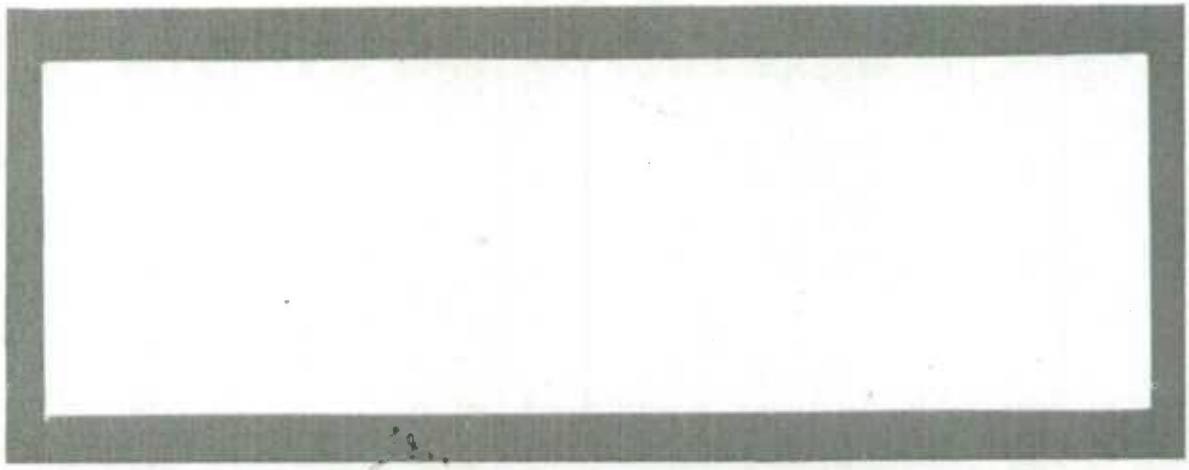
Statistics Canada Statistique Canada



Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

CATEGORICAL MATCHING AND CONSTRAINED REMATCHING OF SURVEY DATAFILES

HSMD-98-008E

Tzen-Ping Liu and Milorad S. Kovačević

Household Survey Methods Division
and
Social Survey Methods Division

Statistics Canada

October 1998

CATEGORICAL MATCHING AND CONSTRAINED REMATCHING OF SURVEY DATAFILES

Tzen-Ping Liu[1] and Milorad S. Kovačević[2]

ABSTRACT

This is a comprehensive article on statistical matching as a technique for integration of data from multiple sources at the micro level by identifying and linking records that correspond to similar individuals. In general, the matched file is obtained from two non-overlapping matching files, a host and a donor file, and is aimed at inference about the population that these files represent. The variables are divided into three groups: the common variables, the additional variables in the host file, and the additional variables in the donor file which one would like to impute onto the host file. However, the joint distribution of non-common variables that appear in two matching files is identifiable from the corresponding marginal distributions only under the assumption of independence of these sets of variables given the distribution of common variables. This assumption can be relaxed if auxiliary information on all variables (or on the additional variables only) is available. In this article we are interested in the use of two types of auxiliary information: the micro data and the categorical distributions of the variables in the matching files. Also we look at the impact of the reference periods of auxiliary information on the quality of matching by dealing with outdated and current information. We develop the methods that include the record weights (i.e., survey or sampling weights) in the matching process. Since the donor and the host record weights are usually different the methods for determining the 'survey' weights for the matched records are developed. A set of practical constraints, such as to use all records from both files, or to keep the size of the matched file within given limits, are fully respected. A method for dealing with non-overlapping ranges of common variables on the matching files, called 'backward' imputation, is also given. Once a matched file is obtained it still can be improved by imposing categorical constraints through an additional rematching or just by an adjustment of weights. A new algorithm for rematching called 'shift-and-share' is developed along with a series of procedures for estimating a 'look-up table' in the form of a joint categorical distribution.

Different methods and related techniques for statistical matching of survey data files are empirically investigated in a large scale simulation study based on the Public Use Micro Files from the 1986 and the 1991 Census for the province of Quebec, Canada. Pairs of non-overlapping matching files are generated by sampling independently from one or both censuses data files. In this simulation study, a complete set of forty-two possible combinations of matching and rematching methods, the method of ratio adjustment of record weights, and current and outdated auxiliary information are considered. In order to evaluate different aspects of the quality of statistical matching, several evaluation measures were developed and applied. It is shown that the quality of the matching lies in a fine classification of the records into the matching classes; in the use of rich and accurate auxiliary information; and in the appropriate use of survey weights. It is also shown that an already matched file can be improved again by some of the rematching techniques under additional categorical constraints. The usual ratio adjustment of record weights according to the categorical constraints could perform poorly. When auxiliary information is available, the modified distance matching method with backward imputation and reexamination by the rematching algorithm is recommended.

Keywords & Phrases: Nearest Available Matching; Pooling; Raking; Shift-and-Share Algorithm; Structural and Unexpected Empty Cells; Weight-Split.

---

1 Tzen-Ping Liu, Methodologist, Survey and Analysis Methods Development Section, Household Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6, Internet: tzenliu@statcan.ca

2 Milorad S. Kovačević, Methodology Research Advisor, Data Analysis Research, Social Survey Methods Division, Statistics Canada, Ottawa, Canada K1A 0T6, Internet: kovamil@statcan.ca

APPARIEMENT CATÉGORIQUE ET RÉ-APPARIEMENT AVEC CONTRAINTE DE FICHIERS DE DONNÉES D'ENQUÊTES

Tzen-Ping Liu[1] and Milorad S. Kovačević[2]

RÉSUMÉ

L'article qui suit fait état d'une étude exhaustive sur l'appariement statistique en tant que technique pour l'intégration de données au niveau micro par l'identification et le jumelage d'enregistrements correspondant à des individus possédant des caractéristiques semblables. En général, le fichier jumelé est obtenu à partir de deux fichiers d'appariement qui ne se chevauchent pas, un fichier donneur et un fichier hôte, en vue d'analyser la population représentée par ces fichiers. Les variables sont divisées en trois groupes: les variables communes, les variables additionnelles sur le fichier hôte, ainsi que les variables additionnelles sur le fichier donneur qui feront l'objet d'une imputation sur le fichier hôte. Cependant, la distribution conjointe des variables non communes qui apparaissent sur les deux fichiers d'appariement n'est identifiable à partir des distributions marginales correspondantes que sous l'hypothèse d'indépendance de ces ensembles de variables étant donné la distribution des variables communes. Cette hypothèse peut être simplifiée si de l'information auxiliaire sur toutes les variables (ou sur les variables additionnelles seulement) est disponible. Dans cet article, nous nous intéressons à deux types d'information auxiliaire: les micro-données et les distributions catégorielles des variables dans les fichiers d'appariement. Nous examinons aussi l'impact de la période de référence de l'information auxiliaire sur la qualité de l'appariement en faisant usage d'informations périmées et d'informations courantes. Nous développons des méthodes qui incluent les poids d'enregistrements (c.à.d. les poids d'enquête ou d'échantillonnage) dans le processus d'appariement. Comme les poids des enregistrements donneurs et hôtes sont habituellement différents , des méthodes pour déterminer les "poids d'enquête" des enregistrements jumelés sont mises au point. Un ensemble de contraintes pratiques telles que tous les enregistrements des deux fichiers sont utilisés, ou pour garder la taille du fichier apparié à l'intérieur de certaines limites, ont été suivies. Une procédure pour tenir compte des variables communes sur les fichiers à apparier dont les domaines ne se recoupent pas est également proposée, appelée imputation "à rebours" ("backward imputation"). Une fois qu'un fichier apparié a été obtenu il peut encore être amélioré en imposant des contraintes catégorielles par sur-appariement ("rematching") ou simplement en ajustant les poids. Un nouvel algorithme de sur-appariement appelé "modifier et partager" ("shift-and-share") est développé en même temps qu'une série de procédures pour l'estimation d'un tableau sommaire sous la forme d'une distribution conjointe de catégories ("joint categorical distribution").

Différentes méthodes et techniques pour l'appariement statistique de fichiers de données d'enquête sont examinées de façon empirique dans une étude de simulation étendue fondée sur les fichiers de micro-données à grande diffusion provenant des recensements de 1986 et 1991 pour la province de Québec, Canada. Des paires de fichiers d'appariement qui ne se chevauchent pas sont générées par échantillonnage indépendant à partir d'un seul ou des deux fichiers de données censitaires. Dans cette étude de simulation, un ensemble complet de quarante-devx combinaisons possibles des méthodes d'appariement et de sur-appariement, de la méthode d'ajustement par le quotient du poids des enregistrements ainsi que de l'information auxiliaire courante et périmée est considéré. Afin d'évaluer différents aspects de la qualité de l'appariement statistique, plusieurs mesures d'évaluation ont été développées et appliquées. Les auteurs montrent que la qualité du jumelage repose sur un arrangement pertinent des enregistrements en classes d'appariement; l'utilisation d'une information auxiliaire riche et précise; ainsi que l'usage approprié des poids d'enquêtes. On constate également qu'un fichier déjà apparié peut être amélioré par l'utilisation de certaines techniques de sur-appariement faisant appel à des contraintes catégorielles additionnelles. L'ajustement par le quotient habituel des poids des enregistrements selon les contraintes catégorielles pourrait ne pas donner les résultats escomptés. Lorsque de l'information auxiliaire est disponible, la méthode d'appariement à fonction de distance modifiée avec imputation à rebours et vérification additionnelle par l'algorithme de sur-appariement est recommandée.

---

1 Tzen-Ping Liu, Méthodologist, Section de développement des méthodes d'enquêtes et d'analyse, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Ottawa, Canada K1A 0T6, Internet: tzenliu@statcan.ca

2 Milorad S. Kovačević, Conseiller en recherche méthodologique, Recherche en analyse de données, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa, Canada K1A 0T6, Internet: kovamil@statcan.ca

TABLE OF CONTENTS

1. INTRODUCTION

Policy relevant analyses of tax and transfer programs, public health and welfare, educational attainment, etc., require comprehensive databases that are usually constituted from datafiles from different sources. These files typically contain very few or no individuals in common. Therefore, exact matching (record-linkage) which establishes the linkage of records from different files that belong to the same individual (unit) is not appropriate. Statistical matching of files, where records that correspond to similar individuals are identified and linked, is frequently used to produce comprehensive files of data from multiple sources.

For example the Canadian Social Policy Simulation Database (SPSD) at Statistics Canada was constructed to support micro analytic modelling by combining data from four major sources: survey data on family incomes and expenditures from the Canadian Survey of Consumer Finances (SCF) and the Canadian Family Expenditure Survey (FAMEX), with administrative data from the Canadian Personal Income Tax Returns (three percent sample of T1 returns) and the Canadian Unemployment Insurance Claim histories (one percent sample), (see Wolfson *et al.* 1987).

In general, the matched file is aimed at inference about the true joint distribution of all variables in it, so we expect that it represents the underlying population, and that the matching error induced by the matching procedure is within the sampling variation.

From a pair of non-overlapping matching files, the conditional joint distribution of the non-common variables given the common variables is identifiable from the corresponding marginal distributions only under the assumption of their conditional independence (CI) (Sims 1972). This assumption is often stressed in the literature on statistical matching. Ruggles, Ruggles and Wolf (1977), Barr, Stewart and Turner (1981), Rodgers and DeVol (1982) and Rubin (1986) give empirical evidence that violation of the conditional independence assumption may result in large errors. In order to overcome the CI assumption, Paass (1986) suggested using additional information in the form of an auxiliary micro data file and applying a certain iterative imputation procedure until some convergence criterion is met. Rubin (1986) proposed a regression method for statistical matching based on either macro or micro information about the relationship between variables involved in matching. Singh *et al.* (1993) considered both Rubin's and Paass's method when the auxiliary information is available in the form of a categorical distribution and proposed a loglinear modification of these methods based on a loglinear method of imputation as introduced by Singh (1988, 1989). The categorical distribution approach is a non-parametric treatment and can potentially recover a relationship between variables and weights. Previous research only used hypothetical data and generally ignored record weights. However, in most of synthetic databases, there are some source micro files that contain survey weights. Thus, a problem is how to weight records in a matched file when the matching records originally have different weights.

The objectives of this study are: (i) to examine whether the earlier findings with synthetic data, (Paass 1986 and Singh *et al.* 1993), hold in the real data case, that is, whether introducing additional constraints in terms of auxiliary categorical tables improves the quality of the matched file when the CI assumption is not valid; (ii) to examine whether the auxiliary information (variables or categorical tables) imposed on the imputation procedure generally improve the performance of all methods; (iii) to investigate the impact of using outdated auxiliary files on different matching strategies; (iv) to modify, adjust and develop the methodology for the categorical matching of records from sample survey files that contain different record weights; (v) to design and develop the methodology for categorically constrained rematching of records from already matched files that contain record weights; and (vi) to modify the methodology for imputation to improve the information contained in the resulting file.

The present simulation study uses data from the Public Use Micro File (PUMF), which created from the 1986 and 1991 Census 2B samples of Households/Housing for the province of Quebec. With respect to the four basic elements of a data set - *units*, *variables*, *weights*, and *reference periods* - the features of these initial data sets seem to be typical of files that have been used in actual statistical matching in the framework of SPSD.

The general framework for statistical matching of survey files is reviewed in Section 2. A number of specific

requirements and restrictions imposed on statistical matching for SPSD are listed and described. Sections 3 and 4 contain a variety of matching methods with the appropriate algorithms for easy implementation. A complete description of the empirical study along with the results and their interpretation is given in Section 5. The problem of evaluation of statistical matching is addressed in Section 6 and several different evaluation measures are presented. Analysis of the matched file, obtained in a matching procedure, is compared with analysis of the original file to evaluate and compare different matching procedures. Some specific remarks and conclusions are made in Section 7.

2. PROBLEM DEFINITION, CONCEPTS, ASSUMPTIONS AND CONSTRAINTS

We assume that a finite population $\wp$ has three groups of characteristics (variables) of interest, X, Y and Z. For unit i in $\wp$, the record $u_i = [X_i, Y_i, Z_i]$ (*i.e.*, a row vector) has a multivariate distribution with the mean $\mu = [\bar{X}, \bar{Y}, \bar{Z}]$ and the variance-covariance matrix partitioned as

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix} \tag{2.1}$$

In general, if we want to estimate a parameter θ which is an expectation of a function of the record u, say $g(u)$, we need information on the joint distribution or density function of the record u, $F(u)$ or $f(u)$:

$$\theta = E\{g(u)\} = \int g(u)dF(u) = \int g(u)f(u)du, \tag{2.2}$$

where $dF(u) = f(u)$. The density function $f(u)$ can be factorized as

$$f(u) = f(X,Y,Z) = \begin{cases} f(X,Y)f(Z|X,Y) & \text{or} \\ f(X,Z)f(Y|X,Z). \end{cases} \tag{2.3}$$

Now, suppose that we are unable to observe the record $u_i = [X_i, Y_i, Z_i]$ for any unit i in $\wp$. Suppose instead that two probability samples, A and B, from $\wp$ are available. One sample contains observations on the (X,Y) and the other on the (X,Z) variables. For practical purposes, we assume that these samples are obtained independently. In statistical matching terminology these samples are micro files and the sampled units are records. So, we have two files, $A = (X_i^A, Y_i^A, w_i^A)$, $i = 1,..., n^A$ and $B = (X_j^B, Z_j^B, w_j^B)$, $j = 1,..., n^B$, where the w's are the corresponding survey weights. Using these two files we can estimate the unknown mean vector μ, and all the components of the population variance-covariance matrix except for the component Σ_{YZ}.

Terms, $f(X,Y)$ and $f(X,Z)$, in the expression (2.3) for the probability density $f(X,Y,Z)$, can be estimated from files A and B, respectively. Terms, $f(Z|X,Y)$ and $f(Y|X,Z)$, however, cannot be estimated based on available information from A or B alone. One possibility is to assume that

$$f(Z|X,Y) = f(Z|X) \quad \text{or} \quad f(Y|X,Z) = f(Y|X). \tag{2.4}$$

This is equivalent to the assumption of CI of Y and Z given X, *i.e.*,

$$f(Y,Z|X) = f(Y|X)\,f(Z|X), \tag{2.5a}$$

which is equivalent to

$$F(Y,Z|X) = F(Y|X)\,F(Z|X). \tag{2.5b}$$

Under this assumption statistical matching involves only the common variables X. The matched datafile can then be used to estimate the unknown $F(u)$ which is required for estimation of θ. However, under the CI assumption the relationship between Y and Z in the matched file, when controlled for X, does not necessarily represent their true relationship in the population.

Similarly, when information on categories of samples A and B are available, we can estimate the categorical distributions (or tables for short) by $\{W_{X^*}\}$, $\{W_{Y^*}\}$, $\{W_{Z^*}\}$, $\{W_{X^*Y^*}\}$ and $\{W_{X^*Z^*}\}$, where stars indicate categories of the categorically transformed variables, and W_{Y^*} is, for example, the sum of weights of records in the category Y^*. Categorical distributions can be good approximations of the distributions $F(X)$, $F(Y)$, $F(Z)$, $F(X,Y)$ and $F(X,Z)$, respectively. However, the categorical distribution $\{W_{Y^*Z^*}\}$ or the distribution $F(Y,Z)$ is impossible to obtain from the separate files A and B.

The strong CI assumption can be dropped if auxiliary information on the (Y,Z) relationship is available. The role of auxiliary information is only to reduce the distortion of the joint distributions in the matched file. It is not aimed at any change of original values observed in the matching files.

There are two general types of auxiliary information, auxiliary micro data and macro level information. In the first case we assume the existence of an additional file C which contains either the full set of variables (X, Y, Z), or the reduced set (Y, Z). The idea is to incorporate information on the true relationship between Y and Z through some sort of a nonparametric regression. If macro level information on the relationship is available as a correlation coefficient then the regression matching method can be used (see Rubin 1986, Little and Rubin 1987, and Singh *et al.* 1993). This study does not include regression matching. Another form of macro level auxiliary information may be a categorical distribution (X^*, Y^*, Z^*) or (Y^*, Z^*) of the categorically transformed original variables (X, Y, Z) or (Y, Z). In that case, the auxiliary categorical distribution can be imposed to the matched file by raking. Our study is especially concerned with this type of auxiliary information.

Remark 2.1: When we deal with partially overlapped files, it is possible to combine a true linkage of the records from the overlapped parts and statistical matching of the non-overlapped parts, *i.e.*, a proxy linkage. As auxiliary information for statistical matching we may use the already linked overlapped part.

Usually there are some additional requirements that the matched file has to fulfil. For example, there are three specific requirements placed on the statistical matching in the SPSD:

(i) maintain the conditional distribution $\hat{F}(Z \mid X)$ as it is *estimated* by the donor file B (or with the least possible amount of distortion);
(ii) use all records from both files;
(iii) keep the size of the matched file under control, *i.e.*, allow the minimal possible inflation of the host file A.

The first task faces difficulties when the weights in the two files are different and when distortion of the distribution functions in the matched file is very likely. There are three general types of distortion: distortion in the marginal distribution of the Z variables; distortion in the joint distribution of the XZ and distortion in the joint distribution of XYZ. The first two affect the targeted conditional distribution $\hat{F}(Z \mid X)$ directly. On the other hand, the file B is a sample and $\hat{F}(Z)$ obtained from B is just an estimate of the population marginal distribution, so we allow the distortion of marginal distributions in the matched file to be within the sampling variation.

The second requirement is not usual in statistical matching where the primary objective is to complete the host file A, which implies that each A-record is matched with one or more B-records. In the case of the SPSD, however, we have to add the other direction as well, *i.e.*, that each B-record must be assigned to one or more A-records. It is important to mention that neither Rubin (1986), nor Paass (1986), nor Singh *et al.* (1993) considered matching under requirement (ii). This requirement comes from the actual matching for the SPSD and the

importance of information from the FAMEX file (in our study simulated as the B file), with the idea to transfer all available variability from the B file, (see e.g. Wolfson *et al.* 1987).

The third requirement, preservation of the size of the host file, comes from cost concerns: any further enlargement of the data base would increase costs of its maintenance and manipulation.

The second requirement, to use all records from the file B, is fully satisfied if the first one is; but the size of the matched file may be overinflated and this would contradicts the third requirement. Hence, an additional procedure is required to reduce the matched file when necessary. However, such a procedure might distort the distributions of the Z and Y variables. Fulfilling the third requirement is apparently at the cost of preservation of Z and Y distributions.

It is assumed that records from both files A and B, have *large positive weights*, and are classified into K matching classes ("pockets" or "blocks" in record linkage terminology or imputation classes in the practice of statistical imputation) identically defined for both files. The classification is made according to the common X variables which are either of the categorical type or categorically transformed. Records within the same class are to be matched. This two step strategy of forming matching classes first and then matching within the class is more efficient than matching on the whole data set. The numbers of records in a given matching class is generally not equal in the two matching files. Also, the corresponding totals of record weights are not equal.

3. MATCHING METHODS

Matching of survey data files consistent with requirements (i) - (iii) given in Section 2 is a multi-stage process consisting of imputation, weight assignment, file reduction and weight adjustment. We also consider an additional preliminary stage of pooling of record weights for files A and B before matching.

Imputation is commonly viewed as a technique for completing an incomplete data set so that standard data analysis methods can be applied. The purpose of imputation in the statistical matching procedure is the creation of a new file which contains X and Y values from A-records and Z values from B-records. The Z values are thought of as imputed. For matching methods based on the X distance, the Y variables may also be imputed into some of B-records (see Sections 3.1.1 and 3.2.1).

After imputation we assign a weight to the new records. Different imputation methods may imply different weight assignment procedures. The main criterion is the preservation or a minimal distortion of the distribution of Z variables from the file B.

This effort frequently results in the increased size of the host file. Any reduction of the file size necessarily leads to a redistribution of the total weight and results in weight adjustments.

In this report we discuss several different methods for statistical matching. Methods are classified into two main groups depending on whether they utilize auxiliary micro files or not. Within these groups there are methods with and without additional categorical constraints imposed on the Y and Z variables. The common characteristic of all of these methods is that the imputation procedure is of the hot-deck type. We define a hot-deck imputation procedure as one where the value imputed to a host record comes as a "live" value from a donor record that satisfies certain criteria, for instance having the minimum distance from the host record or belonging to the same class. Kovačević and Liu (1994) suggested some early ideas on these matching methods and weight assignment procedures. A new idea called backward imputation from host record to donor record will be introduced in Section 3.1.1.

In addition, we study possibilities for improving the quality of the matched file using additional categorical constraints derived from the matching files themselves. We may also use auxiliary categorical information on the variables of interest, if it is available. We assume that the original matching files are available together with the matched file. The idea is to improve the categorical distribution of the matched file $\{W^M_{X\cdot Y\cdot Z\cdot}\}$ by the iterative

adjustment of its margins to the categorical tables $\{W^A_{X^*Y^*}\}$ and $\{W^B_{X^*Z^*}\}$ of the matching files and then to a partial auxiliary categorical table $\{W^C_{Y^*Z^*}\}$. In this way we keep the categorical associations X^*Y^* and X^*Z^* from the matching files, and (if available) the categorical association Y^*Z^* from the partial auxiliary categorical table. We present two different ways of doing this: by ratio adjustment of record weights only, and by an additional partial rematch through the application of the "shift-and-share" algorithm introduced in Section 4 (also see Liu and Kovačević 1996b).

3.1. Matching Methods Based on the Conditional Independence Assumption

In the absence of auxiliary information, matching is based on comparison of values of the common variables X in files A and B, assuming conditional independence of the Y and Z variables for records in the same matching class, as defined earlier. Within each class, a distance function between recipient and donor records may take into account the X variables and, in addition, the record (survey) weights, w. For the sake of simplicity we will omit the class notation with understanding that everything is done at the matching class level.

If only X variables are considered, the distance functions may include either normalized X values or their absolute values. The use of different distance functions will lead to different matched files. The matching can be done using the 'fixed distance tolerance' or the 'nearest available' matching. In the first case, an upper bound for distance is prespecified and the closest record within that bound is the matching record for a given A-record. However, it may happen that there is no record within the bound and that some of the records may remain unmatched. Since we want to use all records from both files, the 'nearest available matching' is more appropriate for our study.

To allow record weights play a role in matching we propose imputation 'on rank' where the distance is defined between the relative cumulative weights (RCW), *i.e.*, the estimated cumulative distributions (see Sections 3.1.2 and 3.2.2). The resulting matching method is denoted as the weight-split procedure indicating a possibility of duplicating records and, consequently, splitting of their weights. Imputation 'on rank' is hard to implement when the common variable X is multivariate. A possible solution lies on sorting according to some predefined order of univariate components of variable, or on a suitable categorical transformation of the components of variable (such as the first principal component).

In the following, we describe the imputation methods mentioned above, along with procedures that allow use of a complete B file, the matched file reduction and the weight assignment. An alternative linnear programming approach is also presented.

3.1.1. Generalized X-Distance Method

In general, for each A-record, a B-record (or a set of B-records) is found such that the X-distance between them is minimal. Then, the Z values from that nearest B-record (from the 'nearest neighbour') are imputed into the corresponding A-record.

Remark 3.1: If there are more than one 'nearest neighbour' record we select one of them at random with either equal probability or proportionally to the record weight. In the experimental part of this study we used equal probability selection.

The X-distance between record $u^A_i = [X^A_i, Y^A_i, .., w^A_i]$ in file A and record $v^B_j = [X^B_j, .., Z^B_j, w^B_j]$ in file B is defined as

$$d_{ij} = ||u^A_i - v^B_j||_X = \begin{cases} \sqrt{[(X^A_i - X^B_j)V_X^{-1/2}]\,[(X^A_i - X^B_j)V_X^{-1/2}]'} & \text{or} \\ |\,[(X^A_i - X^B_j)V_X^{-1/2}]\,|, \end{cases} \tag{3.1}$$

where $||\,.\,||_X$ denotes the distance function based on X, V is a positive semi-definite matrix of the same dimension

as X, and $V_X^{-1/2}$ is the inverse of its Cholesky decomposition. The first formula gives the Euclidean distance, whereas the second describes the absolute distance which is equal to the sum of absolute values of all transformed elements. In particular, the Euclidean distance between records based on the common X variables is given as

$$d_{ij}^{e} = \sqrt{(X_i^A - X_j^B) S_x^{-1} (X_i^A - X_j^B)'}\,, \tag{3.2}$$

where S_x is the variance-covariance matrix for the X variables. It can be estimated from a pooled sample from A and B.

It may happen that some of the B-records are not used in the imputation phase which is in contradiction with the requirement (ii). To overcome this problem, from each leftover B-record we impute its Z value onto the nearest A-record. *Alternatively*, we may impute the Y value from the nearest A-record onto each leftover B-record. The second way is better because the resulting matched file contains all different values of X (sometimes many of the X values in the B file do not have close X values in the A file, or ranges of X values for files A and B are significantly different). The latter type of imputation we term the *backward* imputation, while imputation of a Z value from the B file onto the A file is called the *forward* imputation.

Remark 3.2: To enlarge the neighbourhood of a record i, a small tolerance t can be added to the observed minimum distance d_{ij}, so that other records j, that are within the distance $d_{ij}+t$ from i can become candidates for selection. In the experimental part of this study we let $t=0$. Alternatively, we can fix the precision level of the numerical process to increase the number of candidates for selection.

The weights of the matched records are those from the host file A with an adjustment for multiple (forward and also backward) imputed records. The multiple imputation here means that the Y_i from the same i^{th} A-record attended two or more matched records and each record included a Z_j from different j^{th} B-records, practically replicating it, say J_i times. The original weight of the A-record, w_i^A, has to be allocated proportionally to the corresponding B-records weights $\{w_{i_j}^B\}$, $j=1,...,J_i$. It gives the final weights $\{w_{ij}\}$ as

$$w_{ij} = w_i^A \; w_{i_j}^B \; / \sum_{k=1}^{J_i} w_{i_k}^B, \quad j=1,...,J_i\,. \tag{3.3}$$

After the imputation and weight adjustment processes the matched file is $M = \langle X_i^A \text{ or } X_j^B,\ Y_i^A,\ Z_j^B,\ w_{ij} \rangle$. Note that there was no need to reduce the size of the matched file since it takes the smallest size possible for given files A and B. This method preserves the distributions of X and Y variables from file A. The marginal distribution of Z and the conditional distribution of Z given X may be distorted. Note that we generated the weights of the matched records by equation (3.3) to keep the sum of weights and conditional distribution $Y|X$ of the host file A, and in an attempt to embed the conditional distribution $Z|X$ of donor file B into matched file M.

3.1.2. Generalized X-Rank Weight-Split Method

Let us assume that the records in the matching files are sorted with the respect to the X variable. The original record weights, $\{w_i^t\}$, $i=1,...,n^t$, $t\in\{A,B\}$, are normalized at the imputation class level, so that $\bar{w}_i^t = w_i^t / \sum_{j=1}^{n^t} w_j^t$, $i=1,...,n^t$, $t\in\{A,B\}$. The RCW of a record u_i is $F(u_i^t) = F_i^t = \sum_{j=1}^{i} \bar{w}_j^t$, $i=1,...,n^t$. It is calculated and attached to each record in the class. Records from both files in the same imputation class are ordered jointly according to the corresponding values of the cumulative function regardless of the file. Records in A which have the same X value, or records in B which have the same Y value, can further be ordered at random, if desired. The resulting sequence of RCW values, $\{F_i^t\}$, $t\in\{A,B\}$ is the RCW sequence for the matched file.

A method that uses information contained in both, the X values and the record weights, for matching is the weight-split (WS) matching method. The name comes from the fact that this method usually replicates some of the records from A and B, and, accordingly, splits their weights. The WS method produces simultaneously the matched records and their weights. However, this method uses only information about X-ranks of A and B records respectively, not their values. This method is based on imputation 'on rank' where a Z value from a B-record is imputed onto the A-record with the nearest value of relative cumulative weight (RCW). When records are sorted in ascending order with the respect to X the RCW is the cumulative distribution function (CDF).

The marginal distributions of X, Y and Z, as well as XY are preserved in the resulting matched file. The WS matching procedure has two stages. The first stage is imputation 'on rank'. The next is the computation of relative weights for the records in the matched file. We add two more steps in order to satisfy conditions about the size of the matched file. The third step includes file reduction and is followed by weight adjustment.

In the following we present the algorithm of the WS method.

3.1.2.1. Weight-Split Algorithm

The modified imputation 'on rank' is done in two steps.

Step 1 (Downward step): For each B-record v_j^B impute the Z_j values to all A-records u_i^A for which $F_{j-1}^B < F_i^A \le F_j^B$, $i=1,...,n^A$ and $j=1,...,n^B$.

Step 2 (Upward step): For each B-record v_j^B for which there is no corresponding A-record u_i^A with $F_i^A = F_j^B$, impute the Z_j values to the first A-record u_i^A for which $F_j^B < F_i^A$.

The 'downward' step imputes a Z value from one B-record onto several A-records. The 'upward' step makes possible the imputation of Z values from several B-records to the same A-record.

The total number of records in the matched file is $n = n^A + n^B - T$, where T denotes the number of ties, *i.e.*, number of records with $F_i^A = F_j^B$.

After imputation, each matched record has a pair of RCWs, one coming from A, the other from B. The RCW assigned to a matched record is determined as $F_{ij} = min\{F_i^A, F_j^B\}$, *i.e.*, the RCW of a matched record is the smaller of two corresponding original cumulative weights. The relative weight of the i^{th} record in a matched file is the difference between two successive F values: $\tilde{w}_i = F_i - F_{i-1}$, where $\tilde{w}_1 = F_1$.

To show that the marginal distributions of XY and Z, as well as values of (X,Y) are preserved in the matched file we do the following:

Let F_i^A be the RCW corresponding to (X_i, Y_i) in the i^{th} record of the A file. We will show that for the same X and Y values in the matched file the value of the RCW stays the same. The last imputation of the Z value is always downward imputation, that is $F_i^A \le F_j^B$, and therefore $F_{ij}^M = min\{F_i^A, F_j^B\} = F_i^A$.

Similarly, we show that the marginal distribution of Z is preserved in the matched file. Let F_j^B be the RCW value for the (X_j, Z_j) record. The imputation of a Z value always ends up in the A record with the same or larger RCW value. Thus, if Z_j is assigned to the u_i^A record then $F_j^B \le F_i^A$ and the corresponding RCW of the matched record is $F_{ij}^M = min\{F_i^A, F_j^B\} = F_j^B$.

Although the imputation 'on rank' preserves the marginal distributions it has some practical drawbacks. It may happen that, as a result of numerical rounding errors during the computation of cumulate weights, a relative weight difference is very small (say, w_{ij}<<0.1) instead of being zero. In such a case we discard a "light" record. This "sifting" procedure may be performed later in the final weight adjustment. The size of the matched file in this stage is usually very large. If it is larger than needed, we perform the file size reduction procedure. Note that a light record is always a replicate of a matched record with normal weight.

3.1.2.2. File Size Reduction Procedure

To describe this procedure, we need a few additional definitions:

A match (matched record) is an A-single match when an A-record involved in matching is matched with one and only one B-record. Otherwise, the match is an A-multiple match.

The kernel of a matched file is its subfile which contains all A-single matches and matches with the minimal values of $d_{ij} = ||u_i^A - v_j^B||_X$ from the A-multiple matches. If there is more than one minimum distanced record we select one randomly. Evidently, the size of the kernel is n^A.

The file reduction procedure forms a kernel of the matched file and takes care of fulfilling the requirement of using all records from B:

(i) Discard (delete) all *light* records w_{ij}(<<0.1) from the A-multiple records providing that none of the A-records is eliminated entirely.

(ii) Form a kernel of the matched file, *i.e.*, for each A-record take a match with the minimum distance d_{ij}.

(iii) Check if all B-records were used for the kernel creation. If they are, the matched file is the kernel itself. If not, for each unused B-record find an A-record such that the distance d_{ij} is minimal. If there is more than one minimum distanced record select one at random. Append these records to the kernel. The new file is the final matched file.

3.1.2.3. Final Weight Adjustment

To obtain the final weights in the matched file, we have to adjust the weights obtained after the imputation stage. Let us assume that after the imputation a matched record is $[X_i^A, Y_i^A, Z_j^B, w_{ij}]$, indicating that the Z value from the j^{th} B-record has been imputed to the i^{th} A-record. Its weight w_{ij} has the property $\sum_{j=1}^{J_i} w_{ij} = w_i^A$, where J_i is the number of matches in which the i^{th} A-record participates. After reducing the intermediate matched file, there are J_i' $(1 \le J_i' \le J_i)$ records left. Therefore, the adjusted relative weight of a record in the matched file is

$$w_{ij}' = w_i^A \, w_{ij} / \sum_{k=1}^{J_i'} w_{ik}, \quad j = 1,...,J_i' \,. \tag{3.4}$$

After imputation, file reduction, and weight adjustment the matched file is $M = (X_i^A, Y_i^A, Z_j^B, w_{ij}')$.

3.1.3. An Linnear Programming Approach

To preserve both sets of marginal distributions in the matched file the following conditions must be met

$$\sum_{j=1}^{n^B} w_{ij} = w_i^A, \; i = 1,...,n^A, \text{ and } \sum_{i=1}^{n^A} w_{ij} = w_j^B, \; j = 1,...,n^B, \tag{3.5}$$

which implies total weights $\sum_{i=1}^{n^A} w_i^A = \sum_{j=1}^{n^B} w_j^B$. A matching strategy that satisfies (3.5) is a solution of a linear programming 'transportation' problem (Goel and Ramalingam 1989) where records of one file are 'producers' and records of another are 'consumers'. The cost of transportation is the distance d_{ij} between records and the weights are 'capacities' of producing and consuming, respectively. The objective function of the problem is the total weighted distance

$$f = \sum_{i,j} w_{ij} d_{ij} \tag{3.6}$$

which has to be minimized under constraints (3.5). Conditions (3.5) allow the distribution of Z variables to be precisely replicated in the matched file as that observed in file B. The matched file M keeps the joint and conditional distributions of XY and $Y|X$ from file A, the joint and conditional distributions of XZ and $Z|X$ from file B, but there is no control on the joint distribution of YZ.

One can use existing algorithms from linear programming to solve the problem of statistical matching. However, the implementation of such algorithms may be difficult when data files are large (as they usually are in matching problems). Also, the number of records in the matched file may be as large as $n^A \cdot n^B$, where n^A and n^B are the original file sizes, respectively.

3.2. Matching Methods When an Auxiliary Data File is Available

The underlying assumption for all matching methods in Section 3.1 was the independence of variables Y and Z given the information on the common variable X. If this assumption is not sustainable then the resulting matched file may represent a biased relationship between variables.

The use of auxiliary information to avoid the CI assumption was proposed by Rubin (1986) in the context of statistical matching via the regression method. There, auxiliary information takes a 'macro' form through the correlation coefficients ρ_{YZ} or $\rho_{YZ|X}$ which enable finding a regression equation of Z on Y or on X, respectively. From the regression equation we obtain the intermediate Z value which is then used to find a live Z by some hot-deck imputation method. Finding the appropriate regression equation for Z may not be an easy task especially when Z is multivariate. Also, the derivation of the correlation coefficients depends on the form of the available auxiliary information. If it is in the form of a micro file we condense it into 'macro' form, but if it is given in the form of a categorical table then Rubin's method is not directly applicable.

Paass's method (1986) is also a two stage imputation procedure. First, an intermediate Z value from the auxiliary data file C is found using some hot-deck method of imputation. Then, a live Z value from the B file is found as a nearest neighbour to the intermediate Z value. Paass originally proposed finding the K nearest neighbours from the auxiliary file and then singling out a live Z value from the B file as a match.

Here we assume availability of an auxiliary data file, say C, which contains records with variables (X,Y,Z), or just (Y,Z), along with their survey weights, $C = \langle X_i^C, Y_i^C, Z_i^C, w_i^C \rangle$ or $C = \langle Y_i^C, Z_i^C, w_i^C \rangle$. We distinguish two general methods depending on whether the distance is measured between the observed values of studied variables in two files or between the corresponding RCW values. Again, we see a matching method as a sequence of procedures: intermediate imputation, imputation, weight assignment and adjustment, and file reduction. Some steps which are identical to those of Section 3.1 will be omitted from the presentation.

Even when auxiliary data are available the quality of information may be a problem. Although Singh *et al.* (1993) suggested that auxiliary information need not be a perfect, a certain caution is necessary when such information comes from an outdated source. Usually the outdated auxiliary data file has to be adjusted to the range of the current data. For example, the data on income or consumption may be multiplied by an appropriate inflation factor.

3.2.1. Generalized (X,Y,Z)-Distance Method

The first step of this distance matching method is to identify the nearest neighbours in files A and C using a distance function. For the available full auxiliary file $C = \langle X_i^C, Y_i^C, Z_i^C, w_i^C \rangle$, the distance is

$$d_{ik} = ||u_i^A - v_k^C||_{XY} = \begin{cases} \sqrt{[([X_i^A,Y_i^A] - [X_k^C,Y_k^C])V_{[X,Y]}^{-1/2}]\,[([X_i^A,Y_i^A] - [X_k^C,Y_k^C])V_{[X,Y]}^{-1/2}]'} & \text{or} \\ |([X_i^A,Y_i^A] - [X_k^C,Y_k^C])V_{[X,Y]}^{-1/2}|\,, \end{cases} \tag{3.7}$$

where $||.||_{XY}$ means a distance function based on (X,Y), and $V_{[XY]}$ is a positive semi-definite matrix with the same dimension as (X,Y). The first row denotes the Euclidean distance while the second is characteristic of the absolute distance which is equal to the sum of absolute values of all transformed elements. When only a partial auxiliary file $(C = \langle Y_i^C, Z_i^C, w_i^C \rangle)$ is available, the distance function is

$$d_{ik} = ||u_i^A - v_k^C||_Y = \begin{cases} \sqrt{[(Y_i^A - Y_k^C)V_Y^{-1/2}]\,[(Y_i^A - Y_k^C)V_Y^{-1/2}]'} & \text{or} \\ |[(Y_i^A - Y_k^C)V_Y^{-1/2}]|\,, \end{cases} \tag{3.8}$$

where $||.||_Y$ means a distance function based only on Y, and V_Y is a positive semi-definite matrix with the same dimension as Y. In both cases we form an intermediate file $I = \langle X_i^A, Y_i^A, Z_k^C, w_i^A \rangle$. We keep the weights and the size of the A file.

The next step is the matching of the intermediate file I and the donor file B. The variables in common are X and Z. The distance function is

$$d_{ijk} = ||u_{ik}^I - v_j^B||_{XZ} = \begin{cases} \sqrt{[([X_i^A,Z_k^C] - [X_j^B,Z_j^B])V_{[X,Z]}^{-1/2}]\,[([X_i^A,Z_k^C] - [X_j^B,Z_j^B])V_{[X,Z]}^{-1/2}]'} & \text{or} \\ |([X_i^A,Z_k^C] - [X_j^B,Z_j^B])V_{[X,Z]}^{-1/2}|\,, \end{cases} \tag{3.9}$$

where $||.||_{XZ}$ means a distance function based on (X,Z), and $V_{[XZ]}$ is a positive semi-definite matrix with the same dimension as (X,Z).

In particular, the variance-covariance matrices S for (X,Y), Y and (X,Z) are used as the corresponding V matrices, in the distance function above.

Final weights are obtained as explained in Section 3.1.1. After intermediate imputation, final imputation, and weight adjustment the matched file is $M = \langle X_i^A \text{ or } X_j^B,\ Y_i^A,\ Z_j^B,\ Z_k^C,\ w_{ij} \rangle$. In general, we may not keep the intermediate Z_k^C in the final version of the matched file M.

Remark 3.3. The number of records matched by the backward imputation in distance matching is rather small. From the empirical study, this number accounts for 10% of the donor file B and 2% of the host file A. However, it contributes to efficient utilization of information on all variables and distributions included X. The minimum possible number of the backward imputations is zero and the maximum possible number is $\{n^B$- (number of clusters of $B)\}$. For further discussion see Section 7 and Table 7.1.

3.2.2. Generalized (X,Y,Z)-Rank Weight-Split Method

Here, as in the previous method, we have a two step procedure. The first step is an intermediate imputation, from C to A, and the second is the matching of the intermediate file and the donor file B.

Both imputations, from C to A and from I to B, are done at the points of the nearest RCW values. The RCW function is obtained as follows: records in both files are ordered by common variables. This means that the first sorting is done by X variables and if there are two or more records with the same value of X we sort them by Y variables. The regular RCW is computed by adding the weights of successive records. The impact of the Y variables can be enhanced by using it in forming the matching classes for the intermediate matching.

The size and the weights of the intermediate file $I = \langle X_i^A, Y_i^A, Z_k^C, w_i^A \rangle$ are kept as those of file A. If the auxiliary file C does not contain X variables, the intermediate matching is accomplished by matching on the Y variables only.

The next step is the WS matching of I and B. The variables in common are X, Z. We first order the files according to the Z variables and then by X variables at the level of matching class. The imputation is done and weights are obtained in the way explained in Section 3.1.2.1. The resulting file is $\langle X_i^A, Y_i^A, Z_j^B, Z_k^C, w_{ij} \rangle$.

In order to reduce the number of records and to make use of all of B-records we perform the file reduction procedure as given in Section 3.1.2.2. The distance we use in I-multiple reduction is $d_{ijk} = ||u_{ik}^I - v_j^B||_{XZ}$. The final weights of the records in the matched file are determined by the adjustment procedure presented in Section 3.1.2.3. The final matched file is $M = \langle X_i^A, Y_i^A, Z_j^B, Z_k^C, w_{ij}^I \rangle$. Here, as in the case of distance matching, we may drop the intermediate Z_k^C in the final version of the matched file M.

4. CATEGORICALLY CONSTRAINED REMATCHING AND ADJUSTMENT METHODS

Singh *et al.* (1993) proposed modifications to Rubin's (1986) and Paass's (1986) methods by imposing categorical constraints on the Z values imputed from the B file. Categorical constraints are aimed at preserving the categorical associations from the matching files in the matched file. In this section we assume that a categorical matched file is obtained, for example, using some of the methods described in Section 2. Imposing of categorical constraints onto a matched file is aimed at its improvement. Liu (1998) used a similiar algorithm for construction of a matched file from two matching files.

In general, categorically constrained improvement of matching consists of the following three steps:

(i) transform the variables involved in matching X, Y, Z, into the categorical variables X^*, Y^*, Z^*, using some criteria for optimal partition (see Singh *et al.* 1988), or according to the available auxiliary categorical information, and then

(ii) estimate the joint categorical distribution of X^*, Y^* and Z^* by raking the categorical distribution of the matched file M to the available and the adjusted distributions $\{ W_{X^*Y^*}^A \}$, $\{ W_{X^*Z^*}^B \}$ and (if available) to the auxiliary distribution $\{ W_{Y^*Z^*}^C \}$. We call the estimated categorical distribution a look-up table. It is important to note that we do not use auxiliary distributions on X^*Y^* or X^*Z^* since we would like to maintain these joint margins as observed in files A and B.

(iii) Once the distribution of $X^*Y^*Z^*$ is estimated, we may adjust the individual weights of the records in the matched file, or first perform a partial rematching to satisfy the imposed constraints and then adjust the individual weights where needed.

4.1. Estimation of the Joint Categorical Distribution (The Look-up Table $\{ W_{X^*Y^*Z^*}^L \}$)

We assume that a suitable and a unique categorization of the X, Y, and Z variables is done for all data files involved. Due to a possibly large number of categories, an iterative procedure for estimation of the joint categorical distribution of $X^*Y^*Z^*$ may be lengthy and may require extra computer efficiency. To make the procedure

convergent and fast we propose the following steps. First, to balance the X^* margins of the participating files A and B. Second, if auxiliary categorical information is available, to equalize its Y^* and Z^* margins to the corresponding adjusted margins of the matching files. Third, the 'unexpected' empty cells in the matched file M will cause a non-convergence problem unless treated appropriately. We provide the algorithm for doing so. Finally, the look-up table is obtained by the 'raking' (iterative proportional adjustment) of the margins of the matched file M, modified for unexpected empty cells, to the balanced margins of A, B and C (if available).

4.1.1. Balancing the Categorical Margins of A, B and C

After categorization of the matching files A and B, it is likely that the sums of weights in the corresponding X^* categories are not the same, *i.e.*, $W^A_{X^*} \neq W^B_{X^*}$, and that convergence of the raking procedure is not possible. We investigated two principal ways of initial marginal balancing: pooling the sums of weights of the two files at the level of the X^* category, or alternatively, marginal adjustment by means of raking.

The idea of 'pooling' sums of weights of two files at the level of a X^* category lies essentially in a combination of the sums of weights $W^A_{X^*}$ and $W^B_{X^*}$ to obtain $W^{A_p}_{X^*} = W^{B_p}_{X^*}$, ($p$ stands for 'pooled').
First:

$$W'^{A_p}_{X^*} = W'^{B_p}_{X^*} = \alpha_{X^*} W^A_{X^*} + (1 - \alpha_{X^*}) W^B_{X^*}, \tag{4.1a}$$

and then

$$W^{A_p}_{X^*} = W^{B_p}_{X^*} = W'^{A_p}_{X^*} W / \sum_{\{X^*\}} W'^{A_p}_{X^*}, \tag{4.1b}$$

where $0 \le \alpha_{X^*} \le 1$. An extra ratio adjustment in (4.1b) is needed so that the pooled categorical sums of weights add up to the original total weight, $W = W^A = W^B$. Note that if the pooling coefficient α_{X^*} is constant over the $\{X^*\}$ categories the ratio in (4.1b) is equal to 1.

There are several options for α_{X^*}: $\alpha_{X^*} = n^A (n^A + n^B)^{-1}$, $\alpha_{X^*} = n^A_{X^*} (n^A_{X^*} + n^B_{X^*})^{-1}$, $\alpha_{X^*} = 0$, $\alpha_{X^*} = 1/2$ or $\alpha_{X^*} = 1$, to mention a few. Here n^A and $n^A_{X^*}$ denote the size of the file A and the size of the category X^* of the file A, respectively (similarly for n^B and $n^B_{X^*}$). In the experimental part of this study we used pooling according to a category size.

Further modification of sums of weights is at the level of the X^*Y^* categories for the file A and of the X^*Z^* categories for the file B:

$$W^{A_p}_{X^*Y^*} = W^A_{X^*Y^*} W^{A_p}_{X^*} / W^A_{X^*} \quad \text{and} \quad W^{B_p}_{X^*Z^*} = W^B_{X^*Z^*} W^{B_p}_{X^*} / W^B_{X^*}. \tag{4.2}$$

If auxiliary information is available as a distribution $\{W^C_{X^*Y^*Z^*}\}$ or $\{W^C_{Y^*Z^*}\}$ we need to adjust its margins: $\{W^C_{Y^*}\}$ to the $\{W^{A_p}_{Y^*}\}$, and $\{W^C_{Z^*}\}$ to the $\{W^{B_p}_{Z^*}\}$, so that $W^{C_p}_{Y^*} = W^{A_p}_{Y^*}$ and $W^{C_p}_{Z^*} = W^{B_p}_{Z^*}$:

$${}_1W^{C_p}_{Y^*Z^*}(i) = W^{C_p}_{Y^*Z^*}(i-1) / W^{C_p}_{Y^*}(i-1) \; W^{A_p}_{Y^*} \tag{4.3a}$$

and

$$W^{C_p}_{Y^*Z^*}(i) = {}_1W^{C_p}_{Y^*Z^*}(i) / {}_1W^{C_p}_{Z^*}(i) \; W^{B_p}_{Z^*} \tag{4.3b}$$

with $W^{C_p}_{Y^*}(i-1) = \sum_{\{Z^*\}} W^{C_p}_{Y^*Z^*}(i-1)$, ${}_1W^{C_p}_{Z^*}(i) = \sum_{\{Y^*\}} {}_1W^{C_p}_{Y^*Z^*}(i)$ and $W^{C_p}_{Y^*Z^*}(0) = W^C_{Y^*Z^*}$. We repeat steps (4.3) until $\max_{Y^*} \{ | W^{C_p}_{Y^*} - W^{A_p}_{Y^*} | \} \le \epsilon$ and $\max_{Z^*} \{ | W^{C_p}_{Z^*} - W^{B_p}_{Z^*} | \} \le \epsilon$. The threshold ϵ has to be small in order not to disturb

later steps. In the experimental part of this study we used $\epsilon = 10^{-5}$. Note that the margins of the A and B files used in the adjustments (4.3) are already modified by pooling or by raking.

So far we prepared matching files and an auxiliary table for the future raking of the matched file.

4.1.2. Structural and Unexpected Empty Cells

We observe that some cells in all three categorical distributions (tables) $\{W^A_{X^*Y^*}\}$, $\{W^B_{X^*Z^*}\}$ and $\{W^M_{X^*Y^*Z^*}\}$ may be empty which directly leads to an empty cell after adjustment, but may also cause non-convergence of the algorithm. In that sense we distinguish two possible types of empty cells in the matched file M: the 'structural' empty cell, and the 'unexpected' empty cell.

The structural empty cell refers to the situation where $W^M_{X^*Y^*Z^*} = 0$ and at least one of the corresponding $W^A_{X^*Y^*}$ and $W^B_{X^*Z^*}$ is equal to zero. It doesn't cause any problem during the raking procedure. The second type is a more difficult case. Here $W^M_{X^*Y^*Z^*} = 0$ but none of the corresponding cells in A and B is empty. It may lead to non-convergence of the raking algorithm.

In order to overcome the problem of unexpected empty cells and provide convergence of the raking procedure we do the following: First, increase all cell sums of weights in the matched file, except the structural zeros, by a positive small number δ, so that

$$W^{M'}_{X^*Y^*Z^*} = \begin{cases} 0, \text{ if } X^*Y^*Z^* \text{ is a structural empty cell,} \\ W^M_{X^*Y^*Z^*} + \delta, \text{ elsewhere.} \end{cases} \tag{4.4}$$

In the experimental part of this study we used a minimal record weight 33.333, (which is the self-weighting factor of 1991 Census PUMF records) as δ. Second, since we added this positive (small) number to almost all cells in the matched file, the total weight of the matched file is increased and has to be adjusted back to the original weight. We apply ratio adjustment at the level of $X^*Y^*Z^*$:

$$W^I_{X^*Y^*Z^*} = W^{M'}_{X^*Y^*Z^*} \; W^M \; / \sum_{(X^*Y^*Z^*)} W^{M'}_{X^*Y^*Z^*} \; . \tag{4.5}$$

In this way the total weight remains the same as in the original M file and the raking procedure converges. The new sum of weights $W^I_{X^*Y^*Z^*}$ is the 'initial' sum of weights for the next step.

4.1.3. Raking of the Joint Categorical Distribution of the Matched File

The last step in providing the look-up table is the raking of the margins of the categorized matched file, already corrected for the unexpected zeros, to the balanced X^*Y and Y^*Z^* margins of the matching files:

$${}_1W^L_{X^*Y^*Z^*}(i) = W^L_{X^*Y^*Z^*}(i-1) \; / \; W^L_{X^*Y^*}(i-1) \; W^{A_p}_{X^*Y^*} \tag{4.6a}$$

and

$$W^L_{X^*Y^*Z^*}(i) = {}_1W^L_{X^*Y^*Z^*}(i) \; / \; {}_1W^L_{X^*Z^*}(i) \; W^{B_p}_{X^*Z^*} \tag{4.6b}$$

with $W^L_{X^*Y^*}(i-1) = \sum_{(Z^*)} W^L_{X^*Y^*Z^*}(i-1)$, ${}_1W^L_{X^*Z^*}(i) = \sum_{(Y^*)} {}_1W^L_{X^*Y^*Z^*}(i)$ and $W^L_{X^*Y^*Z^*}(0) = W^I_{X^*Y^*Z^*}$. We repeat this process until $\max_{X^*Y^*}\{ | W^L_{X^*Y^*} - W^{A_p}_{X^*Y^*} | \} \le \epsilon_1$, and $\max_{X^*Z^*}\{ | W^L_{X^*Z^*} - W^{B_p}_{X^*Z^*} | \} \le \epsilon_1$. In the experimental part of this study we used $\epsilon_1 = 10^{-3}$. If auxiliary categorical information is available we add the third step to the iteration process above which also rakes the margins of the categorized matched file to the balanced Y^*Z^* margins of the auxiliary

file:

$$W^{L}_{X^{*}Y^{*}Z^{*}}(i) = {}_2W^{L}_{X^{*}Y^{*}Z^{*}}(i) \,/\, {}_2W^{L}_{Y^{*}Z^{*}}(i)\; W^{C_p}_{Y^{*}Z^{*}}, \tag{4.6c}$$

where left index 2 refers to the result of (4.6b) and ${}_2W^{L}_{Y^{*}Z^{*}}(i) = \sum_{\{X^{*}\}} {}_2W^{L}_{X^{*}Y^{*}Z^{*}}(i)$. We repeat this process until $\max_{Y^{*}Z^{*}}\{|\, W^{L}_{Y^{*}Z^{*}} - W^{C_p}_{Y^{*}Z^{*}}\,|\} \le \epsilon_1$. In the experimental part of this study we used $\epsilon_1 = 10^{-3}$.

Note that, the third step (4.6c) is unique for both types of auxiliary information: full $\{W^{C}_{X^{*}Y^{*}Z^{*}}\}$ and partial $\{W^{C}_{Y^{*}Z^{*}}\}$, since only the relationship between Y^{*} and Z^{*} is used, as explained in the introduction to Section 4 (see (ii)).

If auxiliary information refers to a different time period, *i.e.*, if it is outdated, all weights that come from the outdated file are multiplied by the ratio of total weights $W \,/\, W^{C}$. (Note that $W = W^{A} = W^{B} = W^{M}$.) In that way we preserve the distribution from the auxiliary file and make the total weights in the two years equal.

Remark 4.1: When we build a real social policy simulation database from datafiles, the structural and unexpected empty cells exist and can not be avoided by simple reduction of the number of categories in some dimension. The unexpected empty cells will cause the raking procedure not to converge. The structural empty cells come either from the real population or from the matching datafiles (*i.e.*, the samples). In the first case they are permanent, in the second their appearance is random. However, the both types will increase the number of constraints and reduce the degrees of freedom of the raking procedure. That is, too many structural empty cells, or a small number of categories in each dimension will cause non-convergence of the raking procedure. To carry the initial information into the raked table and to deal with zero cells, one has to arrange the number of marginal categories in such a way that convergence of raking is possible. For example, the number of categories $Y^{*}Z^{*}$ of non-common variables Y and Z must be greater than or equal to 4x2, 2x4 or 3x3, and the number of categories X^{*} of common variables X must be greater than or equal to 4 (also see Deming and Stephan 1940).

4.1.4. Preadjustment of Record Weights

For a given $X^{*}Y^{*}$ category, the sum of weights in the matched file M and in the look-up table $\{W^{L}_{X^{*}Y^{*}Z^{*}}\}$ may not be equal

$$\sum_{\{Z^{*}\}} W^{M}_{X^{*}Y^{*}Z^{*}} \neq \sum_{\{Z^{*}\}} W^{L}_{X^{*}Y^{*}Z^{*}}, \quad \text{for some categories } X^{*}Y^{*}. \tag{4.7}$$

To correct it, we suggest

i) the ratio adjustment of the weights w^{A}_{i} and w^{B}_{j} of records in the matching files A and B according to the pooled sums of weights $W^{A_p}_{X^{*}Y^{*}}$ and $W^{B_p}_{X^{*}Z^{*}}$ before matching by

$$w^{A_p}_{i} = w^{A}_{i}\; W^{A_p}_{X^{*}Y^{*}} \,/\, W^{A}_{X^{*}Y^{*}}, \quad i \in X^{*}Y^{*} \text{ of } A \tag{4.8a}$$

and

$$w^{B_p}_{j} = w^{B}_{i}\; W^{B_p}_{X^{*}Z^{*}} \,/\, W^{B}_{X^{*}Z^{*}}, \quad j \in X^{*}Z^{*} \text{ of } B, \tag{4.8b}$$

or alternatively,

ii) the ratio adjustment of the weights w^{M}_{ij} of records from the matched file M (but before rematching) by

$$w^{M_p}_{ij} = w^{M}_{ij}\; W^{A_p}_{X^{*}Y^{*}} \,/\, \sum_{\{Z^{*}\}} W^{M}_{X^{*}Y^{*}Z^{*}}, \quad ij \in X^{*}Y^{*} \text{ of } M. \tag{4.9}$$

If record weights w_i^A and w_j^B are preadjusted according to i), there is no need for further adjustment of the record weights w_{ij}^M of the matched file M. Our conjecture is that the performance of i) and ii) will be the same, assuming that the categorization Y^* and Z^* of matching files is done in a optimal way.

4.2. Calibration and Ratio Modification

The look-up table $\{W^L_{X^*Y^*Z^*}\}$, obtained as described in Section 4.1 is used for the ratio adjustment of an individual record weight w_{ij} :

$$w_{ij}^R = w_{ij}\, W^L_{X^*Y^*Z^*} / W^M_{X^*Y^*Z^*}, \quad ij \in X^*Y^*Z^* \text{ of } M. \tag{4.10}$$

Evidently, the empty cells of the matched file remain empty. However, in the case of the unexpected empty cells, a loss of weight will occur since a positive weight was allocated to them in the look-up table by the special procedure explained in Section 4.1.3. In order to prevent the loss of weight we do the following calibration of the look-up table before the ratio adjustment of record weights (4.10): (i) put back zeros in the unexpected empty cells, and (ii) adjust the weights over all Z^* categories except those that resulted in unexpected zeros, *i.e.*, the complementary cells of all Z^* of unexpected zeros (distorting slightly the Z^* margin):

$$W^{L_c}_{X^*Y^*Z^*} = W^L_{X^*Y^*Z^*} \sum_{\{Z^*\}} W^L_{X^*Y^*Z^*} \Big/ \sum_{\{Z^*\}\setminus\{\text{unexp. 0 cells}\}} W^L_{X^*Y^*Z^*}, \text{ for each category } X^*Y^*, \tag{4.11}$$

where the subscript c under L stands for 'calibrated'. Within each X^*Y^* category, the summation is taken over all Z^* categories in the numerator, and over all complementary Z^* categories of unexpected zeros in the denominator. Then $W^L_{X^*Y^*Z^*}$ in (4.10) is substituted by the $W^{L_c}_{X^*Y^*Z^*}$ from (4.11). After the application of the ratio adjustment the record weight w_{ij} is transformed to w_{ij}^R. For example, after the ratio adjustment of record weights a new matched file obtained from a (X,Y,Z)-distance matched file M is denoted by $M^R = \langle X_i^A \text{ or } X_j^B,\ Y_i^A,\ Z_j^B,\ Z_k^C,\ w_{ij}^R \rangle$, where only the value of record weights can be different from those in file M.

The advantage of this method is its simplicity since it deals with the record weights only. The disadvantage, however, is a possible distortion of the original distribution on (X,Y,Z). Also, the ratio adjusted weights may be very small or very large. This method does not solve the unexpected empty cell problem although their effect on convergence of the raking algorithm is annulled. Clearly this problem means that information contained in the matching files (A and B) is not fully utilized.

4.3. Partial Rematching Using the "Shift-and-Share" Algorithm

To solve the problem of unexpected empty cells that could not be solved by calibration and ratio adjustment of record weights, we propose a method which uses categorical constraints for an additional rematching (*i.e.*, re-imputation) of the records. Also see Liu and Kovačević (1996b).

We assume that for a given X^*Y^* category there are K (≥ 2) Z^* categories. For each one of them we compute the difference of sums of weights

$$\Delta_{X^*Y^*Z_k^*} = W^M_{X^*Y^*Z_k^*} - W^L_{X^*Y^*Z_k^*}, \quad k = 1,\ldots,K. \tag{4.12}$$

The *goal* is to rearrange the matched file M into M^S such that

$$W^{M^S}_{X^*Y^*Z^*} - W^L_{X^*Y^*Z^*} \approx 0 \tag{4.13}$$

over all categories $X^*Y^*Z^*$. The idea is to reduce the difference $\Delta_{X^*Y^*Z_k^*}$ by shifting the Z_k^* category of one or more records to the complementary Z^* categories, or to force a record to share at least two Z^* categories by replicating it and splitting its weight. "Shifting" or "Sharing" of the Z^* category effectively means finding a new donor record in another Z^* category of the original matching file B. To make sure that rematching doesn't disturb fulfilment of the requirement (ii) in Section 2 for use of all records from both files, we assume that a counter variable q, which counts the A-records that are recipients of the Z value from the same B-record, is known for each record in the matched file M. For example, $[X_i^A, Y_i^A, Z_j^B, w_{ij}^M, q_j=3]$, means that there are two more matched records with the Z value received from the same j-record of file B. In the following we describe the "shift-and-share" algorithm.

We assume that the matched file is categorized appropriately, the look-up table is available, and that the table with the differences is obtained as $\{\Delta_{X^*Y^*Z^*}\}$. The following steps apply within each X^*Y^* category independently.

The "Shift" part:

1. The first step is to check if any difference is greater than the threshold ε (>0) or smaller than $-\varepsilon$. If there is no such difference we end this procedure. In the experimental part of the study we use $\varepsilon=1$.

   We order the Z_k^* categories, $k=1,...,K$, by descending order of $\Delta_{X^*Y^*Z^*}$. Suppose that for the first K_1 categories $\Delta_{X^*Y^*Z_k^*} \geq \varepsilon$, and that for the last K_2 categories $\Delta_{X^*Y^*Z_k^*} \leq -\varepsilon$. Within each category above sort the records by descending value of weight (so that records are selected first by category by descending order of $\Delta_{X^*Y^*Z^*}$ and then by descending value of weight).

2a. Search among the sorted records in the first K_1 categories, starting with the one with the largest positive difference. Until we find the first category k and a record of them, say ij, for which the count $q_j=2$ and the weight satisfies

$$w_{ij} < \min\{\varepsilon+\Delta_{X^*Y^*Z_k^*}; \varepsilon-\Delta_{X^*Y^*Z_K^*}\}, \text{ where } 1 \leq k \leq K_1. \tag{4.14}$$

   If there is no such record in the first K_1 categories we end the "Shift" part of the procedure. Otherwise continue.

3a. Let record ij belong to category Z_k^* and which is the first record with highest weight satisfied (4.14). Replaced its category by the category Z_K^* which has the maximum negative difference.

4a. Rematched ij with another record form the original B file which belongs to category Z_K^*. Do not change the weight of the record w_{ij}.

5a. Update the count variable q and the differences $\Delta_{X^*Y^*Z^*}$:

$$\begin{aligned} q_k &= q_k - 1 \quad \text{and} \quad q_K = q_K + 1, \\ \Delta_{X^*Y^*Z_k^*} &= \Delta_{X^*Y^*Z_k^*} - w_{ik} \quad \text{and} \quad \Delta_{X^*Y^*Z_K^*} = \Delta_{X^*Y^*Z_K^*} + w_{ik}. \end{aligned} \tag{4.15}$$

   We repeat steps 1, 2a-5a until no "shift" is possible.

   The "Share" part: Step 1 is common for both parts.

2b. Again, we look among the sorted records in the first K_1 categories, starting with the one with the largest

positive difference. For each positive (difference) categories, we simultaneously look among the last K_2 negative (difference) categories in inverse sorted order. Until we find the first pair of categories k and t, and records in category k, say ij, for which the weight satisfies

$$w_{ij} > \min\{\Delta_{X^*Y^*Z_k^*};\ -\Delta_{X^*Y^*Z_t^*}\}, \text{ where } 1 \le k \le K_1 \text{ and } K_2 \le t \le K. \tag{4.16}$$

All records are candidates for further processing and we call them as 'shareable'. If there is no shareable record we end the "Share" part of the procedure.

3b. Next, we select a shareable record at random, duplicate it and assign to one of its replicates the category Z_t^* which is the category has smallest negative difference satisfied $w_{ij} \ge -\Delta_{X^*Y^*Z_t^*}$. The other replicate keeps its original category.

4b. The weight of this record is split between the old category Z_k^* and the new Z_t^* in the following way:

Let $\Delta_0 = \min\{\Delta_{X^*Y^*Z_k^*},\ -\Delta_{X^*Y^*Z_t^*}\}$. If $w_{ij} \ge \Delta_0 + \varepsilon$, then Δ_0 is the weight of the replicated record with a new category Z_t^*, and the remainder, $w_{ij} - \Delta_0$ is the new weight of the processed (original) record.
Otherwise, we use $\Delta_0 - \varepsilon/2$ and $w_{ij} - \Delta_0 + \varepsilon/2$, respectively, where $\varepsilon > 0$, $\Delta_0 > \varepsilon$ and $w_{ij} \ge \Delta_0$ as mentioned earlier. In this way we obtain all 'share' record weights *greater than* $\varepsilon/2$.

5b. A replicated record with the newly assigned Z_t^* category and new weight has to be rematched with another record from the original B file which belongs to this category Z_t^*.

6b. Update the differences $\Delta_{X^*Y^*Z_k^*}$ and $\Delta_{X^*Y^*Z_t^*}$ as in step 5a. We repeat steps 1, 2b-6b until no "share" is possible.

Note that in step 4a of the shift part and step 5b of the share part the rematching of another record that belongs to a new category of Z^* from file B can be assisted by information from the auxiliary file C. In this simulation study we keep the intermediate Z_k^C in the matched file M, and use the minimum distance based on variables X and Z to choose the new Z_K^B or Z_t^B.

After the application of the "shift-and-share" rematching algorithm we may need an additional ratio adjustment of individual weights to agree with the look-up table totals. We simply process as explained in Section 4.2 where $\{W^M_{X^*Y^*Z^*}\}$ is replaced by the corresponding value $\{W^{M^S}_{X^*Y^*Z^*}\}$ obtained in the shift-and-share procedure. For example, after the re-matching and the ratio weight adjustment processes the record weight w_{ij} is transformed to w_{ij}^{SR}, a rematched file obtained from a weight-split (X,Y,Z)-rank matched file M, is denoted by $M^{SR} = (X_i^A,\ Y_i^A,\ Z_j^B \text{ or } Z_t^B,\ Z_k^C,\ w_{ij}^{SR})$, where both the record weight, the impute variables Z and the impute category sum of weights could be different from those on file M.

The two categorically constrained procedures described in Sections 4.2 and 4.3 usually result in slightly different matched files. The first procedure is essentially a way to adjust record weights so that the categorical constraints are satisfied. The shift-and-share procedure, in contrast, does not change the weights of most of the records but may have a small number of rematches as results of the shift-and-share procedures aimed at minimal adjustment of the old structure of M. It deals with the unexpected empty cells problem in a straight forward manner by rematching a certain number of records. Therefore, it uses more information from the matching and auxiliary files than the ratio adjustment of record weights only. The second procedure is more complex. Obviously, the more Z^* categories, the longer the procedure. The shift-and-share algorithm complies with the task of fast

reduction of differences between the categorical distributions (tables). That is to do a 'minimum' number of process steps and to change a 'minimum' number of records to balance the categorical distribution of the matched file against a given look-up table. This algorithm does exactly the 'minimum' in the case of two Z^* categories. A summary of the pooling, raking and algorithm functions is shown by an example in Appendix.

Remark 4.2. The shift-and-share rematching algorithm does not depend on the matched file it treats or on the look-up table it uses. When full auxiliary information is available, a *new* look-up table $\{W^{LC}_{X^*Y^*Z^*}\}$ can be built by raking the margins $W^{C}_{X^*Y^*}$, $W^{C}_{X^*Z^*}$ and $W^{C}_{Y^*Z^*}$ of the full auxiliary distribution $\{W^{C}_{X^*Y^*Z^*}\}$ to $\{W^{A_p}_{Y^*Z^*}\}$, $\{W^{B_p}_{X^*Z^*}\}$ and $\{W^{C_p}_{Y^*Z^*}\}$ respectively. Obviously, if the full auxiliary categorical distribution is very good, then the use of this look-up table $\{W^{LC}_{X^*Y^*Z^*}\}$ will be preferred. A new algorithm which uses this new kind of a look-up table $\{W^{LC}_{X^*Y^*Z^*}\}$ in construction of a matched file is suggested, Liu (1998).

5. SIMULATION STUDY

The objective of the simulation study was to compare the matching methods described in Sections 3 and 4 using real survey data. This was done by applying these methods on a large number of independent pairs of samples as the matching files and then evaluating their performances over all samples. Details on the design of the study are given in Section 5.1. In Section 5.2 we provide some particulars of the matching methods considered in the study.

5.1. Design of the simulation study

5.1.1. Initial Data

Matching files A and B, and the auxiliary datafile C were generated from the Public Use Micro Files (PUMF's) from the 1986 and 1991 Census on Households/Housing for the province of Quebec. The PUMF's are samples themselves obtained by subsampling from the Census 2B samples (twenty precent of all Canada households that responded to the long questionnaire of census) in the respective years. The applied sampling procedures were different in the two Censuses (for reference see the Documentation and Users's Guide for the PUMF on Households and Housing, 1989, and 1994) resulting in different types of weights for each year. In 1986 the weights were dependent on the geographic area whereas in 1991 weights were constant over all of Canada and equal to 33.333 (*i.e.*, the self-weighting factor of 1991 Census PUMF records).

Essentially, we considered three different matching settings. In all three, matching files A and B were drawn from the 1991 PUMF. The first one did not assume any auxiliary information. For the second setting we assumed that the current auxiliary information was available and used the 1991 PUMF itself. Finally, the third one combined matching files A and B from the 1991 PUMF with the outdated auxiliary datafile C drawn from the 1986 PUMF.

5.1.2. Variables

In taking variables from the Census PUMF's as X, Y, Z variables, the objective was to define three sets of variables that are similar to those encountered in actual matching for the SPSD. These variables may be highly skewed, long tailed mixtures with discrete components.

As matching variables X we considered variables that provide details on urbanization, residential tenure, presence of mortgage, total household income (HHTOTINC) categorized into five categories, household size, household composition, sex and age of the household maintaineer. They were used as categorical variables for grouping the records into a number of matching classes. The HHTOTINC was also used as a continuous type common variable in evaluations.

Variables on total household investment income (HHNETINV) and total household government transfer

payments (HHGOVINC) were the Y variables in our simulation study. Note that these Y variables are negatively correlated. Variables HHTOTINC and HHNETINV may take negative values, but for the purpose of this study we used their absolute values.

The monthly gross rent (GROSRTH) and, alternatively, the owner's major payments - monthly (OMPH) were chosen to be the imputed variables, Z. Some statistical characteristics for selected variables, based on 28,883 household records from the 1986 Census PUMF for the province of Quebec and 78,027 records from the 1991 Census PUMF for the province of Quebec are given in Table 5.1.

Table 5.1: Descriptive Statistics (Weighted) for Variables Considered in Study

Variables	Mean		Median		Std. Dev.		Skewness		Kurtosis	
	1986	1991	1986	1991	1986	1991	1986	1991	1986	1991
HHTOTINC (X)[1]	30544	40667	26596	34997	22870	30875	1.69	1.97	5.58	8.26
HHNETINV ($Y1$)[1]	4749	6064	1728	2000	9368	12549	5.84	6.53	52.96	70.37
HHGOVINC ($Y2$)[1]	5328	7173	4056	5903	4877	6382	1.42	1.50	3.38	4.44
GROSRTH (Z)[2]	365	483	345	435	162	353	1.52	5.72	3.71	41.09
OMPH (Z)[2]	434	636	389	522	256	490	0.76	1.96	-0.11	5.68

[1] The absolute value of HHTOTINC and HHNETINV are used. Non-zero values of HHNETINV (9,661 households in 1986 and 28,121 households in 1991), and positive values of HHGOVINC (21,992 households in 1986 and 58,938 households in 1991) only were included in the above statistics.

[2] Data on GROSRTH were truncated at $99 and $1000 in 1986 (13,018 households). In 1991 values over $1500 were replaced by the average of all values over $1500 in a particular geographic area (34,385 households)-Quebec had 7 such areas. The difference in the treatment of the tail values is reflected in skewness and kurtosis. Similarly, the data on OMPH were truncated at $99 and at $1100 in 1986 (15,865 households), while in 1991 (43,642 households) all the values over $1650 were replaced by the average in a specific geographic area.

5.1.3. Selection of Study Datafiles (Populations)

The household records in the initial data set were grouped into nine datafiles (populations) according to urbanization (a combination of the Rural/Urban Code with the Census Metropolitan Area Code (CMA)) and residential tenure with the presence of mortgage. Table 5.2 gives an overview of the distribution of records in the initial data set. The difference in sizes comes from the difference in sizes of the PUMF's in these two years.

Table 5.2. Distribution of Records in the Initial Data Sets

Urbanization	Residential Tenure with Presence of Mortgage					
	Rented		Owned with Mortgage		Owned without Mortgage	
	1986	1991	1986	1991	1986	1991
Montréal & Québec City	7338	22848	3889	13119	2389	8616
Other CMAs, CAs & Urban Areas	5128	9530	4273	7092	2933	6329
Rural	554	2007	1043	3700	1361	4786

The following datafiles were chosen (dark shaded cells in Table 5.2).

Table 5.3. List of Study Datafiles

MQR:	Montréal and Québec City, Rented	OTR:	Other CMAs, CAs & Urban Areas,
MQM:	Montréal and Québec City, Owned, Mortgage.	RUO:	Rural, Owned, no Mortgage.

Four datafiles out of these nine were chosen for the simulation study. The selection was made according to the significance of the Pearson partial correlations, ρ, between Y and Z variables, in order to study the impact of conditional independence, and its violation, on the quality of matching. The approach via partial correlations is good for the particular case of the multivariate normal distribution of the X, Y, Z variables because the assumption

of independence of $Y|X$ and $Z|X$ is equivalent to the assumption that the partial correlation between Y and Z, when controlled for X, is equal to 0. The variables denoted as X, Y, Z are, as it was previously mentioned, skewed, truncated and with possible non-linear relationships. Because of this, the Kendall's τ's were calculated as well. Since the number of records in the initial data set was large, the product-moment correlations ρ's and Kendall's τ's were very close.

The values of the partial correlation coefficients ρ's of Y and Z when controlled for X, along with the corresponding p-values and Kendall's τ's, are given in Table 5.4.

The absolute magnitudes of partial correlations were small in all datafiles considered. However, in the datafiles MQR and RUO the correlation between one (of two) Y variable and the Z variable observed in that datafile was significant at the 0.001 level (see Table 5.4). In the datafile MQM, correlations between both Y's and Z were significant. Finally, none of Y variables was significantly correlated with Z in the datafile OTR.

Table 5.4. Partial Correlation Coefficients ρ, their p-Values, and Kendall's τ for Y and Z Variables in the Chosen Datafiles

Datafiles	Y Variables	Z Variables		
		$\rho_{YZ.X}$	p-value	τ
		GROSRTH (Z1)		
MQR	HHNETINV (Y1) HHGOVINC (Y2)	**0.087*** -0.007	0.0001 0.2643	0.072 -0.032
OTR	HHNETINV (Y1) HHGOVINC (Y2)	0.019 0.005	0.0664 0.6236	0.024 -0.050
		OMPH (Z2)		
MQM	HHNETINV (Y1) HHGOVINC (Y2)	**0.046*** **0.037***	0.0001 0.0001	-0.002 -0.010
RUO	HHNETINV (Y1) HHGOVINC (Y2)	0.023 **-0.050***	0.1086 0.0006	0.047 -0.039

* Denotes a significant correlation value.

A statistically significant partial correlation between variables is considered as evidence that the assumption on their conditional independence is unsustainable, and we expected that it would imply the inferiority of the methods based on this assumption.

Remark 5.1. The significance of the partial correlation coefficients ρ's of $Y|X$ and $Z|X$ may come from the very large number of records in each datafile. But, from the social-economic point of view, the relation between variables Y and Z when controlled for X really exists and has an explainable meaning.

5.1.4. Matching Classes

Matching classes were formed within each of the chosen datafiles according to the X categorical variables:

- Type of Household (a variable specially constructed for this study by combining the Census 2B variables on household composition, household size and sex of the household maintaineer, categorized into four categories non-family households, families without kid, families with kid and a male household maintaineer, and family with kid and a female household mountaineer).
- Age of the Household Maintaineer (categorized into three categories by maintaineer's age: $age<25$, $25 \le age<35$ and $35<age$),

and

- Total Household Income (categorized by its absolute value into five quintal categories).

There were 60 possible classes per datafile. Some of them were empty or contained less than six records. In

such cases, classes were redefined and merged. The final number of matching (imputation) classes in each datafile is given in Table 5.5.

So far we have described the initial population that has been used for this simulation study. It is important to emphasize that we treated these datafiles of records as distinctive populations. The next Section presents the sampling method we used for the creation of files A and B.

Table 5.5. Number of Matching Classes, the Range of Number of Records per Class, and the Size of Files

Datafile	Number of Classes	Range of Counts	File Size			
			A	B	C*	
MQR	57	11 - 3693	2856	571	22848	(7338)
MQM	45	10 - 2100	1640	328	13119	(3889)
OTR	54	12 - 1805	1192	238	9530	(5128)
RUO	31	10 - 867	598	120	4786	(1361)

* Size of the outdated auxiliary files are in brackets.

5.1.5. Creation of Matching and Auxiliary Datafiles

Files A and B were obtained as independently drawn random samples from each of the four datafiles. For file C (auxiliary datafile), we used the complete population in these datafiles. Categorical auxiliary information was derived from the complete population in the respective datafiles.

First, a larger sample A (the host file) was drawn as a simple random sample. Its size was about one eighth the size of the initial population. Then, a sample B was selected from the remainder. In this way, we prevent a record from A being matched to itself. The size of B is about one fifth of the size of the file A. This sampling procedure was repeated independently for each simulation. The resulting sizes of samples (files) A and B, and sizes of the auxiliary files C are given in Table 5.5. Consequently, the weights of records in files A and B were obtained by multiplying the original PUMF weights by 8 and 40, respectively. That is the record weight is 266.67 for record in file A, and is 1333.33 for records in file B.

5.1.6. Algorithm for Creation of Files A and B

Files A and B were created according to the following algorithm found to be the most time-efficient among three algorithms considered.

1) Draw a simple random sample without replacement from the population. The size of the sample is n^A.

2) Check on class-saturation. A sample is class-saturated if the sample part coming from the k^{th} class, n_k^A, is equal to the entire class ($n_k^A = N_k$). A class-saturated sample is rejected and Step 1 is to be repeated, otherwise go to Step 3.

3) Suppose that the file A contains records from L different classes. For each class k, ($1 \le k \le L$) that is represented in the file A, select an integer random number n_k such that $1 \le n_k \le N_k - n_k^A$. Compute the values $n_k^* = \lceil n^B \cdot n_k / \sum n_i \rceil$, where n^B is the planned size of the B file and $\lceil x \rceil$ denotes the smallest integer greater than x.

4) Let L_1 be the number of n_k^* such that $n_k^* \ge 2$, and let $D = \sum n_k^* - n^B$.
If $L_1 \le D$ then reduce n_k^* by one in each of L_1 classes. Repeat step 4 until $D = \sum n_k^* - n^B = 0$ is achieved.

If $L_1 > D$ then randomly select D classes with $n_k^* \geq 2$ and reduce their n_k^* by one.

5) Then, from the remaining records in each class k select a simple random sample of size $n_k^B = n_k^*$.
A file B will be the union of these class samples, its size will be $n^B = \sum n_k^B$.

5.2. Matching Methods in the Study

In this study we considered various matching procedures which are variants of the methods described in Sections 3 and 4 based on different combinations of available files, order restrictions, weight adjustment, distance functions and categorical constraints.

Two general matching frameworks were considered: matching without the use of auxiliary file and matching using auxiliary file. In the latter case, we studied two different types of auxiliary file contents: with complete (full) information on all three groups of variables (X,Y,Z), and with incomplete (partial) information, *i.e.*, only (Y,Z). Also, with the respect to the reference period of auxiliary file we considered a current and an outdated auxiliary file.

Note that in this simulation study we use the same groups of variables and the *normalized* Euclidean distance in all matching, size reduction and rematching processes. Depending on if there was any preadjustment of record weight in the original matched files, we had two different schemes: one without any weight adjustment before the rematching or ratio adjustment, and the other with a preadjustment by pooling at the level of the X^* categories.

In order to use an outdated auxiliary file an adjustment of its variables X, Y_1, Y_2, Z was done. We used the ratios of the relevant means for the two years

$$R_x = \bar{X}_{91} / \bar{X}_{86}, \quad R_y = \{ \bar{Y}_{91}^{(1)} + \bar{Y}_{91}^{(2)} \} / \{ \bar{Y}_{86}^{(1)} + \bar{Y}_{86}^{(2)} \}, \quad R_z = \bar{Z}_{91} / \bar{Z}_{86} . \tag{5.1}$$

Then the adjusted value of, for example, the X variable was obtained as $X_{i,86}^0 = R_x \, X_{i,86}, \quad i=1,\ldots,n^C$.

We applied categorically constrained rematching and ratio adjustment (of record weights) to the already matched files. We did not make any additional categorization of the X variables besides the one done for the purpose of formation of matching classes.

The two Y variables were categorized into two categories each, making a total of four combined categories. The four categories of the Y variables were defined as

$$\begin{aligned} Y_1^* &= \{ u_i : \; Y_{1i} = 0, Y_{2i} = 0 \}, \quad Y_2^* = \{ u_i : \; Y_{1i} = 0, Y_{2i} > 0 \}, \\ Y_3^* &= \{ u_i : \; |Y_{1i}| > 0, Y_{2i} = 0 \}, \text{ and } \; Y_4^* = \{ u_i : \; |Y_{1i}| > 0, Y_{2i} > 0 \}, \end{aligned} \tag{5.2}$$

where u_i was the i^{th} record of the matched datafile.

The Z variables was categorized into 2, 3 and 4 categories:

(i) two categories, $\quad Z_1^* = \{ u_i : Z_i \leq med(Z) \}$ and $Z_2^* = \{ u_i : Z_i > med(Z) \}$; (5.3)

(ii) three categories, $\quad Z_1^* = \{ u_i : Z_i \leq Q_1^Z \}, \; Z_2^* = \{ u_i : Q_1^Z < Z_i \leq Q_2^Z \}$
and $Z_3^* = \{ u_i : Z_i > Q_2^Z \}$; (5.4)

(iii) four categories, $\quad Z_1^* = \{ u_i : Z_i \leq Q_1(Z) \}, \; Z_2^* = \{ u_i : Q_1(Z) < Z_i \leq med(Z) \}$,
$Z_3^* = \{ u_i : med(Z) < Z_i \leq Q_3(Z) \}$ and $Z_4^* = \{ u_i : Z_i > Q_3(Z) \}$; (5.5)

where $med(Z)$ was a median, Q_1^Z, Q_2^Z were the lower and upper terciles, and $Q_1(Z)$, $Q_3(Z)$ were the lower and upper quartiles of the Z variables obtained from the entire datafile.

If a matched file was obtained without use of an auxiliary file, then rematching or ratio adjustment is done either without, or with outdated or current auxiliary categorical tables. If a matched file was obtained with the aid of an outdated auxiliary file, then rematching or ratio adjustment is done with outdated or current auxiliary categorical tables. And if a matched file was obtained with the aid of a current auxiliary file, then rematching or ratio adjustment are done only with a current auxiliary categorical table. This combinations are given in Table 5.6 (star and shaded cells). We were aware of other possible combinations, but have not simulated them.

Table 5.6. Combination of Auxiliary File and Auxiliary Categorical Table Considered in the Study

Auxiliary File	Auxiliary Categorical Table		
	Without	Outdated	Current
Without	***	***	***
Outdated	—	***	***
Current	—	—	***

We found that 42 combinations could be considered as well defined matching procedures. They are listed in Table 5.7 using the same notation as in the evaluation plots. For all (Z^*, Y^*) -categorical related methods, the simulations are repeated for Z variable categorized into two, three and four categories.

Table 5.7. List of Matching Procedures Considered in the Study

Distance or weight-split methods without use of an auxiliary file:

M	Minimum X- distance or X-rank weight-split matching.
S_p or S_p(M)	Adjustment of the record weights in the matched file M according to a pooling table of matching file A, then Shift-and-Share rematching with use of a without-auxiliary look-up table.
S or S(M)	Shift-and-Share rematching with use of a without-auxiliary look-up table.
R or R(M)	Ratio adjustment according to a without-auxiliary look-up table.
S_p^o or S_p^o(M)	Adjustment of the record weights in the matched file M according to a pooling table of matching file A, then Shift-and-Share rematching with use of an outdated-auxiliary look-up table.
S^o or S^o(M)	Shift-and-Share rematching with use of an outdated-auxiliary look-up table.
R^o or R^o(M)	Ratio adjustment according to an outdated-auxiliary look-up table.
S_p^c or S_p^c(M)	Adjustment of the record weights in the matched file M according to a pooling table of matching file A, then Shift-and-Share rematching with use of a current-auxiliary look-up table.
S^c or S^c(M)	Shift-and-Share rematching with use of a current-auxiliary look-up table.
R^c or R^c(M)	Ratio adjustment according to a current-auxiliary look-up table.

Distance or weight-split methods with use of an outdated full auxiliary file:

M^o	Minimum (X,Y)- distance to get an intermediate Z then minimum (Z,X)- distance matching, or (X,Y)-rank to get an intermediate Z then (Z,X)-rank matching.
S_p^o or S_p^o(M^o)	Adjustment of the record weights in the matched file M^o according to a pooling table of matching file A, then Shift-and-Share rematching with use of an outdated-auxiliary look-up table.
S^o or S^o(M^o)	Shift-and-Share rematching with use of an outdated-auxiliary look-up table.
R^oor R^o(M^o)	Ratio adjustment according to an outdated-auxiliary look-up table.
S_p^c or S_p^c(M^o)	Adjustment of the record weights in the matched file M^o according to a pooling table of matching file A, then Shift-and-Share rematching with use of a current-auxiliary look-up table.
S^c or S^c(M^o)	Shift-and-Share rematching with use of a current-auxiliary look-up table.
R^c or R^c(M^o)	Ratio adjustment according to a current-auxiliary look-up table.

Categorical Matching and Constrained Rematching

Distance or weight-split methods with use of a current full auxiliary file:

M^c	Minimum (X,Y)-distance to get an intermediate Z then minimum (Z,X)-distance matching, or (X,Y)-rank to get an intermediate Z then (Z,X)-rank matching.
S_p^c or $S_p^c(M^c)$	Adjustment of the record weights in the matched file M^c according to a pooling table of matching file A, then Shift-and-Share rematching with use of a current-auxiliary look-up table.
S^c or $S^c(M^c)$	Shift-and-Share rematching with use of a current-auxiliary look-up table.
R^c or $R^c(M^c)$	Ratio adjustment according to a current auxiliary look-up table.

The *normalized* multivariate Euclidean distance on absolute values $|X|$ and $|Y_1|$ of the variables X (HHTOTINC) and Y_1 (HHNETINV), negative value $-Y_2$ of the variable Y_2 (HHGOVINC), and value of variables Z_1 (GROSRTH) and Z_2 (OMPH) are used in the matching and rematching processes. Note that the non-zero and zero parts of these variables, in general, have different socio-economic attributes.

The weighted mean vector and the weighted variance-covariance matrix, used in these distance measures, are calculated only from the non-zero parts of the $|X|$, $|Y_1|$, $-Y_2$, Z_1 and Z_2 of the four datafiles from the 1991 Census PUMF for the province of Quebec. That is the best estimation of the mean and variance-covariance matrix for the current four populations of 1991 Census, respectively. They are listed in Tables 5.8 and 5.9.

Table 5.8. The Weighted Mean Vector Used in the Study

Datafiles	|HHTOTINC|	|HHNETINV|	-HHGOVINC	GROSRTH	OMPH
MQR	28918	1365	-5139	507	—
OTR	25462	860	-5635	445	—
RUO	35250	2759	-8392	—	193
MQM	59033	1593	-3558	—	1031

Table 5.9. The Weighted Variance-Covariance Matrix Used in the Study

Datafiles	MQR				OTR			
Variable	|HHTOTINC|	|HHNETINV	-HHGOVINC	GROSRTH	|HHTOTINC|	|HHNETINV|	-HHGOVINC	GROSRTH
|HHTOTINC|	492063107				394389231			
|HHNETINV|	64944904	30612462			33102915	12652124		
-HHGOVINC	-36059311	-9207143	27964808		-23257252	-5573969	27923835	
GROSRTH	1770409	771478	-870073	106123	921953	362218	-680681	157809
Datafiles	RUO				MQM			
Variable	|HHTOTINC|	|HHNETINV	-HHGOVINC	OMPH	|HHTOTINC|	|HHNETINV|	-HHGOVINC	OMPH
|HHTOTINC|	720085028				981292707			
|HHNETINV|	129768591	70318259			93164378	29401393		
-HHGOVINC	-23404162	-13556388	45628522		-35014025	-4801181	24615997	
OMPH	629121	371768	-198726	7308	4182687	1221469	-930960	227999

Note that the matched, rematched and ratio adjusted files keep the original records identification (id) and values of all the variables as they are in the original matching files A and B. They also keep the auxiliary record id and values of the auxiliary variables as they are in the auxiliary file C, when available.

In this study, all programming was done in GAUSS version 3.0 and was run under systems Window NT 3.51 on a number of Pentium Pro200 computers. For example, each simulation for the datafiles MQM or RUO takes 6.71 or 0.43 hours, respectively.

6. EVALUATION OF STATISTICAL MATCHING METHODS

To assess the performance of the matching methods four types of evaluation measures were used. They were computed in order to compare

i) the categorical distributions of the true and the matched (X,Y,Z) values;
ii) the matched and the true (X,Y,Z) values in the matched file M;
iii) selected quantiles of the true (X,Y,Z) distribution and corresponding estimates based on the matched file, and the cumulative distribution functions (CDFs) of the true and the matched (X,Y,Z);
iv) the true conditional correlations of Y and Z variables given X, and the conditional correlations estimated from the matched file.

For each evaluation measure, the mean, median, first and third quartile, minimum and maximum values, as well as the Monte-Carlo standard error and coefficient of variation were computed over a number of simulations and for each data set considered. For the comparison of different methods we formed the complete "descriptive vector" consisting of minimum, maximum, first and third quartile, and the median value over all simulations, and graphically presented by the box-plots. We introduced the "rank-plots", a xy-plot type, to present individual simulation results sorted by the values obtained by the 'base' model.

In the following we present in detail the evaluation measures we used, their properties, and a rationale for their utilization in this simulation study.

i) Two measures based on categorical comparisons were considered. Suppose that there are K categories (X^*,Y^*,Z^*) not necessarily the same as those used as matching (imputation) classes or for categorically constrained matching. In this simulation study we used the same classes as formed for categorically constrained matching collapsed to a smaller number of categories to reduce the measurement noise.

(1) The first measure from this group is based on the *Weighted Pearson Chi-Square Statistic*. Let W_k^{ρ} and W_k^{M} be the corresponding (distributions) sum of weights of records u classified into category k with respect to the true and the matched values, respectively. The total number of records of the matched file is n, and the total weight of the matched file is W. That is compared between the true and the matched categorical distributions. The formula for the (weighted) chi-square statistic is

$$\chi^2 = \frac{n}{W}\sum_k [W_k^M - W_k^{\rho}]^2 / W_k^{\rho}. \qquad (6.1)$$

(2) The second measure is the κ coefficient of agreement between two independent classifications of matched records u according to the true and matched values. Let $n_k^{M,\rho}$ be the number of records classified in the same category k with respect to both true and matched values. The corresponding total weight of these records is $W_k^{M,\rho}$. Analogously, let n_k^M and n_k^{ρ} be the numbers of records classified in category k using only matched and only true values, respectively. The coefficient of agreement ($kappa$), introduced by Cohen (1960), takes the form

$$kappa = \{ n\sum_k n_k^{M,\rho} - \sum_k n_k^M n_k^{\rho} \} / \{ n^2 - \sum_k n_k^M n_k^{\rho} \} \qquad (6.2)$$

when based on counts of records only. Taking into account the sum of weights of records this measure becomes:

$$\kappa = \{ W\sum_k W_k^{M,\rho} - \sum_k W_k^M W_k^{\rho} \} / \{ W^2 - \sum_k W_k^M W_k^{\rho} \}. \qquad (6.3)$$

The coefficient of agreement is equal to 0 when the agreement is accomplished by chance and 1 when there is a perfect agreement.

ii) This group of measures is aimed at assessing the impact of matching on the values of variables (X,Y,Z) taken by record u.

(1) To compare the values of variables (matched and true) we use the mean absolute difference between the matched and true individual values, u_i^M and u_i^{ϱ}, respectively. The measure assumes the weights from the matched file, w's, as the common weights for both u values:

$$\bar{d} = \sum_{i \in M} d_i w_i / \sum_{i \in M} w_i , \tag{6.4}$$

where d_i is the distance between u_i^M and u_i^{ϱ},

$$d_i = ||u_i^M - u_i^{\varrho}|| = \begin{cases} \sqrt{[(u_i^M - u_i^{\varrho})V_u^{-1/2}][(u_i^M - u_i^{\varrho})V_u^{-1/2}]'} & \text{or} \\ |[(u_i^M - u_i^{\varrho})V_u^{-1/2}]|, \end{cases} \tag{6.5}$$

where V_u is a positive semi-definite matrix based on (X,Y,Z) variables. In particular, the Euclidean distance between matched and true records is given by

$$d_i^e = \sqrt{(u_i^M - u_i^{\varrho})S_u^{-1}(u_i^M - u_i^{\varrho})'}, \tag{6.6}$$

where S_u is the variance-covariance matrix for the (X,Y,Z) variables. The average taken over all simulations is denoted by $\bar{D}$. In this simulation study we used the Euclidean distance.

(2) Another measure from this group counts the number of records with difference d_i (6.5) within a prespecified range δ. The relevance of a record is weighted by the corresponding record weight. The measure is used in its relative (ratio) form, and is labelled as the δ-difference index

$$G_D(\delta) = \{\sum_{i \in M} I\{d_i \le \delta\} w_i\} / \{\sum_{i \in M} w_i\}, \tag{6.7}$$

where $I\{.\}$ is an indicator variable. Essentially, the δ-difference index is the estimated CDF of the variable d_i at a given δ value. We used δ values equal to 0, 0.5 and 1.

iii) The third group contains measures that compare the quantiles of the distribution of matched u values with the population quantiles. Measures that compare the CDFs are also in this group.

(1) The first measure quantifies the difference between the cumulative distribution functions estimated from the matched file and from the population.

For each matched record u^M we compute $F(u^M)$, the multivariate CDF based on the population ϱ, and $\hat{F}(u^M)$, the estimate of the CDF based on the matched file M. We count the records in the matched file which satisfy

$$h_i = |F(u_i^M) - \hat{F}(u_i^M)| \le \varepsilon \tag{6.8}$$

and compute the ε-difference index

$$G_H(\varepsilon) = \sum_{i \in M} I\{h_i \le \varepsilon\} w_i / \sum_{i \in M} w_i. \tag{6.9}$$

This measure is analogous to the δ-difference index with the topology defined in the space of $F(u)$ values. We considered 0.005 and 0.01 as values of ε.

(2) To introduce the next measure, we first define a finite population quartile for a variable t, $t \in \{X,Y,Z\}$ as $Q_{q_t}(t) = sup\{t_i^{\varrho} \in \varrho \mid F_t(t_i^{\varrho}) \le q_t\}$, where $q_t \in \{0.25, 0.50, 0.75\}$. Next, we define the multivariate quartile lattice as

a vector of quartiles $Q = [Q_{q_X}(X), Q_{q_Y}(Y), ..., Q_{q_Z}(Z)]$, where $q_X, q_Y, ..., q_Z \in \{0.25, 0.50, 0.75\}$. Based on the matched file M, we estimate Q, $\hat{Q} = [\hat{Q}_{q_X}(X), \hat{Q}_{q_Y}(Y), ..., \hat{Q}_{q_Z}(Z)]$, where $\hat{Q}_{q_t}(t) = sup\{t_i^M \in M \mid \hat{F}_t(t_i^M) \le q_t\}$ and $q_t \in \{0.25, 0.50, 0.75\}$. A measure we are proposing has the following form

$$AD(\hat{Q}) = \sum ||\hat{Q} - Q||/L, \tag{6.10}$$

where L, is the total count of possible quartile lattices. For example, if there are four variables $\{X, Y_1, Y_2, Z\}$, the L is equal to $3^4 = 81$.

iv) An evaluation measure aimed at measuring the change in the conditional relationship of Y and Z given X is the absolute difference of correlation coefficients computed for the entire population and estimated from the matched file:

$$ADCorr = |Corr(Y,Z|X) - \hat{Corr}(Y,Z|X)|. \tag{6.11}$$

The average over all simulations is denoted by $\overline{ADCorr}$. A zero value of $\overline{ADCorr}$ means that the true relation between Y and Z variables is preserved in the matched file. In order to compute $ADCorr$ we begin with the equation: $Corr(Y,Z|X) = Cov(Y,Z|X)/\sqrt{Var(Y|X)\,Var(Z|X)}$. Then, the covariance term is computed as a covariance between the residuals of a linear regression of Y on X and the residuals of a linear regression of Z on X: $Cov(Y,Z|X) = Cov(Y,Z) - Cov(X,Y)Cov(X,Z)/V(X)$. The variance is obtained similarly. The weighted covariance and the weighted variance in the expressions are computed from the matched file and the population.

7. RESULTS AND SUMMARY

In this article we investigated the nearest neighbour matching using the distance or weight-split matching. A possible additional categorical improvement by the shift-and-share rematching or ratio adjustment of record weights only was studied in detail. We considered all three possible scenarios regarding auxiliary micro file: without, with a current or with an outdated file. Our simulation study was done for all four data sets (see Section 5.1.3). Since we did not find a big difference in the methods performances among these datafiles, we presented results only for two datafiles MQM and RUO.

First, we found that the additional backward imputation within the distance matching yields a considerable improvement in overall quality. This is due to the more complete exploitation of information on X and on the conditional distributions of Y and Z given X, and thus on the joint distributions (X,Y), (X,Z) and (X,Y,Z). Table 7.1 contains basic statistics on the number of records obtained by the backward imputation of distance matching. These numbers show how much information from file B would be lost otherwise. For example, the mean in the first row of "none auxiliary" case shows that an average of 9.81% of records over 500 simulations on datafile RUO contain X values from the matching file B that are recovered by backward imputation in distance matching. For the weight-split matching a backward imputation is not possible. This implies that information on X, available in both matching files A and B is not fully utilized in the case of the weight-split matching.

Table 7.1. Number of Records (in Percent) Obtained by Backward Imputation of Distance Matching*

Matching File	Statistics	Type of Auxiliary File Used in Matching Process		
		None	Outdated	Current
File B	mean, std	11.05 2.2536 (9.81 2.9153)	12.85 2.1471 (10.48 3.1141)	11.20 2.2295 (11.21 3.0568)
	median	10.67 (10.00)	12.80 (10.00)	10.98 (10.00)
	1st, 3rd quartile	9.45 12.20 (7.50 11.67)	11.59 14.33 (8.33 12.50)	9.45 14.94 (7.50 11.67)
	min, max	5.45 19.51 (1.67 19.17)	7.62 19.21 (1.67 19.17)	6.40 19.51 (2.50 19.17)
File A	mean, std	2.21 0.4507 (1.97 0.5850)	2.57 0.4294 (2.10 0.6249)	2.24 0.4459 (2.25 0.6134)
	median	2.13 (2.01)	2.56 (2.01)	2.20 (2.01)
	1st, 3rd quartile	1.89 2.44 (1.51 2.34)	2.32 2.87 (1.67 2.51)	1.89 2.99 (1.51 2.34)
	min, max	1.10 3.90 (0.33 3.85)	1.52 3.84 (0.33 3.85)	1.28 3.90 (0.50 3.85)

* Computed over 250 simulations for datafile MQM and 500 simulations for datafile RUO, respectively, the values for RUO are in brackets.

Categorical Matching and Constrained Rematching

The further studied categorical matching and constrained adjustment methods were compared using the evaluation measures introduced in Section 6. The main levels for comparison were:

i) Two study datafiles MQM and RUO (different combinations of significance for correlations between Y and Z variables, different sizes and different variability of Z variable);
ii) methods with and without use of an auxiliary file or table;
iii) auxiliary file or table, current or outdated;
iv) methods with and without categorical constrained rematching or ratio adjustment of record weights;
v) with or without record weights adjusted according to pooling table;

Some of findings are as follows:

a) In terms of preserving the original size of the host file we found that all methods perform similarly. The use of an auxiliary file increases the file size by about one percent if the auxiliary file is current, and by about five percent if the auxiliary file is outdated.

The applied methods for constrained rematching change the size of a matched file, while the ratio adjustment of record weights only preserves the original size of the matched file. The largest increase in size occurs when the share part of the shift-and-share rematching algorithm is applied to a file with record weights already adjusted by pooling. The presence and the quality of auxiliary categorical information do not have a significant impact on the change in size. However, the number of Z^* categories used in rematching is directly related to the percentage of increase: 6%, 11%, and 12% in average for two, three and four Z^* categories, respectively. These percentage points represent the combined increase of the matched file when compared to the size of the larger matching file A.

b) Two evaluation measures based on categorical comparisons were considered. In the matching process the number of categories of X was 45 for the MQM and 31 for the RUO datafiles, but for evaluation purposes we re-categorized X into ten classes according to the deciles of the absolute value of the variable X (HHTOTINC). The number of categories for Y was 4, and for the Z variable we tried 2, 3 and 4 categories.

Table 7.2 contains results for the methods before any categorical adjustment. The weight-split method benefits more from the use of auxiliary datafile than the distance method.

Table 7.2. Mean Weighted χ^2-index* Before Categorical Adjustment

Evaluation	Evaluated on 10x4x2 Categories			Evaluated on 10x4x3 Categories		
Matching Method	Type of Auxiliary File Used					
	None	Outdated	Current	None	Outdated	Current
Distance	216.5 (174.6)	224.4 (175.2)	215.1 (172.8)	398.3 (315.5)	416.1 (316.4)	399.2 (312.5)
Weight-Split	179.9 (152.0)	158.9 (146.0)	134.3 (123.5)	328.0 (270.6)	278.3 (256.0)	238.5 (212.8)

* Computed over 250 simulations for datafile MQM and 500 simulations for datafile RUO, respectively, the values for RUO are in brackets.

From Tables 7.3 and 7.4 with average values of the χ^2-index computed from MQM and RUO datafiles, we see that, under the distance method, the shift-and-share rematching algorithm outperforms the ratio adjustment of record weights only. In the case of the weight-split method, the gain from the shift-and-share rematching algorithm is much smaller. This can be explained by the generally better performance of the weight-split matching, so that a smaller 'room' for improvement was left.

A traditional way of presenting simulation results by the box-plots (Figures A1 to A14) hides, somehow, the real behaviour of the compared matching methods. Because of that for χ^2-index we use the rank-plots (Figures B1.1 to B4.12) of simulation results. Simulations are sorted by the value of χ^2-index for the matched file obtained by the matching method without any categorically constrained adjustment. Then for the same matched file (or the same pair of matching files) we plot the χ^2-index value obtained when matching is done by another method or by categorically constrained rematching or ratio adjustment of record weights. Note that for each box-plot of the evaluation measure χ^2-index, we have 6 rank-plots following it. The symbols used in the box-plots are from the

notations in the column one and in rank-plots are from the notations in the column two of Table 5.7. We omit all tables, box-plots and rank-plots of 10x4x4 evaluation results of χ^2-index since they are very similar to 10x4x2 results. We do not present evaluation results for the 10x4x4 matching simulations since they are very similar to 10x4x2 matchings results with an additional noise.

Table 7.3. Mean Weighted χ^2-index* for Categorically Constrained Rematching and Ratio Adjustment (Evaluated on 10x4x2 Categories)

Number of Z' Categories	Rematching / Ratio Adjustment	Type of Auxiliary File					
		None			Outdated		Current
		Type of Auxiliary Table					
		None	Outdated	Current	Outdated	Current	Current
Distance Method							
Two	[1] Shift-and-Share	180.5 (149.5)	175.2 (147.7)	171.7 (144.4)	177.8 (147.9)	175.4 (144.4)	170.6 (143.3)
	[2] Shift-and-Share	186.0 (152.1)	180.3 (150.3)	177.8 (146.7)	184.6 (150.4)	182.0 (147.2)	175.3 (145.8)
	[3] Ratio-Adjustment	203.8 (164.2)	201.1 (166.7)	198.4 (162.9)	208.8 (168.0)	205.9 (164.1)	196.8 (161.5)
Three	[1] Shift-and-Share	190.9 (158.5)	189.7 (156.3)	186.8 (154.3)	194.3 (156.7)	190.0 (155.0)	183.9 (153.8)
	[2] Shift-and-Share	191.8 (157.3)	189.4 (155.5)	186.5 (153.7)	194.5 (156.0)	191.0 (154.4)	183.7 (153.3)
	[3] Ratio-Adjustment	204.8 (167.1)	206.2 (171.4)	203.2 (168.8)	214.6 (172.8)	211.1 (170.5)	201.1 (168.3)
Four	[1] Shift-and-Share	188.3 (157.6)	182.6 (156.0)	179.3 (151.8)	185.6 (154.0)	183.9 (152.1)	178.0 (151.1)
	[2] Shift-and-Share	187.6 (155.4)	182.8 (154.2)	180.6 (150.0)	187.4 (154.6)	183.7 (150.0)	177.7 (148.9)
	[3] Ratio-Adjustment	208.4 (167.5)	209.6 (175.3)	206.1 (169.4)	219.0 (176.3)	215.3 (170.6)	205.2 (168.2)
Weight-Split Method							
Two	[1] Shift-and-Share	177.3 (152.1)	167.9 (147.2)	164.7 (143.8)	154.7 (144.6)	151.8 (141.4)	131.0 (123.3)
	[2] Shift-and-Share	173.0 (147.5)	168.1 (148.0)	165.5 (144.4)	153.6 (144.6)	150.3 (140.8)	128.2 (121.7)
	[3] Ratio-Adjustment	171.3 (146.1)	167.7 (149.5)	165.2 (145.7)	152.4 (146.2)	149.3 (141.9)	126.0 (120.5)
Three	[1] Shift-and-Share	183.5 (158.2)	178.1 (153.2)	174.0 (151.7)	162.5 (150.1)	159.1 (148.6)	138.2 (129.6)
	[2] Shift-and-Share	174.7 (149.1)	174.2 (150.5)	171.1 (148.6)	158.7 (147.0)	155.2 (145.2)	133.5 (125.4)
	[3] Ratio-Adjustment	171.2 (146.0)	171.6 (151.3)	168.6 (148.8)	155.8 (147.6)	152.8 (145.6)	129.5 (123.4)
Four	[1] Shift-and-Share	187.2 (162.1)	176.8 (156.6)	173.7 (152.1)	163.9 (154.1)	159.4 (150.2)	138.8 (131.5)
	[2] Shift-and-Share	175.8 (150.4)	173.0 (153.1)	170.0 (148.5)	158.2 (150.0)	155.0 (145.5)	133.1 (125.8)
	[3] Ratio-Adjustment	171.0 (145.9)	170.8 (155.2)	167.7 (149.3)	154.9 (151.4)	151.5 (145.7)	128.2 (123.1)

* Computed over 250 simulations for datafile MQM and 500 simulations for datafile RUO, respectively, the values for RUO are in brackets.
1 Before the Shift-and-Share rematching, record weights were adjusted according to a corresponding pooling table (see Section 4.1.1)
2 Before the Shift-and-Share rematching, no adjustment on record weights.
3 Only Ratio adjustment of record weights.

From Figures A1 to A4 and B1.1 to B4.12, we see that categorically constrained adjustment via the shift-and-share rematching algorithm in most of the simulations improved the matched file in the sense of reduction of the χ^2-index. The use of a current auxiliary table contributed the most to the improvement. The categorically constrained adjustment via shift-and-share rematching algorithm either with record weights adjusted according to a corresponding pooling table or with no adjustment on record weights resulted in more stable χ^2-indices than when ratio adjustment of record weights was performed only. The performance of shift-and-share rematching with record weights adjusted is slightly better than shift-and-share rematching without record weights adjusted under the distance matching, but they are similar under the weight-split matching. The gain from the adjustment in the case of weight-split matching is slightly smaller than in the case of distance matching.

Categorical Matching and Constrained Rematching

The increase in number of Z^* categories used for rematching slightly improves the quality of the matched file. There is a negative effect of the reduced size of respective categories which may even cancel out the gain from the larger number of categories. Since the more Z^* categories the less records in respective $X^*Y^*Z^*$ cells, an additional noise is generated by the small sample portions, which is reflected in the increased χ^2 value. This is especially evident in the case of the RUO datafile. We obtained the same results for four Z^* categories as for two and three, for the weight-split method and both, rematching via shift-and-share and ratio adjustment records' weights. A small increase in variability was observed for the ratio adjustment. In the case of distance matching, when Z^* is categorized into four categories, both shift-and-share and ratio adjustment resulted in the less stable results, especially ratio adjustment only which can be explained by the reduced number of records in the corresponding matching classes. Therefore, our simulation results show that dealing with only two Z^* categories in rematching will be as efficient as using three or four categories. We also found that the evaluation favours the same number of Z^* categories for evaluation as one used for rematching. The similar findings for the OTR datafile were reported in Liu and Kovačević (1996b), and the partial results for the MQM datafile were presented in Liu and Kovačević (1997).

Table 7.4. Mean Weighted χ^2-index* for Categorically Constrained Rematching and Ratio Adjustment (Evaluated on 10x4x3 Categories)

Number of Z^* Categories	Rematching / Ratio Adjustment	Type of Auxiliary File					
		None			Outdated		Current
		Type of Auxiliary Table					
		None	Outdated	Current	Outdated	Current	Current
Distance Method							
Two	[1]Shift-and-Share	351.9 (284.9)	348.4 (283.4)	345.9 (280.7)	361.4 (283.6)	358.6 (281.8)	347.7 (279.0)
	[2]Shift-and-Share	355.6 (286.0)	352.3 (284.6)	349.6 (282.0)	366.9 (285.4)	364.1 (283.5)	350.9 (280.9)
	[3]Ratio-Adjustment	372.1 (297.3)	371.9 (301.5)	369.5 (298.9)	392.1 (303.4)	389.4 (301.0)	372.0 (296.8)
Three	[1]Shift-and-Share	327.6 (267.1)	314.3 (263.0)	309.3 (254.6)	326.7 (263.9)	320.3 (256.0)	311.5 (254.1)
	[2]Shift-and-Share	333.1 (267.5)	319.8 (263.6)	315.0 (256.5)	332.2 (265.0)	327.7 (257.3)	316.5 (255.4)
	[3]Ratio-Adjustment	363.9 (290.6)	358.4 (299.1)	353.5 (289.7)	375.7 (301.2)	370.3 (292.4)	353.4 (288.6)
Four	[1]Shift-and-Share	345.1 (282.5)	335.6 (279.4)	332.0 (273.7)	345.8 (278.8)	344.7 (274.1)	334.0 (272.0)
	[2]Shift-and-Share	343.4 (278.6)	335.2 (276.0)	333.1 (270.6)	347.1 (275.5)	343.9 (270.9)	332.3 (268.5)
	[3]Ratio-Adjustment	375.3 (296.3)	376.4 (308.9)	372.1 (301.0)	393.6 (308.9)	389.1 (301.9)	371.5 (298.4)
Weight-Split Method							
Two	[1]Shift-and-Share	321.2 (268.9)	313.5 (264.2)	311.7 (262.3)	273.2 (252.0)	269.6 (250.4)	236.4 (214.2)
	[2]Shift-and-Share	312.3 (259.8)	310.8 (262.5)	308.5 (260.4)	269.8 (250.9)	266.5 (248.6)	232.0 (210.7)
	[3]Ratio-Adjustment	308.9 (257.0)	308.0 (262.7)	305.5 (260.5)	267.3 (252.9)	264.0 (250.0)	227.8 (208.5)
Three	[1]Shift-and-Share	329.0 (275.6)	307.0 (264.3)	301.1 (256.8)	272.1 (256.0)	266.2 (248.3)	234.3 (214.9)
	[2]Shift-and-Share	314.4 (260.9)	303.9 (262.0)	299.4 (254.1)	265.8 (251.8)	260.6 (244.3)	226.5 (208.8)
	[3]Ratio-Adjustment	308.3 (256.3)	302.5 (266.2)	297.7 (256.9)	261.7 (254.9)	256.6 (246.7)	220.4 (206.3)
Four	[1]Shift-and-Share	337.3 (284.3)	319.5 (274.8)	316.5 (269.8)	283.0 (263.3)	277.7 (258.5)	245.2 (224.6)
	[2]Shift-and-Share	316.9 (264.0)	312.3 (268.3)	309.5 (263.3)	272.8 (256.5)	269.3 (251.0)	235.3 (215.6)
	[3]Ratio-Adjustment	308.2 (256.5)	308.3 (271.7)	304.8 (264.0)	266.1 (260.1)	262.3 (252.5)	226.4 (211.5)

*, 1, 2 and 3, same as in Table 7.3.

The evaluation by the weighted κ-coefficient is presented only in the form of box-plots for two and three Z^*

categories in Figures A5 to A8. From Figures A5 and A7, for the datafile MQM, we see that the little gain from the distance matching comes from the use of a current auxiliary file. However, categorically constrained adjustments via the shift-and-share algorithm or via the ratio adjustment record weights did not improve matching in the sense of the κ-coefficient. For the weight-split matching there was no gain from the categorically constrained adjustments either from the use of any auxiliary information. From Figures A6 and A8, for the datafile RUO, we see that there is no difference between the considered methods.

c) The quality of matching is also evaluated by the δ-difference ratio. For $\delta = 0$, the special case of the full agreement between imputed and true values, no difference was observed between the methods. The higher average agreement found for the datafile RUO indicated the lower variation of the variable Z (see Figures A9 and A10). For $\delta > 0$ with the datafile MQM, methods based on the minimum distance matching and rematching using the current auxiliary table were slightly better than others. Surprisingly, methods based on the weight-split matching and the rematching by outdated auxiliary table were better than other combinations with the weight-split method (see Figure A11). For the datafile RUO, there was no difference between methods (see Figure A12). We omit box-plots for $\delta = 1$ since they show very similar results as $\delta = 0.5$. For this simulation study, the weighted mean absolute difference $\bar{D}$ is not available.

d) To assess the preservation of the distribution from the true population $\wp$, we used average absolute difference of quartile lattice $AD(\hat{Q})$. These measures compared the quartiles from the matched file M with the quartiles from the population $\wp$. Results obtained for both datafiles are very similar. Rematching did not yield any gain.

Another measure used to assess the preservation of the distribution from the true population was the ε-difference index, for $\varepsilon = .005$ and .01. Although this measure appeared to be the least stable, no clear patterns are found (see Figures A13 to A14). We omit box-plots for $\varepsilon = .01$ since they show very similar results to $\varepsilon = .005$.

e) The preservation of the original relationship between Y and Z was measured by the absolute difference of the correlation coefficients. We computed the correlations between Y and Z when controlled for X in order to quantify the change of the original relationship of Y and Z given X in the matched file. The smaller value of this correlation, the better preservation of the original relationship. The absolute difference of correlation coefficients between Y variables and Z did not show enough sensitivity to discriminate different methods. We found that all methods over the two study datafiles performed similarly regarding this measure. For the datafile RUO bigger variations were generated by small sample sizes. One of the reasons is the magnitude of the partial correlations between Y and Z variables (see Table 5.4). Although some of these correlations appeared as significant, their values were still too small to be considered as changed by a matching procedure.

We summarize our findings along the listed levels for comparison:

i) We did not find significant difference in the performance of methods between the two datafiles. The principal reason is that these datafiles were very similar with the respect to the study variables.
ii) The use of an auxiliary file (variables) only in the matching process does not necessarily improve the quality of the matching.
iii) There is a benefit in using full information from an auxiliary file and a table, especially when auxiliary information is current and used in combination with the shift-and-share rematching.
iv) Categorical constraints utilized by raking improved the distributional aspects of statistical matching (measured by the χ^2-index) in most cases. However, categorical constraints, derived from the matching files only, did not improve matching significantly.
v) Surprisingly, the record weight adjustment according to a pooling table did not effect matching results.
vi) An additional backward imputation (see Section 3.1.1) increases the quality of distance matching, especially when the range of X values in two matching files is different.

Overall, the great similarities in performance of the methods can be attributed to the very fine initial classification of records into matching classes. A large number of matching classes implies small differences between records within classes. Since the methods were applied independently at the level of matching class there was small room available for them to result with different results.

Our general finding is that the quality of any matching procedure can be improved by additional categorical constraints, especially when they are implemented via the shift-and-share rematching algorithm. However, both the rematching and ratio adjustment procedures rely on a look-up table. Hence, the quality of a look-up table is an important issue. It is important to emphasize that the shift-and-share rematching algorithm is oriented to a minimum change in the matched file assuming that it was obtained by an acceptable good matching procedure. The changes affect just a small number of records through a new imputation and the weight assignment. If there are unexpected zero cells in the matched file, the ratio adjustment of record weights according to the categorical constraints could perform poorly. When auxiliary information is available, the modified distance matching method with backward imputation and reexamination by the shift-and-share rematching algorithm is recommended. The weight-split method can be used when the matching files are overlapping enough on common variables and when a good auxiliary file is available.

ACKNOWLEDGMENTS

This work was sponsored by the Social and Economic Studies Division of Statistics Canada. The authors wish to thank Jean-Louis Tambay, Harold J. Mantel, Geoff Rowe, Shiying Wu, Zhaoguo Chen, Mingyu Yu , and René Boyer for their useful discussions and comments, and a lot of the colleagues in all three Methodology Divisions who let the authors used their computers after work hours and on weekends for the simulation studies over a three year period.

REFERENCES

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

Deming, W.E., and Stephan, F.F., (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematics Statistics*, Vol 11, 427-444.

Goel, P.K., and Ramalingam, T., (1989). *The Matching Methodology: Some Statistical Properties* Lecture Notes in Statistics, Springer-Verlag, New York

Kovačević, M.S., and Liu, T.P., (1994). Statistical matching of survey data files: A simulation study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Vol. I, 479-484.

Liu, T.P., (1998). A categorical constraints guided matching algorithm. *Proceeding of the Survey Methods Section, Statistical Society of Canada*, To appear.

Liu, T.P., and Kovačević, M.S., (1996a). Statistical matching of survey datafiles. Presented at the 23rd Meeting of the Advisory Committee on Statistical Methods, Statistics Canada, Ottawa.

Liu, T.P., and Kovačević, M.S., (1996b). Categorically constrained matching. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 123-134.

Liu, T.P., and Kovačević, M.S., (1997). An empirical study on categorically constrained matching. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 167-178.

Paass, G., (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In *Micro-analytic Simulation Models to Support Social and Financial Policy* (Eds. Orcutt, Merz and Quinke) Elsevier Science, Amsterdam.

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

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

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

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

Singh, A.C., (1989). Log-linear imputation. *Proceeding of the Fifth Annual Research Conference, Bureau of the Census*, US Department of Commence, 118-132.

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

Singh, A.C., Mantel, H.J., Kinack, M.D., and Rowe, G., (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59-79.

Wolfson, M., Gribble, S., Bordt, M., Murphy, B., and Rowe, G., (1987). The social policy simulation database: An example of survey and administration data integration. *Proceedings of "Statistics Canada Symposium on Statistical Use of Administrative Data"*, Statistics Canada. 201-229.

APPENDIX

ILLUSTRATION OF CONSTRUCTING THE LOOK-UP TABLE AND REMATCHING BY "SHIFT-AND-SHARE" ALGORITHM

We demonstrate our procedures on the small data sets. Variable X is categorized into 4 matching categories, two Y variables are categorized into 2 categories each, and the Z variable is categorized into 2 categories.

First we present the counts (number of records) and the sums of weights in categorically transformed matching files (samples) A and B, and the matched file M:

Tables A1-A6: Counts and Sums of Weights of the Categorized Matching Files A and B, and the Matched File M

Counts $\{n^{A}_{X^{\cdot}Y^{\cdot}}\}$ and $\{n^{B}_{X^{\cdot}Z^{\cdot}}\}$:

$n^{A}_{\cdot\cdot}$	$Y_1^{\cdot}$	$Y_2^{\cdot}$	$Y_3^{\cdot}$	$Y_4^{\cdot}$	
$X_1^{\cdot}$	39	5	23	2	69
$X_2^{\cdot}$	18	5	4	0	27
$X_3^{\cdot}$	6	5	11	2	24
$X_4^{\cdot}$	43	10	43	5	101
	106	25	81	9	221

$n^{B}_{\cdot\cdot}$	$Z_1^{\cdot}$	$Z_2^{\cdot}$	
$X_1^{\cdot}$	19	3	22
$X_2^{\cdot}$	1	0	1
$X_3^{\cdot}$	4	3	7
$X_4^{\cdot}$	8	6	14
	32	12	44

Sums of weights $\{W^{A}_{X^{\cdot}Y^{\cdot}}\}$ and $\{W^{B}_{X^{\cdot}Z^{\cdot}}\}$:

$W^{A}_{\cdot\cdot}$	$Y_1^{\cdot}$	$Y_2^{\cdot}$	$Y_3^{\cdot}$	$Y_4^{\cdot}$	
$X_1^{\cdot}$	10376.4	1330.3	6119.4	532.1	18358.2
$X_2^{\cdot}$	4789.1	1330.3	1064.2	0	7183.6
$X_3^{\cdot}$	1596.4	1330.3	2926.7	532.1	6385.5
$X_4^{\cdot}$	11440.6	2660.6	11440.6	1330.3	26872.1
	28202.4	6651.5	21550.9	2394.5	58799.4

$W^{B}_{\cdot\cdot}$	$Z_1^{\cdot}$	$Z_2^{\cdot}$	
$X_1^{\cdot}$	25390.7	4009.1	29399.7
$X_2^{\cdot}$	1336.4	0	1336.4
$X_3^{\cdot}$	5345.4	4009.1	9354.5
$X_4^{\cdot}$	10690.8	8018.1	18708.9
	42763.2	16036.2	58799.4

Counts $\{n^{M}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}\}$:

$n^{M}_{\cdot\cdot\cdot}$	$Y_1^{\cdot}Z_1^{\cdot}$	$Y_1^{\cdot}Z_2^{\cdot}$	$Y_2^{\cdot}Z_1^{\cdot}$	$Y_2^{\cdot}Z_2^{\cdot}$	$Y_3^{\cdot}Z_1^{\cdot}$	$Y_3^{\cdot}Z_2^{\cdot}$	$Y_4^{\cdot}Z_1^{\cdot}$	$Y_4^{\cdot}Z_2^{\cdot}$	
$X_1^{\cdot}$	38	4	6	1	23	2	2	0	76
$X_2^{\cdot}$	18	0	5	0	4	0	0	0	27
$X_3^{\cdot}$	0	6	3	2	7	4	1	1	24
$X_4^{\cdot}$	23	21	6	4	27	16	5	0	102
	79	31	20	7	61	22	8	1	229

Sums of weights $\{W^{M}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}\}$:

$W^{M}_{\cdot\cdot\cdot}$	$Y_1^{\cdot}Z_1^{\cdot}$	$Y_1^{\cdot}Z_2^{\cdot}$	$Y_2^{\cdot}Z_1^{\cdot}$	$Y_2^{\cdot}Z_2^{\cdot}$	$Y_3^{\cdot}Z_1^{\cdot}$	$Y_3^{\cdot}Z_2^{\cdot}$	$Y_4^{\cdot}Z_1^{\cdot}$	$Y_4^{\cdot}Z_2^{\cdot}$	
$X_1^{\cdot}$	9312.1	1064.2	1241.6	88.7	5587.3	532.1	532.1	0	18358.2
$X_2^{\cdot}$	4789.1	0	1330.3	0	1064.2	0	0	0	7183.6
$X_3^{\cdot}$	0	1596.4	798.2	532.1	1862.4	1064.2	266.1	266.1	6385.5
$X_4^{\cdot}$	5986.4	5454.2	1596.4	1064.2	7183.6	4257.0	1330.3	0	26872.1
	20087.6	8114.9	4966.5	1685.1	15697.6	5853.3	2128.5	266.1	58799.4

The categorically transformed matched file M contains both types of empty cells. Cells $X_2^{\cdot}Y_1^{\cdot}Z_2^{\cdot}$, $X_2^{\cdot}Y_2^{\cdot}Z_2^{\cdot}$, $X_2^{\cdot}Y_3^{\cdot}Z_2^{\cdot}$, $X_2^{\cdot}Y_4^{\cdot}Z_1^{\cdot}$ and $X_2^{\cdot}Y_4^{\cdot}Z_2^{\cdot}$ are the structural empty cells. Cells (underlined) $X_1^{\cdot}Y_4^{\cdot}Z_2^{\cdot}$, $X_3^{\cdot}Y_1^{\cdot}Z_1^{\cdot}$, and $X_4^{\cdot}Y_4^{\cdot}Z_2^{\cdot}$ are the unexpected empty cells.

Evidently, the marginal sums of weights of $X^{\cdot}$ categories in A and B do not agree. After pooling at the level of the $X^{\cdot}$ categories by category size, we have the following situation:

Tables A7-A8: Sums of Weights $\{W^{A'}_{X^{\cdot}Y^{\cdot}}\}$ and $\{W^{B'}_{X^{\cdot}Z^{\cdot}}\}$ After Pooling

$W^{A'}_{\cdot\cdot}$	$Y_1^{\cdot}$	$Y_2^{\cdot}$	$Y_3^{\cdot}$	$Y_4^{\cdot}$	
$X_1^{\cdot}$	11468.3	1470.3	6763.4	588.1	20290.1
$X_2^{\cdot}$	4486.8	1246.3	997.1	0	6730.2
$X_3^{\cdot}$	1702.1	1418.4	3120.5	567.4	6808.4
$X_4^{\cdot}$	10631.1	2472.3	10631.1	1236.2	24970.7
	28288.3	6607.4	21512.1	2391.7	58799.4

$W^{B'}_{\cdot\cdot}$	$Z_1^{\cdot}$	$Z_2^{\cdot}$	
$X_1^{\cdot}$	17523.3	2766.8	20290.1
$X_2^{\cdot}$	6730.2	0	6730.2
$X_3^{\cdot}$	3890.5	2917.9	6808.4
$X_4^{\cdot}$	14269.0	10701.7	24970.7
	42412.9	16386.5	58799.4

The initial categorical distribution $\{W^{I}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}\}$ for the raking procedure is obtained from the categorical distribution of the matched file after correction for the unexpected empty cells.

Table A9: The Initial Distribution $\{ W^{I}_{X^{\bullet}Y^{\bullet}Z^{\bullet}} \}$ for the Raking Procedure

$W^{I}_{\dots}$	$Y_1^{\bullet}Z_1^{\bullet}$	$Y_1^{\bullet}Z_2^{\bullet}$	$Y_2^{\bullet}Z_1^{\bullet}$	$Y_2^{\bullet}Z_2^{\bullet}$	$Y_3^{\bullet}Z_1^{\bullet}$	$Y_3^{\bullet}Z_2^{\bullet}$	$Y_4^{\bullet}Z_1^{\bullet}$	$Y_4^{\bullet}Z_2^{\bullet}$	
$X_1^{\bullet}$	9204.6	1081.0	1255.7	120.2	5535.9	556.9	556.9	32.8	18344.1
$X_2^{\bullet}$	4749.7	0	1343.1	0	1081.0	0	0	0	7173.8
$X_3^{\bullet}$	32.8	1605.1	819.0	556.9	1867.2	1081.0	294.9	294.9	6551.8
$X_4^{\bullet}$	5928.9	5404.8	1605.1	1081.0	7108.2	4225.6	1343.1	32.8	26729.7
	19916.1	8091.0	5022.9	1758.1	15592.3	5863.6	2194.9	360.5	58799.4

For the purpose of this illustration we assume that (partial) auxiliary categorical distribution $\{ W^{C}_{Y^{\bullet}Z^{\bullet}} \}$ is available.

Tables A10-A11: Original and After Raking Auxiliary Categorical Distributions $\{ W^{C}_{Y^{\bullet}Z^{\bullet}} \}$ and $\{ W^{C_p}_{Y^{\bullet}Z^{\bullet}} \}$

Original auxiliary table

$W^{C}_{\bullet\bullet}$	$Z_1^{\bullet}$	$Z_2^{\bullet}$	
$Y_1^{\bullet}$	21066.5	9699.9	30766.4
$Y_2^{\bullet}$	3366.6	1733.3	5099.9
$Y_3^{\bullet}$	11533.2	9166.6	20699.8
$Y_4^{\bullet}$	1500.0	733.3	2233.3
	37466.3	21333.1	58799.4

After raking to $W^{A_p}_{Y^{\bullet}}$ and $W^{B_p}_{Z^{\bullet}}$

$W^{C_p}_{\bullet\bullet}$	$Z_1^{\bullet}$	$Z_2^{\bullet}$	
$Y_1^{\bullet}$	21637.5	6650.8	28288.3
$Y_2^{\bullet}$	4917.3	1690.1	6607.4
$Y_3^{\bullet}$	14054.9	7457.1	21512.1
$Y_4^{\bullet}$	1803.2	588.5	2391.7
	42412.9	16386.5	58799.4

Finally, the *look-up* table is obtained by raking of the $W^{I}_{X^{\bullet}Y^{\bullet}Z^{\bullet}}$ to the $X^{\bullet}Y^{\bullet}$ margin of A_p, the $X^{\bullet}Z^{\bullet}$ margin of B_p and the $Y^{\bullet}Z^{\bullet}$ categorical distribution $\{ W^{C_p}_{Y^{\bullet}Z^{\bullet}} \}$ is $\{ W^{L}_{X^{\bullet}Y^{\bullet}Z^{\bullet}} \}$. Calibration modifies the look-up table only in six cells (underlined) by setting back to zero the unexpected empty cells, and adding their contents to the complementary $Z^{\bullet}$ cells. The calibrated look-up table represents the categorical distribution of the matched file M after the ratio adjustment of individual weights :

Tables A12-A13: The Look-up Table $\{ W^{L}_{X^{\bullet}Y^{\bullet}Z^{\bullet}} \}$ and Its Calibrated Version $\{ W^{L_c}_{X^{\bullet}Y^{\bullet}Z^{\bullet}} \}$ According to $\{ W^{M}_{X^{\bullet}Y^{\bullet}Z^{\bullet}} \}$

$W^{L}_{\dots}$	$Y_1^{\bullet}Z_1^{\bullet}$	$Y_1^{\bullet}Z_2^{\bullet}$	$Y_2^{\bullet}Z_1^{\bullet}$	$Y_2^{\bullet}Z_2^{\bullet}$	$Y_3^{\bullet}Z_1^{\bullet}$	$Y_3^{\bullet}Z_2^{\bullet}$	$Y_4^{\bullet}Z_1^{\bullet}$	$Y_4^{\bullet}Z_2^{\bullet}$	
$X_1^{\bullet}$	10370.4	1097.9	1254.2	216.1	5460.3	1303.1	438.4	149.7	20290.1
$X_2^{\bullet}$	4486.8	0	1246.3	0	997.1	0	0	0	6730.2
$X_3^{\bullet}$	154.9	1547.2	1110.0	308.1	2379.9	740.6	245.3	322.1	6808.4
$X_4^{\bullet}$	6625.4	4005.7	1306.5	1165.8	5217.6	5413.5	1119.5	116.7	24970.7
	21637.5	6650.8	4917.3	1690.1	14054.9	7457.1	1803.2	588.5	58799.4

$W^{L_c}_{\dots}$	$Y_1^{\bullet}Z_1^{\bullet}$	$Y_1^{\bullet}Z_2^{\bullet}$	$Y_2^{\bullet}Z_1^{\bullet}$	$Y_2^{\bullet}Z_2^{\bullet}$	$Y_3^{\bullet}Z_1^{\bullet}$	$Y_3^{\bullet}Z_2^{\bullet}$	$Y_4^{\bullet}Z_1^{\bullet}$	$Y_4^{\bullet}Z_2^{\bullet}$	
$X_1^{\bullet}$	10370.4	1097.9	1254.2	216.1	5460.3	1303.1	588.1	0	20290.1
$X_2^{\bullet}$	4486.8	0	1246.3	0	997.1	0	0	0	6730.2
$X_3^{\bullet}$	0	1702.1	1110.3	308.1	2379.9	740.6	245.3	322.1	6808.4
$X_4^{\bullet}$	6625.4	4005.7	1306.5	1165.8	5217.6	5413.5	1236.2	0	24970.7
	21482.6	6805.7	4917.3	1690.1	14054.9	7457.1	2069.6	322.1	58799.4

For an application of rematching by the shift-and-share algorithm on M, we first make a table A14 with differences $\Delta_{X^{\bullet}Y^{\bullet}Z_k^{\bullet}}$. The number of moved and replicated records is given in the next table A15. We also provide a table with total weights that were moved from one category to another.

Table A14: Table with Differences $\Delta_{X^{\cdot}Y^{\cdot}Z^{\cdot}_k} = W^{M}_{X^{\cdot}Y^{\cdot}Z^{\cdot}_k} - W^{L}_{X^{\cdot}Y^{\cdot}Z^{\cdot}_k}$

$\Delta_{\cdots}$	$Y^{\cdot}_1Z^{\cdot}_1$	$Y^{\cdot}_1Z^{\cdot}_2$	$Y^{\cdot}_2Z^{\cdot}_1$	$Y^{\cdot}_2Z^{\cdot}_2$	$Y^{\cdot}_3Z^{\cdot}_1$	$Y^{\cdot}_3Z^{\cdot}_2$	$Y^{\cdot}_4Z^{\cdot}_1$	$Y^{\cdot}_4Z^{\cdot}_2$
$X^{\cdot}_1$	-1058.3	-33.7	-12.6	-127.4	+127.0	-771.0	+93.7	-149.7
$X^{\cdot}_2$	+302.3	0	+84.0	0	+67.1	0	0	0
$X^{\cdot}_3$	-154.9	+49.2	-312.1	+224.0	-517.5	+323.6	+20.8	-56.0
$X^{\cdot}_4$	-639.0	+1448.5	+289.9	-101.6	+1966.0	-1156.5	+210.8	-116.7

Table A15: Number of 'Shift-and-Share' Records and Corresponding Total Weights

		$Y^{\cdot}_1Z^{\cdot}_1$	$Y^{\cdot}_1Z^{\cdot}_2$	$Y^{\cdot}_2Z^{\cdot}_1$	$Y^{\cdot}_2Z^{\cdot}_2$	$Y^{\cdot}_3Z^{\cdot}_1$	$Y^{\cdot}_3Z^{\cdot}_2$	$Y^{\cdot}_4Z^{\cdot}_1$	$Y^{\cdot}_4Z^{\cdot}_2$	$Y^{\cdot}_1Z^{\cdot}_1$	$Y^{\cdot}_1Z^{\cdot}_2$	$Y^{\cdot}_2Z^{\cdot}_1$	$Y^{\cdot}_2Z^{\cdot}_2$	$Y^{\cdot}_3Z^{\cdot}_1$	$Y^{\cdot}_3Z^{\cdot}_2$	$Y^{\cdot}_4Z^{\cdot}_1$	$Y^{\cdot}_4Z^{\cdot}_2$
$X^{\cdot}_1$	Shift	0	0	0	0	0	0	0	0								
	Share	0	0	0	0	-1	+1	-1	+1	0	0	0	0	-127.0	+127.0	-93.7	+93.7
$X^{\cdot}_2$	Shift	0	0	0	0	0	0	0	0								
	Share	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$X^{\cdot}_3$	Shift	0	0	0	0	+1	-1	0	0								
	Share	+1	-1	+1	-1	+1	-1	-1	+1	+49.2	-49.2	+224.0	-224.0	+323.7	-323.7	-20.8	+20.8
$X^{\cdot}_4$	Shift	+2	-2	0	0	-4	+4	0	0								
	Share	+1	-1	-1	+1	-1	+1	-1	+1	+639.0	-639.0	-101.6	+101.6	-1156.5	+1156.5	-116.7	+116.7

A negative sign for "shift" means that this category lost records for a complementary category where we have a plus sign. A negative sign for "share" means that some records in this category are replicated, their weights are split, and that replicates are moved to a complementary category, in which we have a plus sign. A zero value means that there was no change.

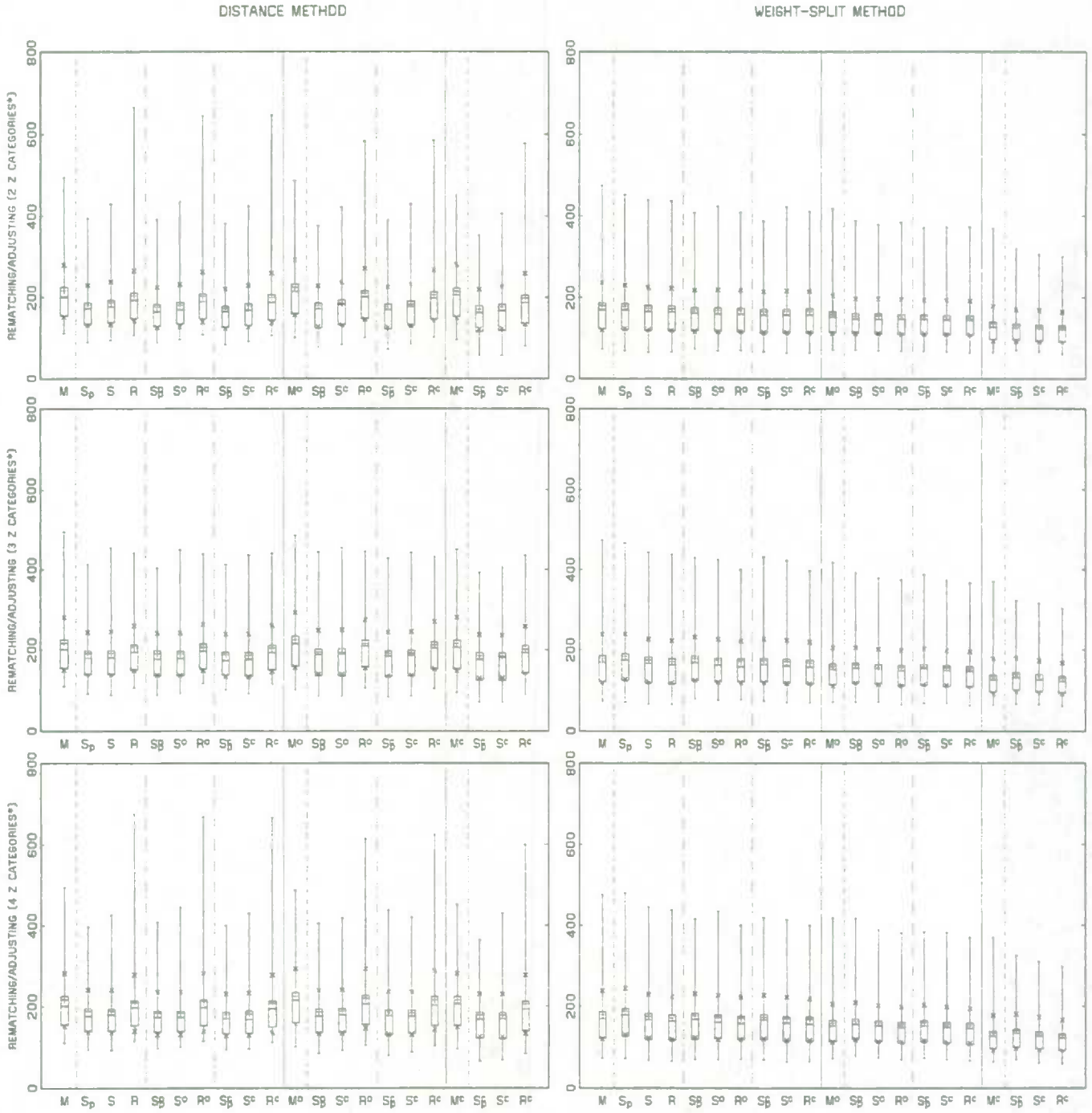
After the application of the "shift-and-share" algorithm we have a revised matched file M^{S} with a new categorical distribution.

Table A16: Categorical Distribution $\{W^{M^S}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}\}$ of the Matched File M After Rematching by the 'Shift-and-Share' Algorithm

$W^{M^S}_{\cdots}$	$Y^{\cdot}_1Z^{\cdot}_1$	$Y^{\cdot}_1Z^{\cdot}_2$	$Y^{\cdot}_2Z^{\cdot}_1$	$Y^{\cdot}_2Z^{\cdot}_2$	$Y^{\cdot}_3Z^{\cdot}_1$	$Y^{\cdot}_3Z^{\cdot}_2$	$Y^{\cdot}_4Z^{\cdot}_1$	$Y^{\cdot}_4Z^{\cdot}_2$	
$X^{\cdot}_1$	9312.1	1064.2	1241.6	88.7	5460.3	659.1	438.4	93.7	18358.2
$X^{\cdot}_2$	4789.1	0	1330.3	0	1064.2	0	0	0	7183.6
$X^{\cdot}_3$	49.2	1547.2	1022.2	308.1	2186.1	740.6	245.3	286.8	6385.5
$X^{\cdot}_4$	6625.4	4815.2	1494.8	1165.8	6027.1	5413.5	1213.6	116.7	26872.1
	20775.8	7426.6	5088.9	1562.6	14737.7	6813.2	1897.3	497.2	58799.4

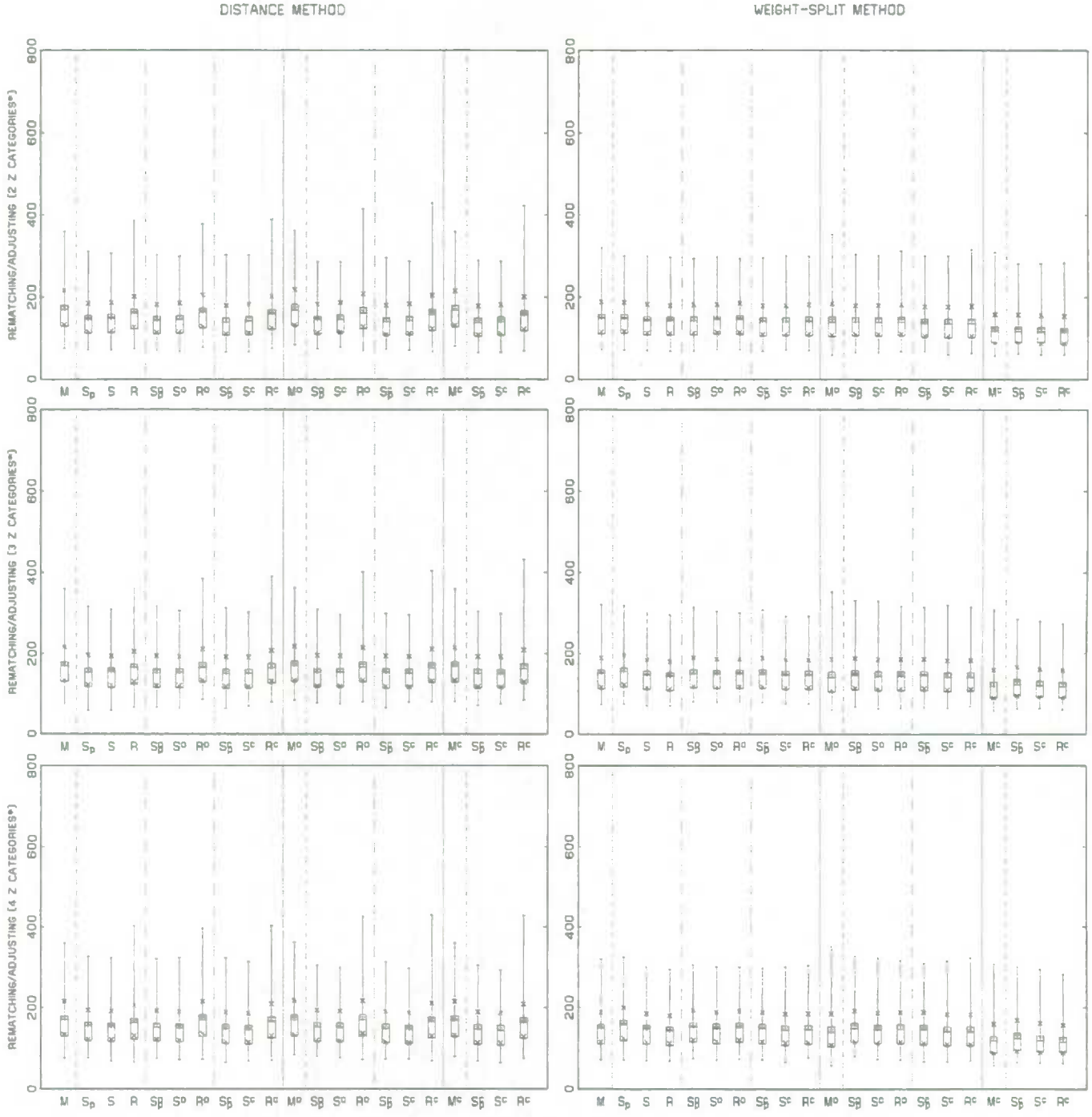
We use this new distribution along with the *new* calibrated version of the look-up table for the ratio adjustment of individual weights of the revised matched file M^{S}. Note that in this illustration the new calibrated look-up table according to $W^{M^S}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}$ is exactly equal to the original look-up table $W^{L}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}$, since all unexpected empty cells in matched file M are filled after shift-and-share rematching, and in $W^{L}_{X^{\cdot}Y^{\cdot}Z^{\cdot}}$ no cell needs setting back to zero.

Figure A1. Weighted χ^2 Evaluated over 10x4x2 Categories of the Matched File
(Computed over 250 Simulations for MQM datafile)



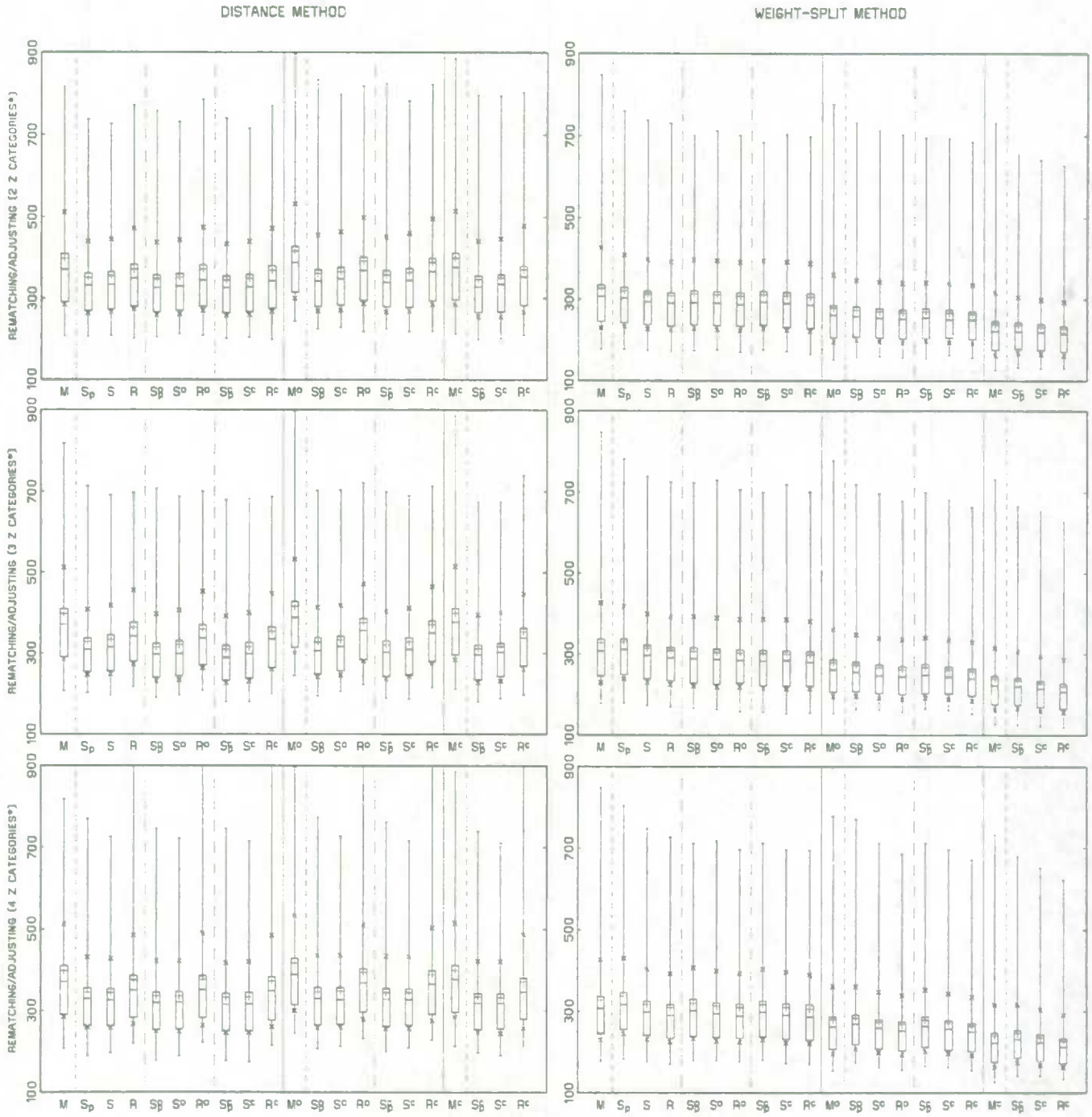
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A2. Weighted χ^2 Evaluated over 10x4x2 Categories of the Matched File
(Computed over 500 Simulations for RUO datafile)



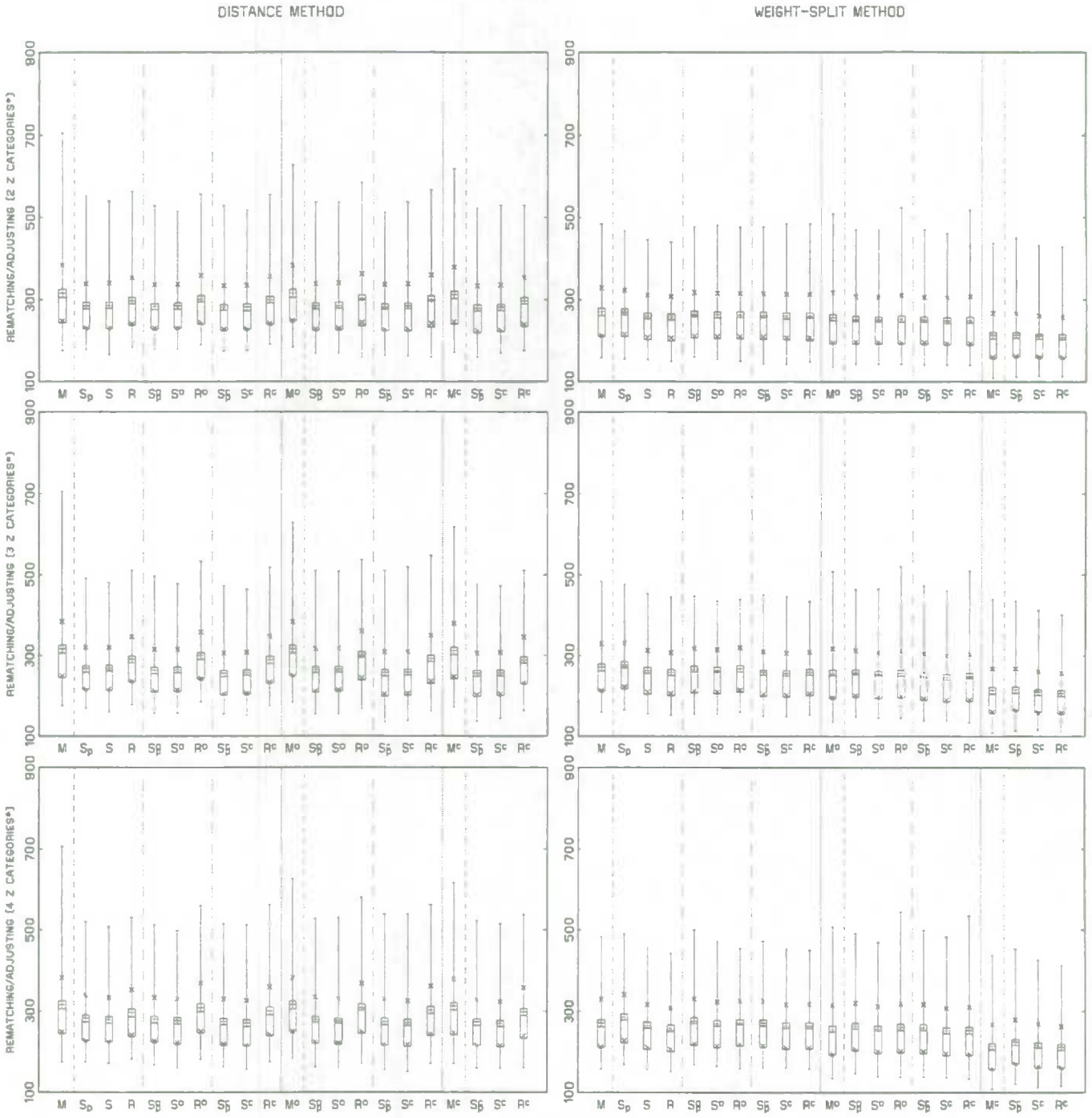
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile · : minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A3. Weighted χ^2 Evaluated over 10x4x3 Categories of the Matched File
(Computed over 250 Simulations for MQM datafile)



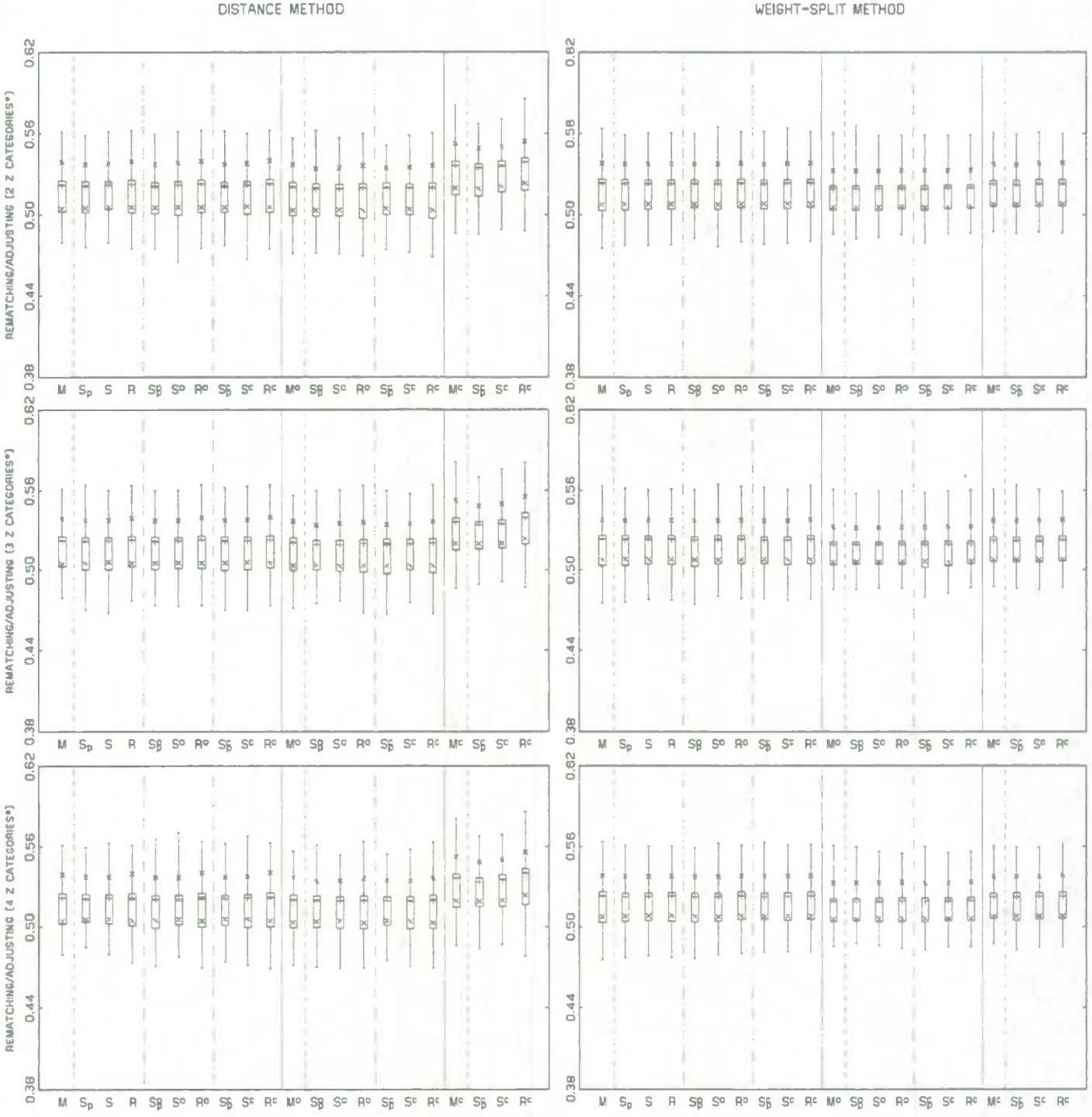
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A4. Weighted χ^2 Evaluated over 10x4x3 Categories of the Matched File
(Computed over 500 Simulations for RUO datafile)



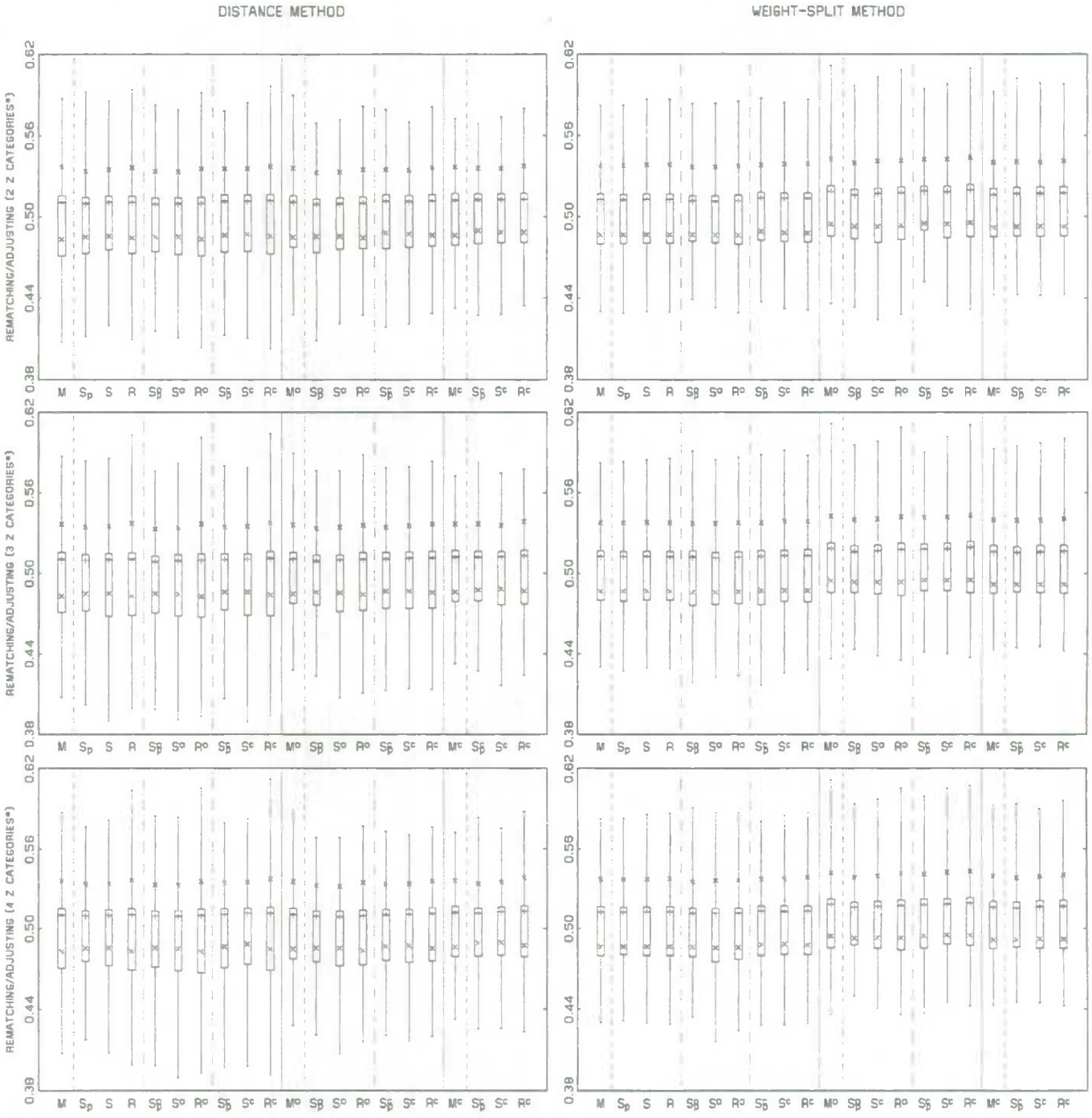
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile · : minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A5. Weighted κ Evaluated over 10x4x2 Categories of the Matched File
(Computed over 250 Simulations for MQM datafile)



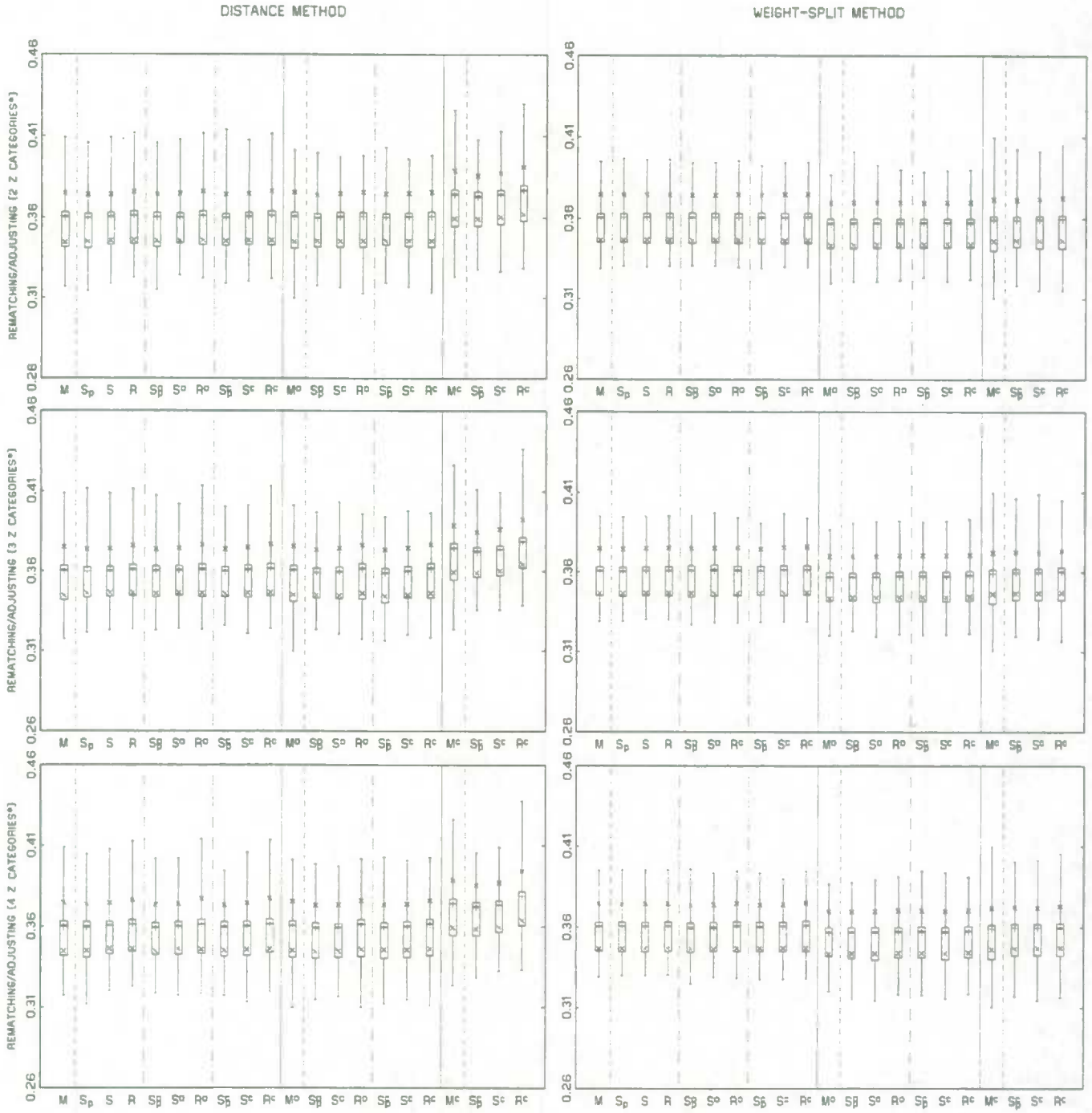
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A6. Weighted κ Evaluated over 10x4x2 Categories of the Matched File
(Computed over 500 Simulations for RUO datafile)



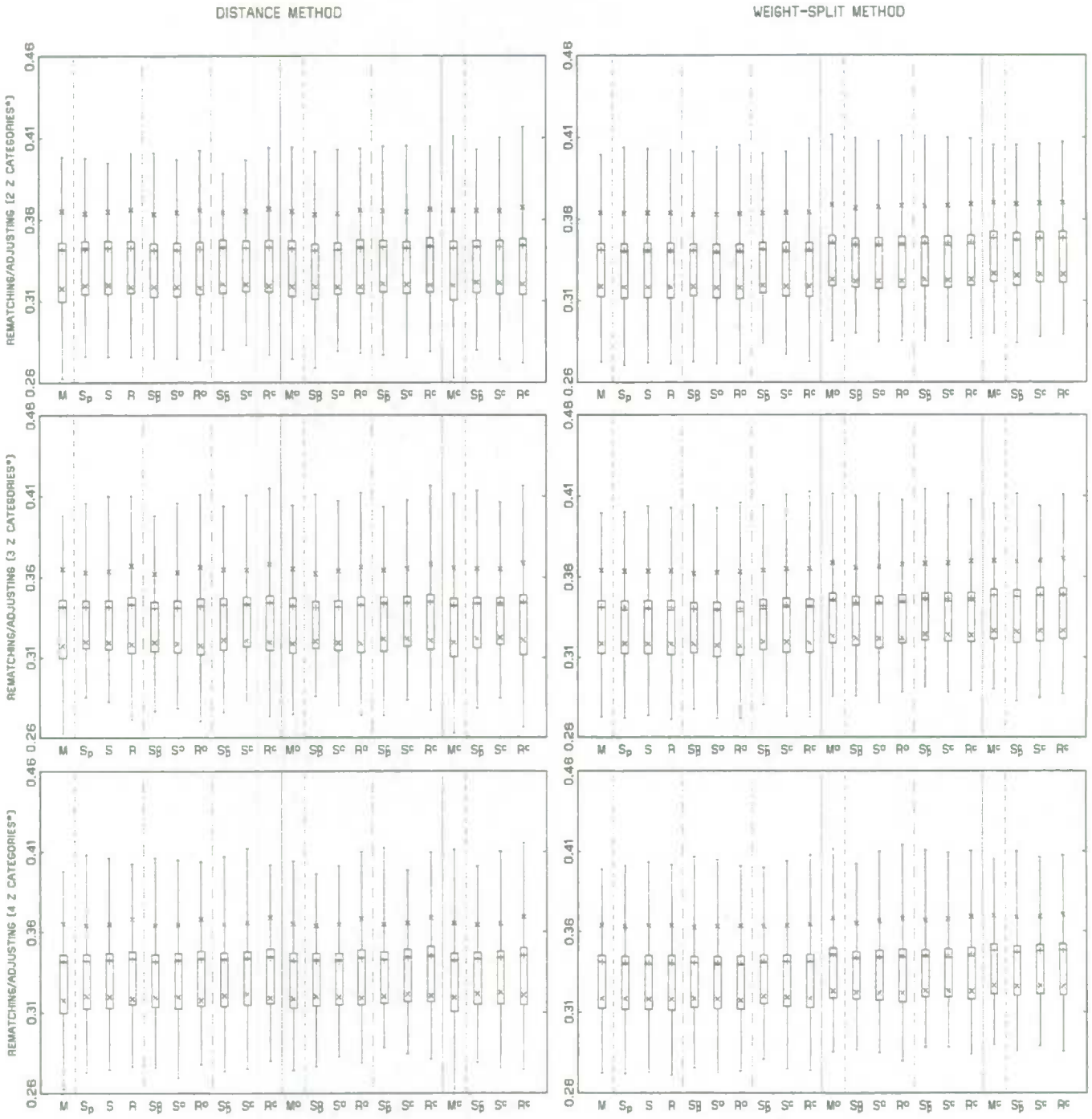
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A7. Weighted κ Evaluated over 10x4x3 Categories of the Matched File
(Computed over 250 Simulations for MQM datafile)



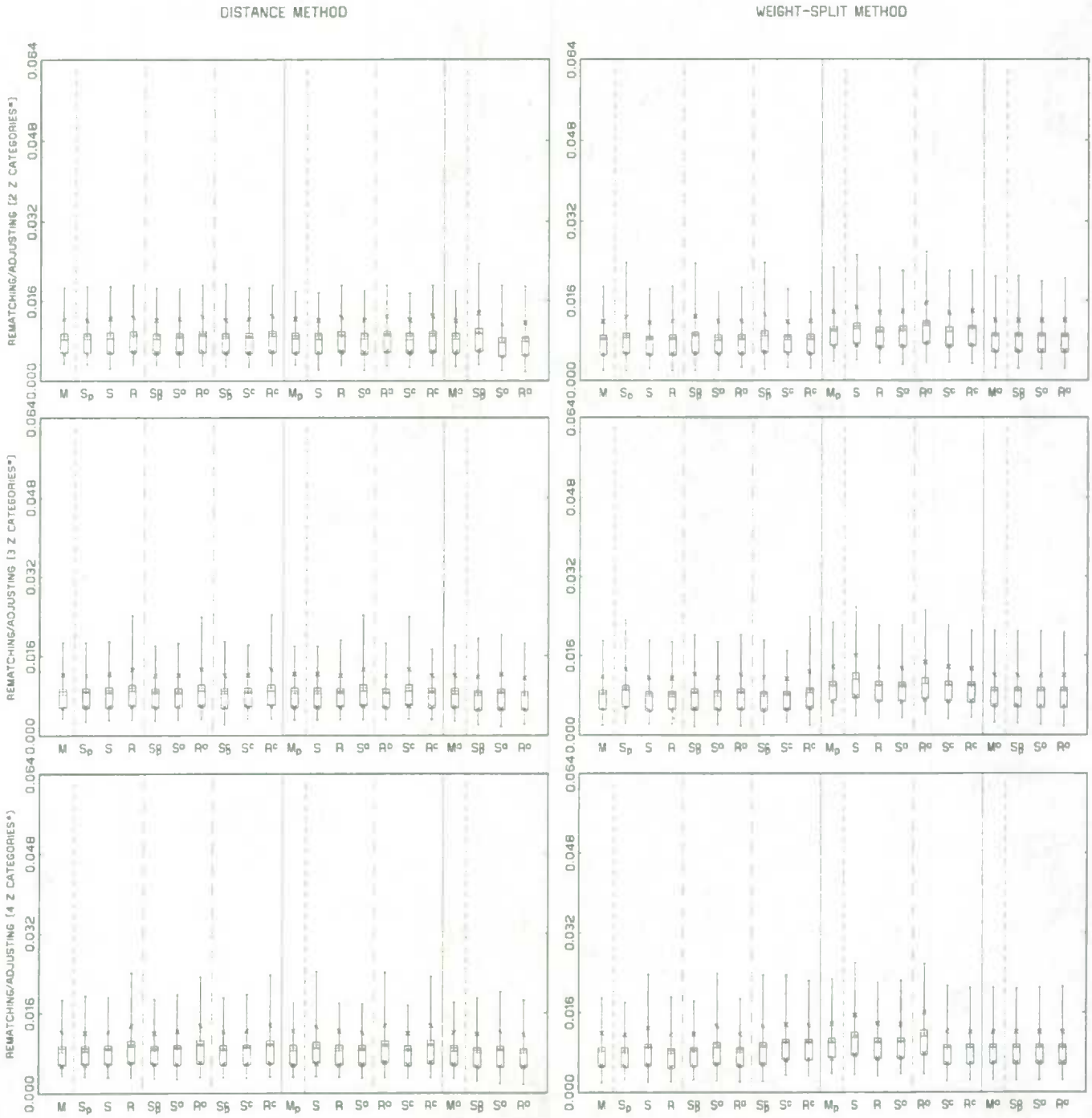
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A8. Weighted κ Evaluated over 10x4x3 Categories of the Matched File
(Computed over 500 Simulations for RUO datafile)



+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile · : minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A9. Weighted δ-difference Index ($\delta = 0$)[1] of the Matched File
(Computed over 250 Simulations for MQM Datafile)

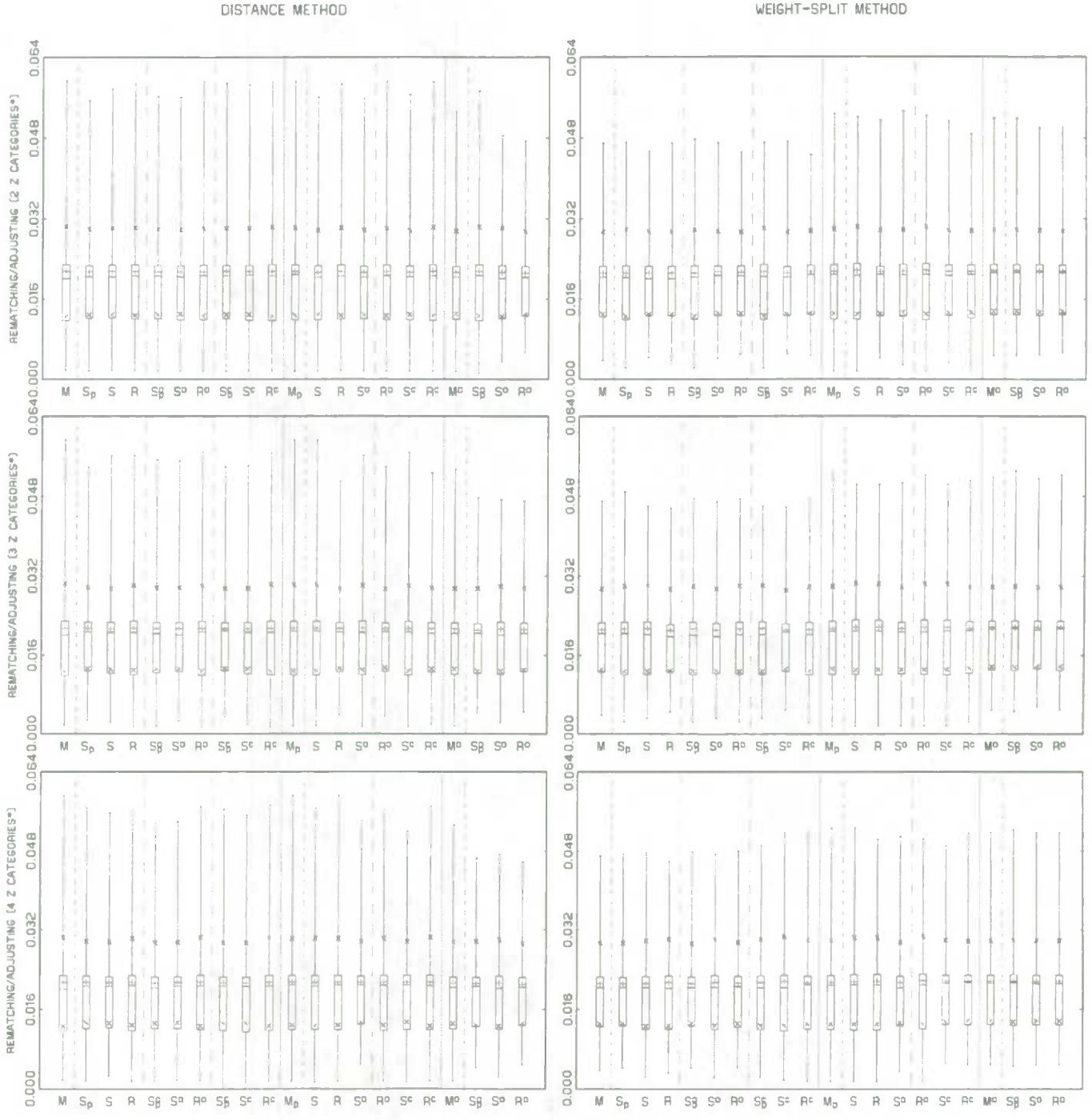


+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ◦: minimum or maximum

* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching and ratio adjustment.

[1] $\delta = 0$ means the imputed and the true values of the variables are identical.

Figure A10. Weighted δ-difference Index ($\delta = 0$)[1] of the Matched File
(Computed over 500 Simulations for RUO Datafile)

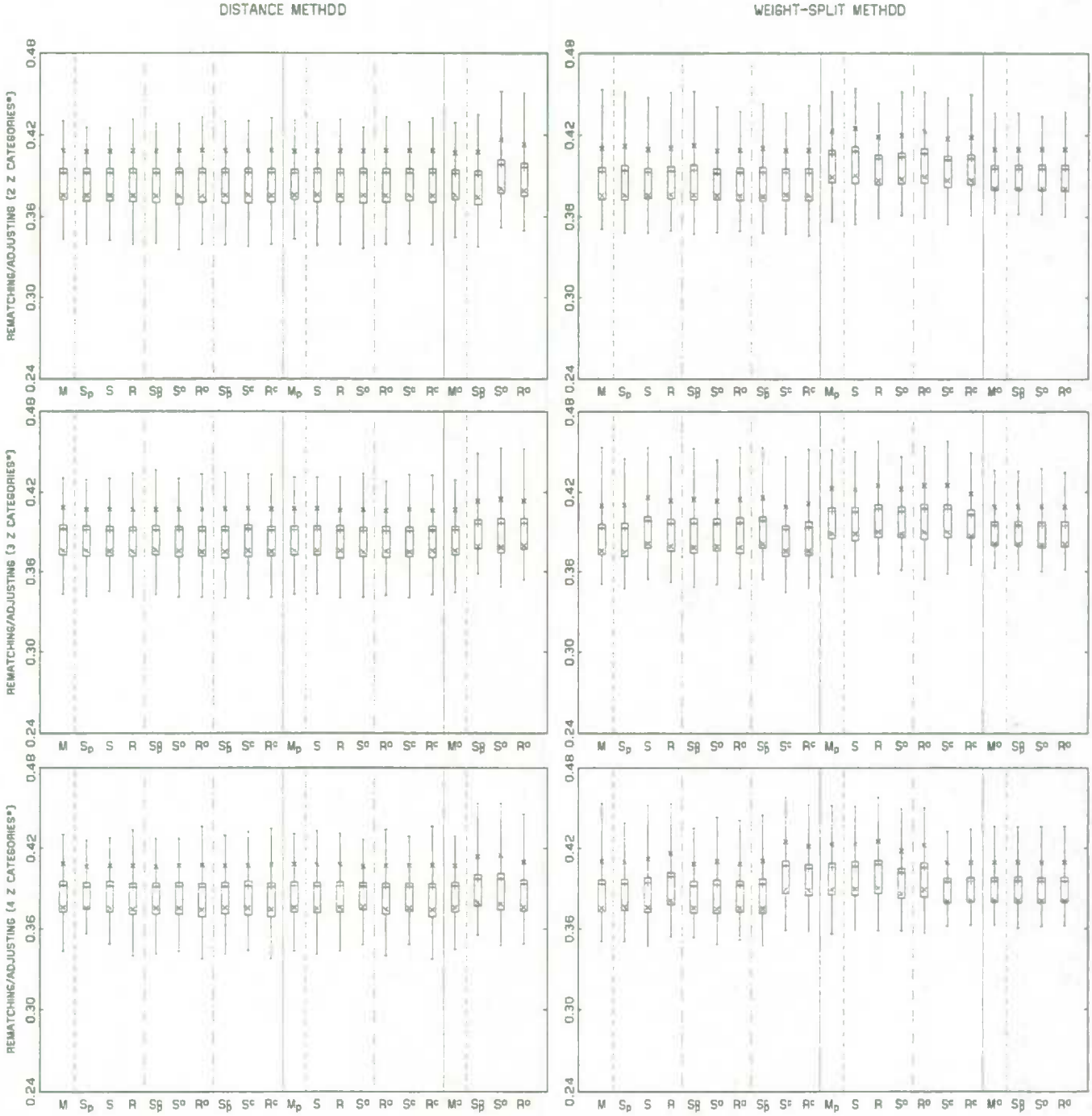


+: mean ·: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile · : minimum or maximum

* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching and ratio adjustment.

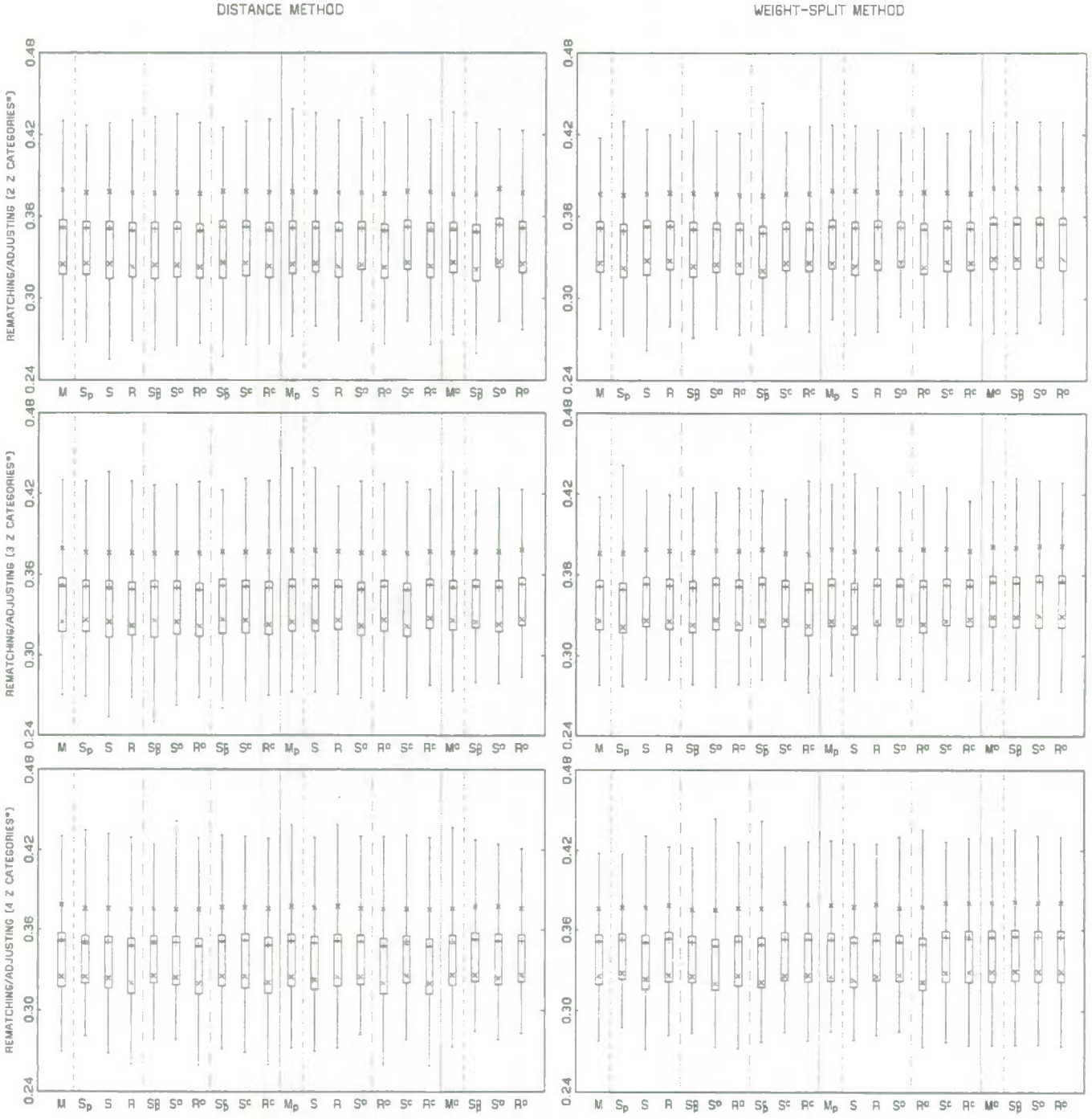
[1] $\delta = 0$ means the imputed and the true values of the variables are identical.

Figure A11. Weighted δ-difference Index (δ = 0.5) of the Matched File
(Computed over 250 Simulations for MQM Datafile)



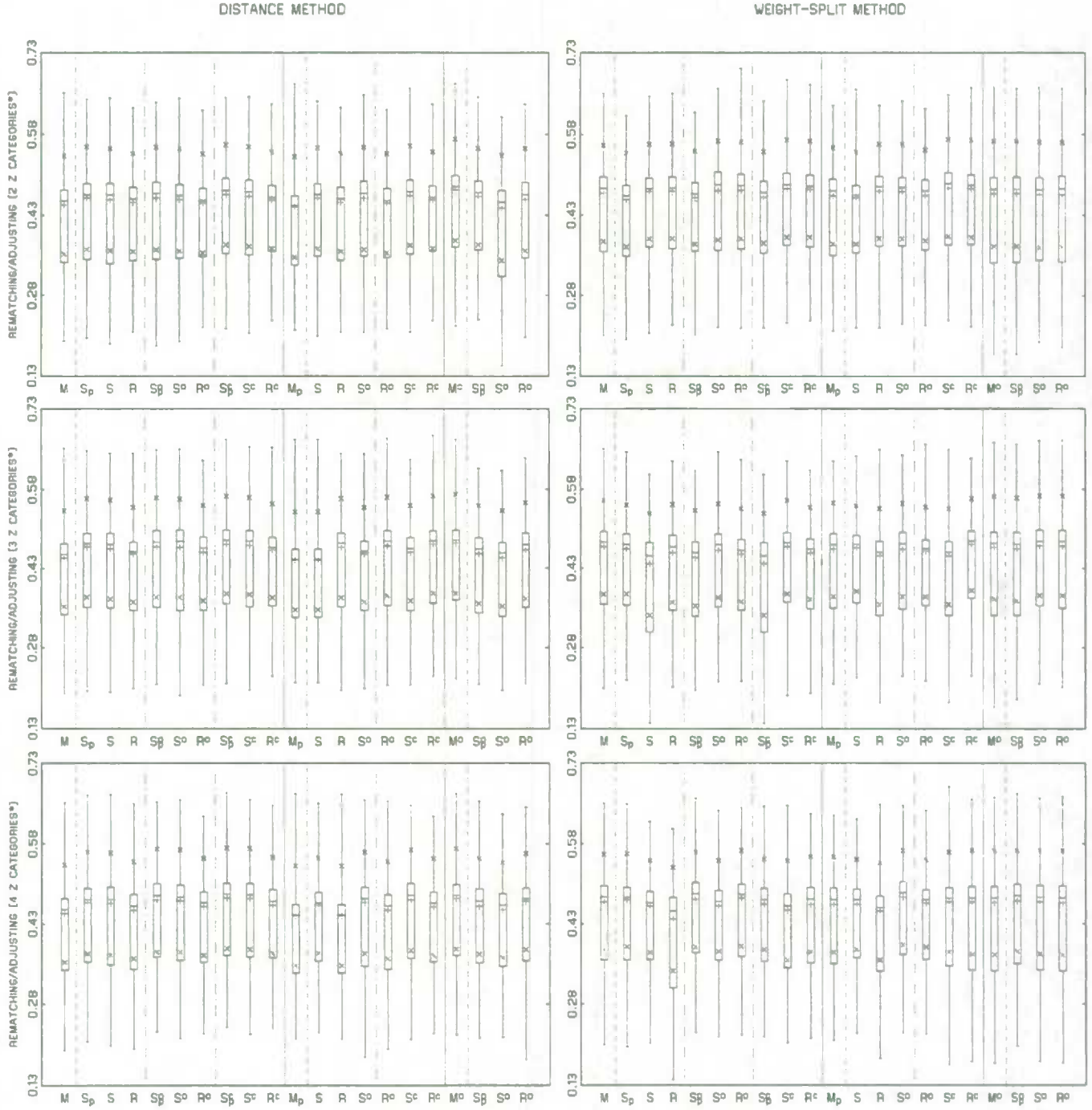
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A12. Weighted δ-difference Index ($\delta = 0.5$) of the Matched File
(Computed over 500 Simulations for RUO Datafile)



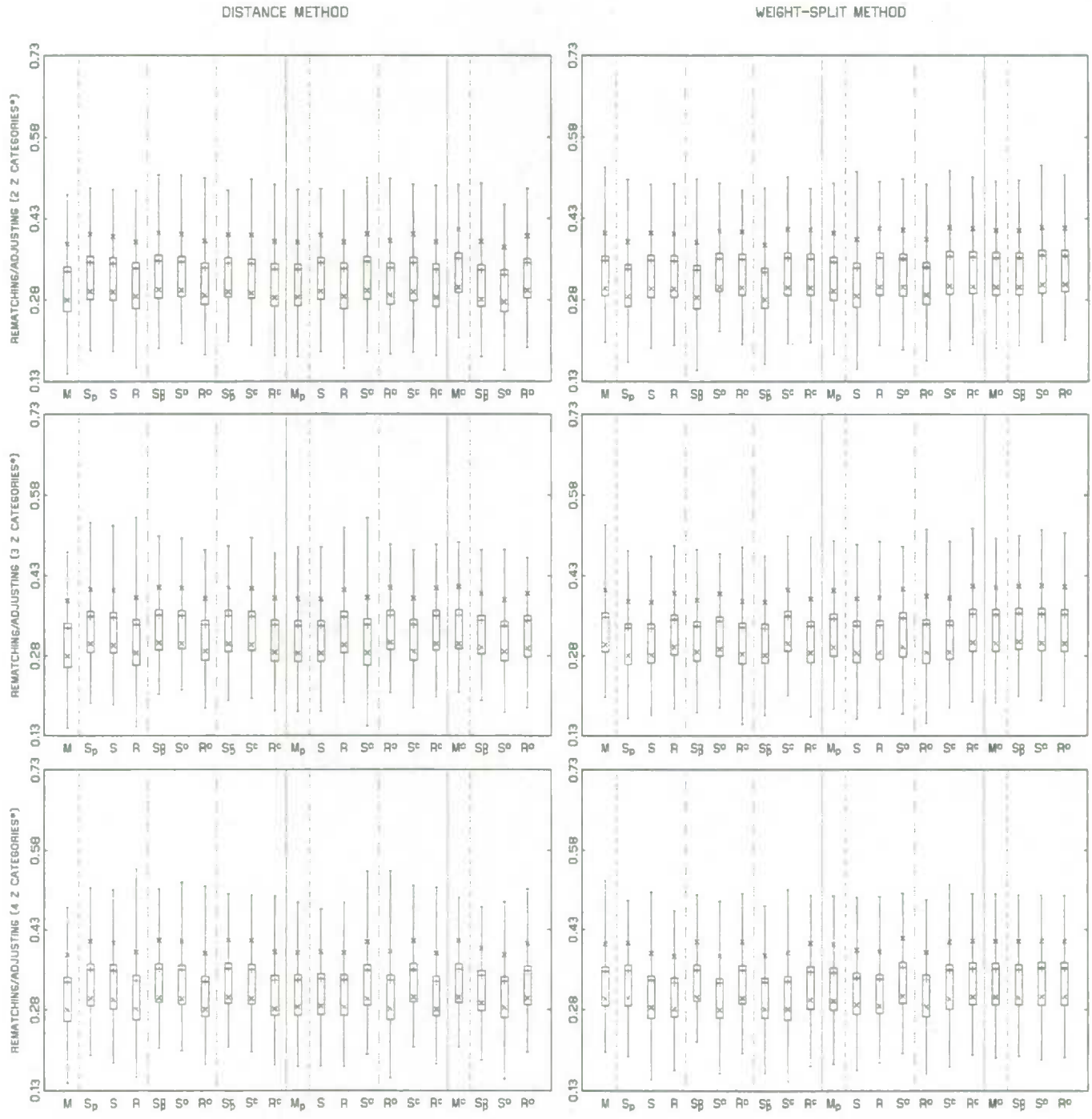
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A13. Weighted ε-difference Index (ε = 0.005) of the Matched File
(Computed over 250 Simulations for MQM Datafile)



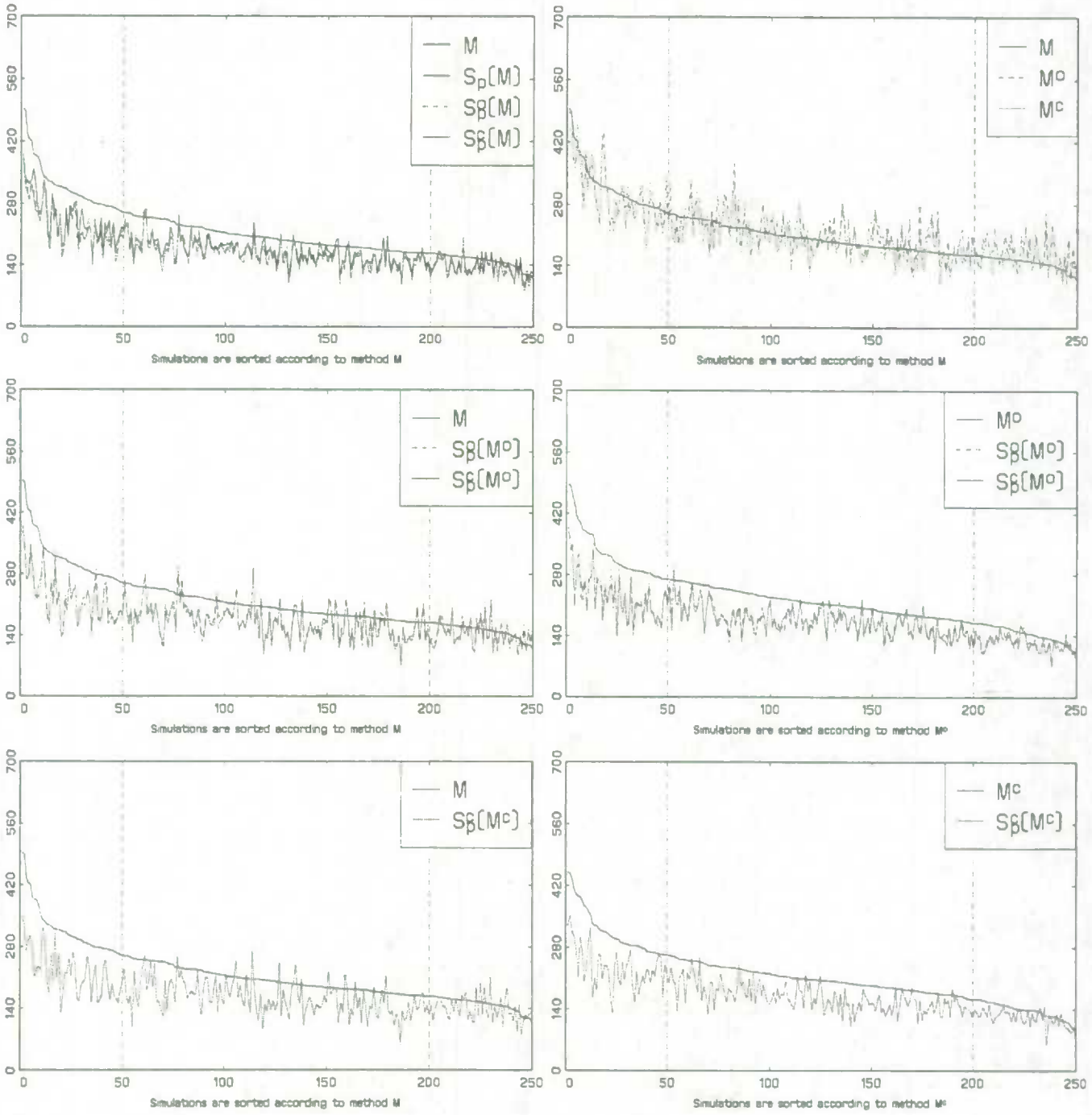
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure A14. Weighted ε-difference Index (ε = 0.005) of the Matched File
(Computed over 500 Simulations for RUO Datafile)



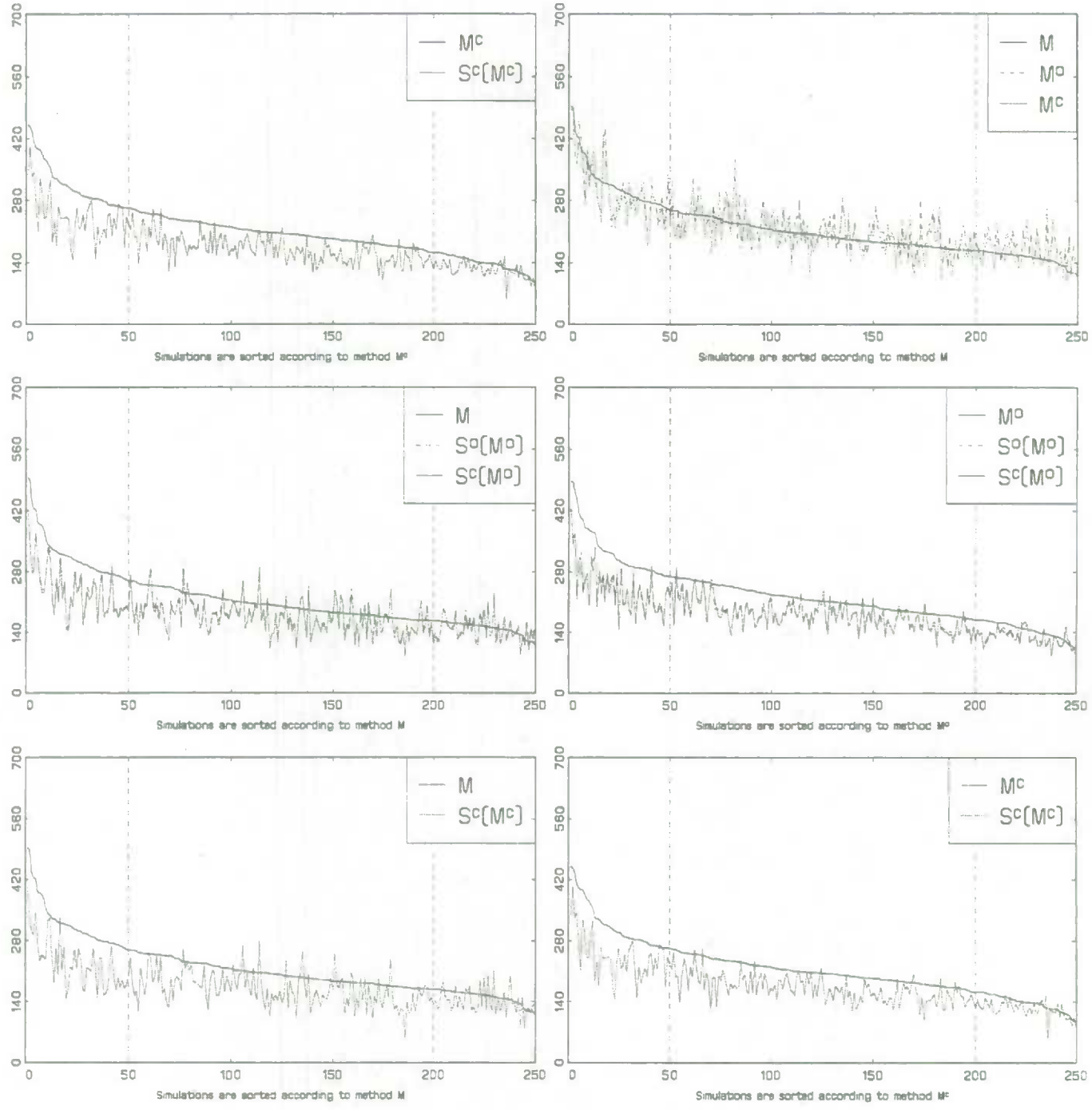
+: mean ×: ±std -: median ⊤: 1st quartile ⊥: 3rd quartile ·: minimum or maximum
* "The number of Z categories" refers to a number of categories of Z used for pooling and for a look-up table construction, and accordingly for rematching or ratio adjustment.

Figure B1.1. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 250 Simulations for MQM Datafile)



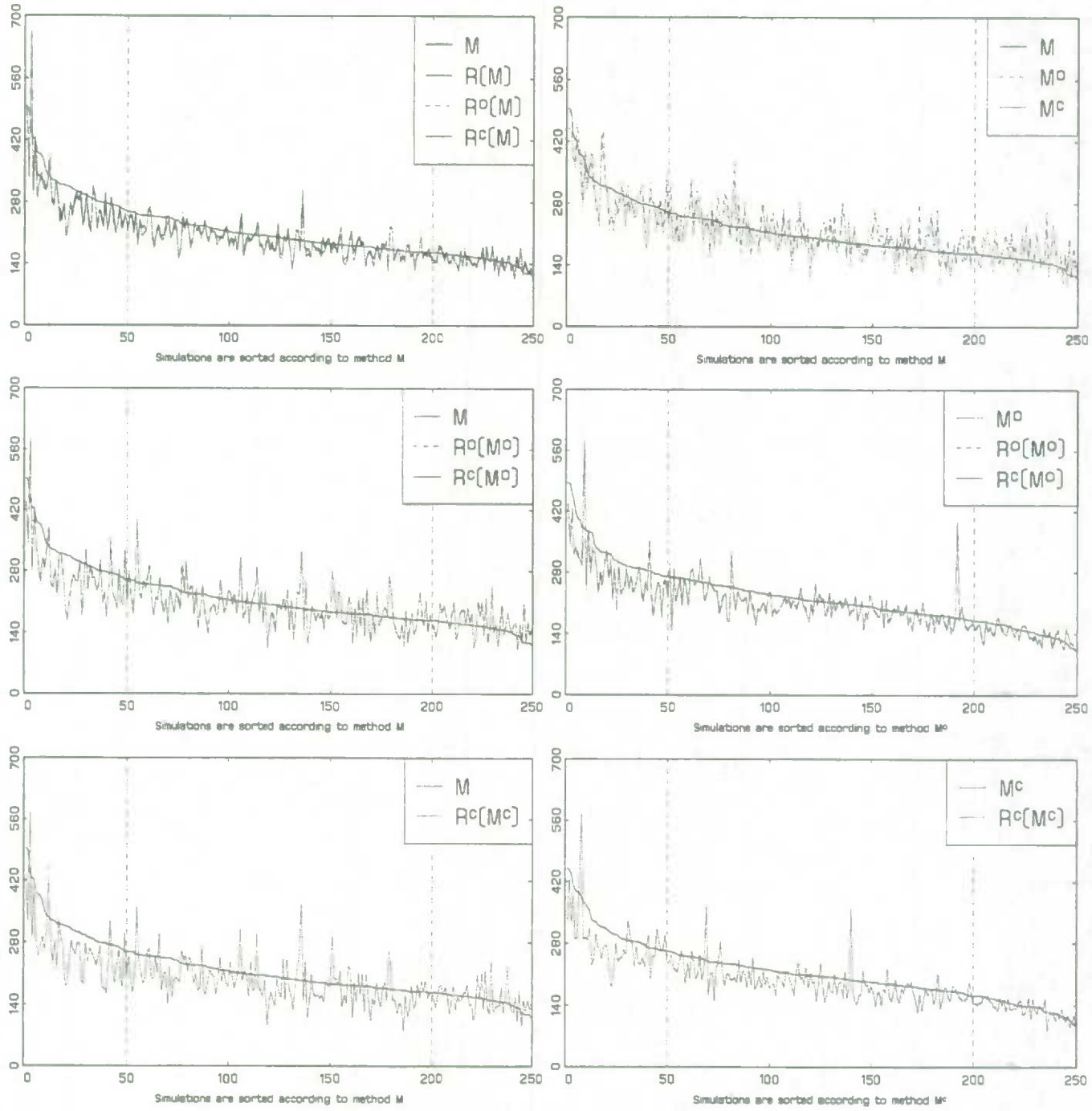
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.1-B1.3 and B1.7-B1.9.

Figure B1.2. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 250 Simulations for MQM Datafile)



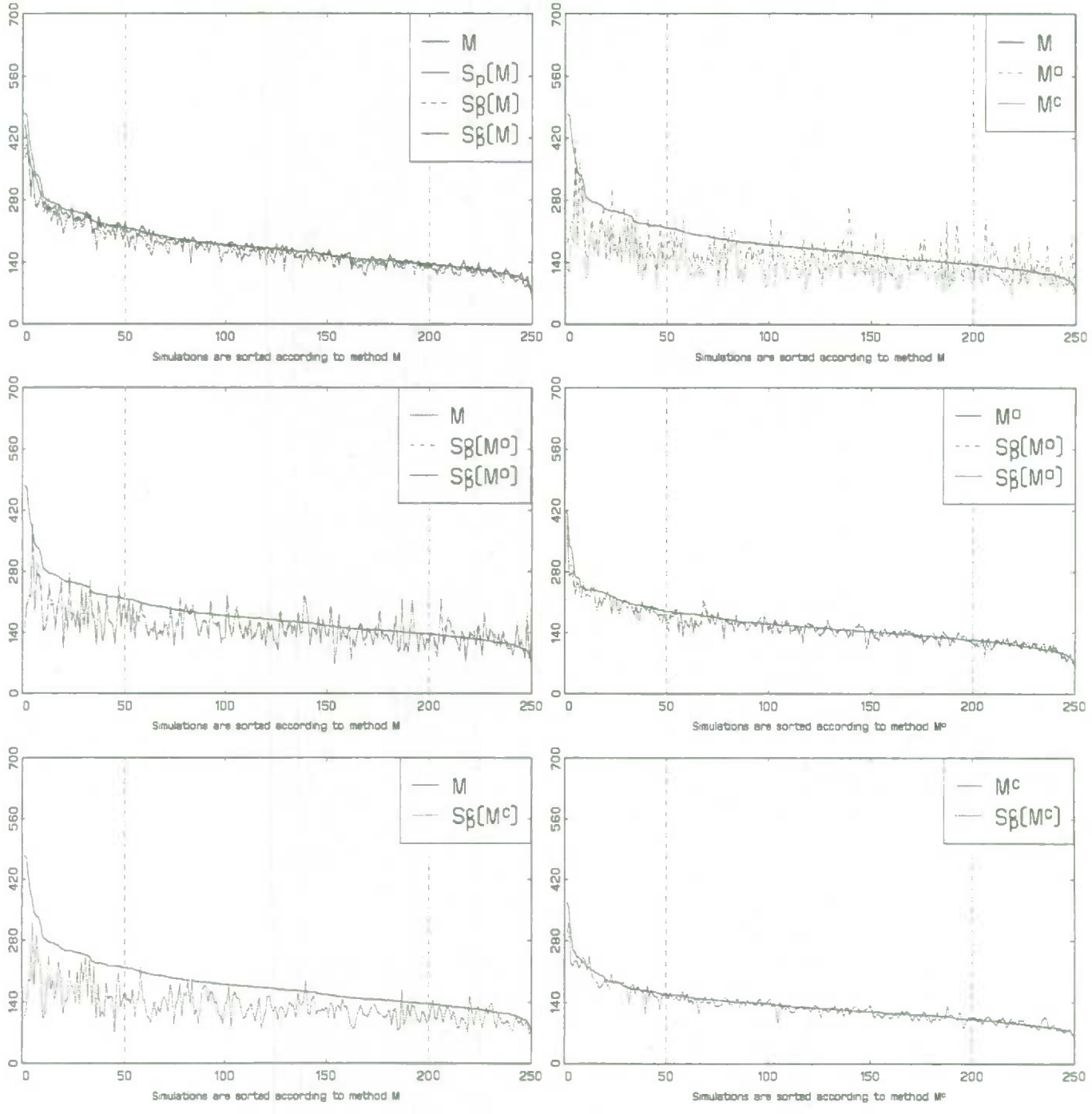
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.1-B1.3 and B1.7-B1.9.

Figure B1.3. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Ratio Adjustment Based on 2 Z Categories: 250 Simulations for MQM Datafile)



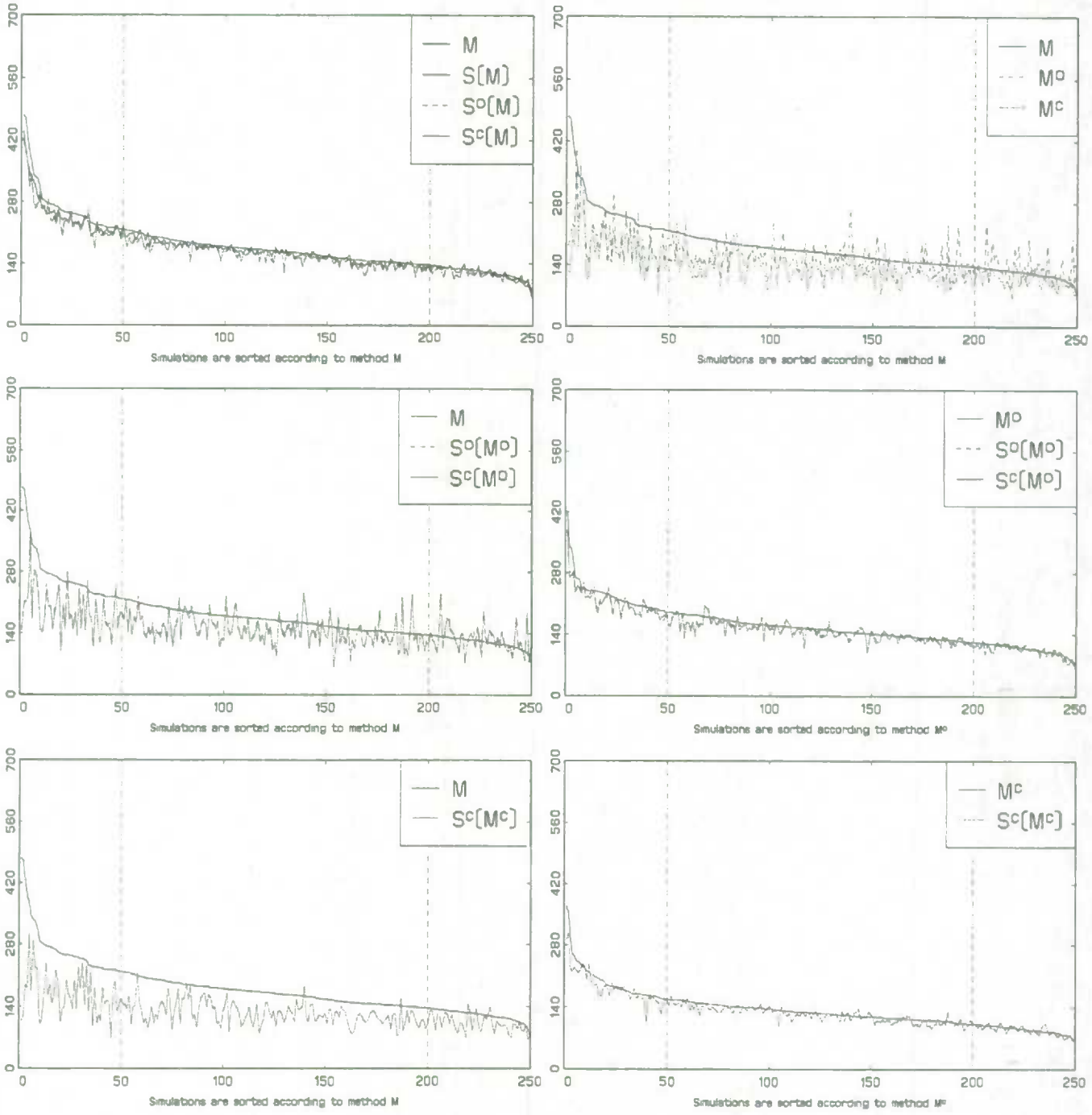
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.1-B1.3 and B1.7-B1.9.

Figure B1.4. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 250 Simulations for MQM Datafile)



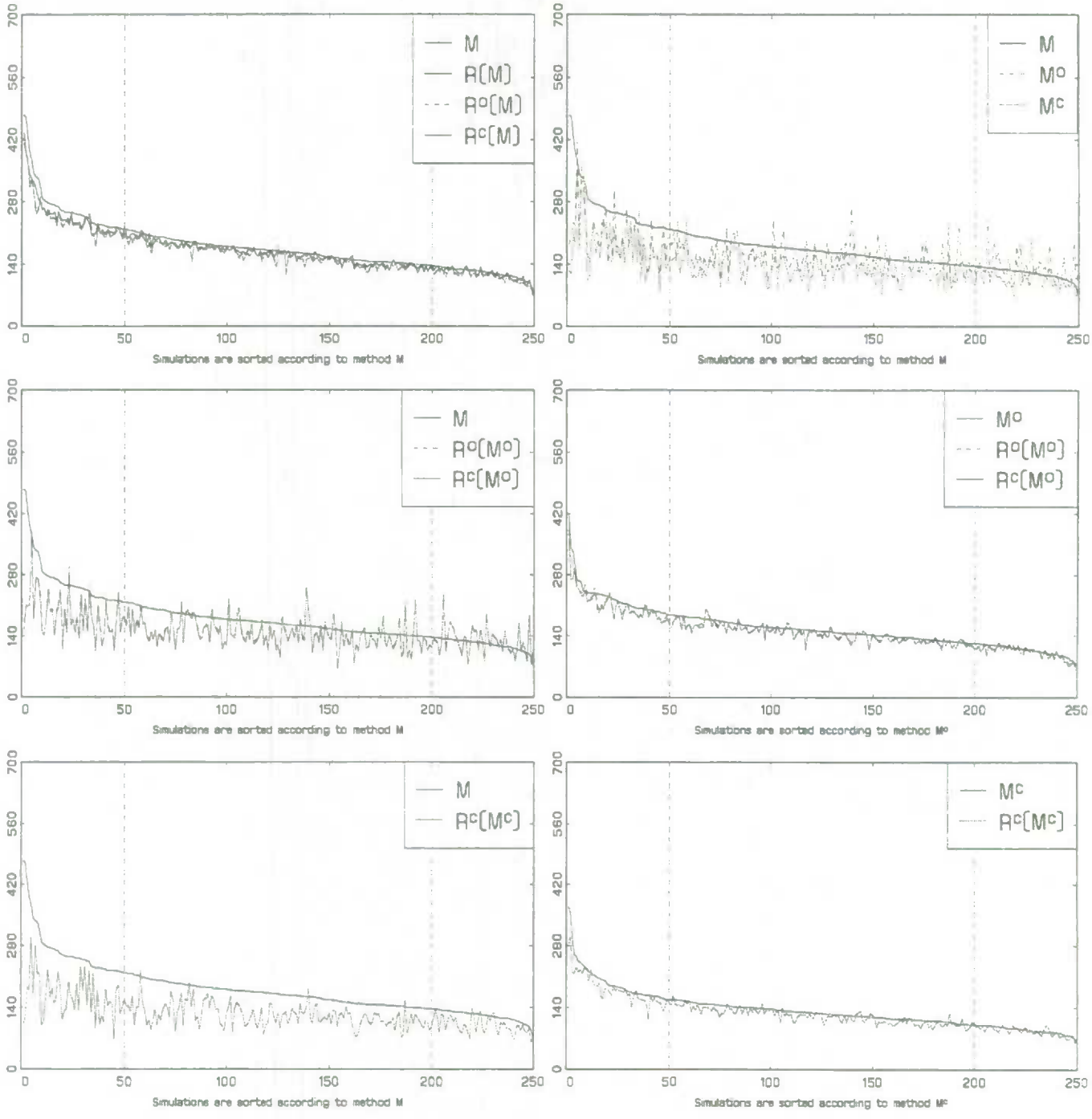
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B1.4-B1.6 and B1.10-B1.12.

Figure B1.5. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 250 Simulations for MQM Datafile)



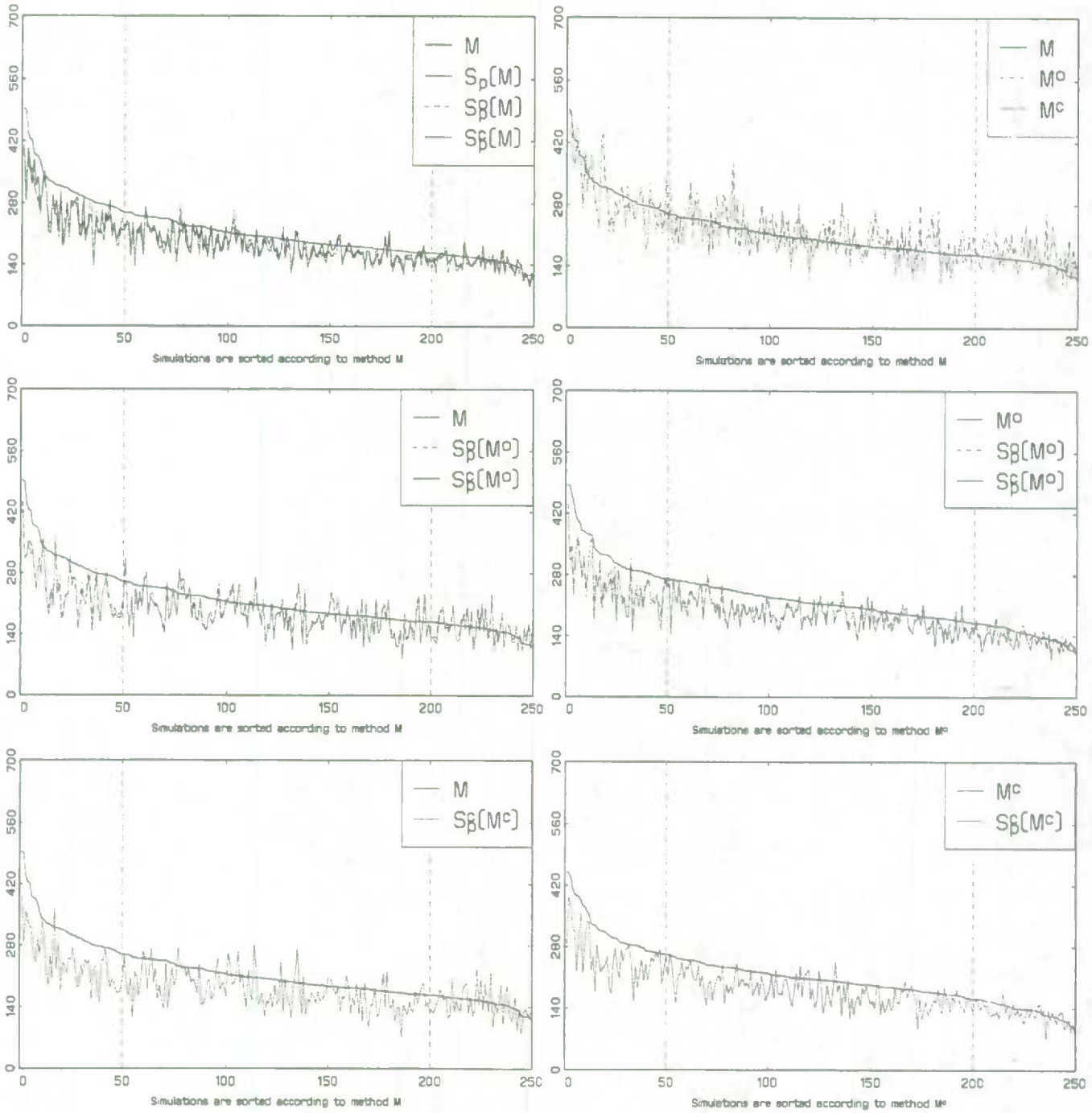
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.4-B16 and B1.10-B1.12.

Figure B1.6. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 2 Z Categories: 250 Simulations for MQM Datafile)



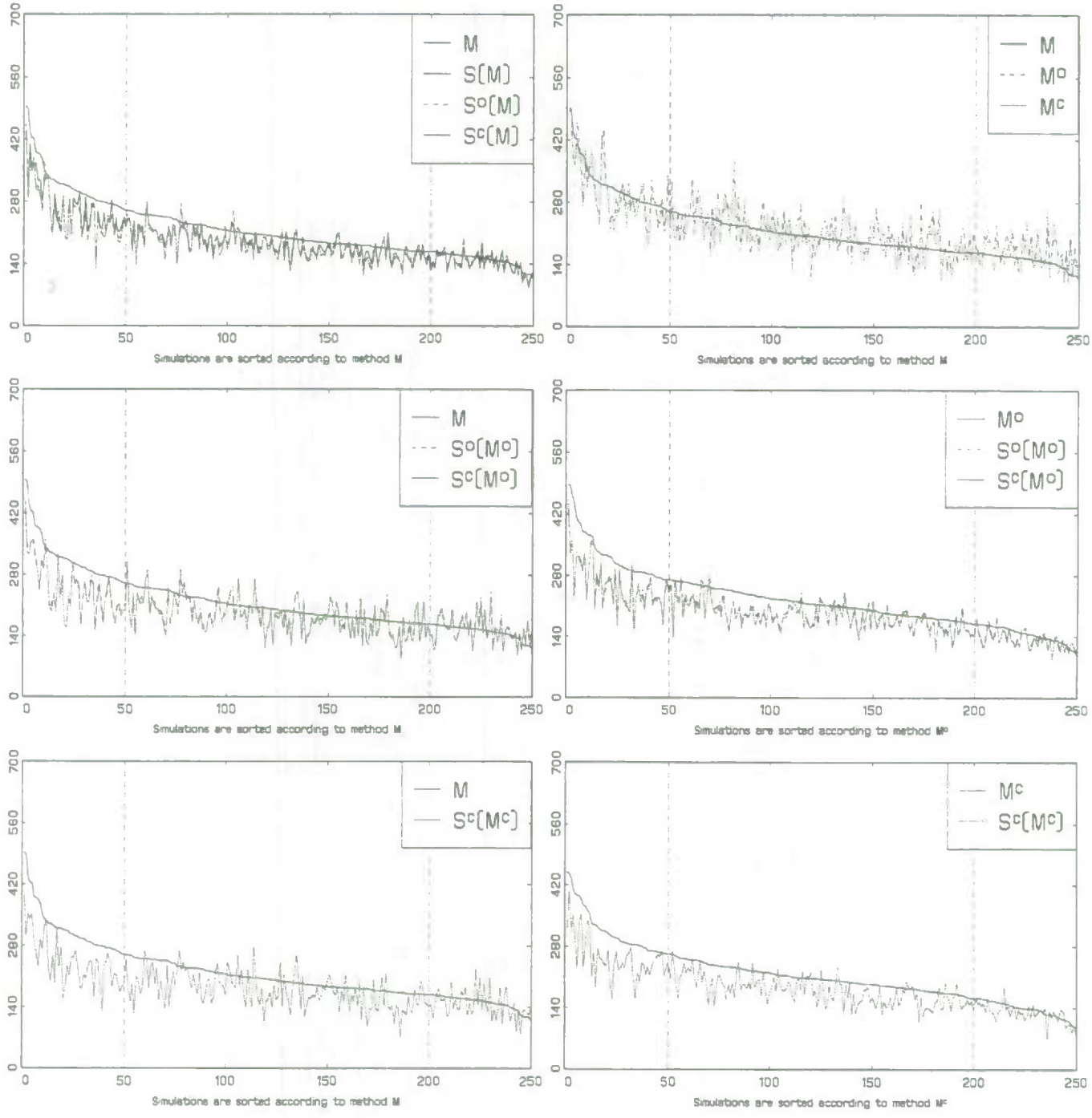
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B1.4-B1.6 and B1.10-B1.12.

Figure B1.7. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 3 Z Categories: 250 Simulations for MQM Datafile)



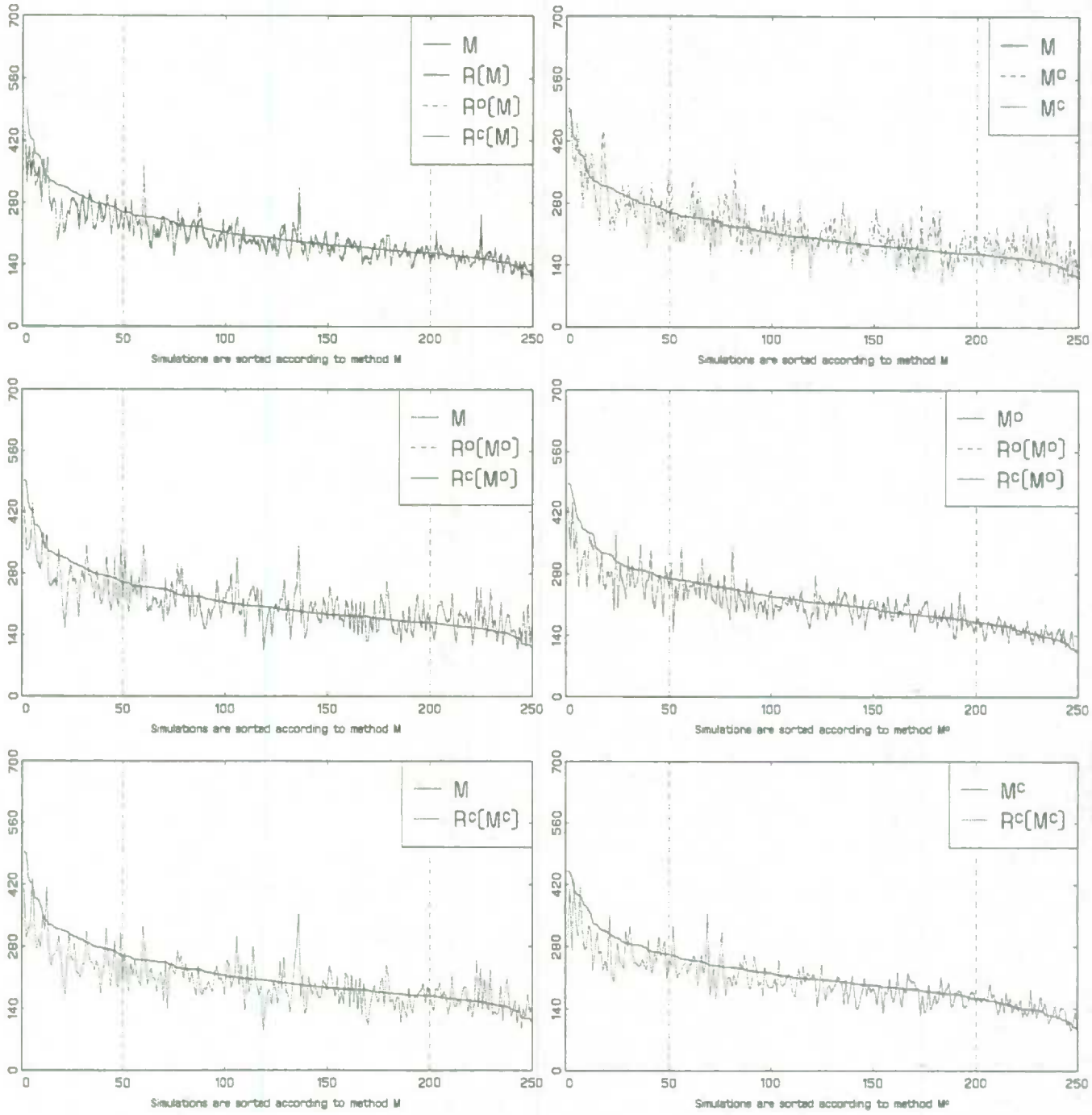
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.1-B1.3 and B1.7-B1.9.

Figure B1.8. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 3 Z Categories: 250 Simulations for MQM Datafile)



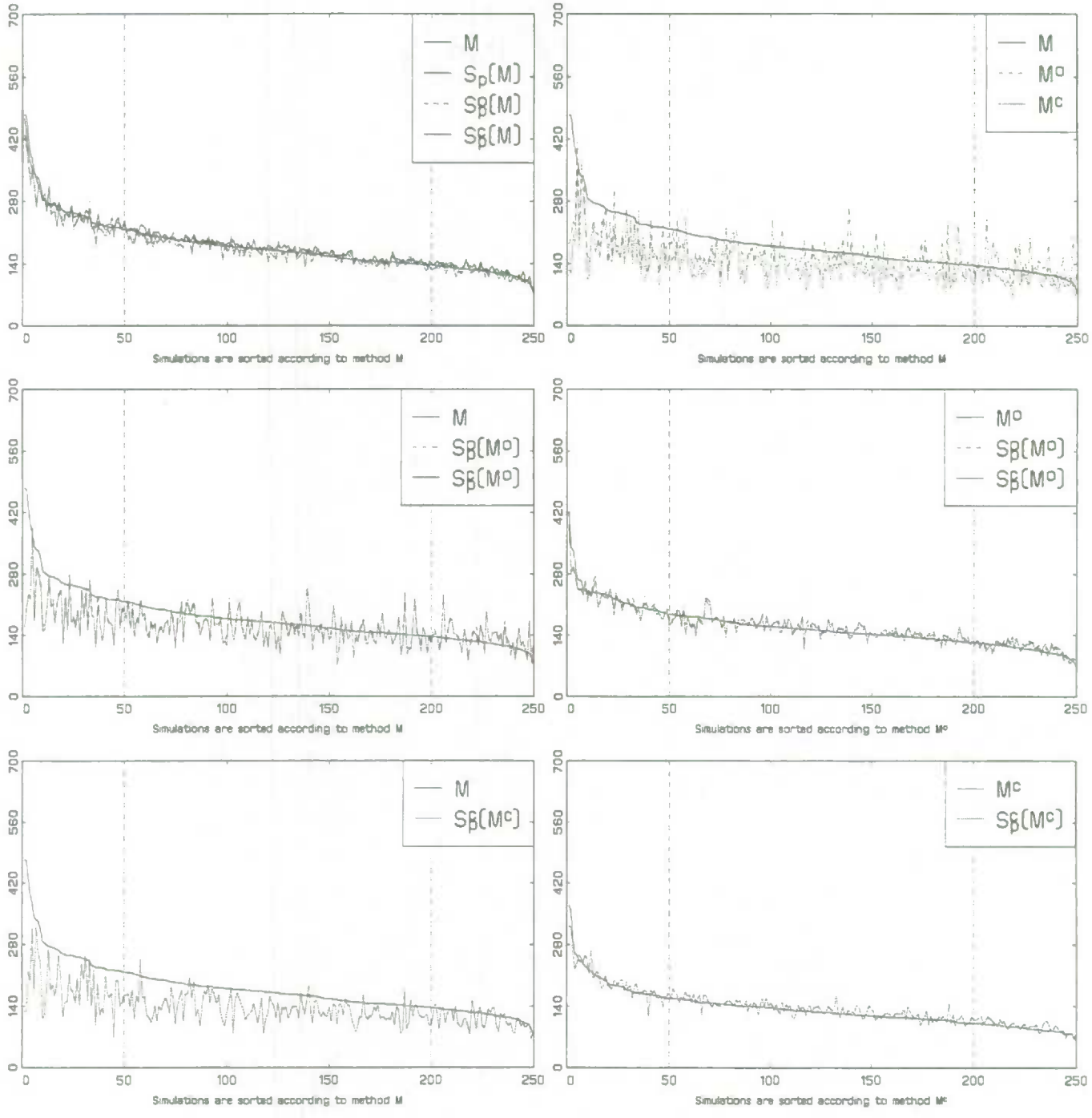
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.1-B1.3 and B1.7-B1.9.

Figure B1.9. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Ratio Adjustment Based on 3 Z Categories: 250 Simulations for MQM Datafile)



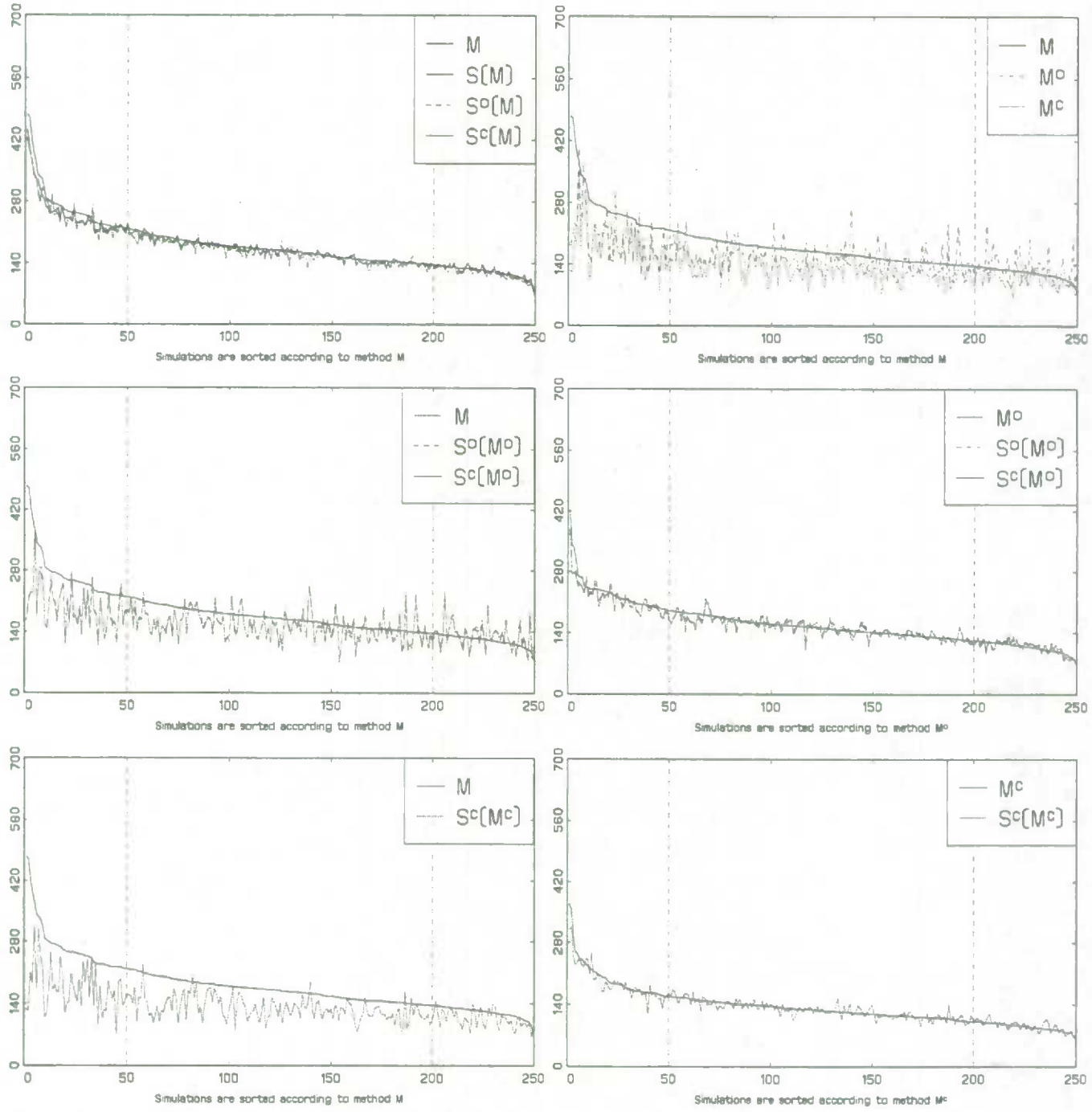
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.1-B1.3 and B1.7-B1.9.

Figure B1.10. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 3 Z Categories: 250 Simulations for MQM Datafile)



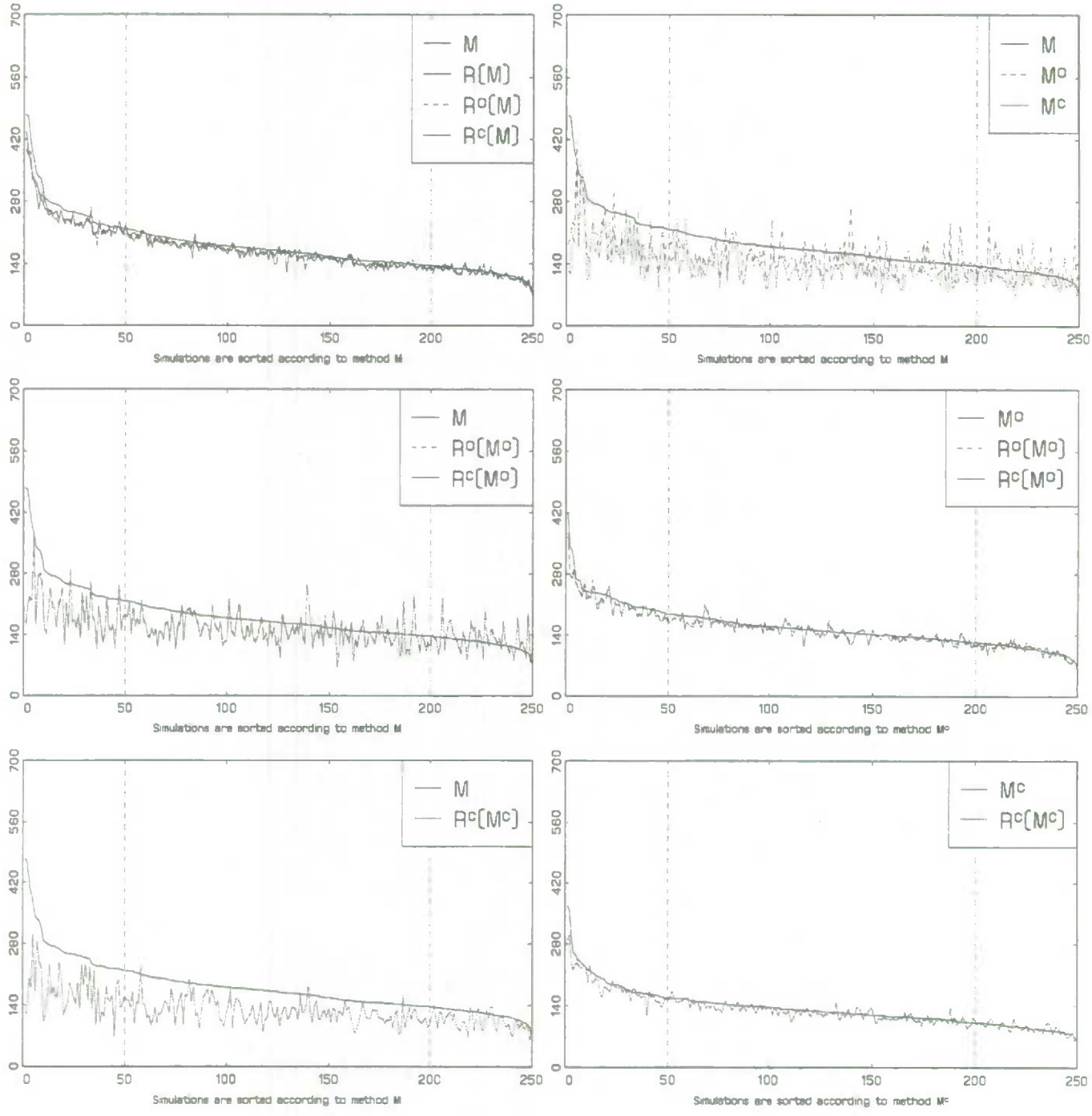
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B1.4-B1.6 and B1.10-B1.12.

Figure B1.11. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 3 Z Categories: 250 Simulations for MQM Datafile)



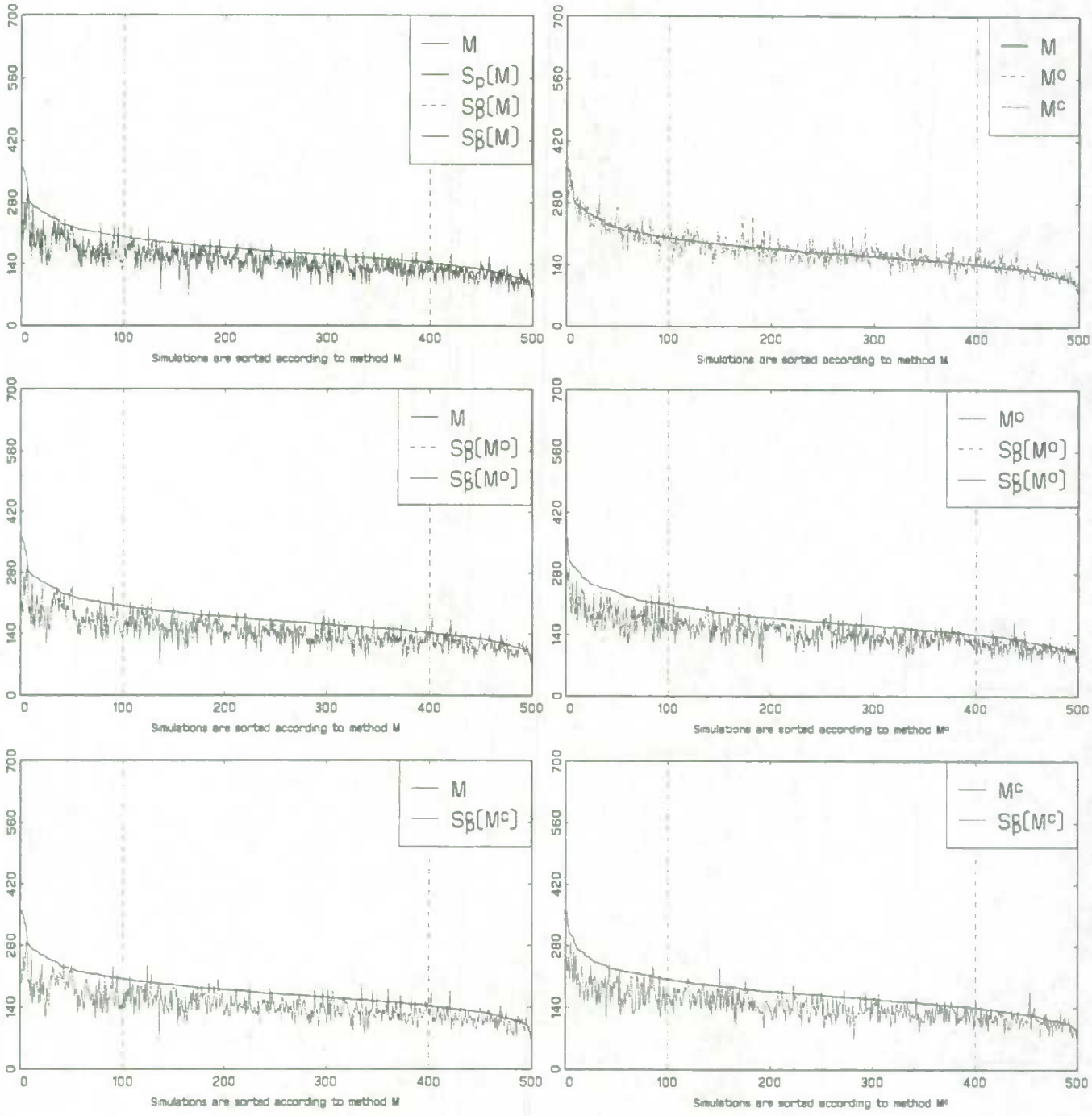
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.4-B1.6 and B1.10-B1.12.

Figure B1.12. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 3 Z Categories: 250 Simulations for MQM Datafile)



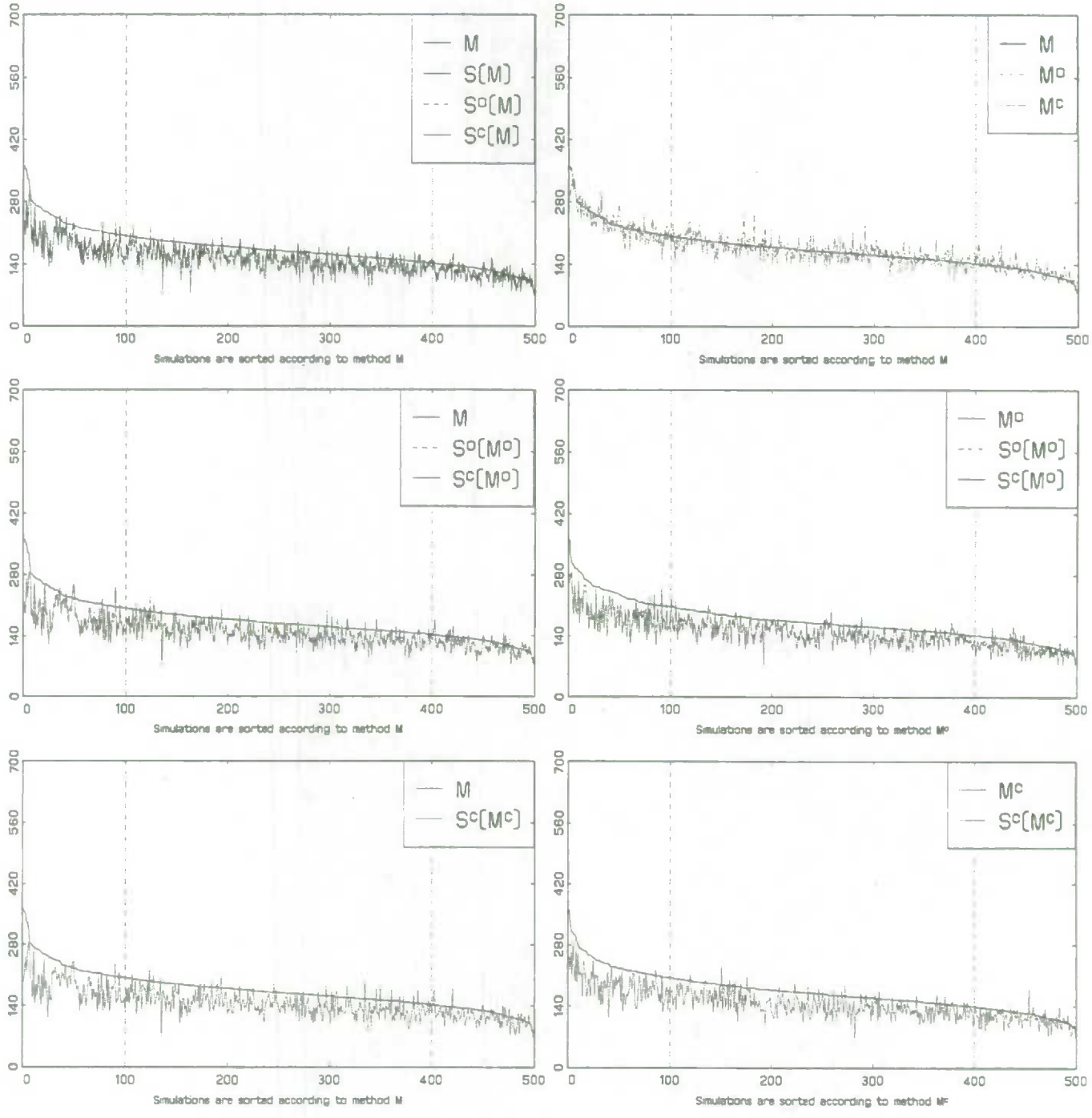
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B1.4-B1.6 and B1.10-B1.12.

Figure B2.1. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 500 Simulations for RUO Datafile)



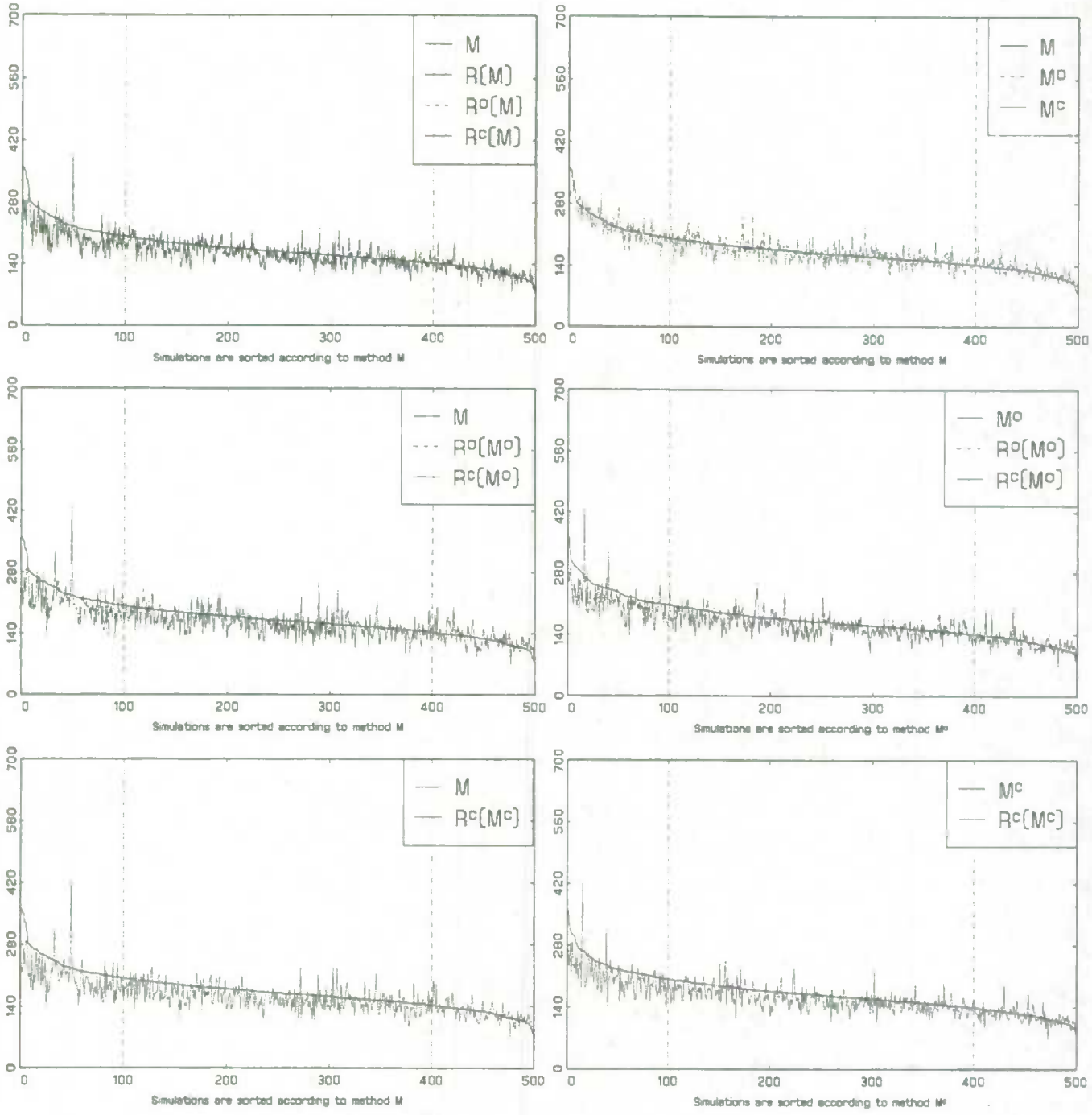
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.1-B2.3 and B2.7-B2.9.

Figure B2.2. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 500 Simulations for RUO Datafile)



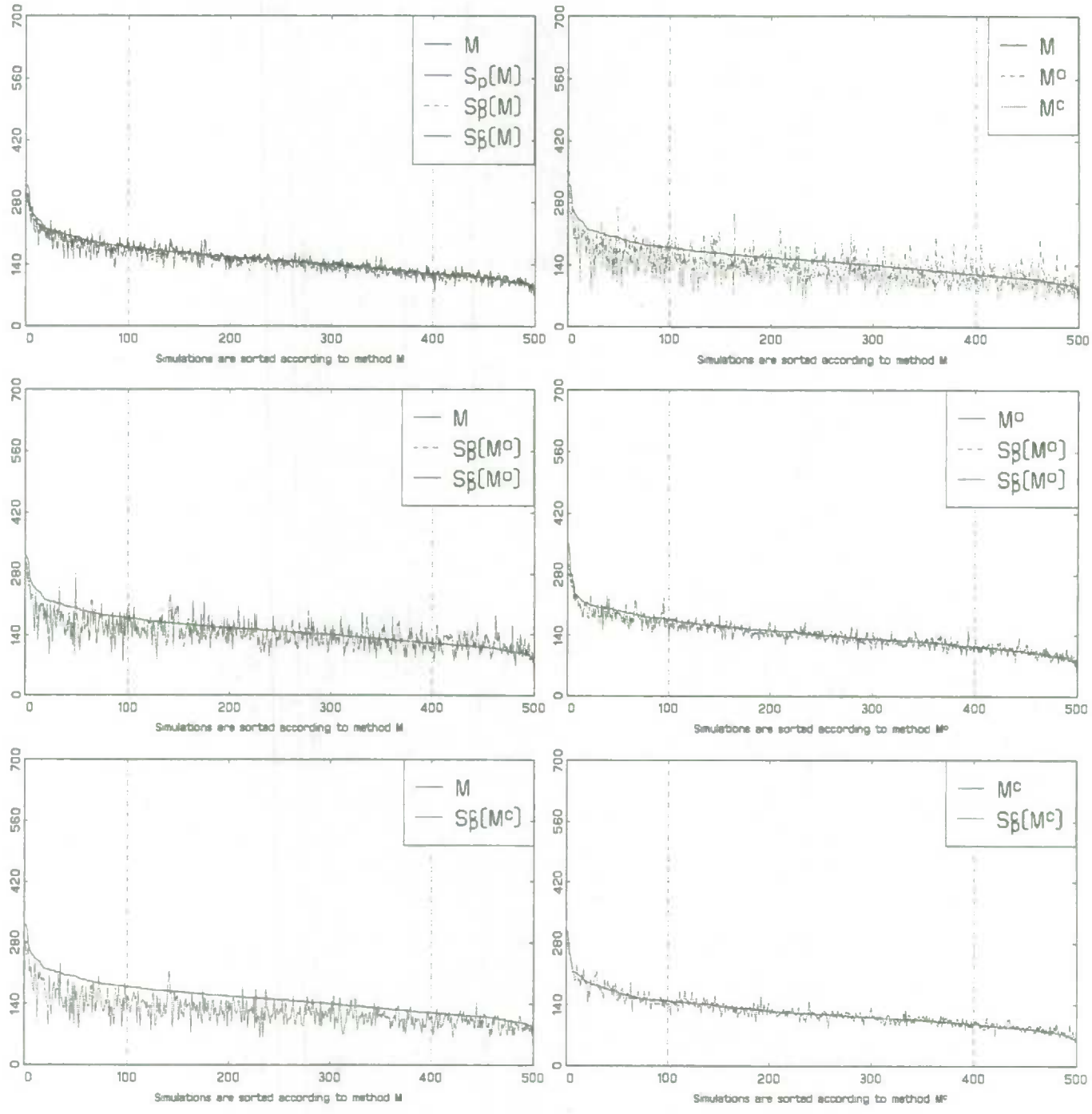
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.1-B2.3 and B2.7-B2.9.

Figure B2.3. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Ratio Adjustment Based on 2 Z Categories: 500 Simulations for RUO Datafile)



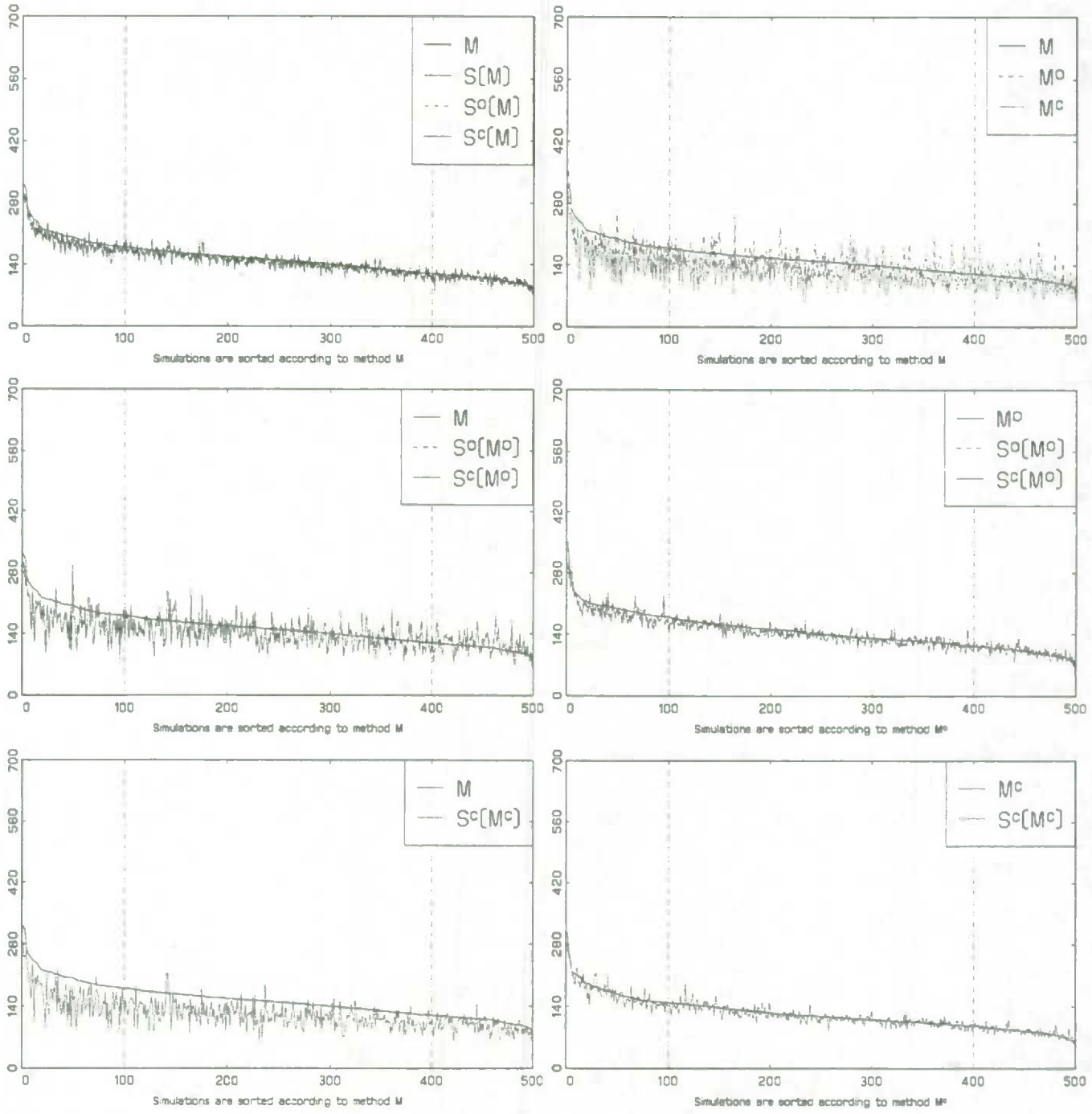
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.1-B2.3 and B2.7-B2.9.

Figure B2.4 Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 500 Simulations for RUO Datafile)



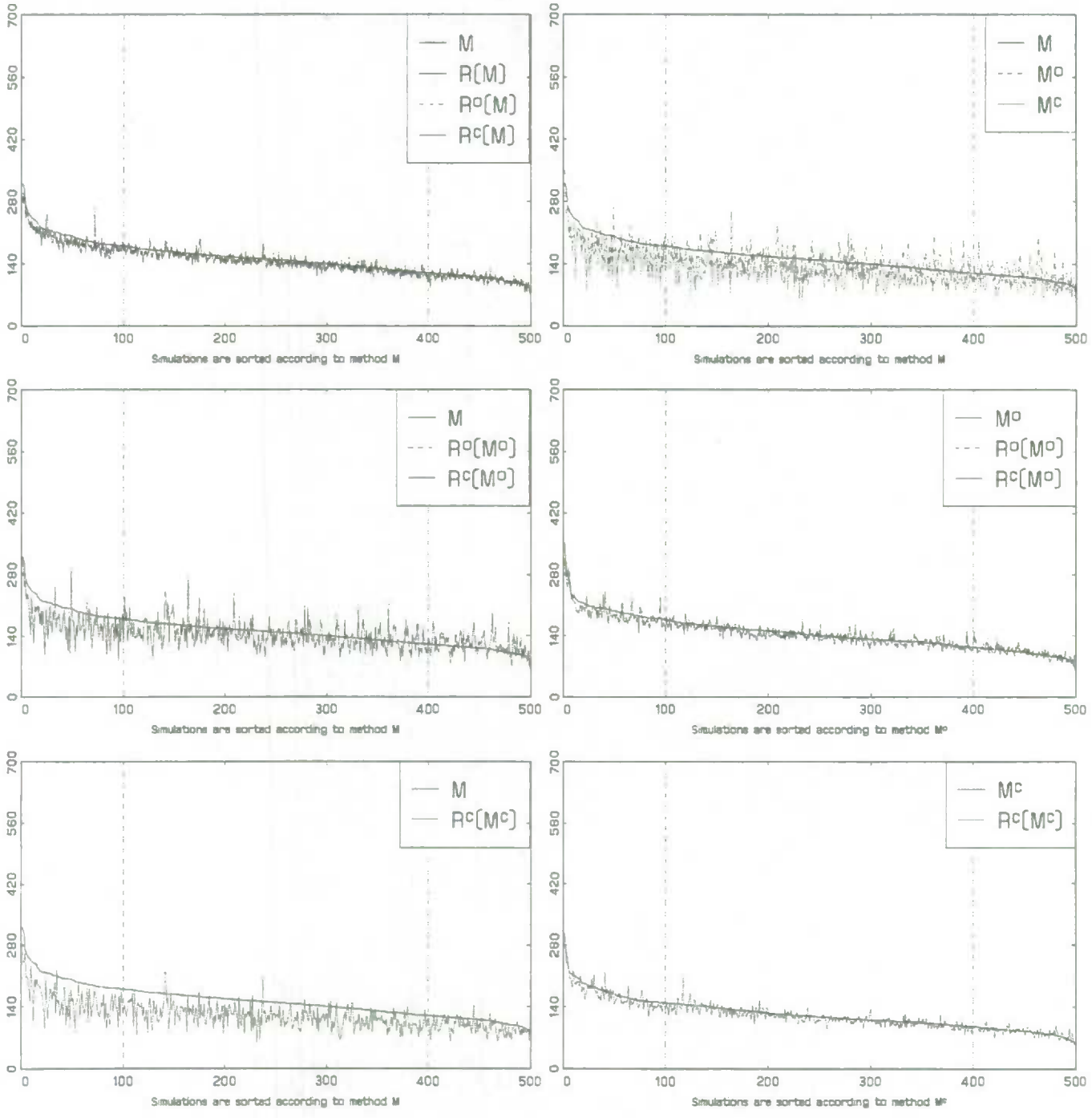
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.4-B2.6 and B2.10-B2.12.

Figure B2.5 Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 500 Simulations for RUO Datafile)



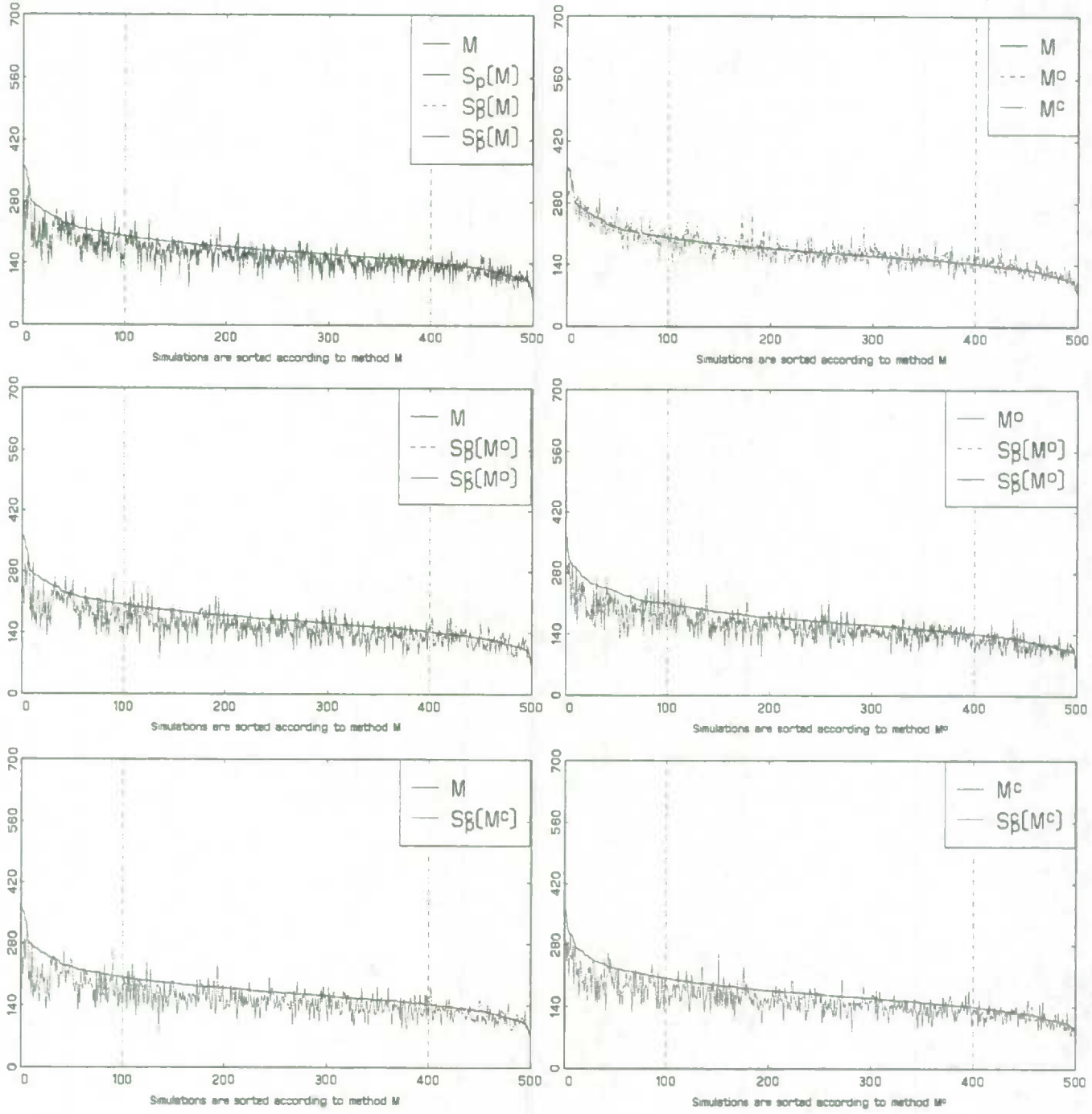
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B2.4-B2.6 and B2.10-B2.12.

Figure B2.6 Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 2 Z Categories: 500 Simulations for RUO Datafile)



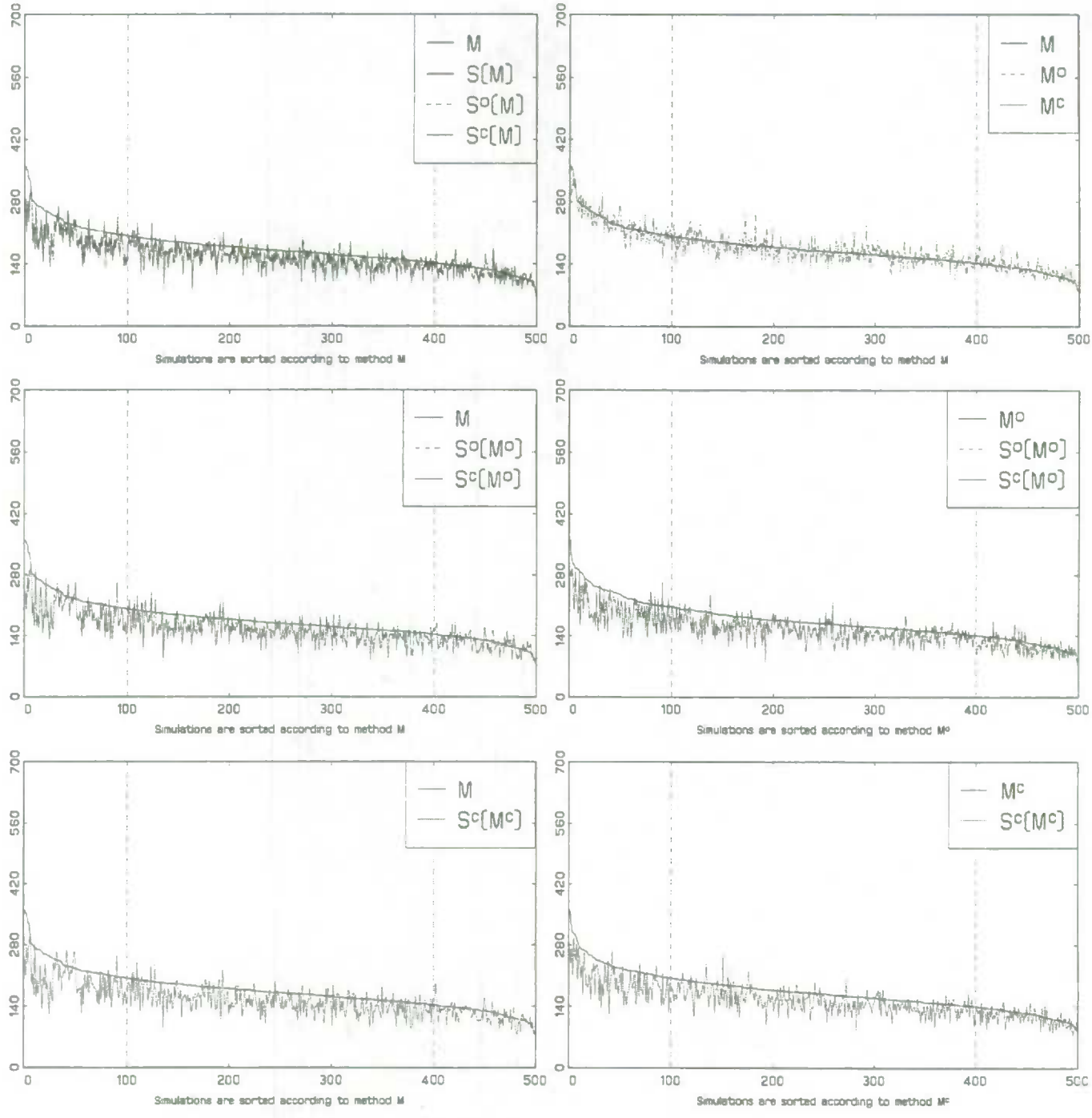
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.4-B2.6 and B2.10-B2.12.

Figure B2.7. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 3 Z Categories: 500 Simulations for RUO Datafile)



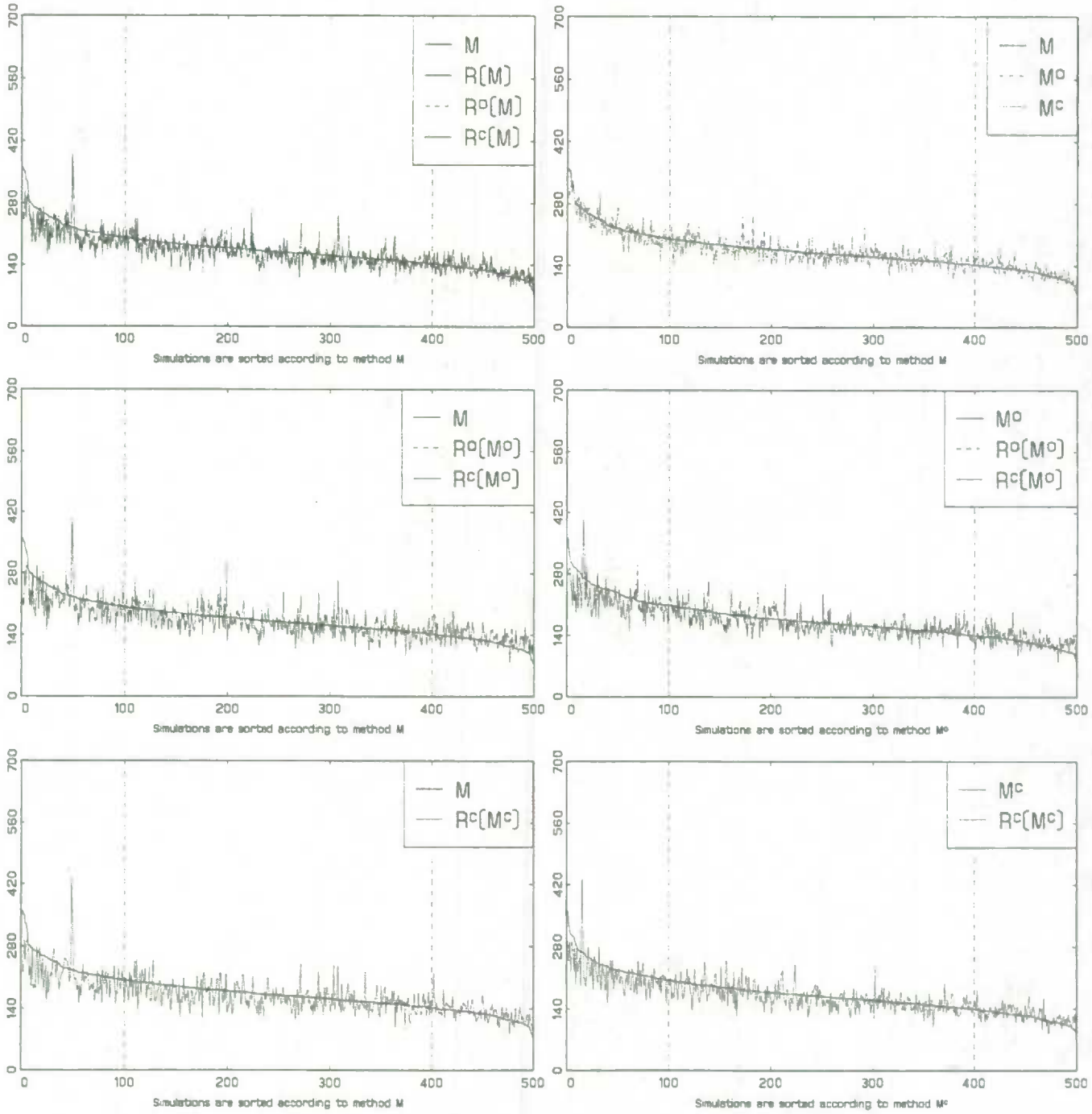
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.1-B2.3 and B2.7-B2.9.

Figure B2.8. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 3 Z Categories: 500 Simulations for RUO Datafile)



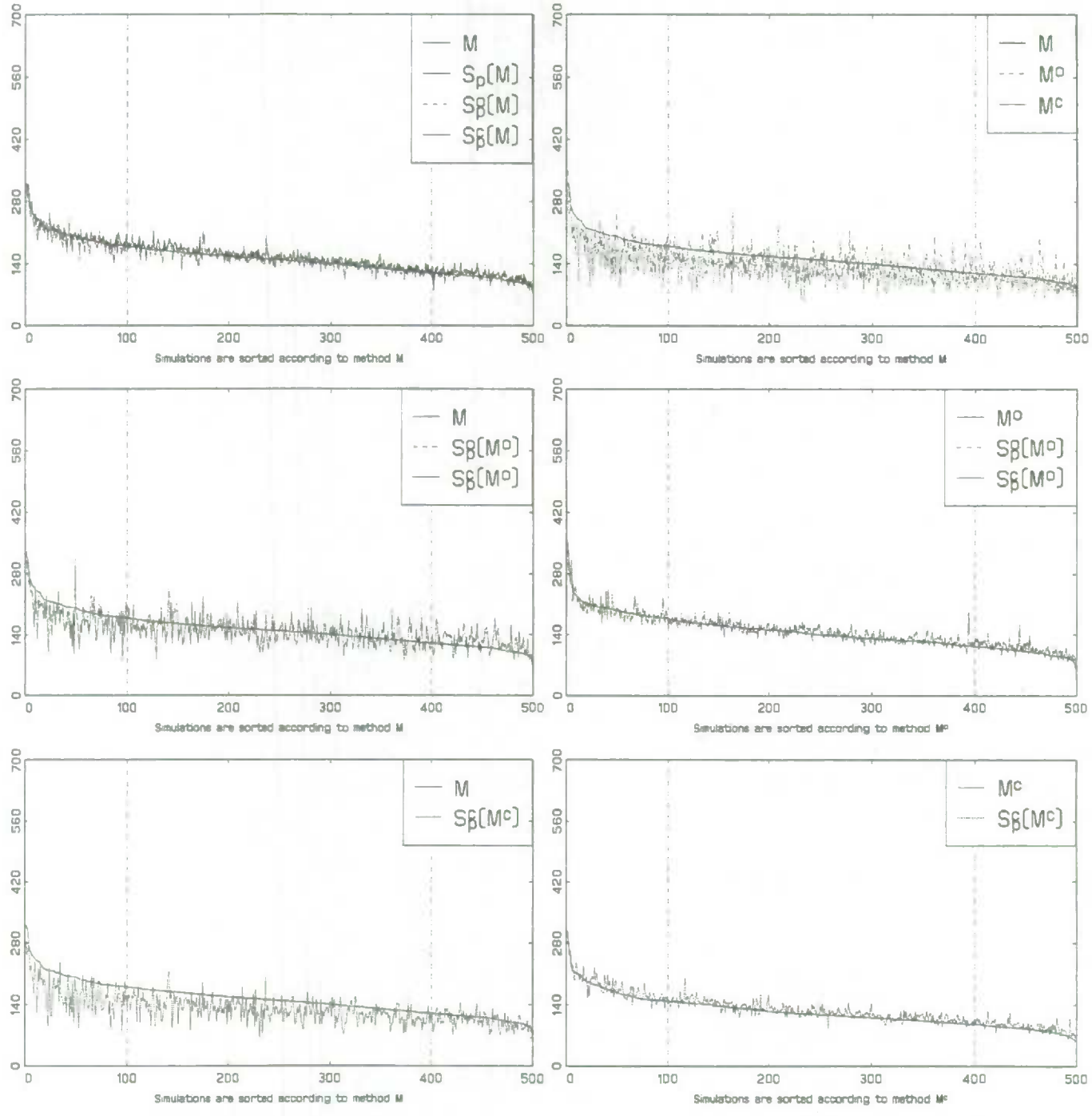
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.1-B2.3 and B2.7-B2.9.

Figure B2.9. Weighted χ^2 Evaluated over 10x4x2 Categories of the Distance Matched File
(Ratio Adjustment Based on 3 Z Categories: 500 Simulations for RUO Datafile)



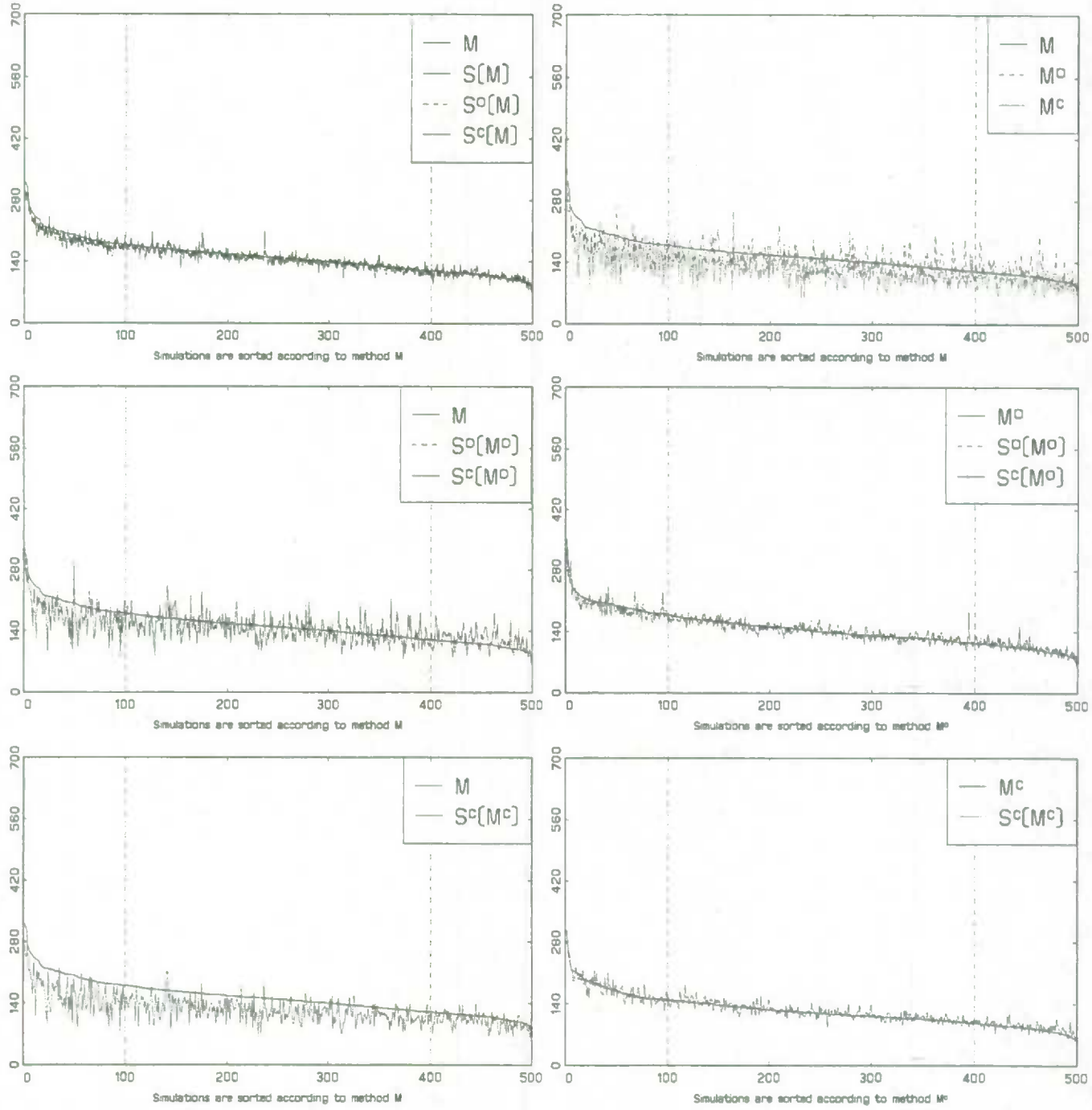
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.1-B2.3 and B2.7-B2.9.

Figure B2.10. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 3 Z Categories: 500 Simulations for RUO Datafile)



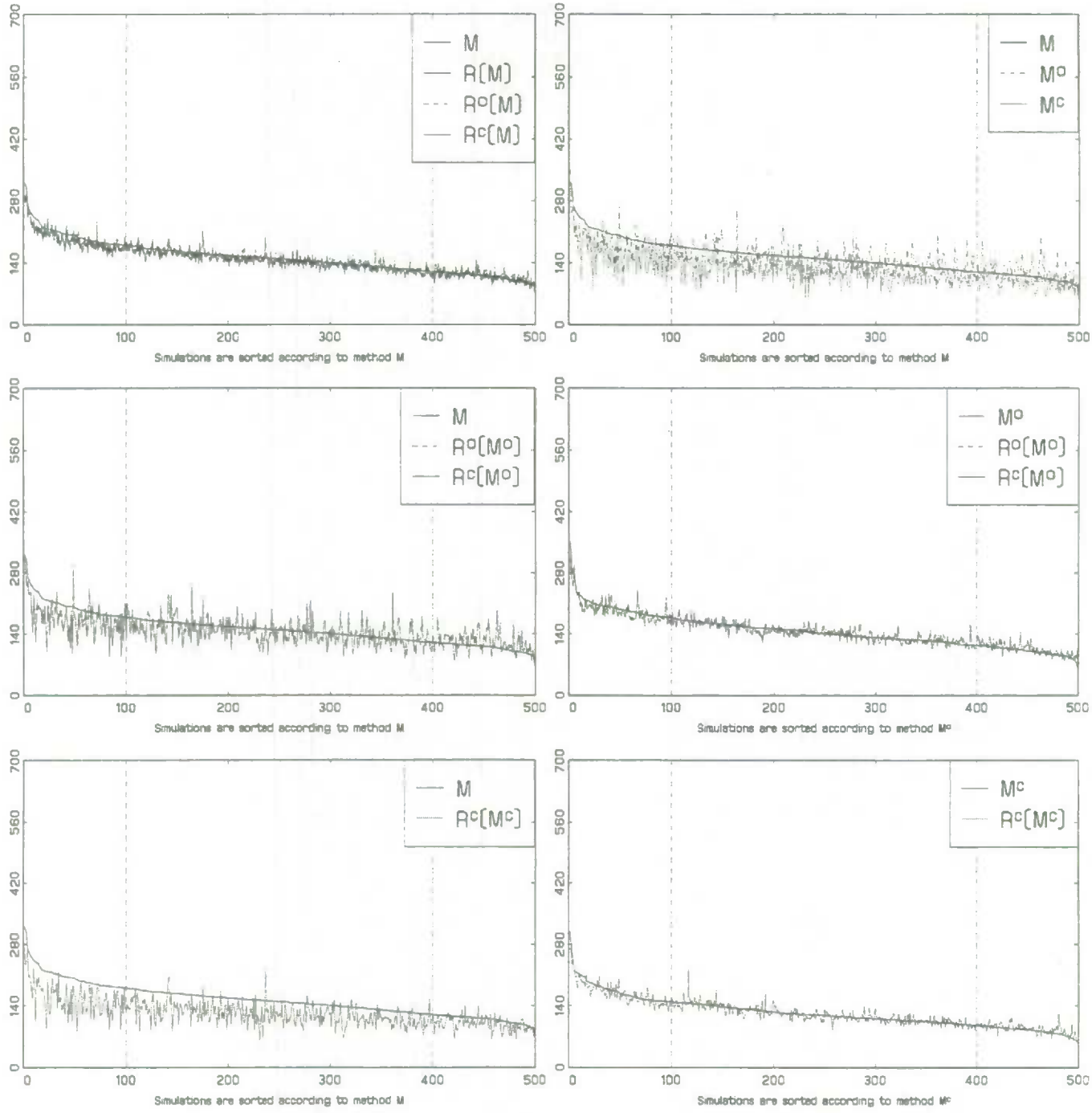
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.4-B2.6 and B2.10-B2.12.

Figure B2.11. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 3 Z Categories: 500 Simulations for RUO Datafile)



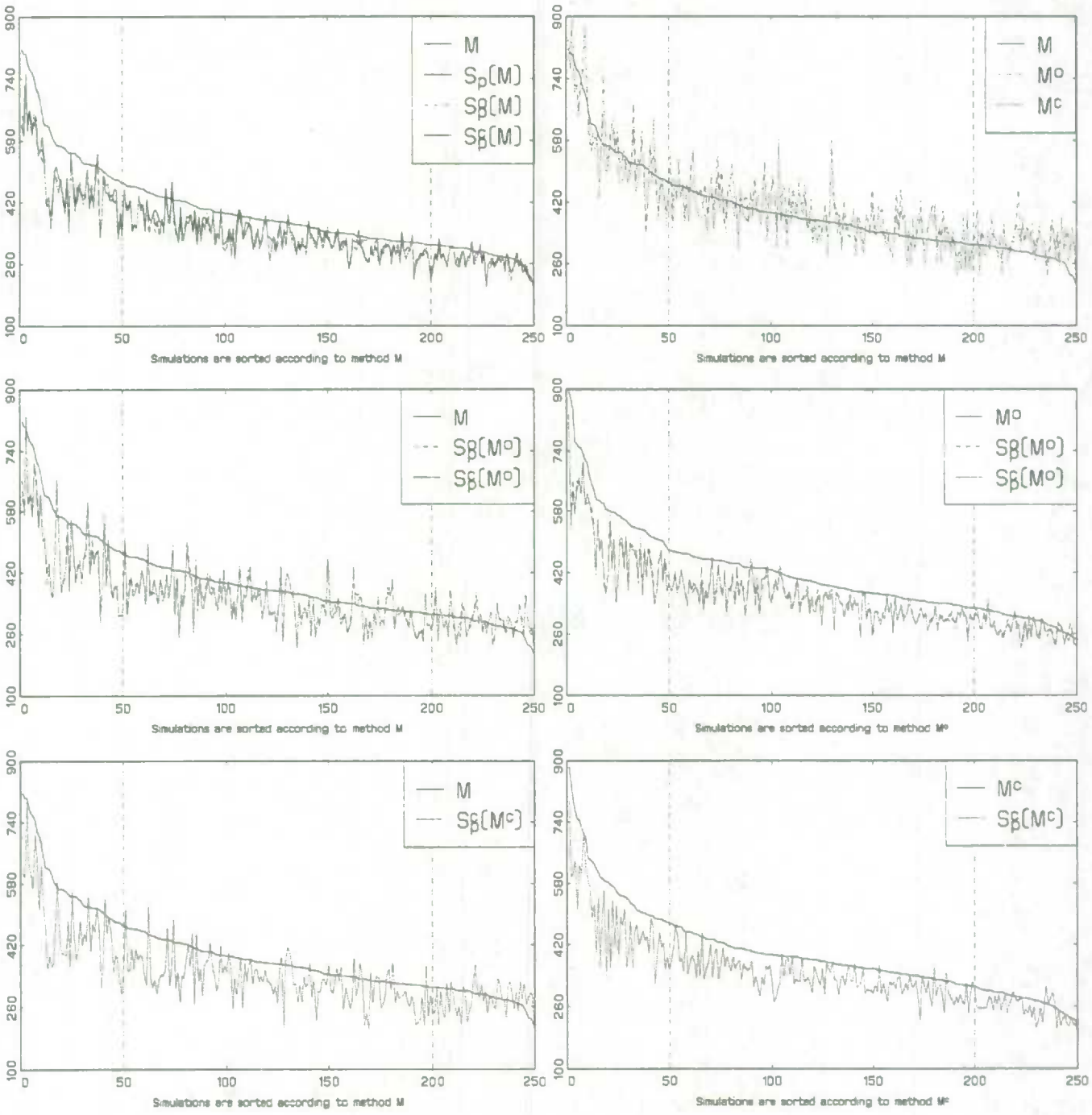
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B2.4-B2.6 and B2.10-B2.12.

Figure B2.12. Weighted χ^2 Evaluated over 10x4x2 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 3 Z Categories: 500 Simulations for RUO Datafile)



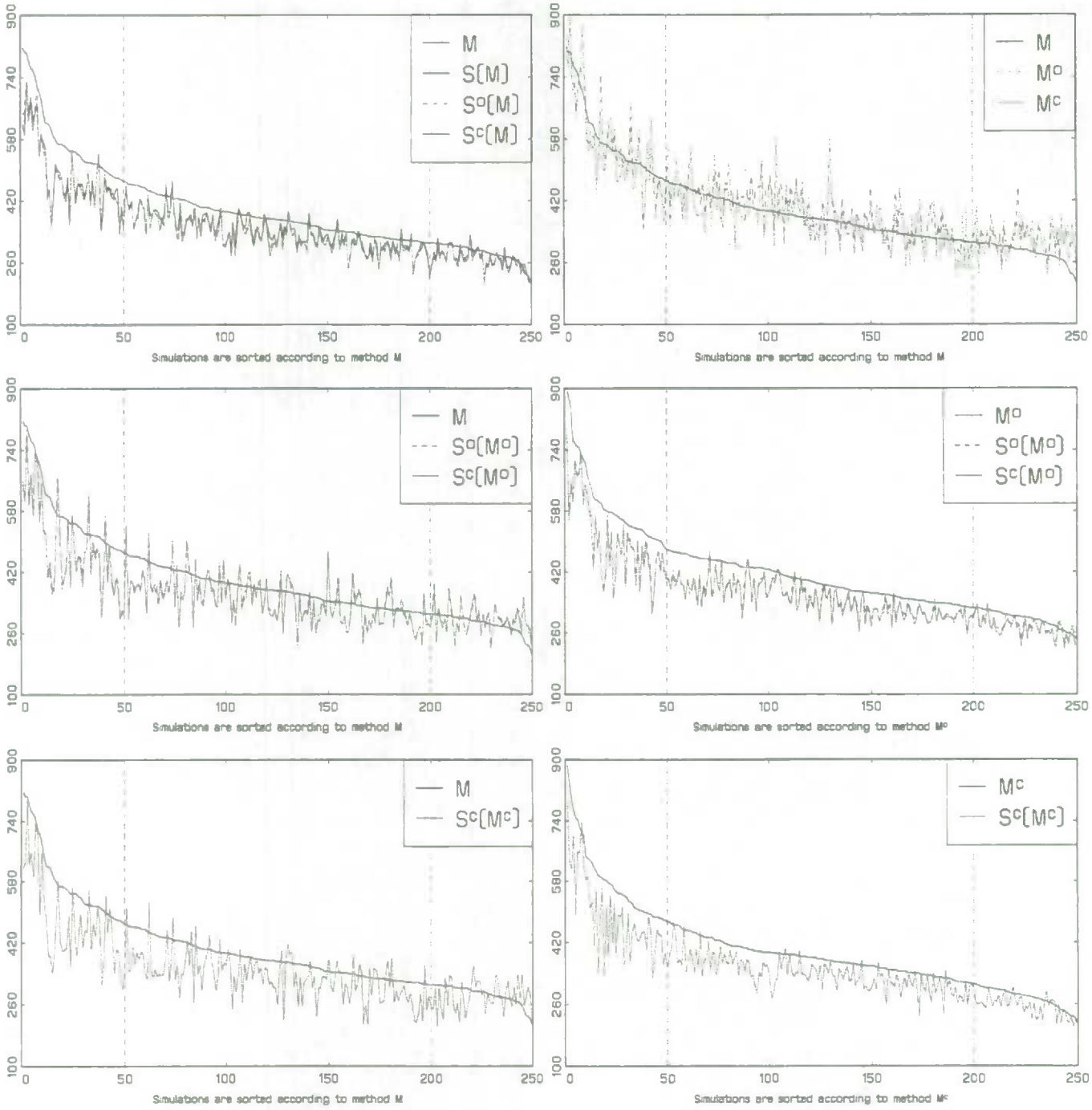
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B2.4-B2.6 and B2.10-B2.12.

Figure B3.1. Weighted χ^2 Evaluated over 10x4x3Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 250 Simulations for MQM Datafile)



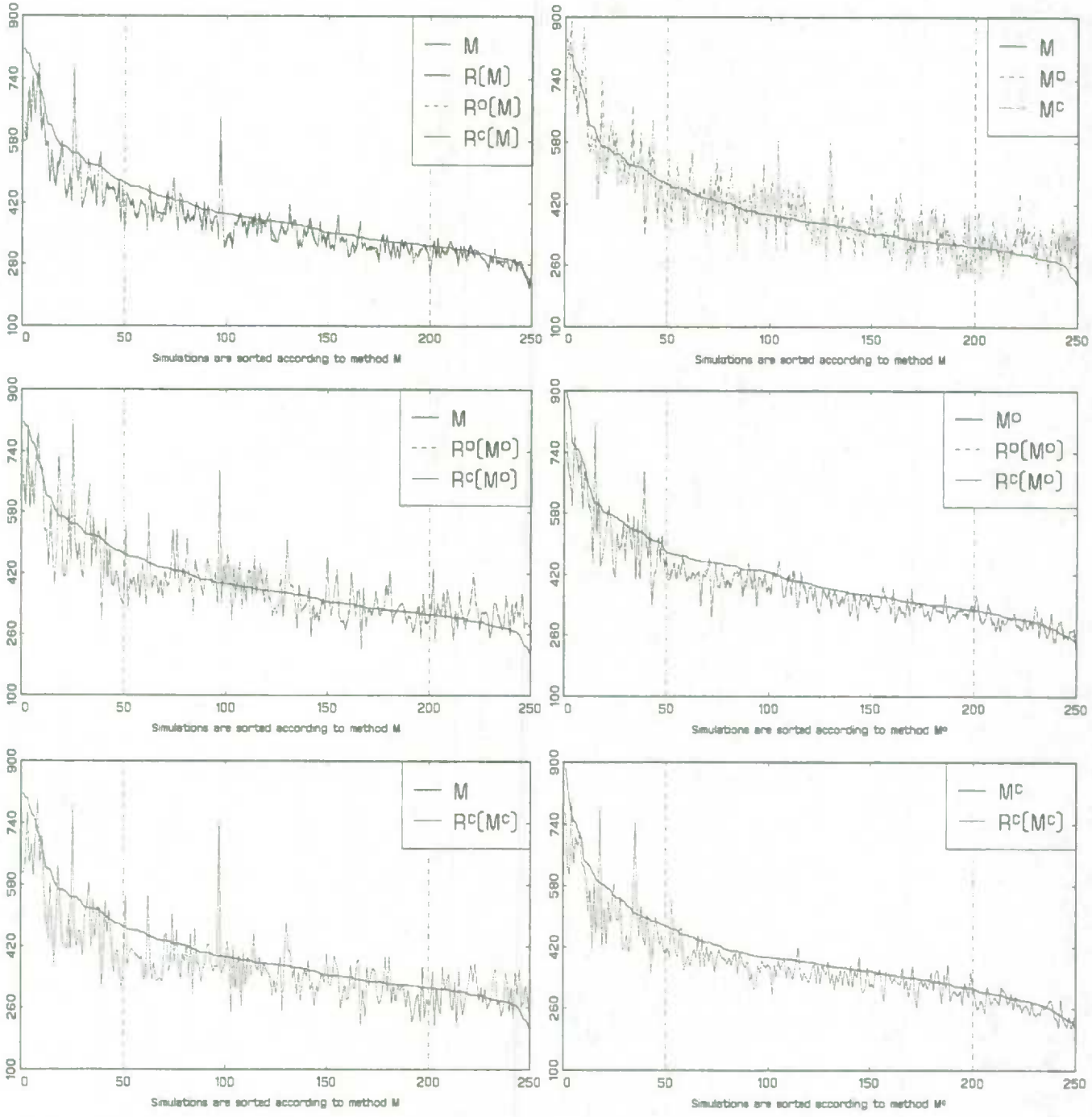
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B3.1-B3.3 and B3.7-B3.9.

Figure B3.2. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 250 Simulations for MQM Datafile)



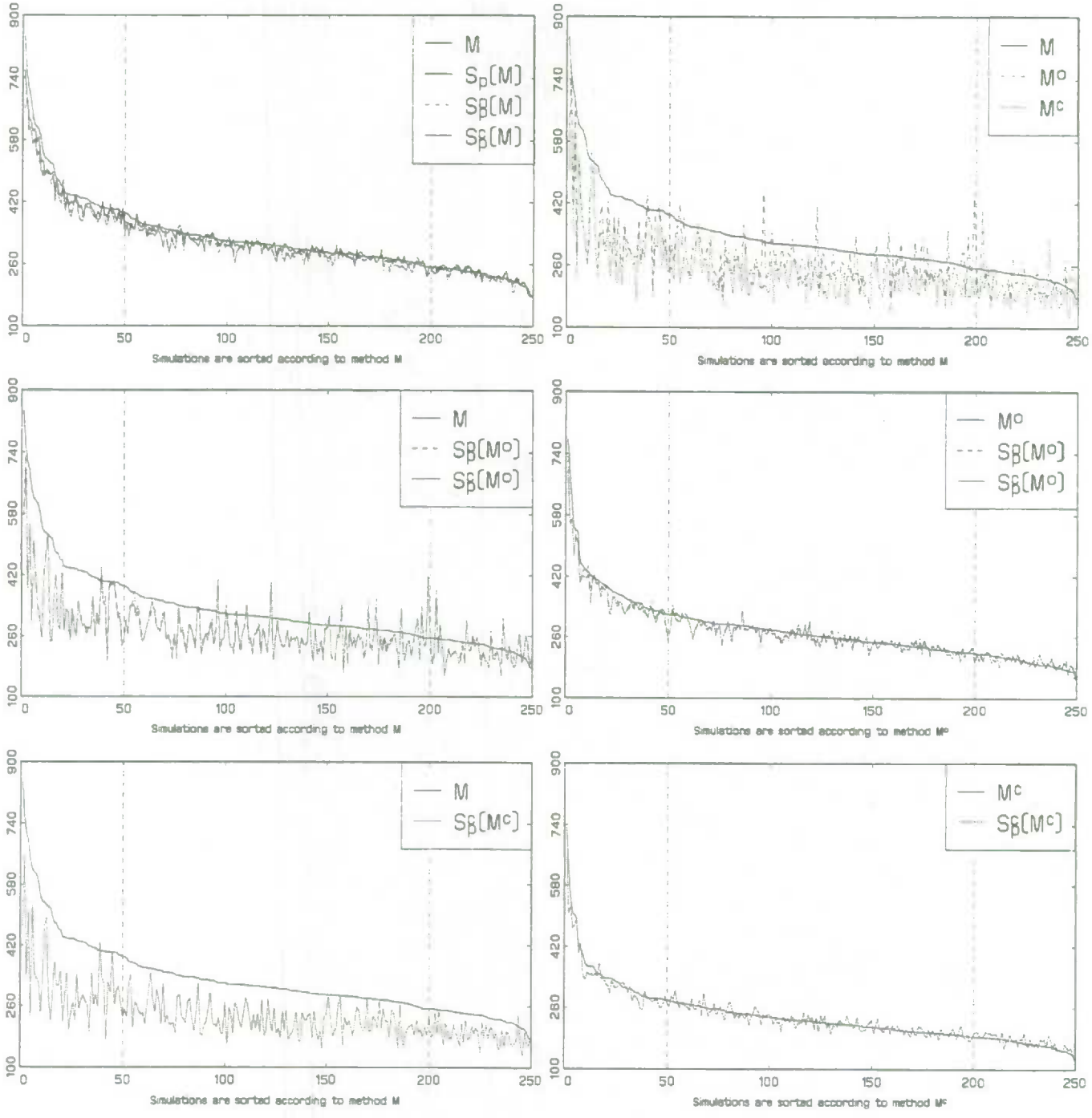
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.1-B3.3 and B3.7-B3.9.

Figure B3.3. Weighted χ^2 Evaluated over 10x4x3Categories of the Distance Matched File
(Ratio Adjustment Based on 2 Z Categories: 250 Simulations for MQM Datafile)



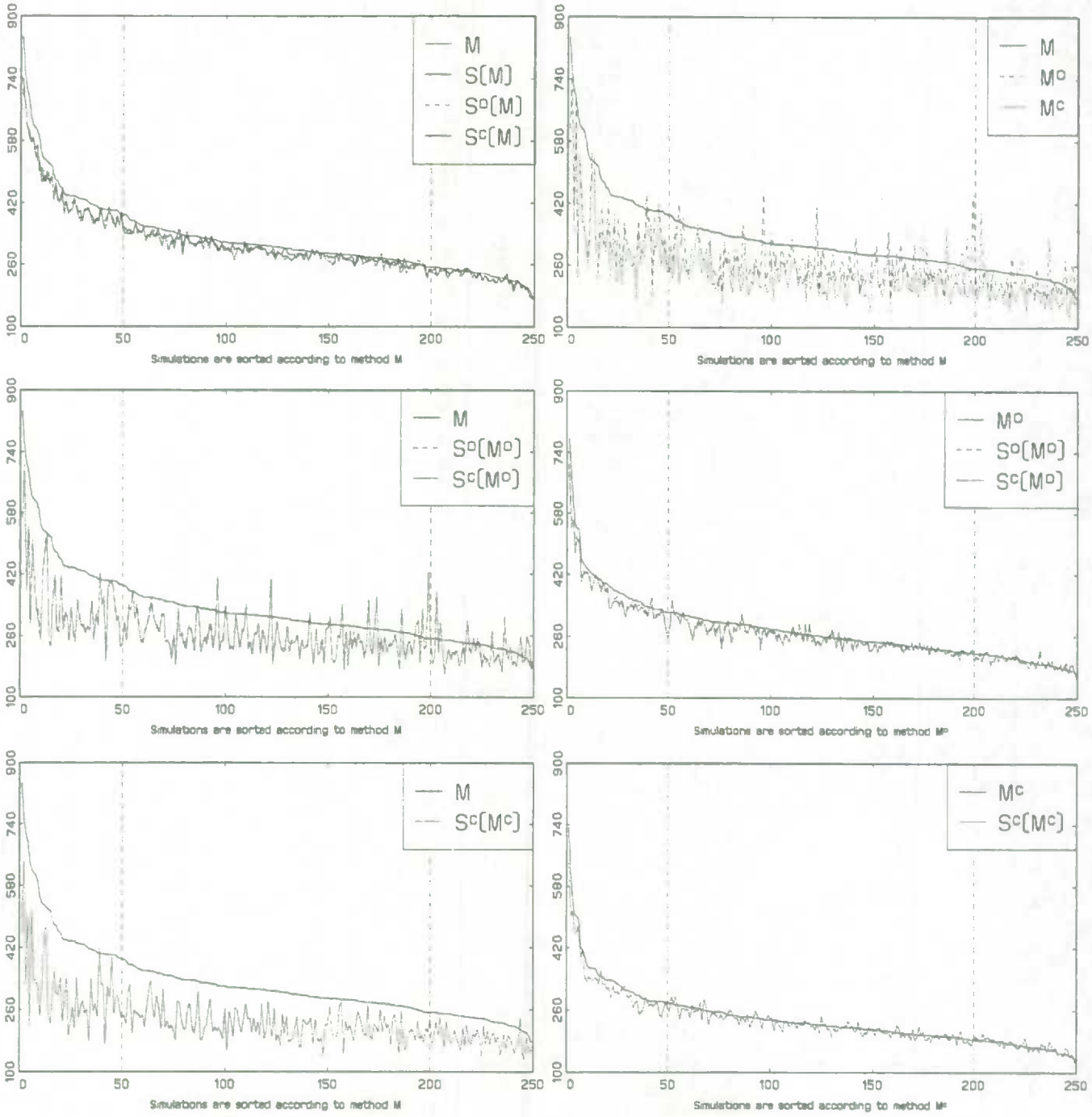
Remark: The rank-plot of the same matched files M, M^D and M^C is repeated in Figures B31-B33 and B37-B39.

Figure B3.4 Weighted χ^2 Evaluated over 10x4x3Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 250 Simulations for MQM Datafile)



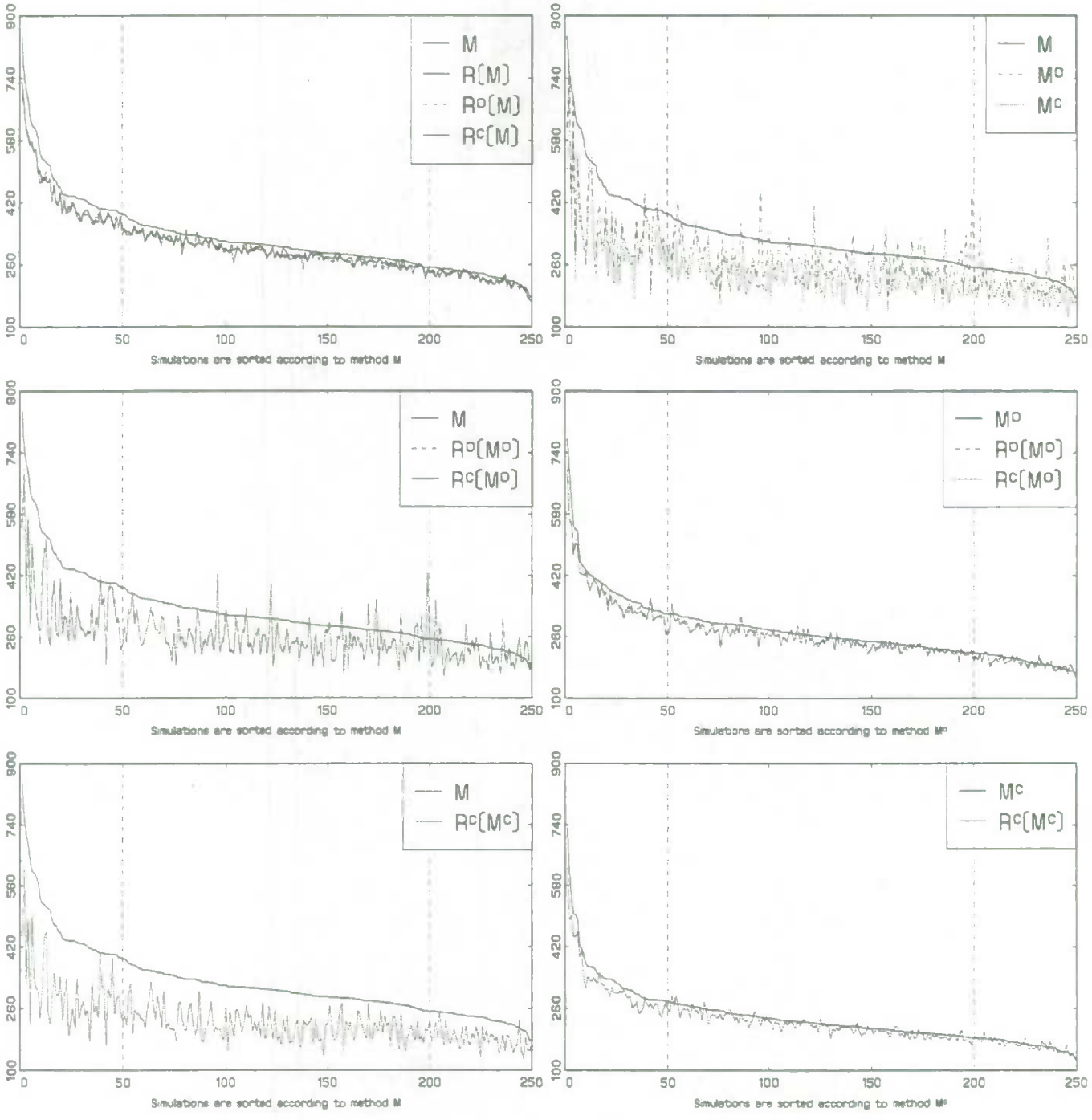
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.4-B3.6 and B3.10-B3.12.

Figure B3.5. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 250 Simulations for MQM Datafile)



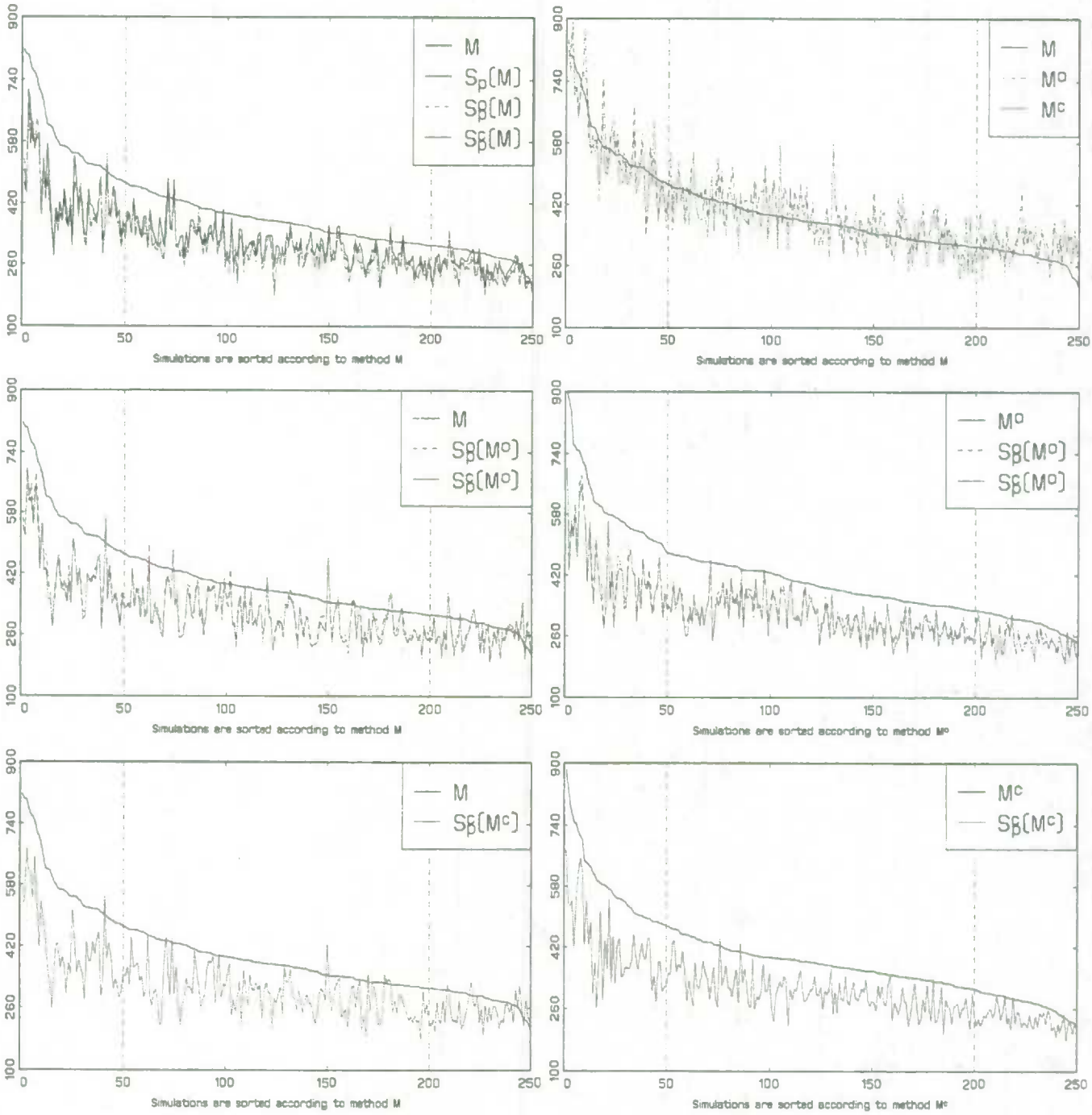
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.4-B3.6 and B3.10-B3.12.

Figure B3.6. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 2 Z Categories: 250 Simulations for MQM Datafile)



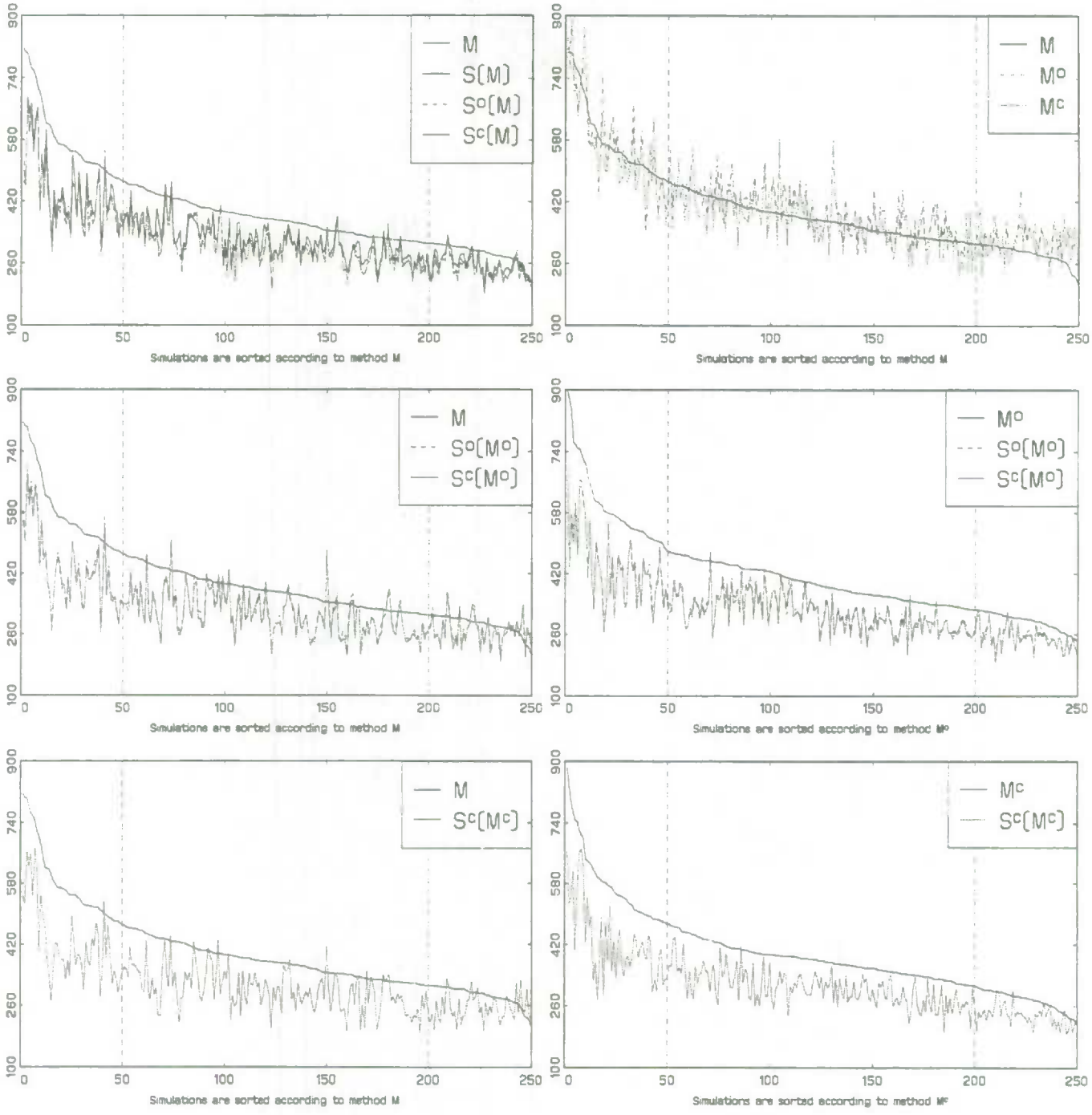
Remark: The rank-plot of the same matched files M, M^D and M^C is repeated in Figures B3.4-B3.6 and B3.10-B3.12.

Figure B3.7. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 3 Z Categories: 250 Simulations for MQM Datafile)



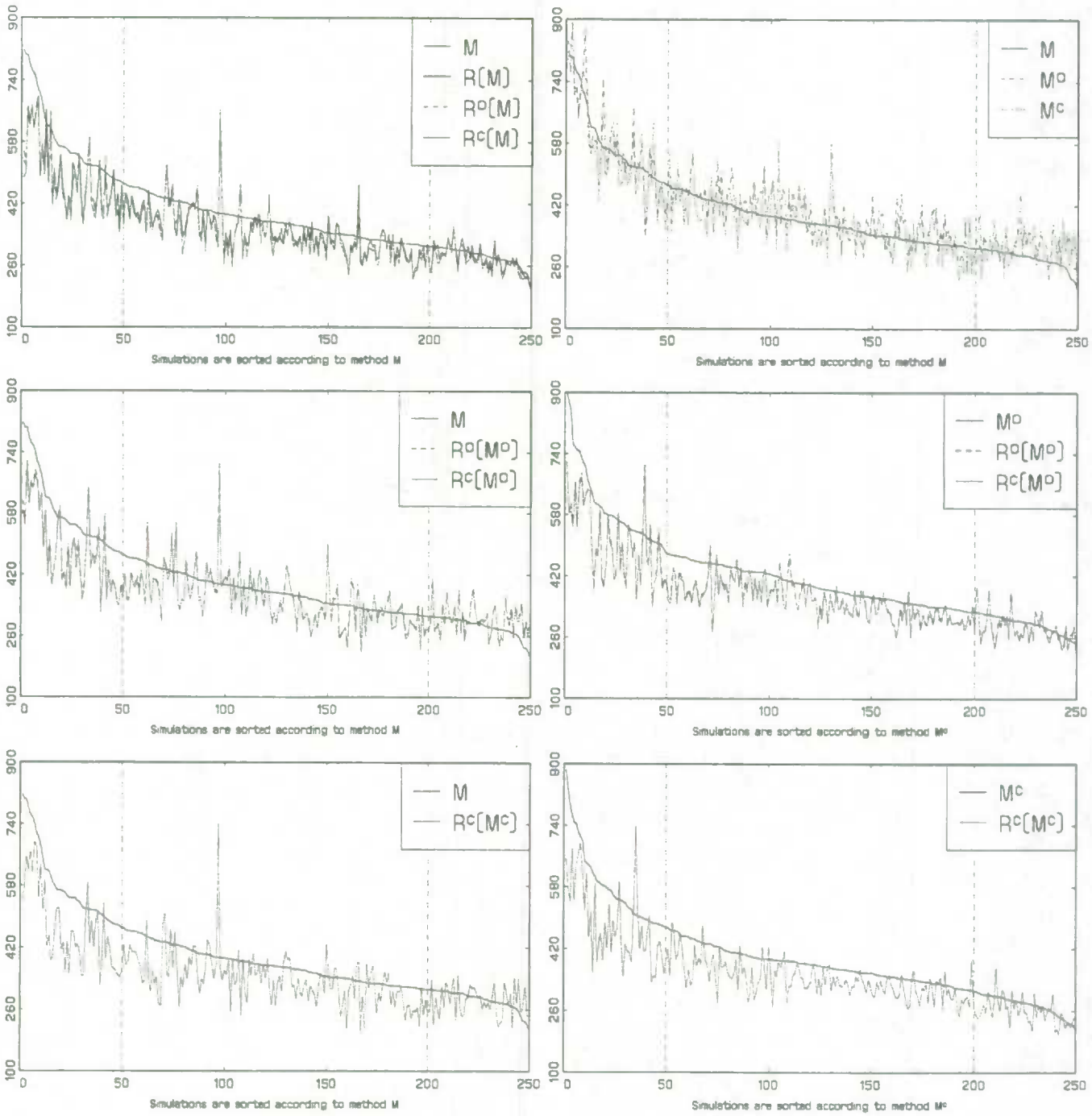
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.1-B3.3 and B3.7-B3.9.

Figure B3.8. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 3 Z Categories: 250 Simulations for MQM Datafile)



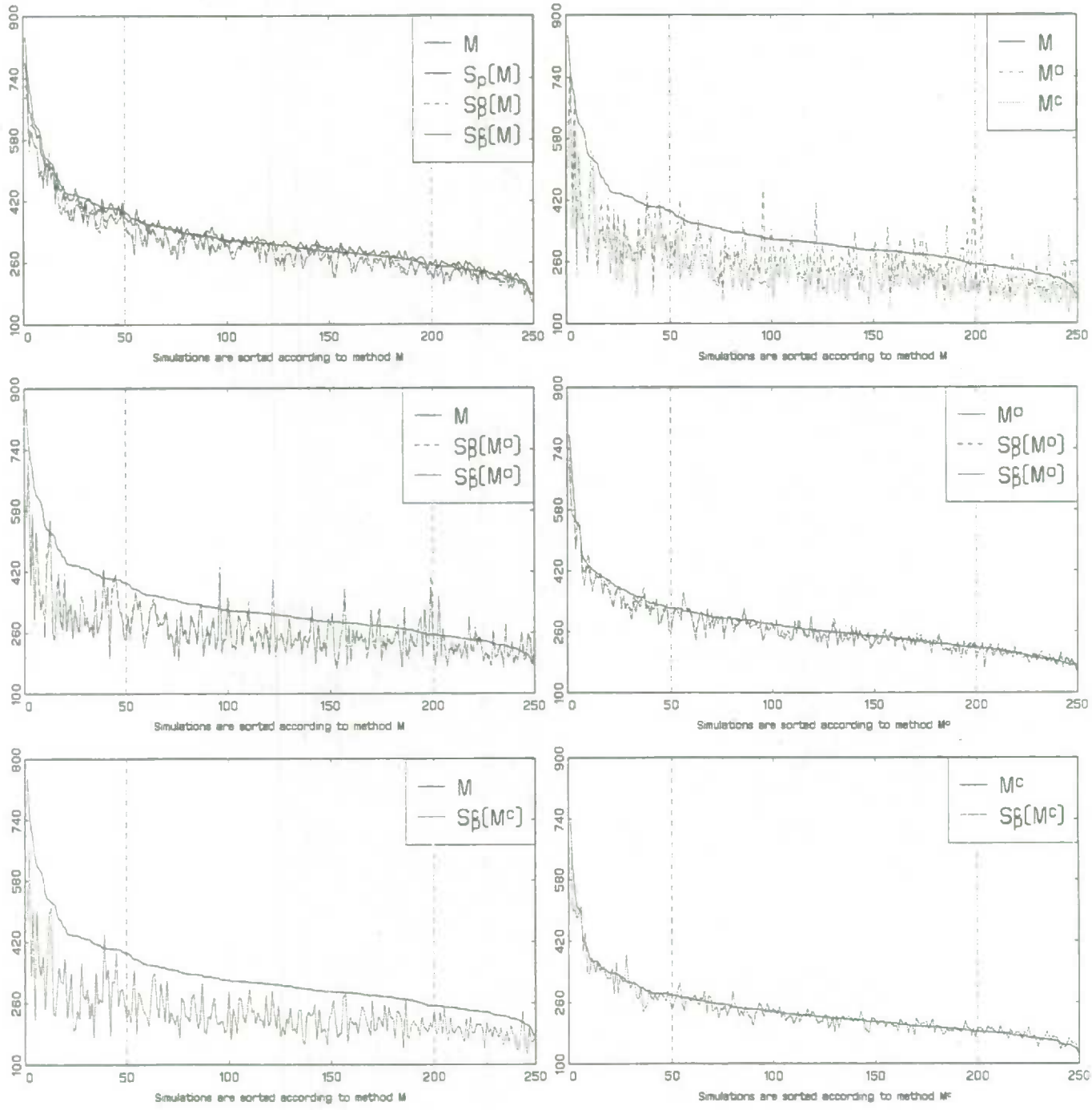
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.1-B3.3 and B3.7-B3.9.

Figure B3.9. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Ratio Adjustment Based on 3 Z Categories: 250 Simulations for MQM Datafile)



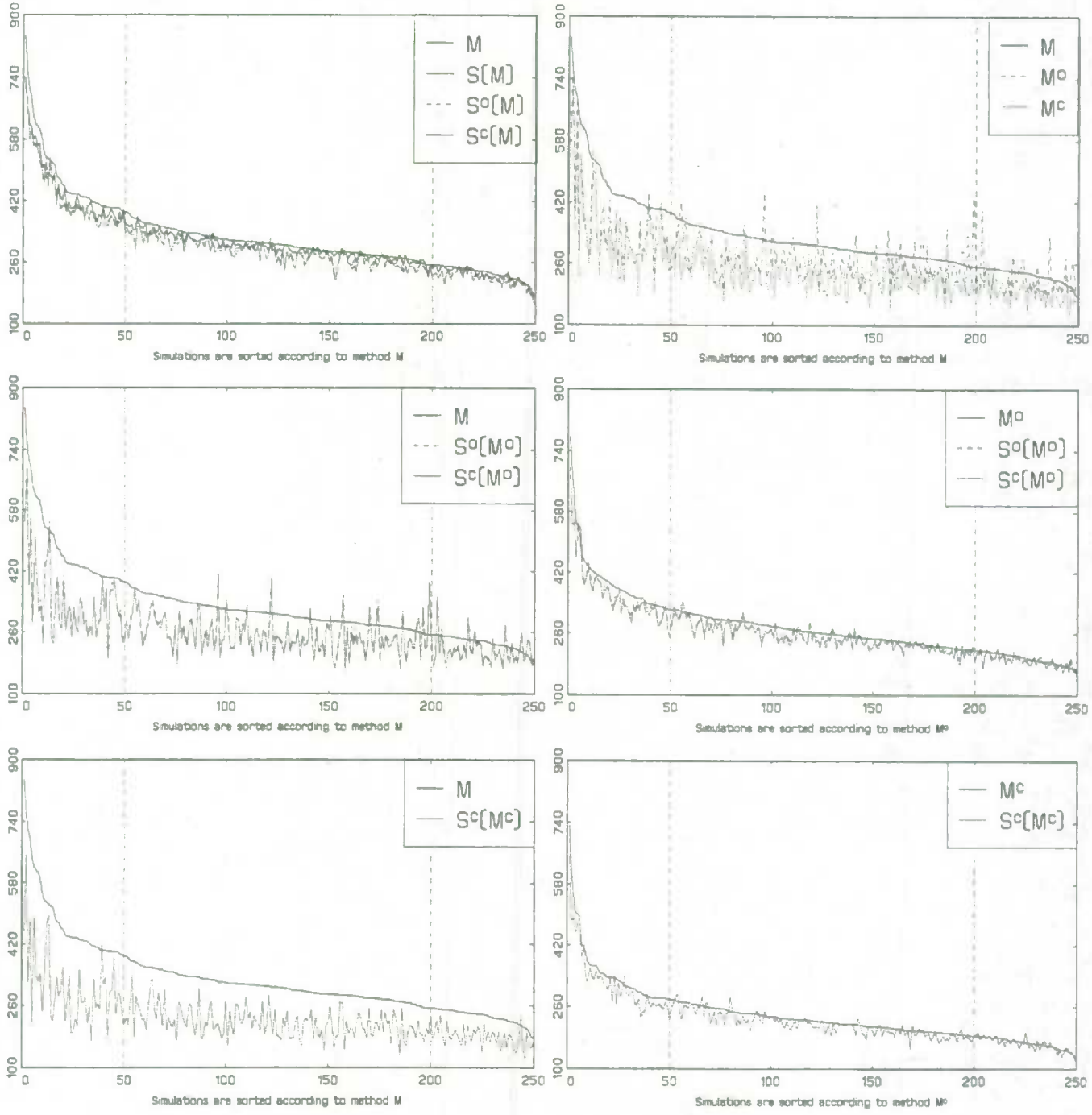
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.1-B3.3 and B3.7-B3.9.

Figure B3.10 Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 3Z Categories: 250 Simulations for MQM Datafile)



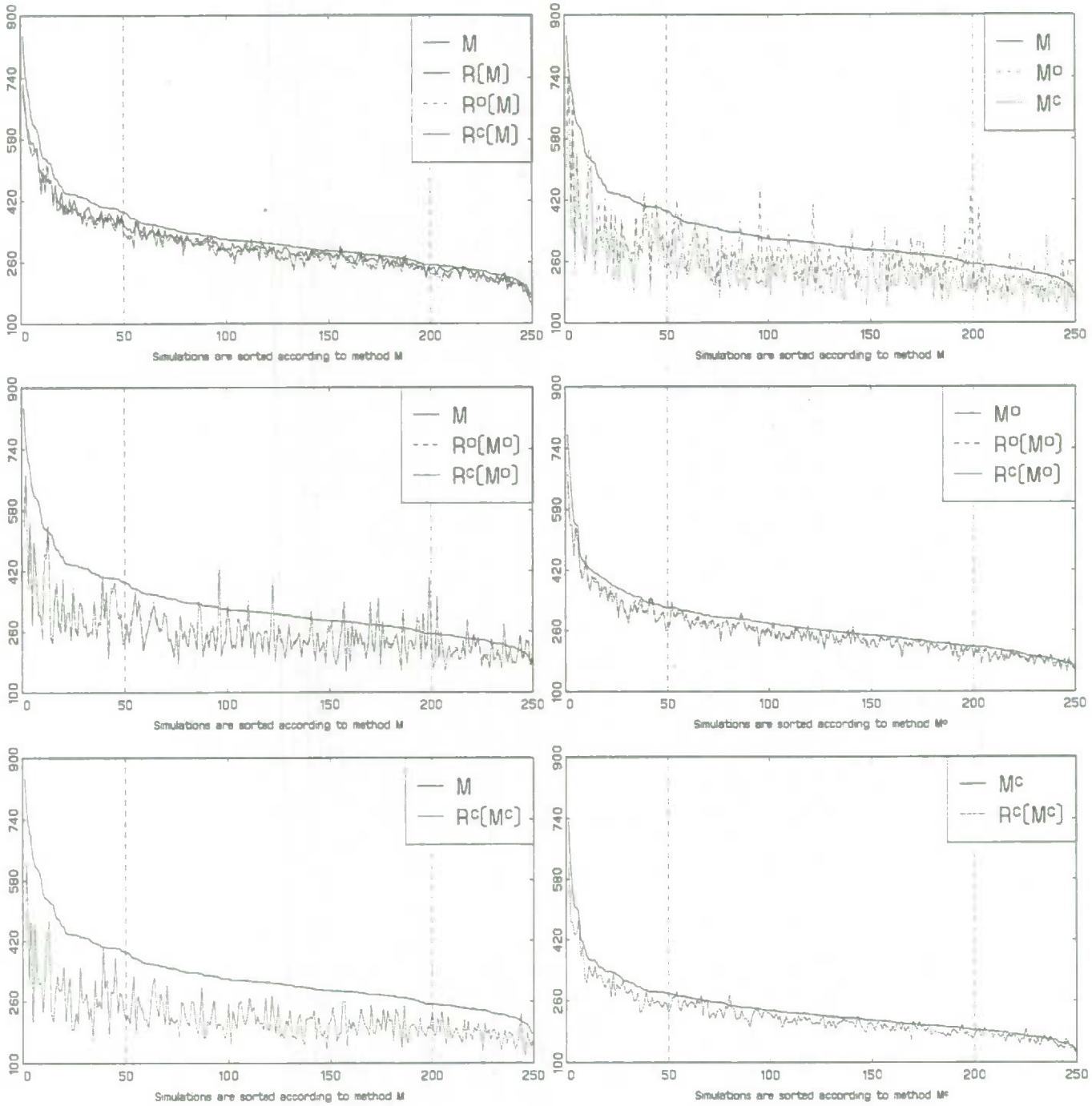
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B3.4-B3.6 and B3.10-B3.12.

Figure B3.11 Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 3Z Categories: 250 Simulations for MQM Datafile)



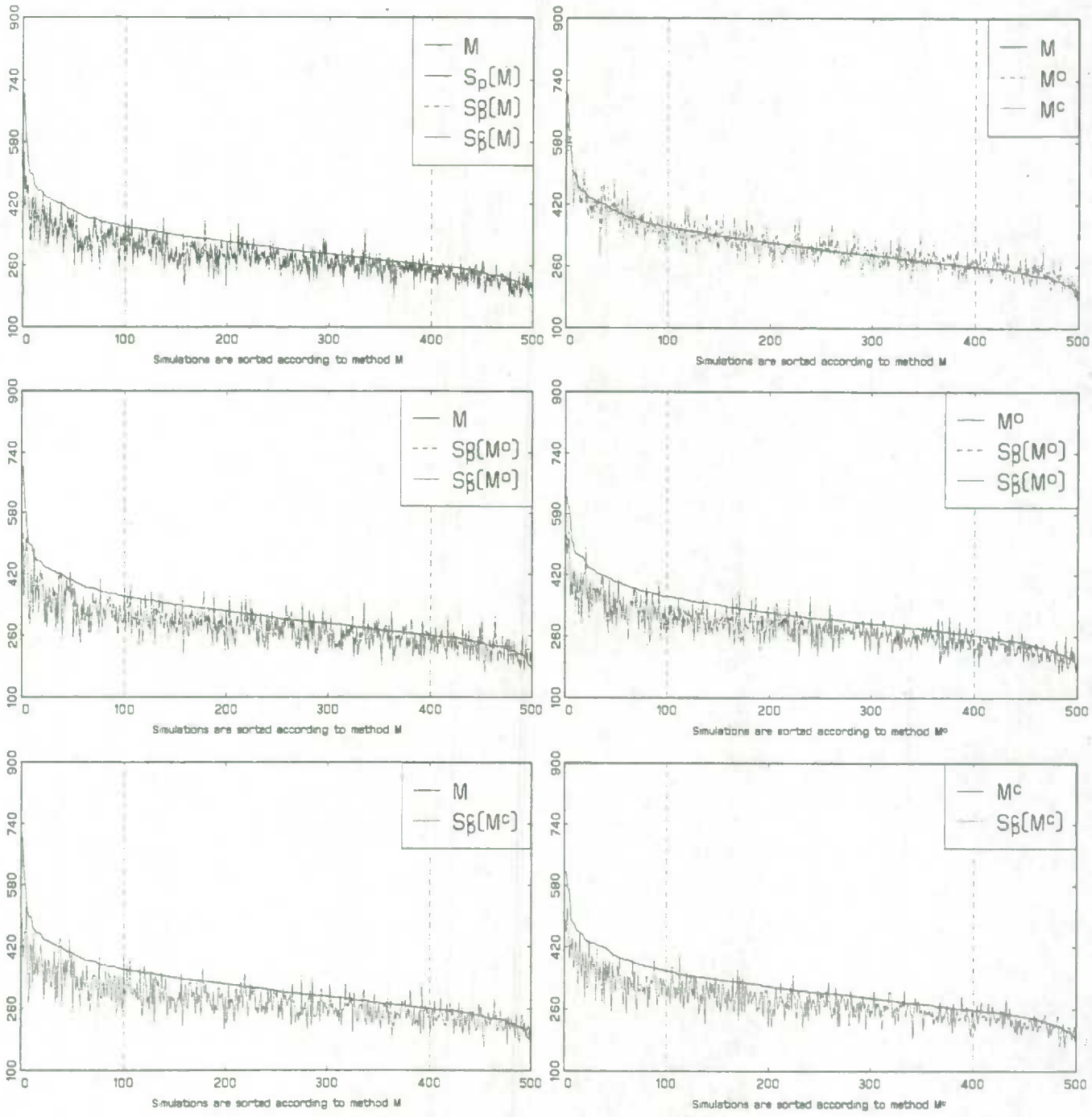
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.4-B3.6 and B3.10-B3.12.

Figure B3.12 Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 3Z Categories: 250 Simulations for MQM Datafile)



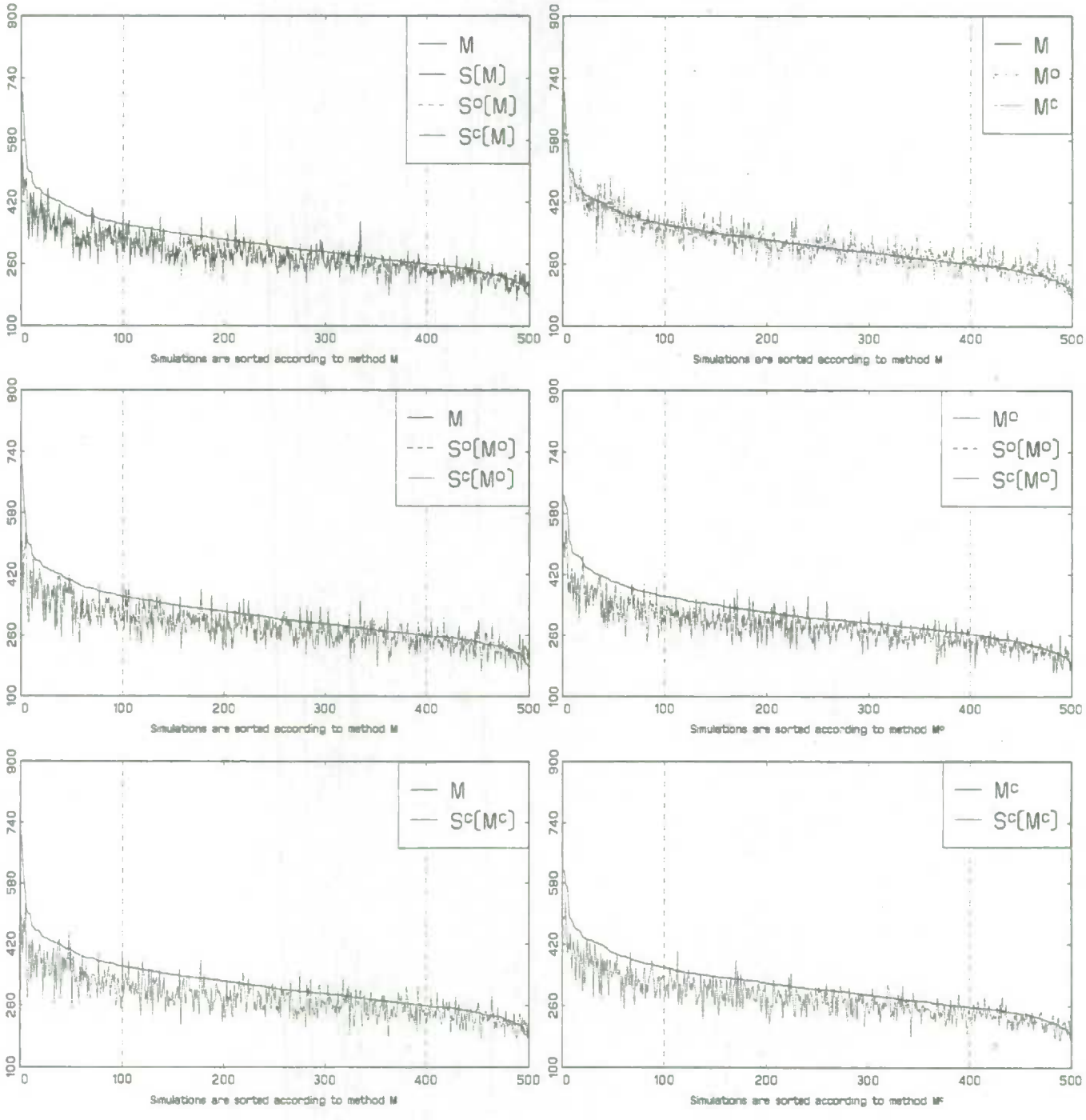
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B3.4-B3.6 and B3.10-B3.12.

Figure B4.1. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 500 Simulations for RUO Datafile)



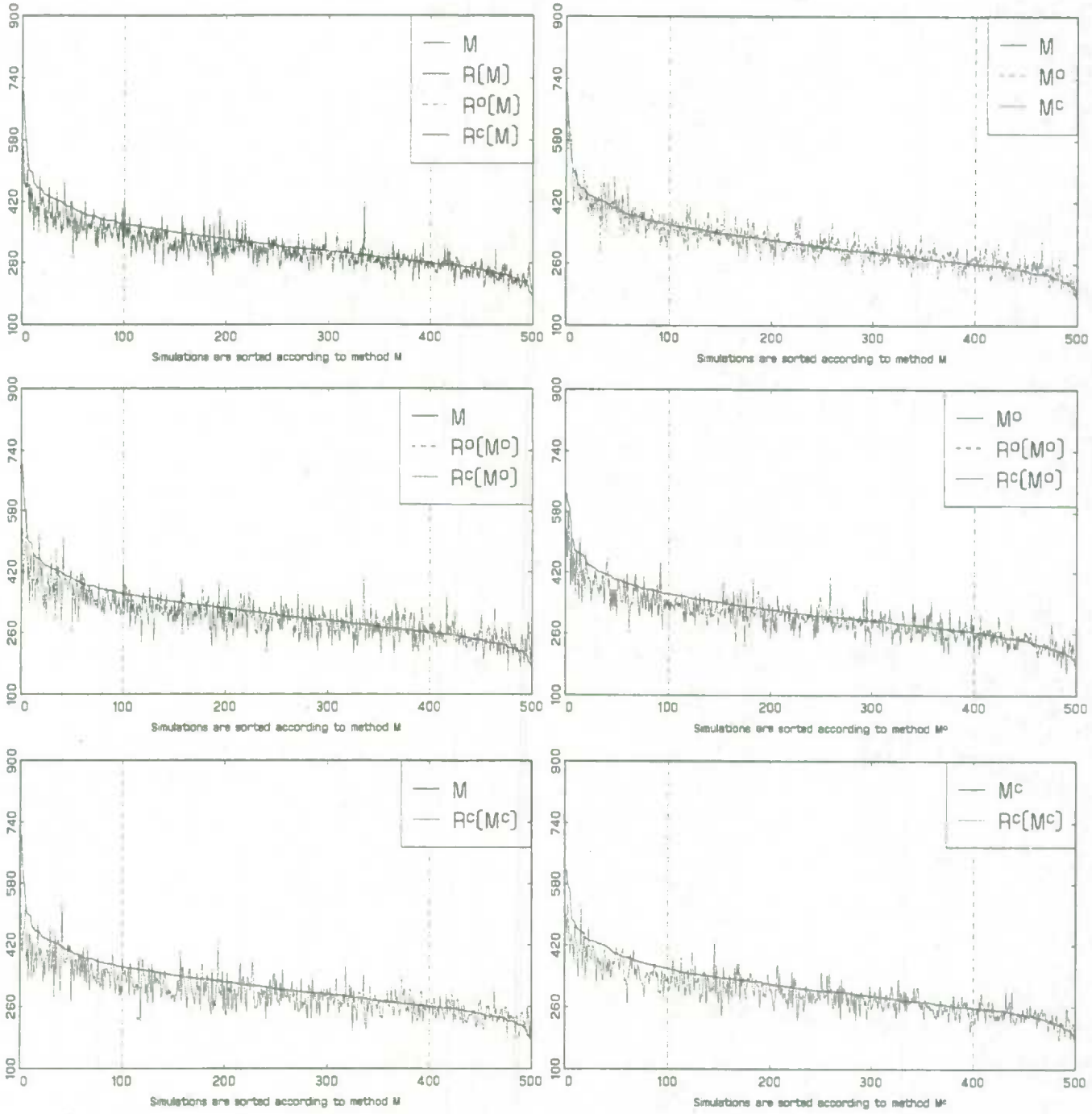
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.1-B4.3 and B4.7-B4.9.

Figure B4.2. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 500 Simulations for RUO Datafile)



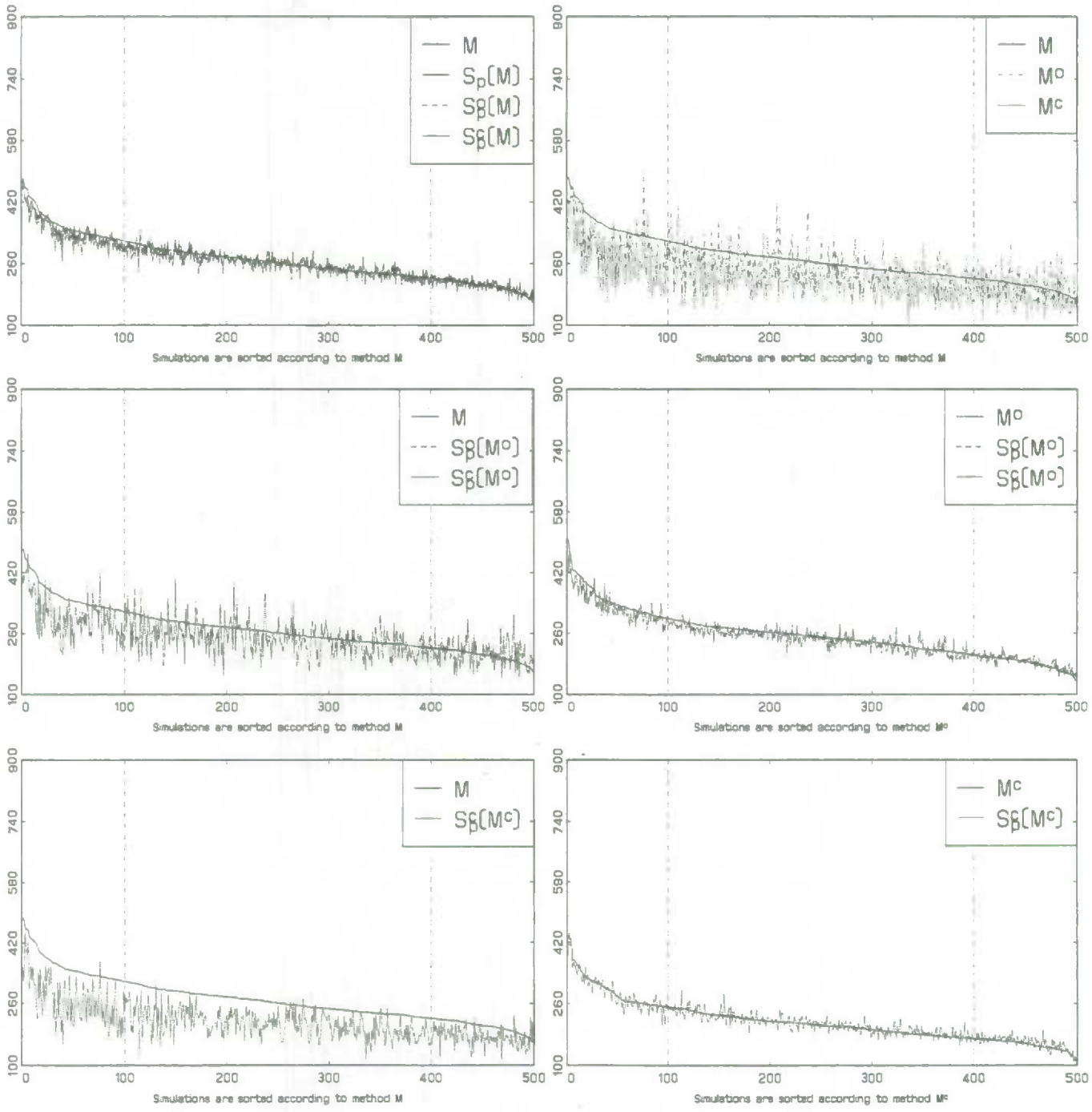
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B4.1-B4.3 and B4.7-B4.9.

Figure B4.3. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Ratio Adjustment Based on 2 Z Categories: 500 Simulations for RUO Datafile)



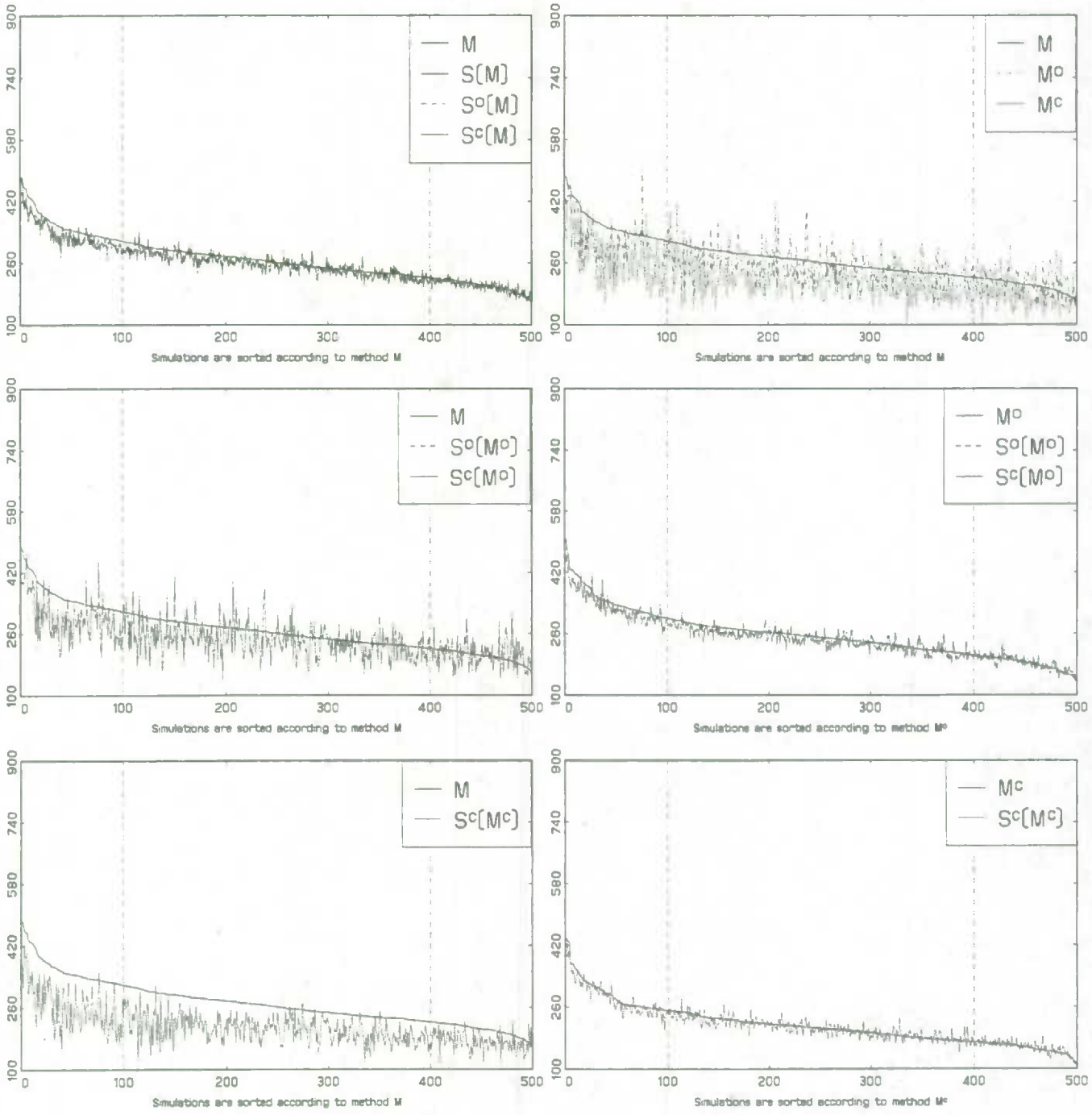
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.1-B4.3 and B4.7-B4.9.

Figure B4.4 Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 2 Z Categories: 500 Simulations for RUO Datafile)



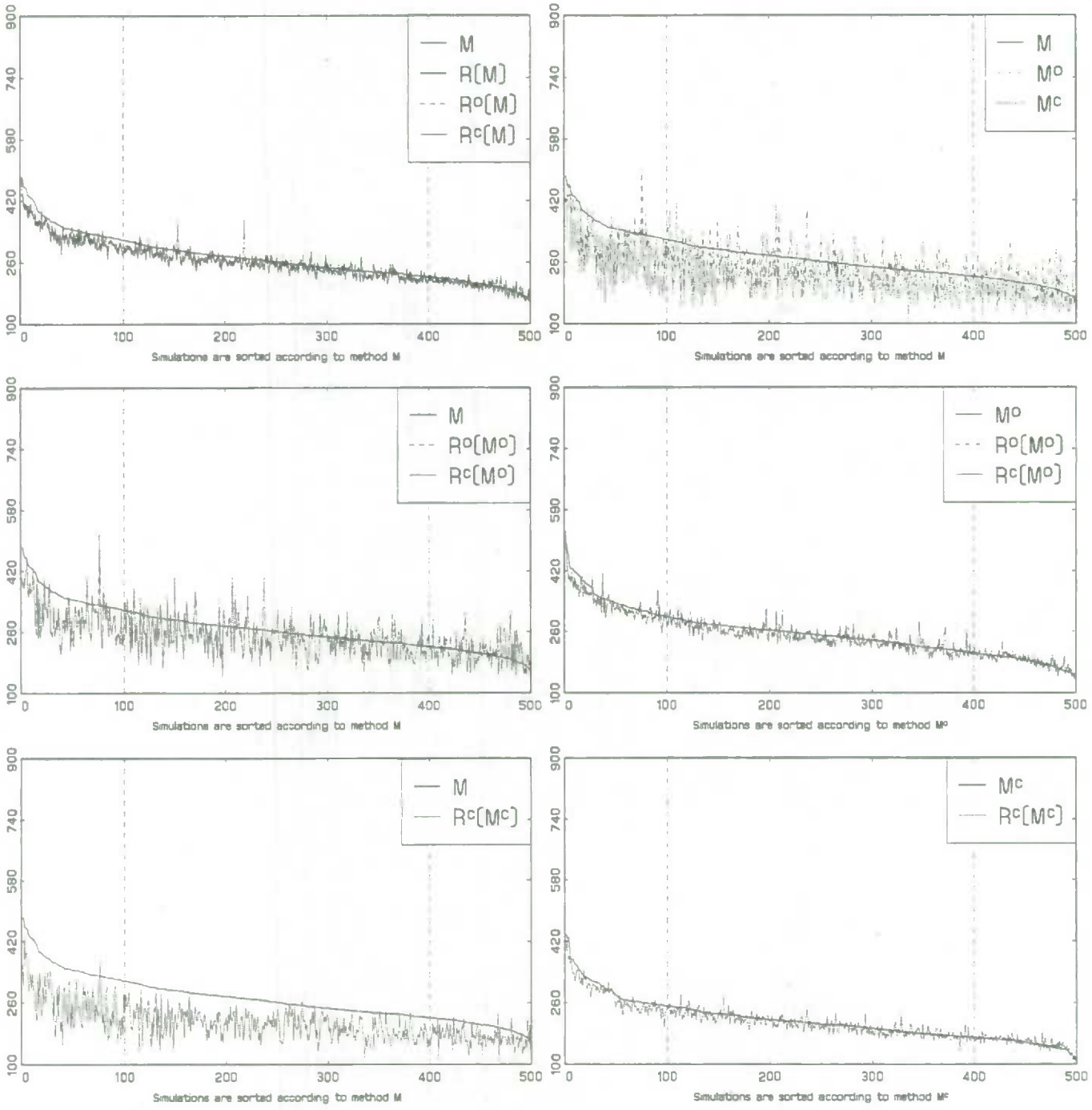
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.4-B4.6 and B4.10-B4.12.

Figure B4.5. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 2 Z Categories: 500 Simulations for RUO Datafile)



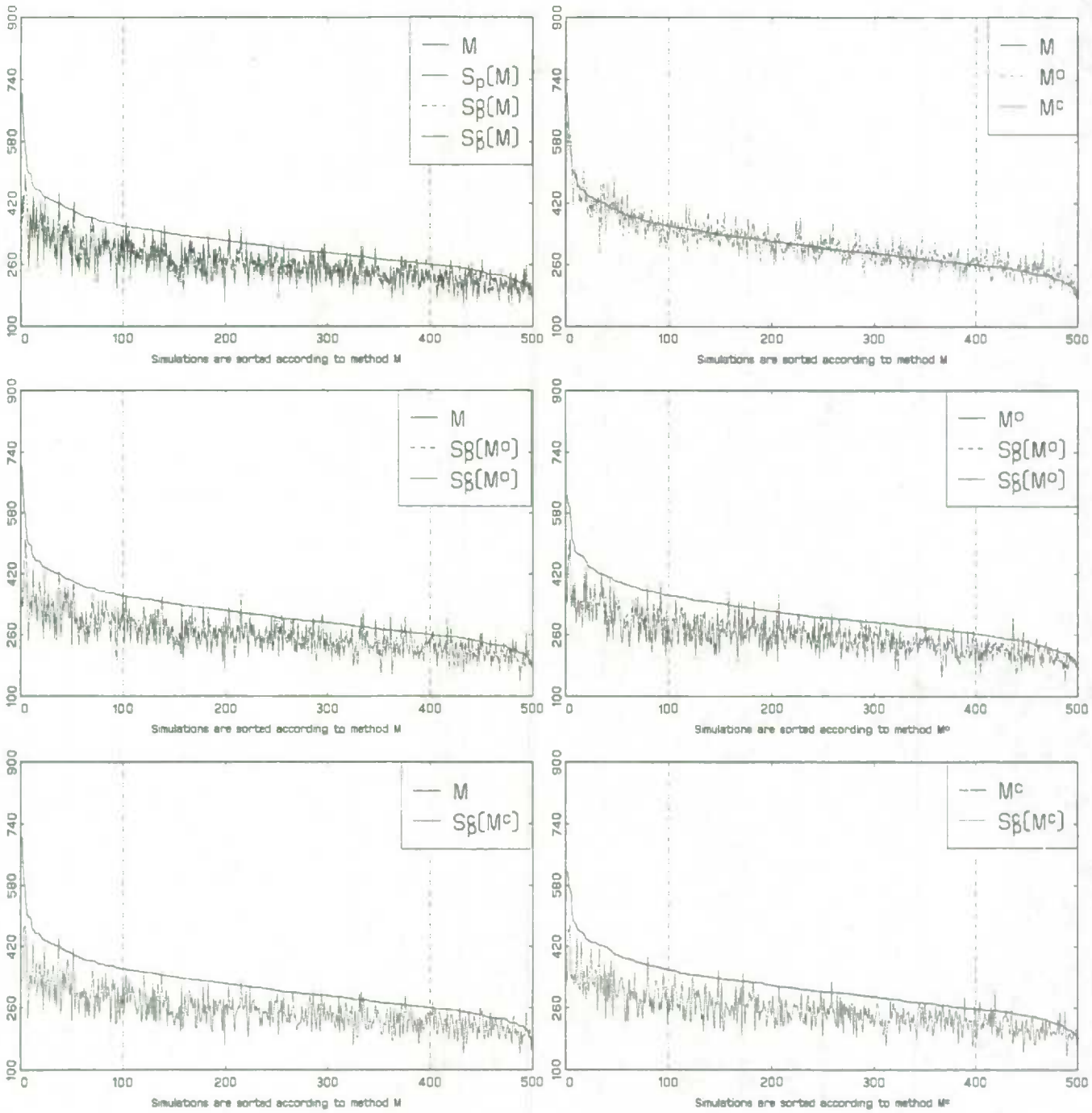
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.4-B4.6 and B4.10-B4.12.

Figure B4.6. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 2 Z Categories: 500 Simulations for RUO Datafile)



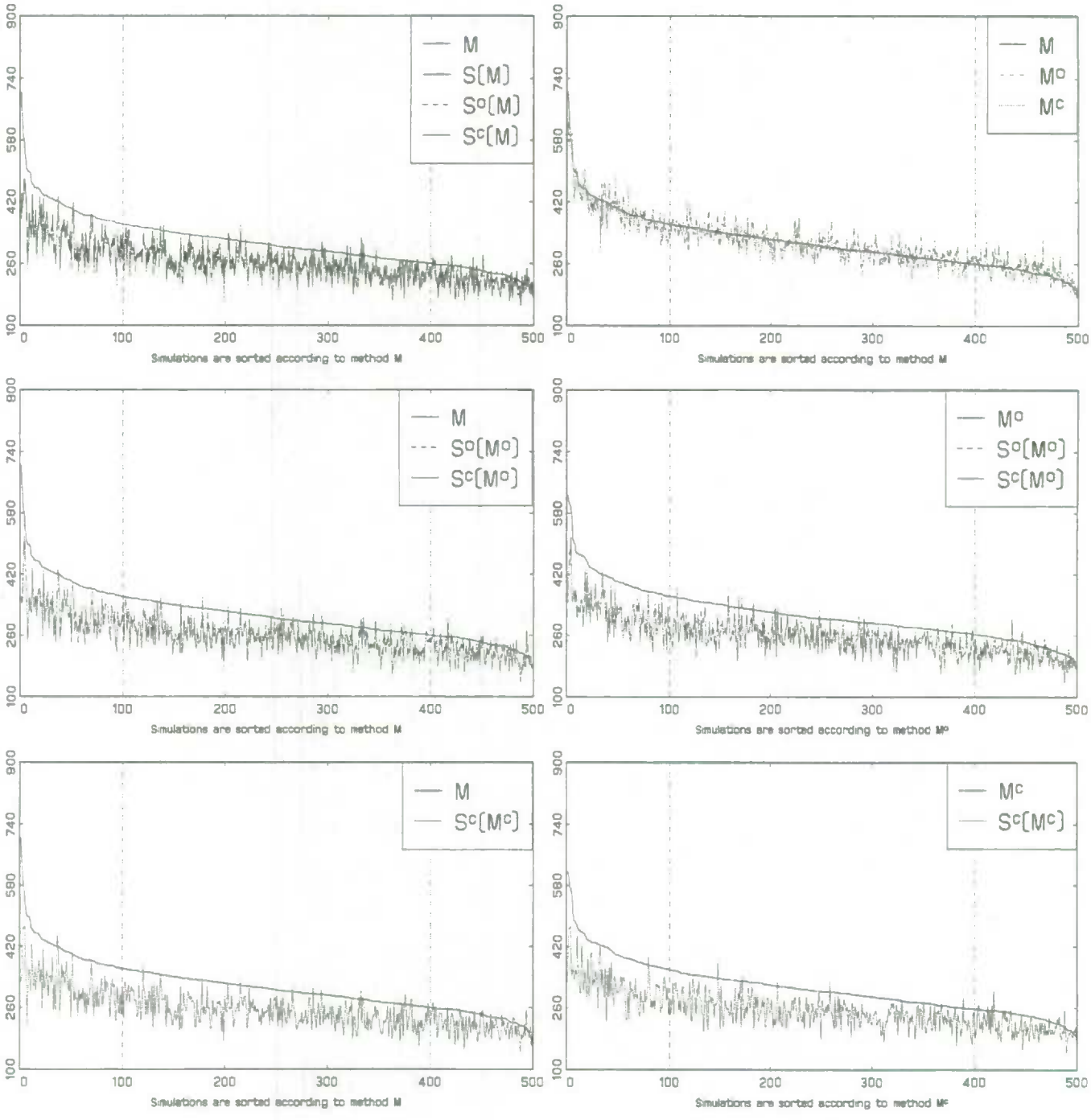
Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B4.4-B4.6 and B4.10-B4.12.

Figure B4.7 Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching with Pooling Based on 3Z Categories: 500 Simulations for RUO Datafile)



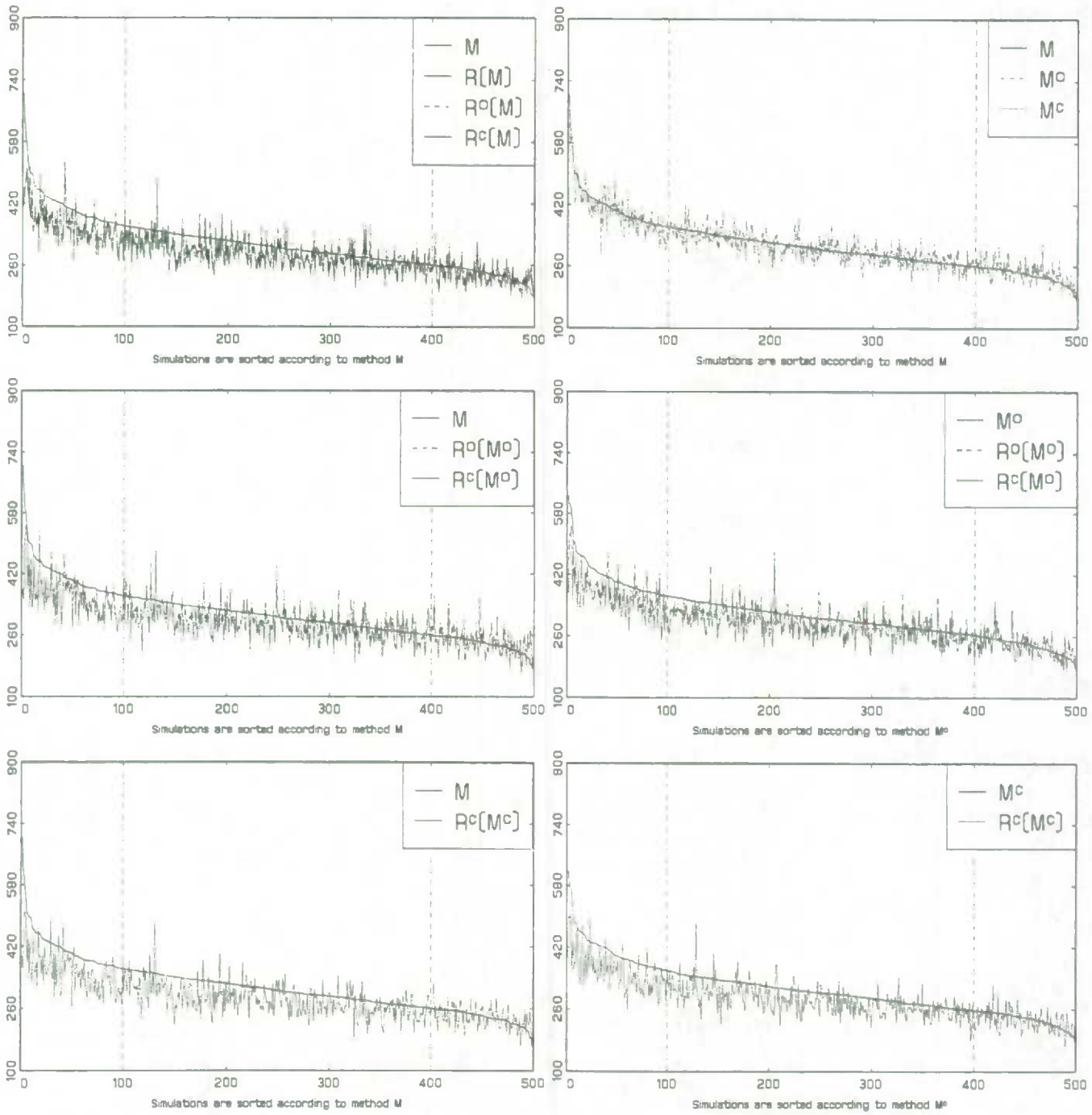
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.1-B4.3 and B4.7-B4.9.

Figure B4.8 Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Shift-and-Share Rematching Based on 3Z Categories: 500 Simulations for RUO Datafile)



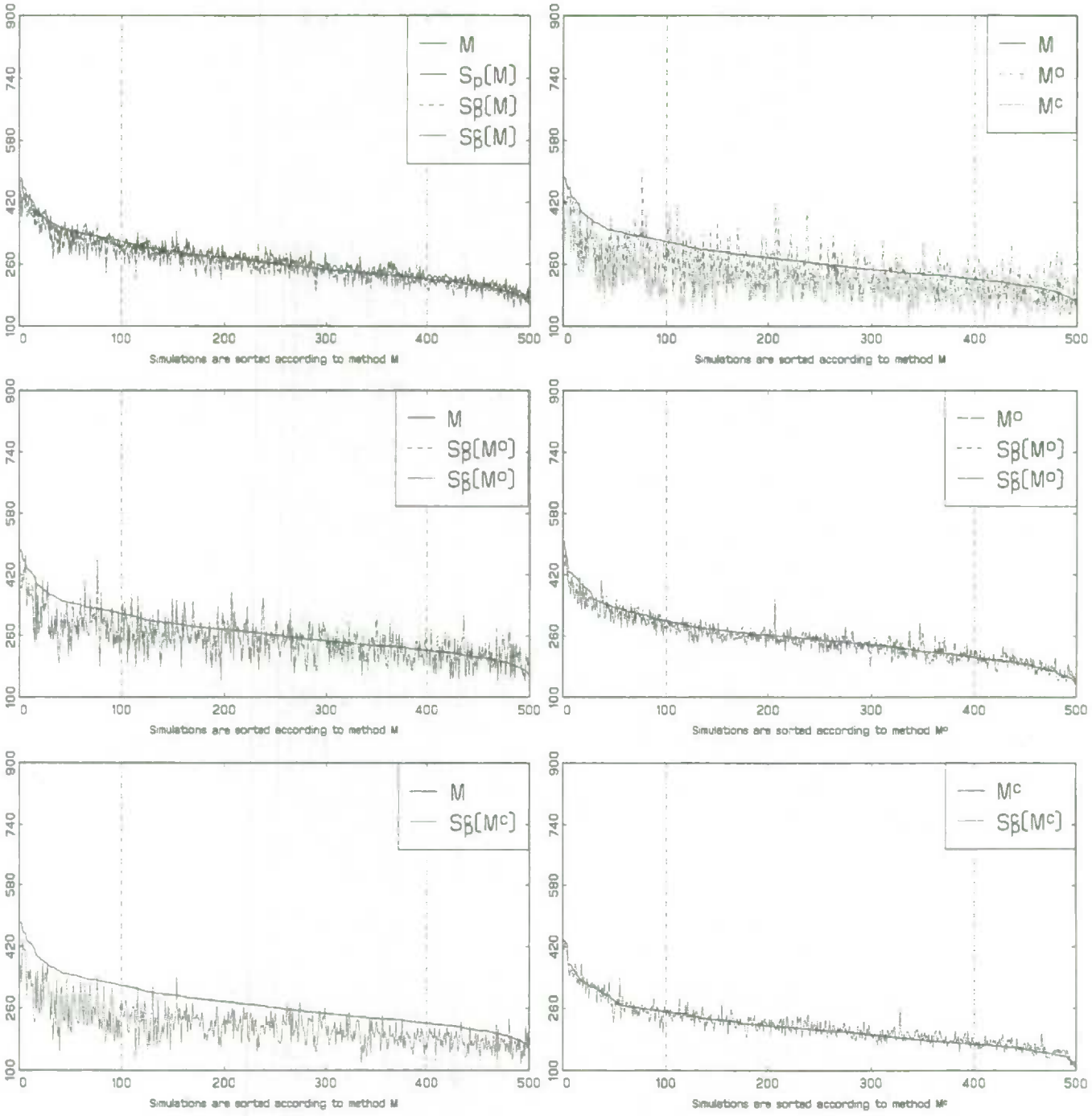
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.1-B4.3 and B4.7-B4.9.

Figure B4.9. Weighted χ^2 Evaluated over 10x4x3 Categories of the Distance Matched File
(Ratio Adjustment Based on 3 Z Categories: 500 Simulations for RUO Datafile)



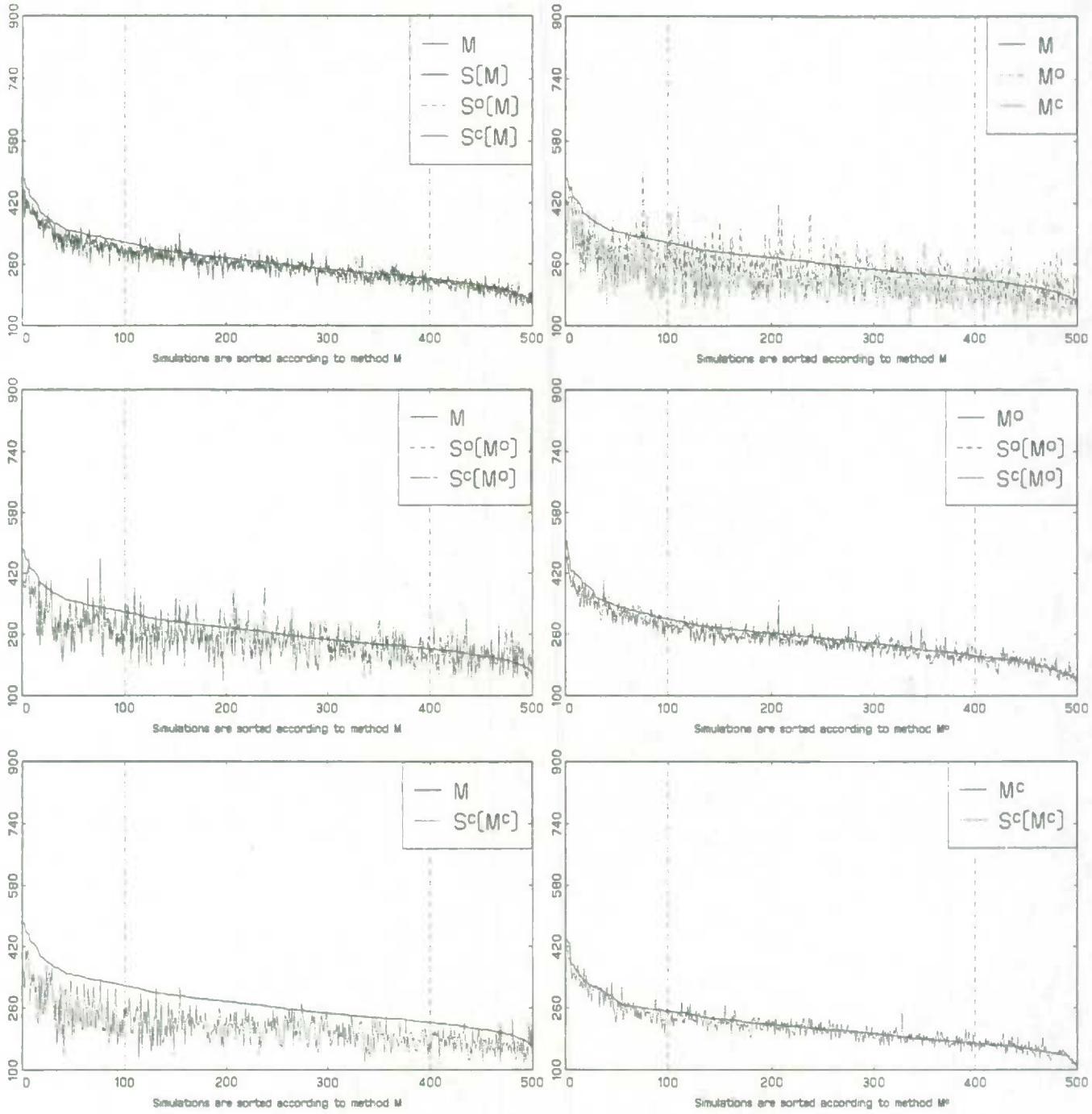
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.1-B4.3 and B4.7-B4.9.

Figure B4.10. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching with Pooling Based on 3 Z Categories: 500 Simulations for RUO Datafile)



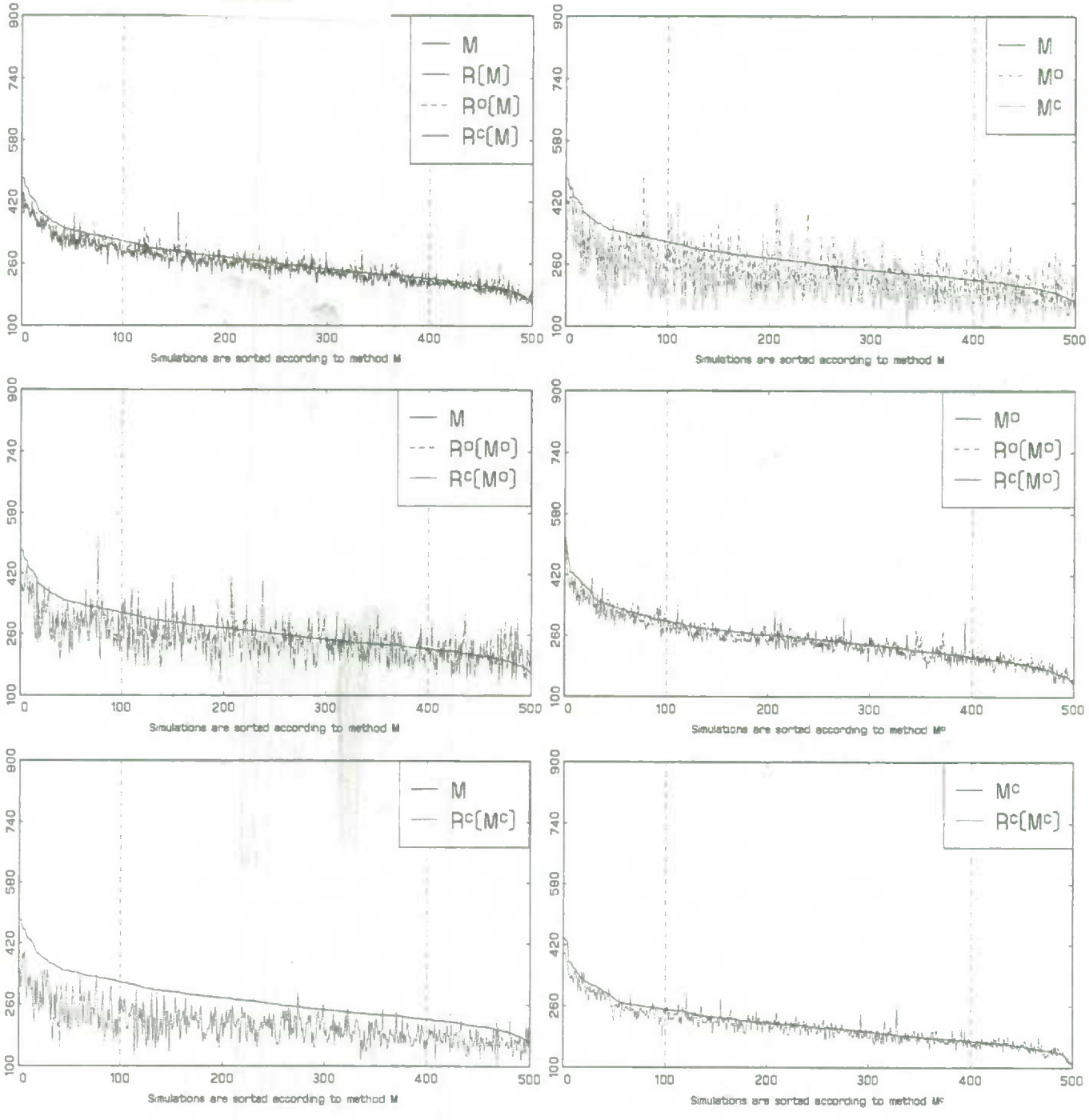
Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.4-B4.6 and B4.10-B4.12.

Figure B4.11. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Shift-and-Share Rematching Based on 3 Z Categories: 500 Simulations for RUO Datafile)



Remark: The rank-plot of the same matched files M, M° and M^{c} is repeated in Figures B4.4-B4.6 and B4.10-B4.12.

Figure B4.12. Weighted χ^2 Evaluated over 10x4x3 Categories of the Weight-Split Matched File
(Ratio Adjustment Based on 3 Z Categories: 500 Simulations for RUO Datafile)



Remark: The rank-plot of the same matched files M, M^o and M^c is repeated in Figures B4.4-B4.6 and B4.10-B4.12.