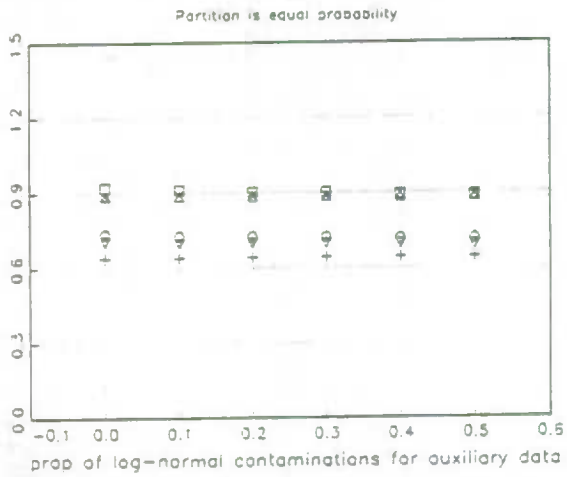


H: Methods using xyz auxiliary information

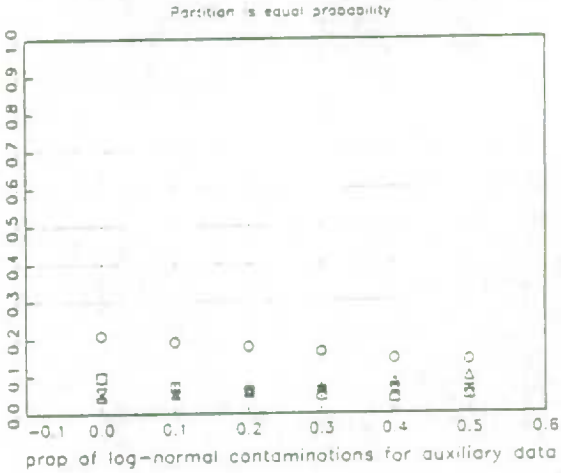
× HOD(auxmicxyz,xz−dist)
□ HOD.LOGLIN(auxcatxyz,x−dist)
△ HOD.LOGLIN(auxmicxyz,auxcatxyz,xz−dist)
\+ REG(auxcorrxyz,xz−dist)
○ REG.LOGLIN(auxcatxyz,xz−dist)
▽ REG.LOGLIN(auxcorrxyz,auxcatxyz,xz−dist)

Proxy auxiliary information, true distribution is normal, rho(y,z|x)=.4

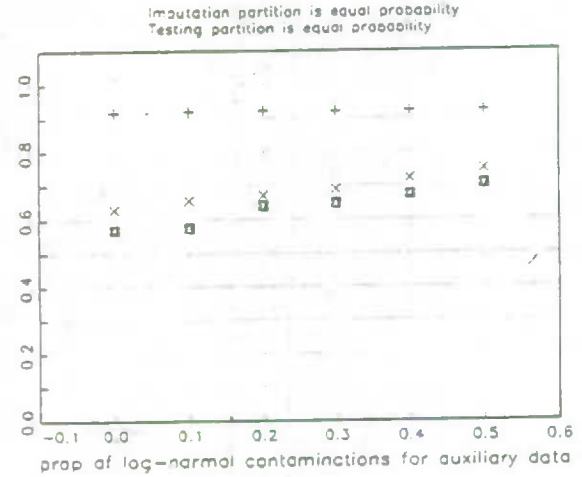
Plot H.1: MAD of z−values



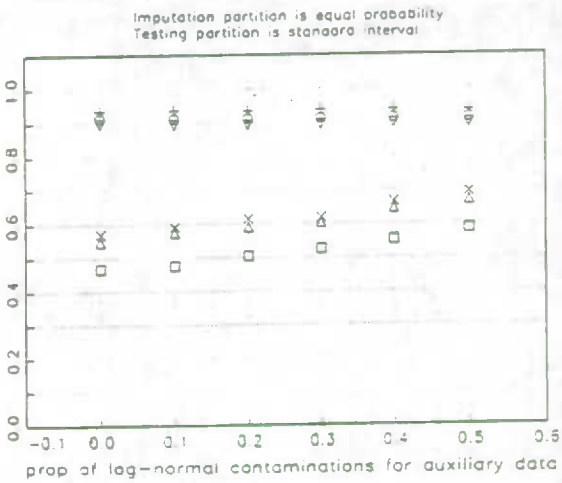
Plot H.2: MAD of covariances of y,z|x



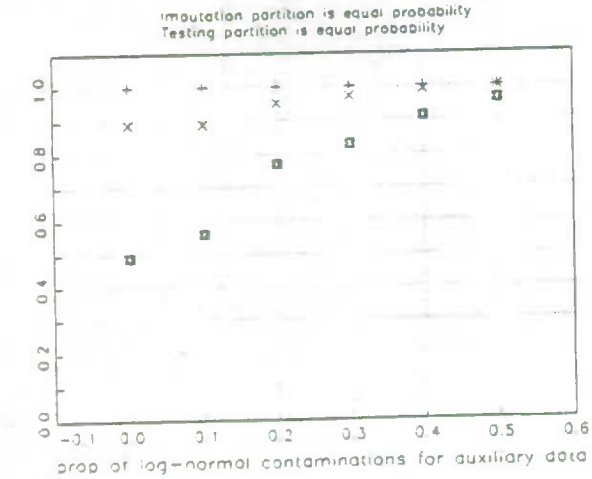
Plot H.3: Transformed mean chi−square statistics



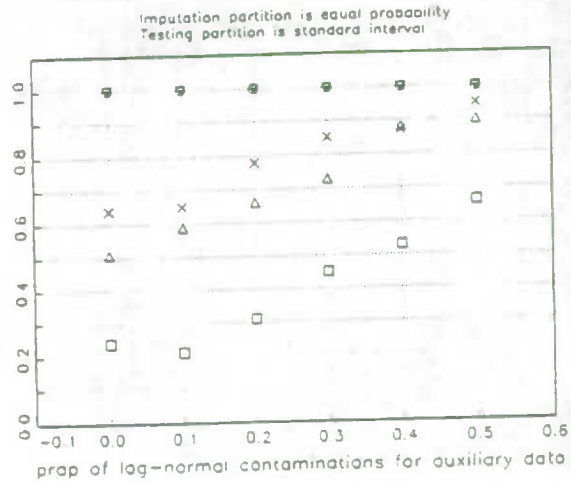
Plot H.4: Transformed mean chi−square statistics



Plot H.5: Proportion of CLRT p−values < 0.05



Plot H.6: Proportion of CLRT p−values < 0.05



H: Methods using yz auxiliary information

× HOD(auxmicyz,xz−dist)
□ HOD.LOGLIN(auxcatyz,x−dist)
△ HOD.LOGLIN(auxmicyz,auxcatyz,xz−dist)
\+ REG(auxcorryz,xz−dist)
○ REG.LOGLIN(auxcatyz,xz−dist)
▽ REG.LOGLIN(auxcorryz,auxcatyz,xz−dist)

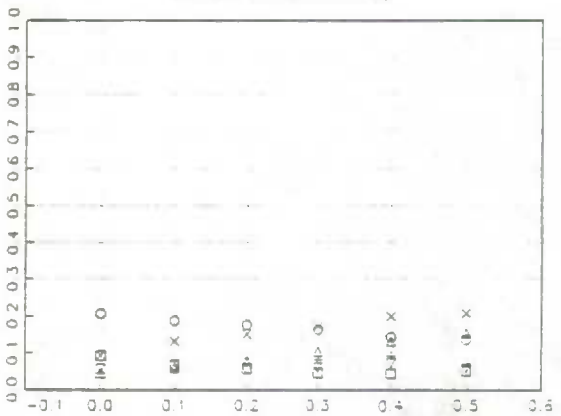
Proxy auxiliary information, true distribution is normal, rho(y,z|x)=.4

Plot H.7: MAD of z−values

Partition is equal probability

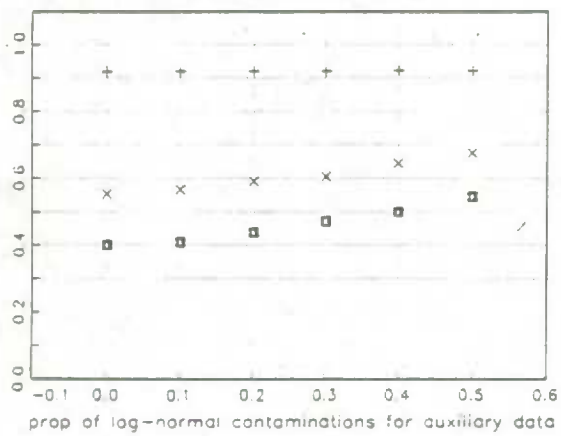
Plot H.8: MAD of covariances of y,z|x

Partition is equal probability



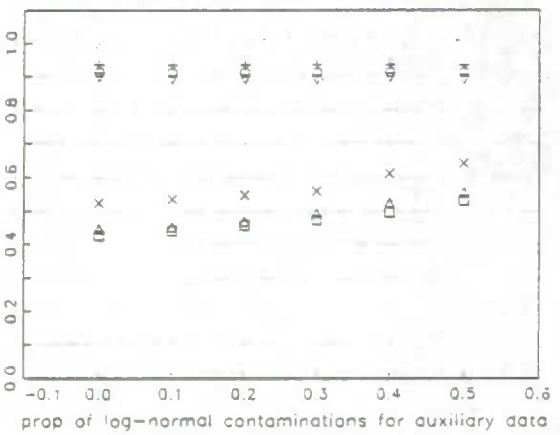
Plot H.9: Transformed mean chi−square statistics

Imputation partition is equal probability
Testing partition is equal probability



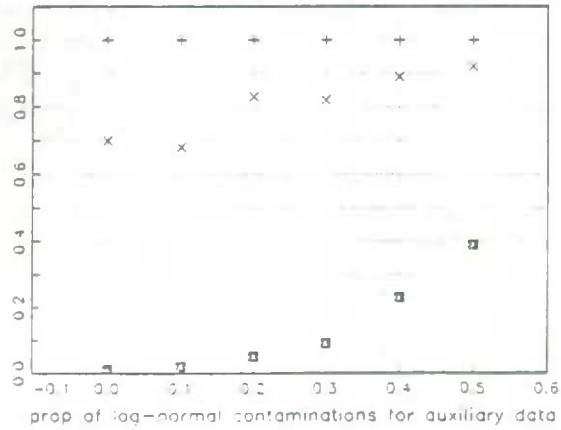
Plot H.10: Transformed mean chi−square statistics

Imputation partition is equal probability
Testing partition is standard interval



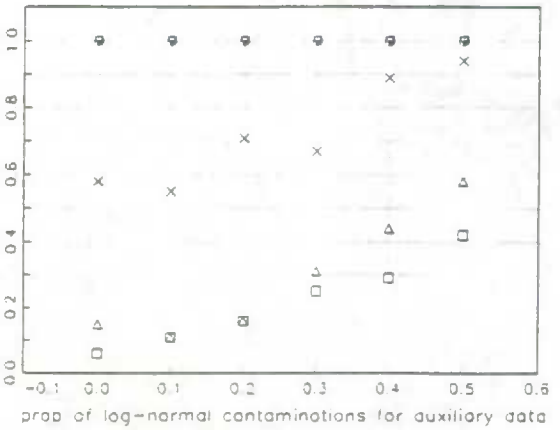
Plot H.11: Proportion of CLRT p−values < 0.05

Imputation partition is equal probability
Testing partition is equal probability



Plot H.12: Proportion of CLRT p−values < 0.05

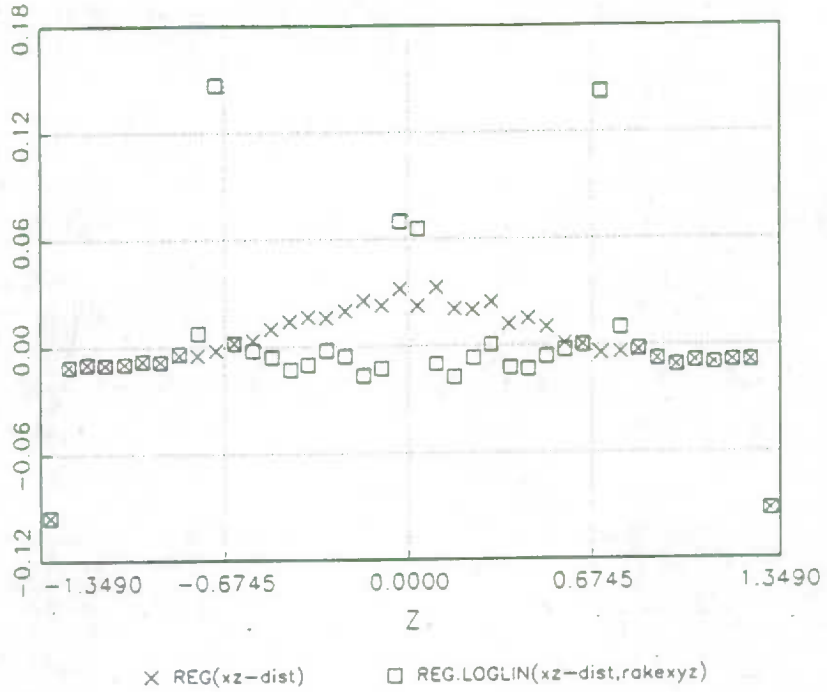
Imputation partition is equal probability
Testing partition is standard interval



I: Differences of Matched and Suppressed Z-histograms

Partition is equal probability, Data are normal, rho(y,z|x)=.4

Plot I.1: Methods with no Auxiliary Information



Plot I.2: Methods with (X,Y,Z) Auxiliary Information