

CATEGORICALLY CONSTRAINED MATCHING

T.P. Liu and M. S. Kovačević¹

ABSTRACT

The idea of categorically constrained matching is to preserve the categorical associations in the matched file under a suitable partition of the variables involved in the matching. However, its implementation is difficult, especially when the 'imputing' variable Z is categorized into more than two categories. The most common way of imposing categorical constraints is to modify the weights in the already matched file by means of raking, so that the category weight total equals a given quantity. In this way categorical constraints affect only the weighting not the matching itself.

We are proposing a new method which leads to an iterative rematching of a small number of records until the categorical constraints are fully satisfied. The proposed method preserves the categorical association in a more flexible way. It allows the reexamination of some matches without changing the original weights of others while keeping the number of 'shifted' or 'shared' records at a minimum. Also, a record with the shifted Z category keeps its original weight whereas the weight of a shared record becomes smaller but controlled by a given threshold. We address the problem of categorically constrained matching in a real situation, when the matching files contain survey weights, and the matched file has to fulfil additional outside requirements on the size and the use of all records from both matching files. The results from the empirical study based on the Public Use Micro File from the Canadian 1986 and 1991 Censuses are presented.

KEY WORDS: Statistical matching; Iterative proportional adjustment (Raking); Survey datafiles; Pooling.

RÉSUMÉ

L'idée derrière l'appariement avec contraintes sur les catégories est de préserver les associations de catégories dans le fichier apparié selon une répartition convenable des variables utilisés pour l'appariement. Cependant, sa mise en oeuvre est difficile, particulièrement dans le cas où la variable "d'imputation" Z est classée dans plus de deux catégories. La méthode la plus fréquente d'imposer des contraintes de catégories, est de modifier les poids dans le fichier préalablement apparié en utilisant le ratissage, de telle sorte que le total des poids de la catégorie soit égal à une quantité donnée. De cette façon, les contraintes de catégories affectent uniquement la pondération, et non l'appariement.

Nous proposons une nouvelle méthode qui mène à un réappariement itératif d'un petit nombre d'enregistrements jusqu'à ce que les contraintes de catégories soient entièrement satisfaites. La méthode proposée préserve l'association entre catégories d'une façon plus flexible. Elle permet de redéfinir certains appariements sans changer les poids originaux d'autres appariements tout en gardant le nombre d'enregistrements 'décalés' ou 'partagés' au plus bas. De plus, un enregistrement avec catégorie Z décalée garde son poids original alors que le poids d'un enregistrement partagé devient plus petit, mais reste à l'intérieur d'un seuil donné. Nous considérons le problème de l'appariement avec contraintes sur catégories dans des situations réelles, lorsque les fichiers d'appariement contiennent des poids d'enquête, et le fichier apparié doit satisfaire des conditions extérieures additionnelles sur la taille et l'utilisation de tous les enregistrements des deux fichiers d'appariement. Nous présentons les résultats de l'étude empirique basée sur le fichier de micro-données à grande diffusion des recensements canadiens de 1986 et de 1991.

MOTS CLÉS: Appariement statistique; ajustement proportionnel par itération (ratissage); fichiers de données d'enquêtes; regroupement.

1. INTRODUCTION

Statistical matching is frequently used to produce comprehensive data from datafiles obtained from different surveys. Essentially it is a procedure which

identifies and links records that correspond to similar individuals. In most applications, microfiles contain survey data with the survey weights attached to records, giving rise to an additional problem of weighting in the matched file. In general, the matched

¹ Tzen-Ping Liu and Milorad S. Kovačević, Survey and Analysis Methods Development Section, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.

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.

We assume that a finite population P has three groups of characteristics (variables) of interest, X , Y and Z and that we are unable to observe the vector (X_i, Y_i, Z_i) for any unit i in P . Suppose instead that two probability samples, A and B , from P are available. One sample contains observations on the X and Y the other on the X and Z variables. For practical purposes, we assume that these samples are obtained independently. In statistical matching terminology, these samples are microdata files and the sampled units are records. Thus, we have two datafiles, $A = (X_i^A, Y_i^A, w_i^A)$, $i=1, \dots, n_A$ and $B = (X_j^B, Z_j^B, w_j^B)$, $j=1, \dots, n_B$, where the w 's are the corresponding survey weights. Using these two files one can estimate, for example, the unknown mean vector $(\bar{X}, \bar{Y}, \bar{Z})$, or the marginal histograms of frequencies $\{W_{X^*}\}$, $\{W_{Y^*}\}$, $\{W_{Z^*}\}$, $\{W_{X^*Y^*}\}$, $\{W_{X^*Z^*}\}$, where asterisks indicate categories of the categorically transformed variables, and W_{Y^*} is, for example, the weight of category Y^* . However, the multivariate histogram $\{W_{Y^*Z^*}\}$ is impossible to obtain from these separate files. Also, the component $\Sigma_{Y^*Z^*}$ of the population variance-covariance matrix is not estimable from these files, and consequently some correlations remain unknown.

To overcome these problems we do statistical matching of two datafiles by complete imputation of a set of missing variables from one datafile onto another. One file (B) is a donor, another (A) is a host file. Variables in common, X , are matching variables. Practically, matching is a two-phase procedure, where in the first phase the most similar records in files A and B are determined based on comparison of values of the common variables. In the second phase, Z values are imputed from B . Variable Y is ignored unless an auxiliary source of information about the relationship between Y and Z is available.

Singh *et al.* (1993) considered the situation when auxiliary information is available in the form of a categorical distribution and proposed a modification of the matching methods based on a loglinear method of imputation as introduced by Singh (1988). Our study is especially concerned with this type of auxiliary information.

Very often in practice there are additional requirements imposed on the matching procedure,

resulting in a number of serious modifications to the regular methods. For example, the following three requirements are placed on the statistical matching for creation of the Social Policy Simulation Database (SPSD) at Statistics Canada: (i) maintain the conditional distribution $\hat{F}(Z|X)$ 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.

The methodology exhibited in this paper considers these constraints. For a full account of the constrained matching see Liu and Kovačević (1996).

This paper investigates possibilities for improving the quality of the matched file using additional categorical constraints derived from the matching files themselves. In addition, we may use an auxiliary microfile, or alternatively, auxiliary categorical information on the variables of interest, if they are available. We assume that the matched file, obtained by some matching method, is available as well as the original matching files. The idea is to improve the categorical distribution of the matched file $\{W_{X^*Y^*Z^*}^M\}$ by the iterative adjustment of its margins to the margins of the matching files and the auxiliary table. In such a way we keep the categorical associations from the matching files $\{W_{X^*Y^*}^A\}$ and $\{W_{X^*Z^*}^B\}$, and possibly $\{W_{Y^*Z^*}^C\}$, auxiliary categorical information. We present two different ways of doing this: by weight adjustment only, and by additional partial rematch through the application of the new 'Shift-and-Share' algorithm introduced in section 4. All steps involved are illustrated by a simple example given in section 5. Results from an extensive simulation along with the final remarks and conclusions are presented in section 6.

In general, categorically constrained 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 available and adjusted marginal distributions (tables), $\{W_{X^*Y^*}^A\}$, $\{W_{X^*Z^*}^B\}$ and possibly $\{W_{Y^*Z^*}^C\}$. We call the estimated categorical distribution a look-up table. It is important to note that we don't use auxiliary information on X^*Y^* or X^*Z^* since

we would like to maintain them as observed in 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.

2. 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 efficacy. 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 'iterative proportional adjustment' (IPA) of the margins of the matched file M , modified for unexpected empty cells, to the balanced margins of A, B and C (if available).

2.1 Balancing the categorical margins of A, B and C

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

The idea of 'pooling' weights of two files at the level of a X^* category lies essentially in a combination of the weights $W_{X^*}^A$ and $W_{X^*}^B$ to obtain $W_{X^*}^{A_p} = W_{X^*}^{B_p}$, (p stands for 'pooled').

First:

$$W_{X^*}^{A_p} = W_{X^*}^{B_p} = \alpha_{X^*} W_{X^*}^A + (1 - \alpha_{X^*}) W_{X^*}^B,$$

and then

$$W_{X^*}^{A_p} = W_{X^*}^{B_p} = W_{X^*}^{A_p} \frac{W}{\sum_{(X^*)} W_{X^*}^{A_p}}, \quad (1)$$

where $0 \leq \alpha_{X^*} \leq 1$. An extra ratio adjustment in (1) is needed so that the pooled categorical 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 (1) is equal to 1.

There are several options for α_{X^*} : $\alpha_{X^*} = 0$, $\alpha_{X^*} = 1/2$, $\alpha_{X^*} = 1$, $\alpha_{X^*} = n^A (n^A + n^B)^{-1}$ or $\alpha_{X^*} = n_{X^*}^A (n_{X^*}^A + n_{X^*}^B)^{-1}$, to mention a few. Here n^A and $n_{X^*}^A$ denote the size of the file A and the size of the category X^* of the file A , respectively (similarly for n^B and $n_{X^*}^B$).

Further modification of weights is at the level of the X^*Y^* cell for file A :

$$W_{X^*Y^*}^{A_p} = W_{X^*Y^*}^A \frac{W_{X^*}^{A_p}}{W_{X^*}^A}$$

and similarly for the X^*Z^* cell of the B file.

Alternatively, balancing the X^* margins of A and B can be done by raking the $\{W_{X^*}^B\}$ to $\{W_{X^*}^A\}$ in the two-step iterative process:

$${}_1W_{X^*Z^*}^{B'}(i) = \frac{W_{X^*Z^*}^{B'}(i-1)}{W_{X^*}^{B'}(i-1)} W_{X^*}^A \quad (2a)$$

$$W_{X^*Z^*}^{B'}(i) = \frac{{}_1W_{X^*Z^*}^{B'}(i)}{{}_1W_{Z^*}^{B'}(i)} W_{Z^*}^B \quad (2b)$$

with $W_{X^*Z^*}^{B'}(0) = W_{X^*Z^*}^B$. We repeat steps (2) until $\max_{X^*} \left\{ |W_{X^*}^{B'} - W_{X^*}^A| \right\} \leq \epsilon$ and $\max_{Z^*} \left\{ |W_{Z^*}^{B'} - W_{Z^*}^B| \right\} \leq \epsilon$, where ϵ is a prespecified small number (*e.g.*, $\epsilon = 10^{-5}$). If the threshold ϵ is very small and the number of X^* categories is large then the convergence in subsequent procedures may be threatened.

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

$${}_1W_{Y^*Z^*}^{C_p}(i) = \frac{W_{Y^*Z^*}^{C_p}(i-1)}{W_{Y^*}^{C_p}(i-1)} W_{Y^*}^{A_p} \quad (3a)$$

$$W_{Y^*Z^*}^{C_p}(i) = \frac{{}_1W_{Y^*Z^*}^{C_p}(i)}{{}_1W_{Z^*}^{C_p}(i)} W_{Z^*}^{B_p} \quad (3b)$$

with $W_{Y^*Z^*}^{C_p}(0) = W_{Y^*Z^*}^C$. We repeat steps (3) until

$$\max_{Y^*} \left\{ |W_{Y^*}^{C_p} - W_{Y^*}^{A_p}| \right\} \leq \epsilon \text{ and } \max_{Z^*} \left\{ |W_{Z^*}^{C_p} - W_{Z^*}^{B_p}| \right\} \leq \epsilon.$$

The threshold ϵ has to be very small in order not to disturb later steps. In the simulation part of this study we used $\epsilon = 10^{-5}$. Note that the margins of the A and B files used in the adjustments (3) are already modified by pooling or by raking.

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

2.2 Structural and unexpected empty cells

We observe that some cells in all three categorical distributions (tables) ($W_{X^*Y^*}^A$, $W_{X^*Z^*}^B$ and $W_{X^*Y^*Z^*}^M$) 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 : i) the 'structural' empty cell, and ii) the 'unexpected' empty cell.

The first type refers to the situation where $W_{X^*Y^*Z^*}^M = 0$ and at least one of the corresponding $W_{X^*Y^*}^A$ and $W_{X^*Z^*}^B$ is equal to zero. The structural empty cell doesn't cause any problem during the IPA procedure. The second type is a more difficult case where $W_{X^*Y^*Z^*}^M = 0$ 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 IPA procedure we do the following: i) increase all cell weights in the matched file, except the structural zeros, by a positive small number δ , so that

$$W_{X^*Y^*Z^*}^{M'} = \begin{cases} 0, & \text{if } X^*Y^*Z^* \text{ is structural empty cell,} \\ W_{X^*Y^*Z^*}^M + \delta, & \text{elsewhere.} \end{cases}$$

In our simulation study we used a minimal record weight (33.333) as δ ; ii) then, 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 $X^*Y^*Z^*$:

$$W_{X^*Y^*Z^*}^I = W_{X^*Y^*Z^*}^{M'} \frac{W^M}{\sum_{\{X^*Y^*Z^*\}} W_{X^*Y^*Z^*}^{M'}}.$$

In this way the total weight remains the same as in the original M file and the IPA procedure converges. The new weight $W_{X^*Y^*Z^*}^I$ is the 'initial' weight at the next step.

2.3 IPA of the joint categorical distribution of the matched file

The last step in providing the look-up table is the IPA (raking) of the margins of the categorized matched file, already corrected for the unexpected zeros, to the balanced margins of the matching files:

$${}_1W_{X^*Y^*Z^*}^L(i) = \frac{W_{X^*Y^*Z^*}^L(i-1)}{W_{X^*Y^*}^L(i-1)} W_{X^*Y^*}^{A_p} \quad (4a)$$

$$W_{X^*Y^*Z^*}^L(i) = \frac{{}_1W_{X^*Y^*Z^*}^L(i)}{{}_1W_{X^*Z^*}^L(i)} W_{X^*Z^*}^{B_p}, \quad (4b)$$

and $W_{X^*Y^*Z^*}^L(0) = W_{X^*Y^*Z^*}^I$. We repeat this process until

$$\max_{X^*Y^*} \left\{ |W_{X^*Y^*}^L - W_{X^*Y^*}^{A_p}| \right\} \leq \epsilon_1, \max_{X^*Z^*} \left\{ |W_{X^*Z^*}^L - W_{X^*Z^*}^{B_p}| \right\} \leq \epsilon_1$$

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

$$W_{X^*Y^*Z^*}^L(i) = \frac{{}_2W_{X^*Y^*Z^*}^L(i)}{{}_2W_{Y^*Z^*}^L(i)} W_{Y^*Z^*}^{C_p} \quad (4c)$$

where 2 refers to the results of the (4b).

Note that the third step (4c) is unique for both types of auxiliary information, full $\{W_{X^*Y^*Z^*}^C\}$ and partial $\{W_{Y^*Z^*}^C\}$, since only the relationship between Y^* and Z^* is used, as explained in introduction (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 totals in the two years equal.

3. CALIBRATION AND RATIO MODIFICATION

The look-up table $\{W_{X^*Y^*Z^*}^L\}$, obtained as described in section 2, now, is used for the ratio adjustment of an individual record weight w_i , $i \in M_{X^*Y^*Z^*}$:

$$w_i' = w_i \frac{W_{X^*Y^*Z^*}^L}{W_{X^*Y^*Z^*}^M}, \quad i \in M_{X^*Y^*Z^*} \quad (5)$$

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 2.2. In order to prevent the loss of weight we do the following calibration of the look-up table before the ratio adjustment (5): (i) put back zeros in the unexpected empty cells, and (ii) adjust the weights in the complementary Z^* cells (distorting slightly the Z^* margin):

$$W_{X^*Y^*Z^*}^{Lc} = W_{X^*Y^*Z^*}^L \frac{\sum_{(Z^*)} W_{X^*Y^*Z^*}^L}{\sum_{(Z^*) \setminus \{\text{unexp. 0 cells}\}} W_{X^*Y^*Z^*}^L} \quad (6)$$

where the subscript c stands for 'calibrated'. The summation is taken over all Z^* categories in the numerator, and over all Z^* categories except those that resulted in unexpected zeros, in the denominator.

Then $W_{X^*Y^*Z^*}^L$ in (5) is substituted by the $W_{X^*Y^*Z^*}^{Lc}$ from (6).

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 (X, Y, Z) . Also, the ratio adjusted weights may be very small or very large. This method doesn't solve the unexpected empty cell problem although their effect on convergence of the IPA algorithm is annulled. Clearly this problem means that information contained in the matching files (A and B) is not fully utilised.

4. PARTIAL REMATCHING USING THE NEW "MINIMUM SHIFT AND SHARE" ALGORITHM

To solve the problem of unexpected empty cells that couldn't be solved by calibration and ratio adjustment, we propose a new method which uses categorical constraints for an additional rematching of the records, as well.

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

$$\Delta_{X^*Y^*Z_k^*} = W_{X^*Y^*Z_k^*}^M - W_{X^*Y^*Z_k^*}^L, \quad k = 1, \dots, K.$$

The goal is to rearrange the matched file M into M' such that

$$W_{X^*Y^*Z^*}^{M'} - W_{X^*Y^*Z^*}^L \approx 0$$

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 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, $\langle X_i^A, Y_i^A, Z_j^B, w_i^M, q_j=3 \rangle$, 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 new "Minimum shift and share" algorithm.

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

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

Suppose that for the first K_1 categories $Z_1^*, Z_2^*, \dots, Z_{K_1}^*$ the difference $\Delta_{X^*Y^*Z_i^*} \geq \epsilon$, and that for the last K_2 categories, $Z_{K-K_2+1}^*, Z_{K-K_2+2}^*, \dots, Z_K^*$, the

difference $\Delta_{X \cdot Y \cdot Z_k^*} \leq -\varepsilon$.

The algorithm has two parts: the ‘Shift’ and the ‘Share’ part. We first introduce the ‘Shift’ part:

(2a) We order the Z_k^* categories, $k=1, \dots, K$, into the descending order regarding the difference $\Delta_{X \cdot Y \cdot Z^*}$. Then we search Z_k^* , $k=1, \dots, K_1$, categories, starting with the one with the largest difference, for records $i \in X \cdot Y \cdot Z_k^*$ with the count $q_i \geq 2$, and the weights

$$w_i < \varepsilon + \Delta_{X \cdot Y \cdot Z_k^*}, \quad 1 \leq k \leq K_1 \quad (7a)$$

and $w_i < \varepsilon - \Delta_{X \cdot Y \cdot Z_k^*}$. (7b)

Note that the last K -th difference has the largest negative value. The records which satisfy conditions (7) and are in the category with the largest possible positive difference are candidates for further processing and we designate them as ‘movable’. If there is no movable record, we end the ‘Shift’ part of the procedure.

(3a) Next we sort all ‘movable’ records in Z_k^* into a descending order according to their weights w_i . The first ‘movable’ record (with the maximum weight) replaces its Z_k^* category with the category Z_K^* , which has the maximum negative difference.

(4a) A record with the changed Z^* category, carries its weight over to the new category. However, it has to be rematched with another record from the original B file which belongs to this category Z_K^* .

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

$$q_k = q_k - 1 \text{ and } q_K = q_K + 1,$$

$$\Delta_{X \cdot Y \cdot Z_k^*} = \Delta_{X \cdot Y \cdot Z_k^*} - w_i \text{ and } \Delta_{X \cdot Y \cdot Z_K^*} = \Delta_{X \cdot Y \cdot Z_K^*} + w_i$$

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

The ‘Share’ part of the algorithm is defined similarly. Step (1) is common for both parts.

(2b) Again, we look at Z_k^* categories, $k=1, \dots, K_1$, starting from the one with the largest difference. For each k we are searching for the j -th category

$j = K, K-1, \dots, K-K_2+1$ with the largest possible negative difference and check if there are records with weights

$$w_i \geq \Delta_{X \cdot Y \cdot Z_k^*}, \quad 1 \leq k \leq K_1$$

and

$$w_i \geq -\Delta_{X \cdot Y \cdot Z_j^*}, \text{ for some } j, \quad K \geq j \geq K-K_2+1.$$

These records are candidates for further processing and we name them as ‘sharable’. If there is no sharable record, we end the ‘Share’ part of the procedure.

(3b) Next we select a sharable record at random, duplicate it and assign to one of its replicates the category with the largest possible negative difference such that $w_i \geq -\Delta_{X \cdot Y \cdot Z_j^*}$. 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_j^* , $K \geq j \geq K-K_2+1$ in the following way:

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

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

(6b) Update the weight differences $\Delta_{X \cdot Y \cdot Z_k^*}$ and $\Delta_{X \cdot Y \cdot Z_j^*}$ as in (5a). We repeat steps (1), (2b) - (6b) until no ‘share’ is possible.

After the application of the “Minimum Shift and Share” 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 3 where $\{W_{X \cdot Y \cdot Z_k^*}^M\}$ is replaced by the corresponding value obtained in the ‘Shift-and-Share’ procedure.

The two categorically constrained procedures described in sections 3 and 4 usually result in slightly different matched files. The first procedure is essentially a way to adjust weights so that the

categorical constraints are satisfied. The new 'Shift-and-Share' procedure, in contrast, doesn't change the weights of all records but may have a small number of rematches which are products of the move and split procedure aimed at minimal distortion of the marginal categorical distributions. It deals with the unexpected empty cells problem straightforwardly by rematching a certain number of records. Therefore it uses more information from matching files than the ratio adjustment only. The second procedure is more complex. Obviously, the more Z^* categories, the longer the procedure. The next section contains a simple but illustrative example of procedures described in sections 2, 3 and 4.

5. ILLUSTRATION

We demonstrate our procedures on 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 number of records and the weights in categorically transformed matching files (samples) A and B , and the matched file M :

Tables 1-6: Number of Records and Weights of the Categorized Matching Files A and B and the Matched File M

Counts $\{n_{X^*Y^*}^A\}$ and $\{n_{X^*Z^*}^B\}$:						Weights $\{W_{X^*Y^*}^A\}$ and $\{W_{X^*Z^*}^B\}$:									
$\{n_{X^*Y^*}^A\}$					Y_1^*	Y_2^*	Y_3^*	Y_4^*	69	$\{n_{X^*Z^*}^B\}$			Z_1^*	Z_2^*	44
X_1^*	39	5	23	2	69	X_1^*	19	3	22	X_1^*	10376.4	1330.3	6119.4	532.1	18358.2
X_2^*	18	5	4	0	27	X_2^*	1	0	1	X_2^*	4789.1	1330.3	1064.2	0.0	7183.6
X_3^*	6	5	11	2	24	X_3^*	4	3	7	X_3^*	1596.4	1330.3	2926.7	532.1	6385.5
X_4^*	43	10	43	5	101	X_4^*	8	6	14	X_4^*	11440.6	2660.6	11440.6	1330.3	26872.1
	106	25	81	9	221		32	12	44		28202.4	6651.5	21550.9	2394.5	58799.4

Counts $\{n_{X^*Y^*Z^*}^M\}$:										Weights $\{W_{X^*Y^*Z^*}^M\}$:																
$\{n_{X^*Y^*Z^*}^M\}$									$Y_1^*Z_1^*$	$Y_1^*Z_2^*$	$Y_2^*Z_1^*$	$Y_2^*Z_2^*$	$Y_3^*Z_1^*$	$Y_3^*Z_2^*$	$Y_4^*Z_1^*$	$Y_4^*Z_2^*$	229	$\{W_{X^*Y^*Z^*}^M\}$								
X_1^*	38	4	6	1	23	2	2	0	76	X_1^*	9312.1	1064.2	1241.6	88.7	5587.3	532.1	532.1	0.0	18358.2							
X_2^*	18	0	5	0	4	0	0	0	27	X_2^*	4789.1	0.0	1330.3	0.0	1064.2	0.0	0.0	0.0	7183.6							
X_3^*	0	6	3	2	7	4	1	1	24	X_3^*	0.0	1596.4	798.2	532.1	1862.4	1064.2	266.1	266.1	6385.5							
X_4^*	23	21	6	4	27	16	5	0	102	X_4^*	5986.4	5454.2	1596.4	1064.2	7183.6	4257.0	1330.3	0.0	26872.1							
	79	31	20	7	61	22	8	1	229		20087.6	8114.9	4966.5	1685.1	15697.6	5853.3	2128.5	266.1	58799.4							

Evidently, the marginal totals of X^* categories in A and B don't agree. After pooling these two tables by

category size at the level of the X^* categories we have the following situation:

Tables 7-8: Tables $\{W_{X^*Y^*}^{A_p}\}$ and $\{W_{X^*Z^*}^{B_p}\}$ After Pooling

$\{W_{X^*Y^*}^{A_p}\}$					$\{W_{X^*Z^*}^{B_p}\}$		
X^*	Y_1^*	Y_2^*	Y_3^*	Y_4^*	Z_1^*	Z_2^*	
X_1^*	11468.3	1470.3	6763.4	588.1	20290.1	17523.	2766.8
X_2^*	4486.8	1246.3	997.1	0.0	6730.2	6730.2	0.0
X_3^*	1702.1	1418.4	3120.5	567.4	6808.4	3890.5	2917.9
X_4^*	10631.1	2472.3	10631.1	1236.2	24970.7	14269.0	10701.7
	28288.3	6607.4	21512.1	2391.7	58799.4	42412.9	16386.5

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

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

Table 9: The Initial Table for the Raking Procedure

$\{W_{X^*Y^*Z^*}^I\}$	$Y_1^*Z_1^*$	$Y_1^*Z_2^*$	$Y_2^*Z_1^*$	$Y_2^*Z_2^*$	$Y_3^*Z_1^*$	$Y_3^*Z_2^*$	$Y_4^*Z_1^*$	$Y_4^*Z_2^*$	
X_1^*	9204.6	1081.0	1255.7	120.2	5535.9	556.9	556.9	32.8	18344.1
X_2^*	4749.7	0.0	1343.1	0.0	1081.0	0.0	0.0	0.0	7173.8
X_3^*	32.8	1605.1	819.0	556.9	1867.2	1081.0	294.9	294.9	6551.8
X_4^*	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

auxiliary categorical information is available in the form of $\{W_{Y^*Z^*}^C\}$.

Tables 10-11: Auxiliary Categorical Table

Original auxiliary table				After raking to $W_{Y^*}^{A_p}$ and $W_{Z^*}^{B_p}$			
$\{W_{Y^*Z^*}^C\}$	Z_1^*	Y_2^*		$\{W_{Y^*Z^*}^C\}$	Z_1^*	Z_2^*	
Y_1^*	21066.5	9699.9	30766.4	Y_1^*	21637.5	6650.8	28288.3
Y_2^*	3366.6	1733.3	5099.9	Y_2^*	4917.3	1690.1	6607.4
Y_3^*	11533.2	9166.6	20699.8	Y_3^*	14054.9	7457.1	21512.1
Y_4^*	1500.0	733.3	2233.3	Y_4^*	1803.2	588.5	2391.7
	37466.3	21333.1	58799.4		42412.9	16386.5	58799.4

Finally, the look-up table obtained by raking of the $W_{X^*Y^*Z^*}^I$ to the X^*Y^* margin of A_p , the X^*Z^* margin of B_p and the Y^*Z^* margin of C_p is $\{W_{X^*Y^*Z^*}^L\}$. Calibration modifies the look-up table only in six cells (underlined) by setting back to zero the unexpected empty cells.

Their contents from the look-up table are added to the complementary Z^* cell. The calibrated look-up table represents the categorical distribution of the matched file after the ratio adjustment of individual weights :

Tables 12-13: The Look-up Table and its Calibrated Version

$\{W_{X^*Y^*Z^*}^L\}$	$Y_1^*Z_1^*$	$Y_1^*Z_2^*$	$Y_2^*Z_1^*$	$Y_2^*Z_2^*$	$Y_3^*Z_1^*$	$Y_3^*Z_2^*$	$Y_4^*Z_1^*$	$Y_4^*Z_2^*$	
X_1^*	10370.4	1097.9	1254.2	216.1	5460.3	1303.1	438.4	149.7	20290.1
X_2^*	4486.8	0.0	1246.3	0.0	997.1	0.0	0.0	0.0	6730.2
X_3^*	154.9	1547.2	1110.0	308.1	2379.9	740.6	245.3	322.1	6808.4
X_4^*	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

$\{W_{X^*Y^*Z^*}^L\}$	$Y_1^*Z_1^*$	$Y_1^*Z_2^*$	$Y_2^*Z_1^*$	$Y_2^*Z_2^*$	$Y_3^*Z_1^*$	$Y_3^*Z_2^*$	$Y_4^*Z_1^*$	$Y_4^*Z_2^*$	
X_1^*	10370.4	1097.9	1254.2	216.1	5460.3	1303.1	<u>588.1</u>	<u>0.0</u>	20290.1
X_2^*	4486.8	0.0	1246.3	0.0	997.1	0.0	0.0	0.0	6730.2
X_3^*	<u>0.0</u>	<u>1702.1</u>	1110.0	308.1	2379.9	740.6	245.3	322.1	6808.4
X_4^*	6625.4	4005.7	1306.5	1165.8	5217.6	5413.5	<u>1236.2</u>	<u>0.0</u>	24970.7
	21482.6	6805.7	4917.3	1690.1	14054.9	7457.1	2069.6	322.1	58799.4

For an application of the Shift-and-Share algorithm

we first make a table with differences $\Delta_{X^*Y^*Z^*}$:

Table 14: Table with Differences $\Delta_{X^i Y^j Z^k} = W_{X^i Y^j Z^k}^M - W_{X^i Y^j Z^k}^L$

$(\Delta_{X^i Y^j Z^k})$	$Y_1^j Z_1^k$	$Y_1^j Z_2^k$	$Y_2^j Z_1^k$	$Y_2^j Z_2^k$	$Y_3^j Z_1^k$	$Y_3^j Z_2^k$	$Y_4^j Z_1^k$	$Y_4^j Z_2^k$
X_1^i	-1058.3	-33.7	-12.6	-127.4	+127.0	-771.0	+93.7	-149.7
X_2^i	+302.3	0.0	+84.0	0.0	+67.1	0.0	0.0	0.0
X_3^i	-154.9	+49.2	-312.1	+224.0	-517.5	+323.6	+20.8	-56.0
X_4^i	-639.0	+1448.5	+289.9	-101.6	+1966.0	-1156.5	+210.8	-116.7

The number of moved and replicated records is given in the next table. 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. The zero value means that there was no change. We also provide a table with weights that were moved from one category to another.

Table 15: Number of ‘Shift-and-Share’ Records and Corresponding Weights

	$Y_1^j Z_1^k$	$Y_1^j Z_2^k$	$Y_2^j Z_1^k$	$Y_2^j Z_2^k$	$Y_3^j Z_1^k$	$Y_3^j Z_2^k$	$Y_4^j Z_1^k$	$Y_4^j Z_2^k$		$Y_1^j Z_1^k$	$Y_1^j Z_2^k$	$Y_2^j Z_1^k$	$Y_2^j Z_2^k$	$Y_3^j Z_1^k$	$Y_3^j Z_2^k$	$Y_4^j Z_1^k$	$Y_4^j Z_2^k$
X_1^i Shift	0	0	0	0	0	0	0	0	0	0	0	0	0	-127.0	+127.0	-93.7	+93.7
Share	0	0	0	0	-1	+1	-1	+1		0	0	0	0	0	0	0	0
X_2^i Shift	0	0	0	0	0	0	0	0	0	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_3^i Shift	0	0	0	0	+1	-1	0	0	+49.2	-49.2	+224.0	-224.0	+323.7	-323.7	-20.8	+20.8	
Share	+1	-1	+1	-1	+1	-1	-1	+1									
X_4^i Shift	+2	-2	0	0	-4	+4	0	0	+639.0	-639.0	-101.6	+101.6	-1156.5	+1156.5	-116.7	+116.7	
Share	+1	-1	-1	+1	-1	+1	-1	+1									

After the application of the ‘Shift-and-Share’ algorithm we have a revised matched file with a new categorical distribution. We use this new table along

with the look-up table for the ratio adjustment of individual weights.

Table 16: Categorical Table of the Matched File M After Rematching by the ‘Shift-and-Share’ Algorithm

	$Y_1^j Z_1^k$	$Y_1^j Z_2^k$	$Y_2^j Z_1^k$	$Y_2^j Z_2^k$	$Y_3^j Z_1^k$	$Y_3^j Z_2^k$	$Y_4^j Z_1^k$	$Y_4^j Z_2^k$	
X_1^i	9312.4	1064.2	1241.6	88.7	5460.3	659.1	438.4	93.7	18358.2
X_2^i	4789.1	0.0	1330.3	0.0	1064.2	0.0	0.0	0.0	7183.6
X_3^i	49.2	1547.2	1022.2	308.1	2186.1	740.6	245.3	322.1	6808.4
X_4^i	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

6. EMPIRICAL STUDY AND CONCLUDING REMARKS

A large simulation study on the performance of a variety of matching methods was undertaken in Statistics Canada. 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 Censuses on Households/Housing for the province of Québec. We chose variables from the Census PUMF's as *X*, *Y*, *Z* that were similar to those encountered in actual matching for the SPSD. As matching variables *X*, we considered variables that provide details on urbanization, residential tenure, presence of mortgage, total household income categorized into five categories, household size, household composition, sex and age of the household maintainer. They were used as categorical variables for grouping the records into a number of matching classes. The total household income was also used as a continuous type common variable. The variables on total household investment income and total household government transfer payments were the *Y* variables in our simulation study. Note that these *Y* variables are negatively correlated. The *Y* variables may take negative values, but for the purpose of this study we used their absolute values. The monthly gross rent or the owner's major monthly payments were chosen to be the imputed variable, *Z*.

Matching files *A* and *B* were obtained as independent and nonoverlapping simple random samples from the population. For file *C* (auxiliary data file), we used the complete population. Also, categorical auxiliary information was derived from the complete population. Sample *A* (the host file) was larger and was selected first. Its size was about one eighth the size of the population. A sample *B* was selected from the remainder. In this way, we prevented a record from *A* being matched to itself. The size of *B* was about one fortieth the size of population. This sampling procedure was repeated independently for each simulation. For a complete description of the simulation study, see Liu and Kovačević (1996).

For the purpose of this article we assumed that the matched file *M* was obtained by the nearest neighbour matching using the Euclidean distance measure and without use of any auxiliary information. Files *A* and *B* were drawn from PUMF from 1991 from the subpopulation of households with rented dwellings (22,848 records) in Montreal and Québec City.

Then, we applied categorically constrained matching to the already matched file. We categorized the *X* variables into ten categories, the *Y* variables into

two categories each, making a total of four categories. The *Z* variable was categorized into two, three or four categories, using the median, terciles and quartiles of the *Z* distribution respectively. Pooling as described in 2.1 was done by category size.

We assume three different scenarios: without an auxiliary categorical table, with the auxiliary categorical table but outdated (based on the complete corresponding subpopulation of PUMF records from 1986), and with the current auxiliary categorical table (based on the PUMF records from the corresponding subpopulation in 1991).

The performance of the two methods for categorical matching (ratio adjustment only and shift-and-share followed by the ratio adjustment) was evaluated by comparison of the categorical distributions of the population and the resulting matched files via the Chi-Square Index

$$\chi^2 = \frac{n^M}{W} \sum_k \frac{(W_k^M - W_k)^2}{W_k + 0.5 W / n^M} \quad (8)$$

where n^M denotes the size of the matched file, and W_k^M and W_k the weight of the k -th cell in the matched file and population, respectively, and W denotes the total population size. Classes for computation of the Chi-Square Index were the same as those used for categorically constrained matching. 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 for the Chi-square index over 1000 simulations. Here (Figure 1) we present values of (8) for the original matched file and for files obtained by the two studied methods of categorical constraining. After sorting into a decreasing order by the value of (8) for the original matched file, we are displaying the first 250 simulations for each of the settings considered (no auxiliary categorical information, with the five years out-of-date auxiliary table, and with the current auxiliary table).

Our general finding is that the quality of a matched file can be improved by additional categorical constraints in most of the cases. In particular, if there is no auxiliary information on the Y^*Z^* distribution, there is a very small gain of conducting the categorically constrained matching (mainly for the extreme cases) and there is no difference in performance of the two methods compared. There is a considerable gain when either auxiliary categorical information is available (five years out-of-date or current). We find that both methods perform similarly

and yield more for a finer categorization of the Z variable. From Table 17 with the average values of the Chi-Square Index computed over 1000 simulations we see that the Shift-and-Share algorithm combined with the ratio adjustment slightly outperforms the ratio adjustment only.

Table 17: Mean Chi-Square Index
Computed over 1000 Simulations

Number of Z Categories	Original Matched File	Categorically Constrained Matching	Type of Auxiliary Table		
			None	Outdated	Current
Two	55.6	Shift-and-Share	51.8	42.2	40.0
		Ratio Adjustment	52.6	43.6	41.4
Three	89.5	Shift-and-Share	84.4	70.1	61.3
		Ratio Adjustment	84.8	72.1	64.2
Four	122.9	Shift-and-Share	115.9	89.8	83.9
		Ratio Adjustment	113.5	90.4	85.2

ACKNOWLEDGMENT

This work was supported by the Social and Economic Studies Division of Statistics Canada. The authors benefited from discussions with Harold Mantel and Geoff Rowe.

REFERENCES

- 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., and Kovačević, M.S. (1996). "Statistical matching of survey datafiles", Presented at the 23rd meeting of the Advisory Committee on Statistical Methods, Statistics Canada, Ottawa.
- 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.
- 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.
- Singh, A.C. (1988). "Log-linear imputation", Methodology Branch Working Paper, SSMD, 88-029E. Statistics Canada.
- 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.

Figure 1: CHI-SQUARE INDEX FOR CATEGORICALLY CONSTRAINED MATCHING
 (250 simulations sorted by the Chi-Square Index of the original matched file)

