# STATISTICAL MATCHING OF SURVEY DATAFILES: A SIMULATION STUDY

Milorad S. Kovačević, Tzen-Ping Liu
Milorad S. Kovačević, Statistics Canada, Ottawa K1A 0T6

KEY WORDS: auxiliary datafile, conditional independence, categorical constrains, survey weight

## 1. INTRODUCTION

Statistical matching is frequently used for the production of a comprehensive datafile from data from multiple sources. The idea is to identify and link records from different files that correspond to similar individuals.

This problem coincides with the estimation of a density function based on an incomplete set of marginals. The conditional distribution of variables that appear in two disjunctive files given the set of common variables is identifiable from the corresponding marginals only under the assumption of their conditional independence (CI) (Sims, 1972).

In order to overcome the CI assumption, Paass (1986) was suggesting the use of additional information in form of an auxiliary micro datafile. Rubin (1986) proposed a regression method for statistical matching based on either macro or micro information on relationship of 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 categorical distribution and proposed the loglinear modification of these methods based on the loglinear method of imputation as introduced by Singh (1988).

In most of applications, micro files contain survey data with the survey weight attached to each record. The problem is how to weigh records in the composite file when matching records originally had different weights.

Also, in practice, there are some additional requirements imposed on statistical matching that act as constraints. For example, the following four requirements were placed on the statistical matching used to creating the Social Policy Simulation Data Base (SPSD) at Statistics Canada:

(i) to maintain the conditional distribution $F(Z|X)$ as it is in the donor file $B$, (or with the smallest possible distortion);

(ii) to use all records from both files;

(iii) to keep the size of the matched file under control, to allow the minimal possible inflation to the host file;

(iv) to make the weights of records in matched file to be integers.

(i) The first task faces difficulties when the weights in two files are different and when distortion of the distribution functions in the matched file is high likely. There are three general types of distortion: distortion in the marginal distributions of the $Z$ variables; distortion in the joint distribution of the $(X,Z)$ and distortion in the joint distribution of $(X,Y,Z)$. Each of these distortions is of obvious importance in the context of SPSD. The first two affect the targeted conditional distribution $F(Z|X)$ directly. On the other hand, the file $B$ is a sample taken from the population, so we expect that the distortion of $F(Z|X)$ after matching, is within the sampling variation.

(ii) The second requirement comes from the actual matching for the SPSD and the importance of information from the donor file. This requirement is not an usual one in statistical matching where the primary objective is to complete the host file $A$.

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

(iv) A composite file is considered as a sample from the real population and its weights are supposed to show how many units from the population a particular synthetic record represents.

Our empirical study assumes these constraints and the objectives are set accordingly. The primary objective of this study is to modify, adjust and develop the methodology for statistical matching of records from files obtained in different sample surveys under constraints (i)-(iv).

Also, the simulation is intended to examine whether the earlier findings with synthetic data, Paass (1986), Singh et al, (1993), hold under conditions similar to those in a real matching settings, that is to examine if the CI assumption can be successfully overcome by the use of appropriate auxiliary information.

Section 2 reviews matching methods that were considered in this project. A complete description of the simulation study is given in Section 3. The evaluation of statistical matching is addressed in Section 4. Some results and their interpretation along with specific remarks and conclusions are given.

## 2. MATCHING METHODS

Matching of survey datafiles coherent with requirements (i) - (iv), given in Section 1, evolves in a multistage process consisting of the imputation, weight assignment, file reduction, weight adjustment and weight integerization.

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 a statistical matching procedure is creation of a new file which contains $X$ and $Y$ values from $A$-records and $Z$ values from $B$-records. A $Z$ value is thought of as an imputed value.

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

This effort frequently results by the increased size of the host file. Any reduction of the file size necessarily leads to the redistribution of total weight and results in weight adjustment. Finally, we want to integerize weights in the matched file in order to obtain a file with the meaningful survey-type weights.

In this section we discuss several different methods for statistical matching and describe their algorithms. Methods are classified into two big groups depending on whether they rely on CI assumption or utilize auxiliary information. Within these groups we have methods with and without additional log-linear constraints imposed on $Z$-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 an imputed value comes as a live value from a donor record which satisfies a certain criterion, for instance the minimum distance or belonging to the same class.

### 2.1 Matching Methods Based on CI Assumption

In the absence of an auxiliary datafile, matching is based on comparison of values of the common variables $X$ assuming conditional independence of $Y$ and $Z$ variables. It is assumed that records in both files, are preliminarily classified into $K$ matching classes (pockets in the file linkage terminology or imputation classes in the practice of statistical imputation), according to common $X$ variables which are either of the categorical type or categorically transformed. Within each class $X^*$, a distance function between recipient and donor records may take into account the $X$ variables and, in addition, the record (survey) weight, $W$. For the sake of simplicity we will omit the class notation emphasizing that everything we do is at the matching class level.

If $X$ variables are only considered, the matching can be done using the 'fixed distance tolerance', or as the 'nearest available' matching. In the first case an upper boundary for distance is given and the closest record within defined boundaries is a matching record. However, it may happen that there are no records within the boundaries. Since we have to use all records from both files the nearest available matching is more appropriate for our study.

If record weights play a role we suggest the imputation 'on rank' which means the imputation of a $Z$ value to the point of a given relative cumulative weight value (RCW). The resulting matching method is denoted as the weight-split method indicating a possibility of splitting weights.

**$X$-distance Method** In general, for each $A$-record, a $B$-record (or a set of $B$-records) is found such that their $X$-distance is minimum. Then, a $Z$ value from the minimum distanced $B$-record (from the 'nearest neighbour') is imputed into the corresponding $A$-record. If there are more than one 'nearest neighbour' we select one at random.

It might happen that some of $B$-records are not used in the imputation phase which is in contradiction with the requirement (ii). To overcome this, for each leftover $B$-record we find the nearest $A$-record. This leads to the multiple imputation: $Z$ values from two or more different $B$-records may be imputed to the same $A$-record and thus, the weights need some adjustment. When $Z$ values from $J_i$ different $B$-records are imputed to the same $i$-$th$ $A$-record replicating it $J_i$ times, the original weight, $W_i^A$, has to be adjusted proportionally to the corresponding $B$-records weights $\{W_{i_j}^B\}$, $j=1,...,J_i$, giving the final weights $\{W_{ij}\}$:

$$W_{ij} = W_i^A \cdot W_{i_j}^B \Big/ \sum_{k=1}^{J_i} W_{i_k}^B, \quad j=1,...,J_i.$$

Note that we didn't need to reduce the composite file since it takes the smallest size possible for the given files $A$ and $B$.

This method preserves distributions of $X$ and $Y$ variables from file $A$, but the marginal distribution of $Z$ and the conditional distribution of $Z$ given $X$ from $B$ are distorted. In order to maintain the consistency of the marginal distributions the following conditions must be met:

$$\sum_{i=1}^{n^A} W_{ij} = W_j^B, \quad \sum_{j=1}^{n^B} W_{ij} = W_i^A, \quad \sum_{i=1}^{n^A} w_i^A = \sum_{j=1}^{n^B} w_j^B \qquad (2.1)$$

The optimal weights can be obtained as the solution to a 'transportation' problem (Goel, P.K and Ramalingam, T, 1989) where the objective function is

the total weighted distance $f = \sum_{i,j} w_{ij} d_{ij}$ and has to be minimized under constraints (2.1). The implementation of this approach may be difficult when datafiles are large (as they are in usual matching set-up).

Conditions above allow the multivariate distribution of $Z$ variables to be precisely replicated in the composite file as observed in file $B$.

**Weight-split method** A method that uses the information contained in both, record weights and $X$ values, is the weight-split matching method. The name comes from the fact that this method usually replicates some of records from $A$ and consequently splits their weights. If $X$ values are not sorted we deal with the random weight-split method, otherwise we have the $X$-rank weight-split method. In both cases we apply modified imputation 'on rank' where a $Z$ value from $B$ is imputed to the $A$ record with the nearest value of the RCW.

Assume that both files are sorted with the respect to $X$ variable. The RCW of a record $u_i$ is $F_i = F(u_i) = \sum_{j=1}^{i} w_j$, $i = 1,...,n$ and is attached to each record in the class. Then, the records from both files in the same imputation class are ordered jointly according to the values of the RCW regardless of the file. The resulting sequence of cumulatives, $\{F_i^s\}$, $s \in \{A,B\}$, is the RCW sequence for the matched file.

The modified imputation 'on rank' is as follows: First we impute the $Z$ value from the k-th $B$ record to all $A$ records for which $F_{k-1}^B < F_i^A \le F_k^B$, ('downward' step). Then, from a given $B$ record, if there is no $A$ record with the same $F$ value, impute $Z$ to the first $A$ record with $F^A > F^B$, ('upward' step).

The total number of records in the matched file is $n = n_A + n_B - T$, where $T$ denotes the number of records with $F_i^A = F_j^B$. The RCW assigned to a synthetic record is determined as $F_{ik} = min\{F_i^A, F_k^B\}$.

It can be easily shown that the marginal distributions of $X, Y$ and $Z$ are preserved in the composite file.

Although, the imputation 'on rank' preserves the marginal distributions it has some practical drawbacks. It may happen that the resulting relative weight is too small and gives $W_i < 1$. In such a case we discard this "light" record. Then, the size of the composite file is usually very large. To reduce it we apply the sequential file reduction procedure in which we reduce the size but still use all of records from both files.

To determine the final weights of records in the matched file, we have to adjust the weights obtained by the imputation 'on rank'. We use the same procedure as in case of the distance matching.

### 2.2 Matching Methods When Auxiliary Datafile is Available

Here we assume availability of an auxiliary datafile, say $C$, which contains records with $X,Y,Z$ values or just $Y,Z$, along with their survey weights. Again, we see a matching method as a sequence of procedures. We omit steps which are identical to those in section 2.1.

**$(X,Y,Z)$-distance Method** The first step is to identify the nearest neighbours in files $A$ and $C$ using a distance function of the common variables $X$ and $Y$ or just $Y$, depending on $C$. Then, $Z$ values from $C$ are imputed to $A$ and the intermediate composite file is obtain. In this step we kept the weights and the size of the $A$ file. Next step is matching of the intermediate file and the donor file $B$. The common variables are $X$ and $Z$. Further on, we essentially repeat the $X$-distance matching procedure.

**Weight-split ($X,Y,Z$-rank) method** Assuming that the auxiliary file $C$ contains, beside $X,Y,Z$ or $Y,Z$, survey weights, we can perform the weight-split matching. In this case, we have an additional imputation, the intermediate imputation, from $C$ to $A$. Intermediate imputation is done at the points of the nearest RCW values. Prior to imputation, records in both files were ordered by common variables. It means, that the first sorting is done by $X$ variable and then if there are two or more records with the same value of $X$ we sort them regarding $Y$. Then the RCW is computed, by adding weights of successive records. The RCW is $F_{XY}(u_{(i)})$, where subscript denotes the order of sorting.

In that way we obtain the intermediate file based on the information about $X,Y$ distribution and the distribution of $Y,Z$ or $X,Y,Z$ variables obtained from the auxiliary file. The size and the weights in the intermediate file are the same as in $A$ file. Weights from the auxiliary file were used just in the intermediate imputation and not for an adjustment of the resulting weights.

In the next step we perform weight-split matching of the intermediate and $B$ file. The variables in common are $X, Z$. We first order files according to $Z$ variable and then by $X$. The RCW's are $F_{ZX}(u_i)$. Further on, the imputation is done and weights are obtained in the way explained in 2.1. Also, the file reduction and the final weights adjustment is done in the same way as given in 2.1.

### 2.3 Categorically Constrained Matching

The main idea is to: (i) transform the variables involved in matching, $(X,Y,Z)$, into the categorical variables $(X^*, Y^*, Z^*)$ using some of criteria for optimal partition (see Singh *et al.*, 1988), and then to (ii) esti-

mate the distribution of $(X^*,Y^*,Z^*)$. (iii) Once the distribution of $(X^*,Y^*,Z^*)$ is estimated, a suitable scheme for determining $Z$ values within $B\langle X^*,Z^*\rangle$ and their imputation into $A\langle X^*,Y^*\rangle$ is needed.

The purpose of the method is to preserve categorical associations of the data under a suitable partition of the $(X,Y,Z)$ variables.

Using file $B$ one can estimate the conditional categorical distribution, $F(Z^*|X^*)$. The assumption that $F(Z^*|X^*) = F(Z^*|X^*,Y^*)$ is a categorical version of the conditional independence assumption. This expression provides a starting association structure that assumes proportionality across the $Z^*$ categories:

$$W_{X^*,Y^*,Z^*} = W^A_{X^*,Y^*}.(W^B_{X^*,Z^*}/W^B_{X^*}). \qquad (2.2)$$

The resulting matched file will maintain the marginal and conditional categorical distributions from the original files. If files $A$ and $B$ are without weights, (2.2) accommodates counts instead as proposed by Singh *et al.* (1988).

A potential matching pair is in the same $X^*$ category in $A$ and $B$. The imputation is done for each $(X^*,Y^*)$ cross-category of $A$, independently, using HOD methods explained in 2.1. The file reduction and the weight adjustment are made according to the method used for imputation. In that way we obtain an intermediate matched file $A'\langle X,Y,Z,W'\rangle$.

Let $T'_{(X^*,Y^*,Z^*)}$ be a total weight of a cross-category $(X^*,Y^*,Z^*)$ of the file $A'$. Suppose that $W_{X^*,Y^*,Z^*}$ is the weight obtained by (2.2) adjusting procedure. If the difference $W_{X^*,Y^*,Z^*} - T'_{(X^*,Y^*,Z^*)}$ is less than 1 for any $(X^*,Y^*)$ cross-category, the intermediate matched file $A'\langle X,Y,Z,W'\rangle$ is considered as the final matched file. Otherwise, we perform the minimum move and split procedure in which we move records between $(X^*,Y^*,Z^*_1)$ and $(X^*,Y^*,Z^*_2)$, or we duplicate them, split their weights and move slices until the difference becomes negligible. Here we assume that there are two $Z^*$ categories: $(X^*,Y^*,Z^*_1)$, $(X^*,Y^*,Z^*_2)$. The complete algorithm is given in Kovacevic and Liu (1994).

Another approach to the categorically constrained matching is to modify weights in the matched file already obtained by some of matching procedures described earlier. This file, say $A'\langle X^*,Y^*,Z^*,W'\rangle$, is firstly categorized on the same way as the original files $A$ and $B$, and then adjusted to marginal categorical distributions of the original files, $A\langle X^*,Y^*,W^A\rangle$, $B\langle X^*,Z^*,W^B\rangle$ through the two-step iterative raking procedure. The advantage of this method is its simplicity. The disadvantage, however, is the distortion of the original $X,Y$ distribution. Here we used the first approach.

Categorical constraints may be given as a special auxiliary information about the categorical distribution, or can be extracted from the auxiliary datafile $C$, in the form of categorical distribution $(X^*,Y^*,Z^*)$.

Assume that a partition of the variables of interest which is close to the optimal (Singh *et al.*, 1988), is known. Again, the first step in implementing the log-linear matching is to obtain the estimate of the joint categorical distribution for $X^*,Y^*,Z^*$, by raking the auxiliary file $C$ to meet margins of $A$ and $B$ file.

The imputation of $Z$ values from $B$-records (in the same $X^*$ category), file reduction and the weight assignment are performed using some of the methods explained previously.

## 3. SIMULATION STUDY

The simulation study is based on data from the Public Use Micro File (PUMF) from 1986 Census 2B on Household/Housing for the province Québec (Canada). The Census 2B was conducted on a 20% sample of the Canadian population.

In designating variables from the Census 2B file as $X,Y,Z$ variables, the objective was to define three sets of variables that are similar to the variables encountered in actual matching for the SPSD. These variables may be highly skewed, long tailed mixtures with discrete components.

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 were considered as matching variables, $X$. They were used as categorical variables for grouping the records into the number of matching classes. The total household income was also used as a continuous type common variable.

The total household investment income and total household government transfer payments are $Y$ variables in the simulation. The monthly gross rent and, alternatively, the owner's major payments - monthly were chosen to be imputing variables, $Z$.

Records in the initial data set were grouped into nine groups according to the urbanization (a combination of the Rural/Urban Code with the Census Metropolitan Area Code (CMA) and the residential tenure with the presence of mortgage.

Four groups out of these nine were chosen according to the significance of the Pearson partial correlations, $\rho$, between $Y$ and $Z$ variables when controlled for $X$. The approach via partial correlations is good for the particular case of the multivariate normal distribution of the $X,Y,Z$ variables where the assumption on the 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 are as it was previously mentioned, skewed, truncated with possible nonlinear relationship. Because of that, the Kendall's $\tau$ was calculated as well. The following groups emerged as appropriate:

| | |
|---|---|
| MQR: | Montreal and Québec City, Rented; |
| MQM: | Montreal and Québec City, Owned with Mortgage; |
| OTH: | Other CMAs, CAs & Urban Areas, Owned without Mortgage |
| RUR: | Rural, Owned with Mortgage |

The absolute magnitudes of partial correlations were small in all of groups considered and $\rho$'s and $\tau$'s were very close.. However, in groups MQR and RUR correlation between one (of two) $Y$ and $Z$ was significant at 0.1% level. In MQM group, none of correlations were significant. Finally, both $Y$ variables were significantly correlated with $Z$ in OTH. A statistically significant partial correlation is considered as an evidence of the failure of the assumption on their conditional independence.

Further, records were classified into matching classes to $X$ categorical variables. There were about 60 possible classes per group. When some of them were empty or contained less than six records, classes were redefined.

Datafiles ($A$ and $B$) were created as random samples from each of four groups. As a file $C$ (auxiliary datafile) we used the complete population.

First, a larger sample $A$ was drawn as a simple random sample with the size of about one third of the initial population. Then, a sample $B$ was selected from the remaining two thirds as one fifteenth of the size of the initial population. It was obtained in such way that all matching classes that were represented in the sample $A$ were represented in the sample $B$, as well. Resulting sizes of files are given below.

| Group | File A | File B |
|---|---|---|
| MQR | 2448 | 490 |
| MQM | 1302 | 260 |
| OTH | 978 | 196 |
| RUR | 348 | 70 |

The sampling procedure was independently repeated in each simulation.

We simulated various matching methods which are variants of the methods described in section 2, obtained under different combinations of available files, order restrictions, distance functions and log-linear constrains.

Two general matching frameworks were investigated: matching without auxiliary file and with available auxiliary file. In the later case, we studied two different types of the auxiliary file content: a full information on all of three groups of variables $(X,Y,Z)$, and an incomplete information, only $(Y,Z)$.

Then, we considered matching with and without categorical constraints. We didn't make any additional categorization of $X$ variables besides one done for the purpose of imputation. The $Y$ as well as $Z$ variable were categorized into two categories each. Finally, we used different distance measures: Euclidian and absolute distance, and performed the random imputation as well.

Some combinations didn't make sense and they were excluded, some were redundant and we used them just in one form.

We found that 18 combinations could be considered as well defined matching procedures. They are listed below.

Methods without use of auxiliary file:

| | |
|---|---|
| M1 | Weight-split on random order |
| M2 | Categorically adjusted M1 |
| M3 | Weight-split ($X$-rank) |
| M4 | Categorically adjusted M3 |
| M5 | Minimum Distance |
| M6 | Categorically adjusted M5 |

Methods with use of partial auxiliary file:

| | |
|---|---|
| M11 | Weight-split ($Y1,Y2$-rank) |
| M12 | Categorically adjusted M11 |
| M13 | Euclidian Minimum Distance |
| M14 | Categorically adjusted M13 |
| M15 | Absolute Minimum Distance |
| M16 | Categorically adjusted M15 |

Methods with use of the full auxiliary file:

| | |
|---|---|
| M21 | Weight-split ($X,Y1,Y2$-rank) |
| M22 | Categorically adjusted M21 |
| M23 | Euclidian Minimum Distance |
| M24 | Categorically adjusted M23 |
| M25 | Absolute Minimum Distance |
| M26 | Categorically adjusted M25 |

## 4. EVALUATION, RESULTS AND COMMENTS

The performance of matching methods was evaluated by comparison of matched $z_{m,i}$ and suppressed true $z_{s,i}$ values in the matched file considering their weights as well. The average, minimum and maximum values, the Monte-Carlo standard error and coefficient of variation were computed over 1,000 simulations for each measure and for each data set.

The first group of measures evaluates the marginal distribution of $Z$ variable and the distribution conditional to given $X$. We used the weighted mean of absolute differences between the matched and suppressed individual $Z$ values. Also, we considered the

weighted proportion of records with $z_{m,i}$ value within $\delta$-neighbourhood of the true value and the weighted proportion of records with the cumulative distribution function values (CDF) within $\varepsilon$-neighbourhood of the true CDF. Several different values for $\varepsilon$ and $\delta$ were considered.

When the quality of matching is evaluated by the weighted mean absolute difference between matched and suppressed values or using the $\delta$-concordance ratio, methods based on the minimum distance imputation and full auxiliary information are far the best. This can be explained by the nature of imputation methods constructed to meet the minimum distance requirement. There was no significant difference between Euclidian and absolute distance performance.

The CDFs for matched and suppressed $Z$ values are closest when the weight split methods are used. However, the use of the auxiliary information doesn't dramatically improve the quality of matching as it was in case of the absolute difference.

Two measures based on categorical comparisons were considered as well. One is the $\kappa$-coefficient of agreement of two independent classifications of the matched records according to the suppressed and matched $Z$ values, another is the weighted Pearson chi-square statistic transformed to lie in [0,1).

When the categorical agreement of the matched and suppressed values is considered, the methods based on the full auxiliary information and the minimum distance imputation showed better performance than others.

We computed the partial correlations between $Y$ and $Z_s$-$Z_m$ 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 the better preservation of the original relationship. We found that methods with the use of auxiliary information performed better. However, the use of partial auxiliary information did slightly better job than the use of the full auxiliary. Also, methods based on the minimum distance performed better than the weight split-methods.

Matching methods that rely on CI assumption produced better results under categorical constraints. For methods that use a partial auxiliary file, categorical constraints did slight improvement, too. However, when we compare performance of the methods that use full auxiliary datafile it seems that imposing of categorical constraints is unnecessary and cumbersome procedure that erodes the high quality of matching that has already been achieved.

The size of the host file was inflated by all methods similarly, with the slight advantage of the methods without use of the auxiliary information. Also, categorically constrained matching resulted in larger matched file. In general, methods based on the imputation on 'rank' produce larger matched files.

To summarize, our simulation study based on real survey datafiles confirms that the use of an auxiliary datafile improves the quality of matching and serves as a protection against the possible violation of the CI assumption. However, the impact of the quality of auxiliary file is yet to be discussed. Also, categorically constrained matching increases quality of matched file. The survey weights can be optimally handled through some of presented algorithms, although their adjustment depends on the number and type of constraints that matching process is submitted to.

**Acknowledgment** The authors wish to thank Harold J. Mantel and Geoff Rowe for their useful discussions and comments.

**References:**

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

Kovacevic, M.S. and Liu, T.P. (1994) Statistical Matching of Survey Data Files: A Simulation Study. Unpublished manuscript. Statistics Canada.

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.

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

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.