

## A CATEGORICAL CONSTRAINTS GUIDED MATCHING ALGORITHM

Tzen-Ping Liu<sup>1</sup>

### ABSTRACT

Statistical matching is a technique for combining of data from multiple sources (*i.e.*, the matching files) at the micro level by identifying and linking records that correspond to similar individuals. In a real situation, the matching files may contain survey weights and the resulting matched file has to fulfill additional outside requirements on its size and the use of all records from all matching files. In this article, which considers pairs of matching files, a new categorical constraints guided, minimum distance and maximum weight matching algorithm is introduced. The algorithm iteratively synthesizes matched records from two of matching files in a manner that the nearest and heaviest weight records match first, and the farthest and lightest weight records match last, while satisfying the full auxiliary categorical constraints. The resulting matched file has an architecture which preserves the categorical association of the variables and the weights, satisfies the outside requirements, and keeps the matched records' weight greater than a given threshold.

KEY WORDS: Forward and backward imputations; Pooling; Raking; Shift and share adjustments; Split weight; Survey datafiles.

### RÉSUMÉ

La méthode statistique de jumelage est une technique d'intégration de données provenant de différentes sources (c.-à-d., les fichiers de jumelage) au niveau le plus fin par identification et appariement d'enregistrements qui correspondent à des individus possédant des caractéristiques semblables. En situation réelles, les fichiers de jumelage peuvent contenir des poids de sondage et le fichier apparié obtenu doit satisfaire des contraintes additionnelles externes concernant sa taille et l'utilisation de tous les enregistrements provenant des fichiers de jumelage. Dans cet article, considérant les paires de fichiers jumelés, un nouvel algorithme guidé par les contraintes catégorielles, de distance minimum et de poids maximum est introduit. L'algorithme synthétise itérativement les enregistrements jumelés à partir d'une paire de fichiers de sorte que les enregistrements les plus près et possédant les poids les plus élevés sont appariés en premier alors que les enregistrements les plus éloignés et possédant les poids les plus faibles sont appariés en dernier, tout en satisfaisant à toutes les contraintes catégorielles. Le fichier jumelé qui en résulte a une architecture qui préserve l'association catégorielle des variables et des poids, satisfait aux exigences externes, et fait en sorte que le poids des enregistrements appariés est plus élevé qu'un niveau pré-établi.

MOTS-CLÉS: Imputation en avant et en arrière; groupement; itération; ajustements par déplacement et partage; poids partagé; fichiers de données d'enquête.

### 1. INTRODUCTION

Policy relevant analysis of tax and transfer programs, public health and welfare, educational attainment etc., require comprehensive databases which usually consist of datafiles from different sources (*i.e.*, the matching files). These files typically contain very few or no individuals in common and 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 synthesize comprehensive files of data from multiple sources. In general, the matched file is aimed at inference about the true joint distribution of all variables in it, so it is expected that the underlying population is represented, and that the distortion of marginal distributions of matching files is be within the sampling variation. In most applications matching files contain survey data with survey weights attached to the records. Another problem is how to weight records in a matched file when the matched records originally had different weights.

---

<sup>1</sup> Household Survey Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6, tzenliu@statcan.ca

For example the Canadian Social Policy Simulation Database (SPSD) at Statistics Canada was constructed to support micro-analytic modeling 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 e.g. Wolfson *et al.* 1987)

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 non-overlapping samples are obtained independently. In statistical matching terminology these samples are microdata files and the sampled units are records. Thus, we have two microdata files,  $A = (X_i^A, Y_i^A, w_i^A)$ ,  $i=1, \dots, n^A$  and  $B = (X_j^B, Y_j^B, w_j^B)$ ,  $j=1, \dots, n^B$ , 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  $\{W_{X^*}\}$ ,  $\{W_{Y^*}\}$ ,  $\{W_{X^*Y^*}\}$ , where asterisks indicate categories of the categorically transformed variables, and  $W_{Y^*}$  is, for example, the weight of category  $Y^*$ . However, the multivariate histogram is impossible to obtain from these separate files. Also, the covariance  $\Sigma_{Y^*Z^*}$  of the population variance-covariance matrix can not be estimated from these files, and consequently some correlations remain unknown.

In practice there are additional requirements imposed on the matching procedure, resulting in a number of modifications to the regular methods. For example, the following three requirements are placed on the statistical matching for the SPSD at Statistics Canada: (i) maintain the conditional distribution  $F(Z|X)$  as it is estimated by the smaller matching 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 larger matching file  $A$ .

The conditional distribution of variables that appear in two non-overlapping files given the distribution of common variables is identifiable from the corresponding marginal only under the assumption of their conditional independence (CI) (Sims, 1972). The hurdle of this assumption has often been stressed in the literature on statistical matching. Ruggles *et al.* (1977), Barr *et al.* (1981), Rodgers and DeVol (1982) and Rubin (1986) give empirical evidence that violation of the conditional independence (CI) 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 microdata file and 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. In a different way, Goel and Ramalingm (1989) considered a matching strategy to preserve both side marginal distributions of the matching files into the matched file as a solution of a linear programming transportation problem. Singh *et al.* (1993) consider 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, 1989). The categorical distribution approach is a method which potentially recovers any relationship between variables and weights, while previous research only tested on hypothetical data and did not include techniques on the use of weights.

Liu and Kovačević (1996, 1997 and 1998a,b) had built a two-phase matching procedure to use auxiliary information (variables and tables) and deal with survey weights. The first phase matching is with or without auxiliary variables and the second phase is a minimum adjustment rematching algorithm base on the original categorical weights-sum tables of the matching files and a *partial* (non-common variables only) auxiliary categorical weights-sum table.

This paper investigates possibilities for improving the quality of the purpose matched file using additional categorical constraints derived from an all variables (*full*) auxiliary categorical association (*categorical weights-sum table*) and the matching files themselves. The idea, to synthesize a matched file  $M$  which has improve categorical distribution  $\{W_{X^*Y^*Z^*}^M\}$  by the iteratively proportional matching of the matching files according to a categorical look-up table, is developed in

Section 2. In such a way, the categorical associations  $\{W_{X^*Y^*}^A\}$  and  $\{W_{X^*Z^*}^B\}$  from the two matching files, a possibly full auxiliary categorical association  $\{W_{X^*Y^*Z^*}^C\}$ , and its marginal categorical association  $\{W_{Y^*Z^*}^C\}$  are kept.

## 2. THE CATEGORICAL LOOK-UP TABLE

In the proposed algorithm, to sufficiently use the categorical information contained in a pair of matching files and a full auxiliary weights-sum table, the construction of categorical constraints of the matching procedures consists of the following four steps:

- (i) transformation of 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) identify the  $X^*$  marginal distribution of matching files by pooling process, and adjust the matching record weights simultaneously.
- (iii) for the partial marginal table  $\{W_{Y^*Z^*}^C\}$  of a full auxiliary categorical distribution (tables)  $\{W_{X^*Y^*Z^*}^C\}$ , rate its  $Y^*$  and  $Z^*$  marginal distributions according to the pair of matching files.
- (iv) estimation of the joint categorical distribution of  $X^*, Y^*$  and  $Z^*$  by raking the full auxiliary categorical distribution (tables)  $\{W_{X^*Y^*Z^*}^C\}$  to available and adjusted marginal distributions (tables),  $\{W_{X^*Y^*}^A\}$ ,  $\{W_{X^*Z^*}^B\}$  and keep the new  $\{W_{Y^*Z^*}^C\}$ . We call the estimated categorical distribution (constraints) a look-up table.

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 the raking procedure does not converge. Two principal ways of initial marginal balancing were investigated: pooling the weights of the two files at the level of the  $X^*$  category, or alternatively, marginal adjustment by means of raking, and adjusting the record weights  $w_i^A$  and  $w_j^B$  in all corresponding categories of matching files. After adjusting the "unexpected" empty cells and keeping the "structural" empty cells, balance the margin of the partial auxiliary categorical table  $\{W_{Y^*Z^*}^C\}$  according to the margins  $\{W_{Y^*}^A\}$  and  $\{W_{Y^*}^B\}$ .

The last step in providing the look-up table  $\{W_{X^*Y^*Z^*}^{LC}\}$  is generated by the raking of the margins of the full auxiliary categorized table  $\{W_{X^*Y^*Z^*}^C\}$ , already corrected for its "unexpected" empty cells and keeping its "structural" empty cells, to the balanced margins  $\{W_{X^*Y^*}^A\}$  and  $\{W_{X^*Z^*}^B\}$  of the matching files and the balanced partial margin  $\{W_{Y^*Z^*}^C\}$  of the full auxiliary table itself.

## 3. THE MATCHING ALGORITHM

Without loss of generality, we assume that the matching files  $A$  and  $B$  have *large positive* record weights, the file size of  $A$  are several times larger than the file size of  $B$ , both files are categorized appropriately and with same marginal weights-sums  $W_{X^*}^A$  and  $W_{X^*}^B$  on category  $X^*$ , and a look-up table  $\{W_{X^*Y^*Z^*}^{LC}\}$  is available. The following steps apply within each  $X^*Y^*$  category of matching file  $A$ , and to all  $X^*Z^*$  category of matching file  $B$  for the same category of  $X^*$ . We also assume that for a given  $X^*Y^*$  category there are  $K(\geq 2)$   $Z^*$  categories and for a given  $X^*Z^*$  category there are  $N(\geq 2)$   $Y^*$  categories. For each of them compute the difference based on all categories  $X^*Y^*Z^*$ .

$$\Delta_{X^*Y^*Z^*_i} = W_{X^*Y^*Z^*_i}^M - W_{X^*Y^*Z^*_i}^{LC} \quad (1)$$

Where  $W_{X^*Y^*Z^*_i}^M$  is the weights-sum of matched file  $M$  based on the category  $X^*Y^*Z^*_i$ . Before matching starts, the matched file  $M$  is empty and the *initial* value of  $\Delta_{X^*Y^*Z^*_i}$  is  $-W_{X^*Y^*Z^*_i}^{LC}$ . The *goal* of the matching algorithm is to arrange a matched file  $M$  such that

$$\Delta_{X^*Y^*Z^*_i} \approx 0, \text{ over all categories } X^*Y^*Z^*_i. \quad (2)$$

In the following, a minimum distance, maximum weight and look-up table guided iteratively proportional matching algorithm is used, to build the desired matched file from the given matching files. The matching algorithm has four parts: forward imputation, backward imputation, shift adjustment and share adjustment.

*Forward imputation:*

- 1a. Check first if any record in matching file  $A$  is unmatched. If there is no unmatched record in  $A$

this procedure ends.

- 2a. Generate a possible imputation list  $pL = \langle i, w_i^A, d_i, ct_1, ct_2, \dots, ct_K \rangle$  for all unmatched records  $i$  in the category  $X^*Y^*$  of  $A$ . In  $pL$ ,  $w_i^A$  is the weight of record  $i$  in  $A$ ,  $d_i$  is the evaluated distance, and  $ct_k$  is the count of available records  $j$  within the neighbour  $d_i$  in the category  $X^*Z_k^*$  of  $B$ . The evaluated distance  $d_i$  for the first iteration of imputation is equal to the minimum distance  $d_m$  from the record  $i$  to all records  $j$  in  $B$ . For the  $n^{th}$  iteration of imputation,  $d_i$  is equal to the  $(n-1)^{th}$  extension distance of  $d_m$ .
- 3a. Order the  $Z_k^*$  categories,  $k=1, \dots, K$ , into by sorting the difference  $\Delta_{X^*Y^*Z^*}$  in ascending order.
- 4a. For the first  $k$  category  $Z_k^*$ , order all  $ct_k > 0$  partial lists  $pL_k$  in  $pL$ , into ascending order regarding the distance  $d_i$  and nest descending order regarding the weight  $w_i^A$ . Note that the first  $\Delta_{X^*Y^*Z_k^*}$  always smaller than zero, when there are unmatched records in the matching file  $A$ .
- 5a. From the first record  $i$  (of  $A$ ) list in  $pL_k$ , construct a matched pair (record)  $\langle X_i^A, Y_i^A, Z_j^B, w_i^A, w_j^B, d_{ij}, w_{ij} \rangle$ , where  $d_{ij}$  is the distance measure from the record  $i$  of  $A$  to the record  $j$  of  $B$ . If more then one record  $j$  of  $B$  is available, select one of them randomly with probability proportionally to the weight  $w_j^B$ . Assign the matched weight  $w_{ij}$  equal to the weight  $w_i^A$  of record  $i$  of  $A$ , and update the difference  $\Delta_{X^*Y^*Z_k^*}$  to  $\Delta_{X^*Y^*Z_k^*} + w_i^A$ .
- 6a. Repeat steps 1a-5a until no forward imputation under this defined neighbour is possible, then extend the neighbour and go back to step 1a.

*Backward imputation:*

In the backward imputation process, a record  $i$  in  $A$  may be used multiple times if two or more different records  $j$  in  $B$  are matched with the same  $i^{th}$  record in  $A$ , replicating it, say,  $J_i$  times. To maintain the total weights  $W^A$ ,  $W_{X^*}^A$ ,  $W_{Y^*}^A$  and  $W_{X^*Y^*}^A$  of file  $A$ , the matched weight  $w_{ij}$  will be recalculated proportionally to the corresponding  $B$  records weights  $\{w_j^B\}$ ,  $j=1, \dots, J_i$ . The new *split* (matched) weights  $\{w_{ij}\}$  is given as

$$w_{ij} = w_i^A \cdot w_j^B / \sum_{k=1}^{J_i} w_{ik}^B, \quad j=1, \dots, J_i. \quad (3)$$

Note that this step will keep the total weights and the conditional distribution  $Y|X$  of matching file  $A$ , and attempt to embed the conditional distribution  $Z|X$  of matching  $B$  into the matched file.

- 1b. Check first if any record  $j$  in matching file  $B$  is unmatched. If there is no unmatched record in  $B$  this procedure ends.
- 2b. Generate a possible imputation list  $qL = \langle j, w_j^B, d_j, ct_1, ct_2, \dots, ct_N \rangle$  for all unmatched record  $j$  in the category  $X^*Z^*$  of  $B$ . In  $qL$ ,  $w_j^B$  is the weight of record  $j$  in  $B$ ,  $d_j$  is the evaluated distance, and  $ct_k$  is the count of available records  $i$  within the neighbour  $d_j$  in the category  $X^*Y_k^*$  of  $A$ . The evaluated distance  $d_j$  for the first iteration of imputation is equal to the minimum distance  $d_m$  from the record  $j$  to all records  $i$  in  $A$ . For the  $n^{th}$  iteration of imputation, use  $d_j$  equal to the  $(n-1)^{th}$  extension distance of  $d_m$ .
- 3b. Order the  $Y_l^*$  categories,  $l=1, \dots, N$  by sorting the difference  $\Delta_{X^*Y^*Z^*}$  in ascending order.
- 4b. For the first  $l$  category  $Y_l^*$ , order all  $ct_l > 0$  partial lists  $qL_l$  in  $qL$ , into ascending order regarding the distance  $d_j$  and nest descending order regarding the weight  $w_j^B$ .
- 5b. Begin from the first record  $j$  (of  $B$ ) list in  $qL_l$ ,

and construct a *possible* matched pair (matched record)  $(X_j^B, Y_i^A, Z_j^B, w_i^A, w_j^B, d_{ij}, w_{ij})$ , where  $d_{ij}$  is the distance measure from the record  $j$  of  $B$  to the record  $i$  of  $A$ . If more than one record  $i$  of  $A$  is available, select one of them randomly with probability proportionally to the weight  $w_i^A$ .

- 6b. Next, use equation (3) to calculate the initial values of the matched weight  $w_{ij}$  and updated weights  $w_i^A$  of all already matched record  $i^*$  associated with the record  $i$  of  $A$ .
- 7b. If all new weights  $w_{ij}$  and  $w_i^A$  are less than the threshold  $\epsilon (>0)$ , and if more than one record  $i$  of  $A$  is available, select one with smaller weight and go back to step 6b. Otherwise, go back to 2b and from the list  $qL_i$ , chose the next record  $j$  (of  $B$ ) and repeat steps 5b-7b until no record  $j$  (of  $B$ ) in list  $qL_i$  is available.
- 8b. If all new weights  $w_{ij}$  and  $w_i^A$  are greater than the threshold  $\epsilon$ , the matched pair and the weight  $w_{ij}$  are kept, and all the corresponding matched pairs  $i^*$  are assigned the new updated weights  $w_i^A$ . Update the matched weights-sum by recalculating  $W_{X^*Y^*Z^*}^M = \sum_{ij \in X^*Y^*Z^*} w_{ij}$ . Update the difference  $\Delta_{X^*Y^*Z^*}$  according to equation (1).
- 9b. Repeat steps 1b-8b until no backward imputation under this defined neighbour is possible, then extend the neighbour and go back to step 1b.

Note that after the imputation processes, an *intermediate* matched file had built and a check is needed to make sure all records in both matching files are used before the adjustment processes. If not, a reduction on the threshold  $\epsilon$  may be needed or all processes stopped.

#### Shift adjustment:

For the 'shift adjustment' part, to make sure that rematching doesn't disturb fulfilment of the requirement for the use of all records from both files, a counter variables  $q$  need created previously. The  $q^A$  counts the number of

records  $i$  of file  $A$  that are recipients of the same record  $j$  of file  $B$ , are known for each record in the *intermediate* matched file after forward and backward imputations. For example, a *intermediate* matched record from pair  $i$  and  $j$  is  $(X_i^A, Y_j^B, Z_j^B, w_i^A, w_j^B, d_{ij}, w_{ij}, q_i^B=1, q_j^A=3)$ , means that there are two more matched records received from the same record  $j$  of file  $B$ . The  $q^B$  counts the number of records  $j$  of file  $B$  that are recipients from the same record  $i$  of file  $A$  is defined similarly. Clearly that from the backward imputation approach  $q^B=1$ . That is, no backward imputed records can shifted.

- 1c. Check first if any difference  $\Delta_{X^*Y^*Z^*}$  that is greater than the threshold  $\epsilon (>0)$  or smaller than  $-\epsilon$ . If all  $|\Delta_{X^*Y^*Z^*}| \leq \epsilon$  this procedure ends.
- 2c. Order the  $Z_k^*$  categories,  $k=1, \dots, K$ , by sorting the difference  $\Delta_{X^*Y^*Z^*}$  in descending order. Assuming that for the first  $K_1$  categories  $Z_1^*, Z_2^*, \dots, Z_{K_1}^*$  the difference  $\Delta_{X^*Y^*Z_k^*} \geq \epsilon$ , and that for the last  $K_2$  categories,  $Z_{K-K_2+1}^*, Z_{K-K_2+2}^*, \dots, Z_{K^*}^*$  the difference  $\Delta_{X^*Y^*Z_k^*} \leq -\epsilon$ .
- 3c. Then, search  $k^{th}$  categories  $Z_k^*$ ,  $k=1, \dots, K_1$ , for records  $ij \in X^*Y^*Z_k^*$  with the count  $q_j^A \geq 2$ , and the weights
- $$w_{ij} < \epsilon + \Delta_{X^*Y^*Z_k^*}, \quad 1 \leq k \leq K_1 \quad \text{and}$$
- $$w_{ij} < \epsilon - \Delta_{X^*Y^*Z_k^*}, \quad K-K_2+1 \leq k \leq K. \quad (4)$$

Note that the  $1^{st}$  difference has the largest positive value and the last  $K^{th}$  difference has the largest negative value. The records which satisfy conditions (4) and are in the category with the possible positive difference are candidates for further processing and are designated them as "movable". If there is no movable record, the 'shift' procedure ends.

- 4c. For all candidates  $ij$ , generate a possible move list  $mL = (i, j, w_{ij}, d_i, f_1, \dots, f_{K_1}, f_{K_2}, \dots, f_{K^*})$ . For the current matched record  $ij$ , where  $i$  and  $j$  are the record identifiers in matching files  $A$  and  $B$ , respectively. In  $mL$ ,  $w_{ij}$  is the current matched weight,  $d_i$  is the evaluated distance,  $f_1, \dots, f_{K_1}$

are the can-move-out flags, and  $f_{k_2}, \dots, f_{k_1}$  are the can-move-in flags. Note that there is only one can-move-out flag  $f_k = 1$ , and there is at least one can-move-in flag  $f_r = 1$ . The evaluated distance  $d_i$  for the first iteration of adjustment is equal to the distance  $d_j$  for the matched pair  $(i, j)$ . For the  $n^{\text{th}}$  iteration of adjustment,  $d_i$  is equal to the  $(n-1)^{\text{th}}$  extension distance of  $d_j$ .

5c. Next, order the list  $mL$ , by sorting all "movable records"  $ij$  into the ascending order by the distance  $d_i$  and the weights  $-w_{ij}$  (i.e., descending on  $w_{ij}$ ), where  $-w_{ij}$  is nested within  $d_i$ .

6c. If the first "movable record"  $ij$  is within current distance and with the possible maximum weight, then from its can-move-in flag  $f_{k_2}, \dots, f_{k_1}$ , choose the category  $Z_i^*$  that has minimum  $\Delta_{X^*Y^*Z^*}$  ( $\leq -\epsilon$ ) as the first move in category, and replaces its  $Z_k^*$  category with the move in category  $Z_i^*$ . If more than one record  $ij$  is first "moveable" (each record may belong to different can-move-out categories), then select one of them randomly. If more than one category is first move in, select one of them randomly.

7c. A record with the changed  $Z_k^*$  category, carries its weight  $w_{ij}$  over to the new category. However, it has to be rematched (by similar steps in forward imputation) with another record  $m$  to the original  $B$  file which belongs to this category  $Z_i^*$ , and assign the new rematched weight  $w_{im}$  equal to  $w_{ij}$ , construct a new rematched pair (record)  $\langle X_i^A, Y_i^A, Z_m^B, w_i^A, w_m^B, d_{im}, w_{im} \rangle$ , where  $d_{im}$  is the distance measure from the record  $i$  of  $A$  to the record  $m$  of  $B$ . If more than one record  $m$  of  $B$  available, select one of them randomly with probability proportionally to the weight  $w_m^B$ .

8c. Update the count variable  $q^A$  and the weight differences  $\Delta_{X^*Y^*Z^*}$ :

$$q_j^A = q_j^A - 1 \text{ and } q_m^A = q_m^A + 1, \quad (5a)$$

$$\Delta_{X^*Y^*Z_k^*} = \Delta_{X^*Y^*Z_i^*} - w_{ij} \text{ and}$$

$$\Delta_{X^*Y^*Z_i^*} = \Delta_{X^*Y^*Z_i^*} + w_{ij}. \quad (5b)$$

9c. Repeat steps 1c-8c until no shift adjustment under this defined neighbour is possible, then extend the neighbour and go back to step 1c.

*Share adjustment:*

1d. Steps 1 and 2 are the same as to 1c and 2c in shift adjustment.

2d. Look at  $Z_k^*$  categories,  $k=1, \dots, K_1$ , starting from the one with the largest difference. For each  $k$ , search for the  $t^{\text{th}}$  category  $t = K, K-1, \dots, K-K_2+1$  with the largest possible negative difference and check if there are records with weights

$$w_{ij} \geq \Delta_{X^*Y^*Z_k^*}, \quad 1 \leq k \leq K_1 \text{ and} \quad (6a)$$

$$w_{ij} \geq -\Delta_{X^*Y^*Z_t^*}, \text{ for some } t, K \geq t \geq K-K_2+1. \quad (6b)$$

These records are candidates for further processing and are termed "splittable". If there are no "splittable records", the 'share' procedure ends.

3d. For all candidates  $ij$ , generate a possible split list  $sL = \langle i, j, w_{ij}, d_i, f_1, \dots, f_{K_1}, f_{K_2}, \dots, f_{K_1} \rangle$ . For the current matched record  $ij$ , where  $i$  and  $j$  are the records identifiers in matching files  $A$  and  $B$ , respectively. In  $sL$ ,  $w_{ij}$  is the current matched weight,  $d_i$  is the evaluated distance,  $f_1, \dots, f_{K_1}$  are the can-split-out flags, and  $f_{K_2}, \dots, f_{K_1}$  are the can-split-in flags. Note that there is one can-split-out flag  $f_k = 1$ , and there is at least one can-split-in flag  $f_r = 1$ . The evaluated distance  $d_i$  for the first iteration of adjustment is equal to the distance  $d_j$  for the matched pair  $(i, j)$ . For the  $n^{\text{th}}$  iteration of adjustment,  $d_i$  is equal to the  $(n-1)^{\text{th}}$  extension distance of  $d_j$ .

4d. Next, order the list  $sL$ , by sorting all "splittable records"  $ij$  into ascending order by the distance  $d_i$  and the weights  $-w_{ij}$  (i.e., descending on  $w_{ij}$ ), where  $-w_{ij}$  is nested within  $d_i$ .

5d. The first "split record"  $ij$  is within current distance and with the possible maximum weight. Then from its can-split-in flag  $f_k, \dots, f_K$ , choice the category  $Z_i^*$  has minimum  $\Delta_{X^*Y^*Z_i^*}$  ( $\leq -\epsilon$ ) as the first split in category. If more then one record  $ij$  is first "splittable" (each record may belong to different can-split-out categories), then select one of them randomly. If more then one category is first split in, select one of them randomly.

6d. Duplicate the split record and assign to one of its replicates the category  $Z_i^*$  with the largest possible negative difference such that  $w_{ij} \geq -\Delta_{X^*Y^*Z_i^*}$ , and replace its  $Z_k^*$  category with the first split in category  $Z_i^*$ . The other replicate keeps its original category.

7d. A replicated record with the newly assigned  $Z_i^*$  category has to be rematched (by similar steps in forward imputation) with another record  $m$  from the original  $B$  file which belongs to this category  $Z_i^*$ . If more then one record  $m$  of  $B$  available, select one of them randomly with probability proportionally to the weight  $w_m^B$ .

8d. The weight of this record  $ij$  is split between the old category  $Z_k^*$  and the new  $Z_i^*$ ,  $K \geq t \geq K - K_2 + 1$  in the following way:

Let  $\Delta_0 = \min\{\Delta_{X^*Y^*Z_k^*}, -\Delta_{X^*Y^*Z_i^*}\}$ . If  $w_{ij} \geq \Delta_0 + \epsilon$ , then  $\Delta_0$  is the weight of the replicated record  $im$  with a new category  $Z_i^*$ , and the remainder,  $w_{ij} - \Delta_0$  is the new weight of the processed (original) record  $ij$ . Otherwise, use  $\Delta_0 - \epsilon/2$  and  $w_{ij} - \Delta_0 + \epsilon/2$ , respectively, where  $\epsilon > 0$ ,  $\Delta_0 > \epsilon$  and  $w_{ij} \geq \Delta_0$  as mentioned earlier. Note that, in this way all 'split' weights *greater than*  $\epsilon/2$ .

9d. Construct a new replicated matched pair (record)  $\langle X_i^A, Y_i^A, Z_m^B, w_i^A, w_m^B, d_{im}, w_{im} \rangle$  or  $\langle X_m^B, Y_i^A, Z_m^B, w_i^A, w_m^B, d_{im}, w_{im} \rangle$ , where  $d_{im}$  is the distance measure from the record  $i$  of  $A$  to the record  $m$  of  $B$  and  $w_{im} = \Delta_0$  or  $\Delta_0 - \epsilon/2$ .

10d. Update the weight differences  $\Delta_{X^*Y^*Z_i^*}$  and  $\Delta_{X^*Y^*Z_k^*}$  as equation (5b) in step 8c.

11d. Repeat steps 1d-10d until no share adjustment under this defined neighbour is possible, then extend the neighbour and go back to step 1d.

After forward imputation, backward imputation, shift adjustment and share adjustment, the result matched file building from the matching files  $A$  and  $B$  is  $M = \langle X_i^A \text{ or } X_j^B, Y_i^A, Z_j^B, w_{ij} \rangle$ .

A simple and useful distance function is a Pitman distance  $d_{ij}^p$  based on common variables  $X$ ,

$$d_{ij}^p = \sqrt{[X_i^A - X_j^B] V_X^{-1} [X_i^A - X_j^B]^T}, \quad (7)$$

where  $i$  and  $j$  are the record identifiers in the matching files  $A$  and  $B$ , respectively, and  $V_X$  is the identical matrix  $I$  or the variance-covariance matrix of the variables  $X$ . The distance  $d_i^p$  or  $d_j^p$  for the first iteration of imputation parts is equal to the minimum distance  $d_m^p$ , either from the record  $i$  in  $A$  to all records  $j$  in  $B$ , or from the record  $j$  in  $B$  to all records  $i$  in  $A$ . The distance  $d_i^p$  for the first iteration of adjustment parts is equal to the distance  $d_{ij}^p$  of the matched pair  $(i, j)$ . For the  $n^{\text{th}}$  iteration of imputation,  $d_i^p = d_m^p + (n-1)\delta$  or  $d_j^p = d_m^p + (n-1)\delta$  is used, and for the  $n^{\text{th}}$  iteration of adjustment,  $d_i^p = d_{ij}^p + (n-1)\delta$  is used, where  $\delta$  is an added tolerance for the extension of distance neighbour.

#### 4. CONCLUDING REMARKS AND FUTURE WORKS

Alternatively, the previous shift-and-share rematching algorithm (Liu and Kovačević 1996, 1998a,b) can be used to substitute the two adjustment parts of the present matching algorithm. Similarly, the two adjustment parts of the present matching algorithm can be used in rematching, when (full) auxiliary weights-sum table of all variables is available. A optimal matched file can be built by additional categorical constraints and weights of all matched record beyond a given threshold, especially

when they are implemented via the nearest-and-heaviest-weight first matching algorithm. When (partial) auxiliary weights-sum table of non-common variables are available. The quality of any matched file can be improved by additional categorical constraints, especially when they are implemented via the shift-and-share rematching algorithms in previous or via shift-and-share adjustment parts of current. For detail compare this two algorithms, a completed empirical study will conducted in near future.

### ACKNOWLEDGMENTS

This work was sponsored by divisional research fund of Household Survey Methods Division of Statistics Canada. The author wishes to thank Craig Seko, Harold Mantel, Shiyong Wu, Jack Gambino, Mingyu Yu, and René Boyer for their useful discussions and comments.

### 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.
- Goel, P.K., and Ramalingam, T. (1989). *The Matching Methodology: Some Statistical Properties* Lecture Notes in Statistics, Springer-Verlag, New York
- Liu, T.P., and Kovačević, M.S. (1996). Categorically constrained matching. *Proceeding of the Survey Methods Section, Statistical Society of Canada*, 123-133.
- Liu, T.P., and Kovačević, M.S. (1997). An empirical study on categorically constrained matching. *Proceeding of the Survey Methods Section, Statistical Society of Canada*, 167-178.
- Liu, T.P., and Kovačević, M.S. (1998a). Categorical matching and constrained rematching of survey datafiles. Working Paper HSMD-98-008E, Statistics Canada, Ottawa.
- Liu, T.P., and Kovačević, M.S. (1998b). Categorically constrained rematching of survey datafiles. To be submitted to the Journal of American Statistical Association.
- 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 Commerce*, 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.