WORKING PAPER NO. SSMD-88-030 E

METHODOLOGY BRANCH

STATISTICAL MATCHING USING LOG LINEAR IMPUTATION

SSMD-88-030 E

A.C. Singh, J.B. Armstrong and G.E. Lemaître

Social Survey Methods Division, Statistics Canada, Ottawa

ABSTRACT

In this paper the method of log-linear imputation or LLI (Singh 1988) for the problem of statistical matching is briefly described by means of a simple example taken from Rodgers (1984). A simulation experiment for evaluating LLI in comparison to some other methods of statistical matching is also described. It involves use of synthetic data generated from multivariate normal distributions. Some preliminary empirical results indicate the potential advantage of LLI over hot deck imputation methods used for Statistical matching.

KEY WORDS : Conditional independence; Categorical transformation; Missing at random.

RÉSUMÉ

Dans cet article, on presénte la méthode d'imputation par modèle log-linéaire ou méthode ILL (Singh 1988) appliquée au problème de l'appariement statistique en se servant d'un exemple de Rodgers (1984). On décrit egalement une étude de Monte Carlo, présentement en cours, qui a pour but de comparer la méthode ILL à d'autres méthodes d'appariement statistique. Dans cette étude, on utilise des données artificelles qui proviennent d'une loi multinormale. Quelques résultats empiriques préliminaires indiquent les avantages potentiels de la méthode ILL par rapport à la méthode "hot deck" dans le contexte de l'appariement statistique.

MOTS CLÉS : Indépendance conditionnelle; catégorisation; non-réponse aléatoire

1. THE PROBLEM OF STATISTICAL MATCHING

The problem of statistical matching arises when one is interested in merging two (or more) data files in the absence of unique identifying information at the micro level. This contrasts with the problem of exact matching for file merging via techniques such as record linkage because the set of units in the two files for statistical matching may be completely disjoint or have only a small unknown overlap. Some useful references for statistical matching are Kadane (1978), Sims (1978), U.S. Department of Commerce (1980), Rodgers (1984) and Rubin (1986), among others. The two files may have been collected in two separate surveys using different samples or one file could correspond to an administrative data source. For example, at Statistics Canada, a microeconomic database termed SPSD or Social Policy Simulation Database (Wolfson et al. 1987) was constructed for use in economic policy analysis. It was built in part by statistical matching of information from Revenue Canada with the Survey of Consumer Finance. The present investigation was motivated in part by considerations of evaluating SPSD.

In statistical matching, the problem can be formulated as follows. Consider two microdata files denoted by A and B. The file A contains information on the vectors of variables X and Y, the file B contains information on vectors X and Z and for the purposes of analysis at the microlevel we are interested in constructing file C that contains for each micro-unit on the original file A, information about X,Y and Z.

The process of statistical matching for file merging can be viewed (see Rubin 1986) as a process of imputing Z values for the candidate records (X,Y) in file A using (X,Z) records from file B as donors in a single super file obtained by combining files A and B. As usual, we assume that the Z values are missing at random in the combined file. However, it differs from the usual imputation procedures because there are no donor records containing the complete set of values (X,Y,Z). Therefore some additional assumptions/techniques are required to estimate the conditional distribution $f(Z|X,Y)$ from donor records which in turn could be used for drawing imputed values. Two situations arise.

Case I Y Ignorable This corresponds to the assumption of conditional independence of Y and Z given X i.e. $f(Z|X,Y) = f(Z|X)$. Thus the information in Y can be ignored and the problem of completing records with missing Z values in file A reduces to the usual imputation problem in a single file. Commonly used methods of imputation include class

mean, hot deck (random, distance and sequential) and regression, see e.g. Little and Rubin (1987). A general approach to statistical matching that has often been used in practice is equivalent to distance hot deck imputation (HDI - distance) in the combined file when Y information is ignored. More specifically, cohorts (or imputation classes) are first formed using X variables i.e. divide files A and B into subfiles such that within each subfile all records have the same value for all cohort variables. Next to complete a file A record, one looks in the subfile from file B corresponding to the same cohort and minimizes the value of a distance function defined using X in order to choose a Z value.

The above HDI approach to statistical matching could cause distortion in the marginal distribution of Z in the matched file. This is a minor problem and can be resolved using constrained matching techniques. There is, however, a more serious problem resulting from the assumption of conditional independence. The relationship between Y and imputed Z values in file A may differ substantially from the true relationship between Y and Z; see Rubin (1986) for illustrations of the sensitivity of statistical matching results to departures from the conditional independence assumption. This is a major problem since matching was conducted in first place to analyse the Y,Z relationship. This leads to the following more realistic situation

Case II Y Non-ignorable In this case, the Y information is not ignored in the process of statistical matching. The method of log-linear imputation for multiple files (LLI-M as defined in Singh 1988) can be used for this purpose. The use of the term log-linear reflects the use of log-linear modelling for estimating the conditional distribution for imputation in the categorical framework. The basic idea of LLI-M approach is to transform the statistical matching problem to one involving categorical variables $X_\star, Y_\star, Z_\star$ so that the unavoidable assumption of conditional independence holds approximately in the transformed framework. This of course cannot be checked directly because there is no information on the joint distribution of $X_\star, Y_\star$ and $Z_\star$. However, an important advantage of the categorical approach is that a suitable criterion can be constructed to control possible violation of the conditional independence assumption. This criterion is used to choose categories for $(X_\star, Y_\star, Z_\star)$ and thus Y information is indeed used in the process. After a suitable partition of (X,Y,Z) space into categories for $(X_\star, Y_\star, Z_\star)$ is selected, the LLI method is used to first impute Z up to a $Z_\star$ category using the conditional categorical distribution $f(Z_\star | X_\star)$ within the imputation class $(X_\star, Y_\star)$ and then a value of Z within the $Z_\star$ category is chosen appropriately.

Section 3 contains a brief description of log-linear imputation for the usual imputation problem in a single file (i.e. Case I) by means of a simple example of Rodgers (1984) given in Section 2. This is then used to motivate LLI for statistical matching (i.e. Case II) in Section 4 for the same example. The log-linear imputation methods for the two cases (corresponding to single and multiple files) are denoted respectively by LLI-S and LLI-M. We omit theoretical details which can be found in Singh (1988). In Section 5, a simulation method for evaluating LLI is described and some results from a preliminary phase are reported. Some remarks and directions of further research are outlined in Section 6.

2. RODGERS' EXAMPLE

The following miniature example of Rodgers (1984) on statistical matching will be convenient to describe LLI-S and LLI-M in later sections. The example involves eight records from file A and six records from file B. There are two X variables -- sex (X1) and age (X2). Y contains one variable, log(personal earnings), and Z contains one variable, log(property income). For both files, the data (see Table 1) are simple random samples drawn from populations of 24 units. The weight assigned to each record is the reciprocal of the probability of selection.

Table 1: Data on Files A and B

Case	X1 (A)	X2 (A)	Y	Wt	Case	X1 (B)	X2 (B)	Z	Wt
A1	M	42	9.156	3	B1	F	33	6.932	4
A2	M	35	9.149	3	B2	M	52	5.524	4
A3	F	63	9.287	3	B3	M	28	4.223	4
A4	M	55	9.512	3	B4	F	59	6.147	4
A5	F	28	8.494	3	B5	M	41	7.243	4
A6	F	53	8.891	3	B6	F	45	3.230	4
A7	F	22	8.425	3					
A8	M	25	8.867	3					
Mean			8.97					5.55	
SD			0.38					1.57	

Assuming that Y information is ignorable (i.e. case I of Section 1), Rodgers (1984) obtained the matched file C (see Table 2) both under unconstrained and constrained matching using HDI - distance method for statistical matching. The sex variable X1, was

used as cohort (or imputation class) and the age variable X2 was used for the distance function |X2(A)-X2(B)|. The constrained matching restricted the first and second moments of the distribution of imputed Z-values to be the same as the moments of the distribution of donor values.

Table 2: Statistical Matching by HDI-Distance

Sex	Age	Imputed Values of Z		
		Unconstrained HDI	Constrained	HDI[1]
M	42	7.243	5.524	(1)
			7.243	(2)
M	35	7.243	4.223	(1)
			7.243	(2)
F	63	6.147	6.147	
M	55	5.524	5.524	
F	28	6.932	6.932	
F	53	6.147	6.147	(1)
			3.230	(2)
F	22	6.932	6.932	(1)
			3.230	(2)
F	25	4.223	4.223	
Mean		6.3	5.55	
SD		1.06	1.57[2]	

[1] Numbers in parentheses denote sample weights
[2] Based on 5 degrees of freedom

3. LLI-S FOR STATISTICAL MATCHING (Y IGNORABLE)

This is the usual imputation problem in a single super file as mentioned earlier in Section 1. We are interested in completing the data set (i.e. the single combined file containing 14 records for the Rodgers' example). There are two types of records corresponding to two patterns of response by units in the survey. One type of record contains complete information i.e. response for all variables in vector X and Z. There are six such donor records. The other type of records contains responses for X but Z is missing. There are eight such candidate records. The Y values are totally ignored.

The main ideas of the LLI-S method are:

(i) transform both X and Z to categorical variables $X_\star$ and $Z_\star$ to obtain a subset of X as suitable predictors and to get optimal imputation classes as defined by an instability measure related to coarseness of the categorical partition,

(ii) smooth the conditional categorical distribution $f(Z_\star|X_\star)$ using log-linear modelling,

(iii) use $f(Z_\star|X_\star)$ to impute Z up to a $Z_\star$ category according to proportional allocation within $X_\star$ categories, and

(iv) determine Z values within $Z_\star$ categories in order to complete missing records.

By contrast, in HDI methods, only X is categorized in forming imputation classes and the conditional distribution $f(Z_\star|X_\star)$ is used for imputation. Choice of imputation classes is not based on some optimality criterion but on subject matter considerations. It is easily seen that LLI-S would be equivalent to HDI when the imputation classes are not required to be optimal, a saturated log-linear model is employed (i.e. no smoothing to the empirical distribution) and the condition of proportional allocation is not applied.

For Rodgers' example, the LLI-S method can be described in the following five steps. First define an initial partition P_0 that provides a fairly fine grid of the three dimentional space of (X,Z) values from donor records. Let X_0, Z_0 denote the corresponding categorized X,Z variable. The three dimensional table of weighted counts based on an initial partition, P_0, is given in Table 3.

Table 3: Weighted Counts for Partition P_0 (or P_1) (donor records)

	$Z<4.5$	$4.5\leq Z<6.5$	$Z\geq 6.5$	Row Total
M Age < 45	4	0	4	8
M Age ≥ 45	0	4	0	4
F Age < 45	0	0	4	4
F Age ≥ 45	4	4	0	8
Column Total	8	8	8	24

Step I Choice of X Variables - We need to investigate the strength of the relationship between X_0 and Z_0 and determine whether or not both X variables should be retained. We use I-divergence under pseudo-multinomial assumptions to define a X^2 type measure of association (I_A) between X_0 and Z_0. Let H_1: $Z_0 \perp X_0$ denote the hypothesis

of independence of X_0 and Z_0. Then $I_A(Z_0 \perp X_0) = (\frac{1}{4}) \Sigma_t\, p_t(0) \log (p_t(0)/\hat{p}_t(1))$ where $p(0)$, $p(1)$ denote respectively the observed and expected proportion vectors. The I_A metric is asymptotically equivalent to Hellinger distance and is therefore approximately bounded above by 1. It provides a convenient distance metric that could be used for survey data as well as administrative data. We use $\delta_0 = .05$ as a working threshold value in order to decide when I_A is large enough i.e. when $I_A \geq \delta_0$.

Using the definition $(0)\log(0)=0$, one can calculate $I_A(Z_0 \perp X_0) = 0.1591$, which is greater than the working threshold value of 0.05. Similarly, using a table combined over sex categories and a table combined over age categories, we can calculate, respectively, $I_A(Z_0 \perp X_{02}) = 0.1155$ (association due to age) and $I_A(Z_0 \perp X_{01}) = 0.0$ (association due to sex). By taking the difference (.1591 - .1155), one can determine that $I_A(Z_0 \perp X_{01} \mid X_{02})$ (association due to sex|age) is 0.0436. Although the conditional association due to sex is less than our working value of 0.05, it is close to the threshold, so we decide to keep both age and sex as predictor variables. Consequently, we have $P_1=P_0$ where P_1 denotes the partition that would have been obtained corresponding to the chosen subset X_1 of X_0. In this case X_0 is not reduced. So $X_1=X_0$. Also set $Z_1=Z_0$ for notational convenience.

Step II Choice of Optimal Partition $P_\star$ - Let G denote a class of partitions $P_1, P_2, \ldots$ such that for each P_j the association between the corresponding categorical variables X_j and Z_j is high. The P_j's represent modified versions of P_1 which may be coarser or have different cell boundaries. We then define an instability measure $R(\epsilon)$ related to the coarseness of a partition that will allow us to choose the optimal partition $P_\star$ from the class G. We have $R(\epsilon) = n(\epsilon)/T$, where $n(\epsilon)$ is the number of cells with proportions less than or equal to ϵ, a small predetermined positive constant. Note that cells with zero counts will contribute to $R(\epsilon)$. Now, the optimal partition $P_\star$ is the partition in G for which $R(\epsilon)$ is smallest. The use of $R(\epsilon)$ has an heuristic justification. For a particular choice of X_1 and Z_1, modification of a partition to make it finer will generally increase association between the corresponding categorical variables. Use of $R(\epsilon)$ militates against selection of fine partitions containing many cells with small proportions.

Suppose the class G consists of two partitions -- P_1, given by Table 3, and P_2, given by Table 4. Note that $I_A(Z_2 \perp X_2)=0.0577$, which is greater than our working threshold. Therefore, the partition P_2 does qualify to belong to class G. To determine the optimal

partition, $P_\star$, within G we compute the instability measure $R(\varepsilon)$ for P_1 and P_2 with $\varepsilon=0.01$. We have $R(\varepsilon)=0.5$ for P_1 and $R(\varepsilon)=0.25$ for P_2. Consequently, the optimal partition is $P_\star=P_2$ for this particular illustration.

Table 4: Weighted Counts for Partition P_2 (donor records)

	$Z<6$	$Z\geq 6$	Row Total
M Age < 45	4	4	8
M Age ≥ 45	4	0	4
F Age < 45	0	4	4
F Age ≥ 45	4	4	8
Column Total	12	12	24

Step III Log-linear Model Selection - For the joint categorical distribution $f(X_\star,Z_\star)$ corresponding to $P_\star$, select a log-linear model using the donor data set. While a parsimonious model is desirable, the saturated model can be retained if it is not feasible to reduce it. Using Table 4, one can test the independence of age and sex, conditional on Z. The I_A - measure corresponding to this hypothesis is 0.0435, which although smaller than our working threshold value, is close to it. Consequently, we decide to retain age-sex interaction terms in the model. One could also, of course, compare the saturated model to a model with no three-factor interaction. For illustrative purposes, we decide to retain the saturated model.

Step IV Estimation of the Conditional Categorical Distribution $f(Z_\star|X_\star)$ -Expected counts corresponding to the saturated model are, of course, equal to observed counts. The distribution of $Z_\star$ for each $X_\star$ category, given in Table 5, can be easily obtained from the weighted counts in Table 4.

Table 5: Estimate of $f(Z_\star|X_\star)$ (from donor records)

	$Z<6$	$Z\geq 6$	Row Total
M Age < 45	1/2	1/2	1
M Age ≥ 45	1	0	1
F Age < 45	0	1	1
F Age ≥ 45	1/2	1/2	1

Step V Imputation of Missing Z Values - We allocate the set of candidate records within each imputation class proportionally according to $f(Z_\star|X_\star)$ as shown in Table 6. Hot deck distance over age can be used within each imputation class to assign records with missing values to the two $Z_\star$ categories ($Z<6$ and $Z\geq 6$) as well as to determine Z values to impute within each $Z_\star$ category. The only fractional counts occur for the imputation class defined by sex=M, age<45. In this case the imputed value can be determined as an average of "closest" values from each $Z_\star$ category, weighted according to the non-integral portions of the counts. Since there is only one donor record in each $Z_\star$ category, our imputed value is the average of 4.223 (only value of Z for sex=M, age<45, Z<6) and 7.243 (only value of Z for sex=M, age<45, Z>6).

Table 6: Proportional Allocation in LLI-S

	Z<6	Z≥6	Number Missing	Observed Ages for Candidate Records
M Age <45	1.5	1.5	3	25,35,42
M Age ≥ 45	1	0	1	55
F Age < 45	0	2	2	22,28
F Age ≥ 45	1	1	2	53,63

The imputed values using LLI-S are given in Table 9 along with those for LLI-M (to be described in the next section).

4. LLI-M FOR STATISTICAL MATCHING (Y NON-IGNORABLE)

In this section we consider the same example, except that we do not ignore Y in the formation of imputation classes and the computation of imputed values. The main ideas of the LLI-M method can be summarized as follows.

(i) We transform X,Y,Z to categorical variables $X_\star, Y_\star, Z_\star$ in order to find a suitable subset of X to predict Y and Z, and to obtain optimal imputation classes as defined by a sensitivity measure related to departures from the conditional independence assumption in the categorical framework.

(ii) As in LLI-S, smooth the conditional categorical distribution $f(Z_\star|X_\star,Y_\star)$ using log-linear modelling for the donor data set. Here it is assumed that $f(Z_\star|X_\star,Y_\star)$

is the same as $f(Z_\star|X_\star)$ i.e. categorically, the conditional independence assumption is valid. In terms of log-linear modelling, it implies that the parameters involving $(Y_\star,Z_\star)$ factor effects and $(X_\star,Y_\star,Z_\star)$ factor effects are set equal to zero. In the final Section 6, we suggest ways in which this condition can be relaxed.

(iii) Use $f(Z_\star|X_\star,Y_\star)$ to impute Z up to a $Z_\star$ category according to proportional allocation within $(X_\star,Y_\star)$ categories, and

(iv) specify a suitable imputation scheme for determining Z values within $Z_\star$ categories.

For Rodgers' example, we can describe LLI-M in five steps. Suppose the initial partition P_0 for the multiple file method corresponds to the (X_0,Y_0) counts given in Table 7 and the (X_0,Z_0) counts given in Table 3. Note that Y_0 involves two categories, defined by $Y<9$ and $Y\geq 9$.

Table 7: Weighted Counts for Partition P_0 (candidate records from file A)

	$Y<9$	$Y\geq 9$	Row Total
M Age < 45	3	6	9
M Age ≥ 45	0	3	3
F Age < 45	6	0	6
F Age ≥ 45	3	3	9
Column Total	12	12	24

Step I Choice of X Variables - The measure of association between X_0 and Y_0 corresponding to P_0 is I_A=0.0703, a value greater than the working threshold of 0.05. Thus, the chosen Y partition does provide high association with X_0. In Section 3, we already considered the effects of dropping variables on the association between X_0 and Z_0 and concluded that it was not possible to drop X_0 variables in the prediction of Z_0. Consequently, we set $P_1=P_0$.

Step II Choice of Optimal Partition $P_\star$ - We need to check departures from the conditional independence assumption of $Y_\star$ and $Z_\star$ given $X_\star$. There is of course no information on the joint distribution of $(X_\star,Y_\star,Z_\star)$. However, under pseudo-multinomial assumptions, we can estimate the expected proportion vector q corresponding to the conditional independence hypothesis. This allows us to construct an upper bound on the X^2

distance corresponding to the hypothesis of conditional independence which is defined by $\eta = tr(diag\ (q)^{-1})$.

The measure η is termed as the sensitivity measure and a good choice of grid or partition is one that tends to make η small. However, use of η exclusively leads to a very coarse grid and a trivialization of the problem. For this reason, we introduce a balancing factor that requires high association between $X_\star$ and $Y_\star$, as well as between $X_\star$ and $Z_\star$. Note that the requirement of high association favours use of fine grids. Thus there is a tradeoff between low sensitivity and high association. We first generate a class of grids by defining categories in various ways. Next we restrict attention to the class G of grids for which association measures are above a threshold value. Finally, the optimal partition $P_\star$ is the grid in the class G for which η is minimized. It should be noted that in practice, some elements of q may be zero, leading to computational difficulties. In such cases, zero elements of q can be replaced by some small positive constant, say γ, and all entries rescaled such that their total is unity.

In the present example, consider the simple case of two partitions -- P_1, defined above, and P_2, which involves $X_2=X_1$, $Y_2=Y_1$, and two Z_2 categories ($Z<6$ and $Z\geq 6$) as in Table 4. The significance of associations of X with Y and X with Z have already been established for these two partitions. Using $\gamma=0.005$ for zero cell proportions the values of the sensitivity measure η are 24,452.5 for P_1 and 14,828.5 for P_2. Hence, we choose P_2 as the optimal partition $P_\star$ in the class G.

Step III Log-linear Model Selection - This step involves modelling the joint distribution $f(X_\star,Z_\star)$ using data from file B and was already considered in Section 3. As before, the saturated model is used.

Step IV Estimation of the Conditional Distribution $f(Z_\star|X_\star,Y_\star)$ - Since we assume independence of $Y_\star$ and $Z_\star$ given $X_\star$, the estimate of $f(Z_\star|X_\star)$ can be used as an estimate of $f(Z_\star|X_\star,Y_\star)$ for all categories of $Y_\star$. Estimation of $f(Z_\star|X_\star)$ in this case is identical to the corresponding situation in Section 3.

Step V Imputation of Missing Z Values - Initially, we allocate the set of candidate records within each $(X_\star,Y_\star)$ category according to the conditional distribution $f(Z_\star|X_\star,Y_\star)$ determined in the previous step. The counts are shown in Table 8 and imputed values, obtained using the distance metric over age to assign incomplete records

to Z_* categories and to impute values corresponding to both integral and fractional counts, are given in Table 9. Table 9 also gives the values imputed using the LLI-S method of the previous section.

Table 8: Proportional Allocation in LLI-M

	Y<9			Y>9		
	Z<6	Z≥6	Number Missing	Z<6	Z≥6	Number Missing
M Age < 45	.5	.5	1	1	1	2
M Age ≥ 45	0	0	0	1	0	1
F Age < 45	0	2	2	0	0	0
F Age ≥ 45	.5	.5	1	.5	.5	1

Table 9: Statistical Matching by LLI-S and LLI-M

Sex	Age	Y	Imputed Values of Z	
			LLI-S	LLI-M
M	42	9.156	7.243	7.243
M	35	9.149	5.733	4.223
F	63	9.287	6.147	4.688
M	55	9.512	5.524	5.524
F	28	8.494	6.932	6.932
F	53	8.891	3.23	4.688
F	22	8.425	6.932	6.932
M	25	8.867	4.223	5.743
Mean			5.745	5.746
SD			1.41	1.18

It is seen from Table 9 that the imputed values of Z by LLI methods are somewhere between constrained and unconstrained imputed values obtained by HDI, as measured by mean and SD (standard deviation) of the distribution of imputed values. The propose of this example was only to illustrate computational aspects of LLI. For comparing performance of LLI with HDI, we consider a simulation experiment to be described in the next section.

5. A SIMULATION EXPERIMENT FOR LLI EVALUATION

In this section we present some empirical results from a preliminary phase of the LLI evaluation study based on synthetic data. The data files were created by drawing a random sample of size 1000 from the distribution

$$(X1,X2,Y,Z)' \sim N(0,\Sigma)$$

where the elements σ_{ij}'s of Σ were prescribed as $\sigma_{11}=\sigma_{22}=\sigma_{33}=\sigma_{44}=1$, $\sigma_{12}=.40$, $\sigma_{13}=.50$, $\sigma_{14}=.60$, $\sigma_{23}=.25$, $\sigma_{24}=.40$, and $\sigma_{34}=-.3$, $-.2$, $-.1$, 0, $.3$. Only σ_{34} or Cov(Y,Z) was allowed to take five different values. The Cholesky decomposition ($\Sigma=FF'$) was employed to transform a vector U of four independent N(0,1) variables to obtain (X1,X2,Y,Z)' via FU. Therefore, only Z values are affected when Cov(Y,Z) varies. For each choice of Cov(Y,Z), we create two data files A and B by dividing 1000 sample observations into two equal parts. For data file A, Z values are suppressed and for B, Y values are suppressed. Thus we have five sets of files A and B obtained from the same set of N(0,1) random numbers.

For statistical matching purpose, file B is used to impute Z values (denoted by Z^I) for file A. Since true Z (denoted by Z^T) values are known for our experiment, we can easily compute RMSE(Z^I) as the square root of the sum of squared (Z^I-Z^T). Some other evaluation measures can be obtained by comparing the conditional variance (Z) and Cov(Y,Z) given X for Z^I and Z^T values in file A. For instance one can use the relationship $Cov(Y,Z|X)=Cov(Y,Z) - Cov(Y,X)\ V(X)^{-1}\ Cov(X,Z)$ where each term is computed using the data in file A.

In the preliminary evaluation study, the LLI-M method was compared with the HDI method for statistical matching. The Euclidean distance over X was used in HDI as well as in the step Vof LLI-M. For LLI-M, a proper full scale search for an optimal partition $P_\star$ using I_A and η measures was not done due to time-constraints. A 4x3x3x4 partition of (X1,X2,Y,Z) space was chosen with cut-off points (-.78, -.24, .24) for X1, (-.33, .23) for X2, (-.18, .24) for Y, and (-.80, -.26, .28) for Z when COV(Y,Z)=-.30. The cut-off points define cells for the partition e.g. (-.33,.23) defines three cells namely, $(-\infty, -.33]$, (-.33, .23] and $(.23, \infty)$. These cut-off points were chosen as functions of sample mean and variance such that they correspond to ranges of

approximately equal probability under normality. Only the boundary points for Z are affected when Cov(Y,Z) varies. Furthermore, the saturated log-linear model was used in LLI-M and finally, the imputation classes for HDI were chosen to coincide with the X categories in LLI-M partition. Thus, the differences between HDI and LLI-M are expected to be due to the impact of proportional allocation only.

Table 10 shows the results of the above evaluation study. Both LLI and HDI are marked by (x) to indicate that they were performed under certain limitations, namely,

(i) the partitioning was not optimal,
(ii) no smoothing was done, and
(iii) the hot deck imputation classes were formed from the partition chosen for LLI-M.

Results from this study given in Table 10 although limited, indicate that the effect of LLI-M is in the right direction in comparison to HDI. Methods for detailed investigation for LLI evaluation are currently being considered.

Table 10: Evaluation Measures for LLI^x and HDI^x

Data Set (file A)	RMSE		COV (Y,Z|X1,X2)			V(Z|X1,X2)		
	LLI^x	HDI^x	TRUE	LLI^x	HDI^x	TRUE	LLI^x	HDI^x
1	1.11	1.17	-.61	-.015	.036	.60	.66	.73
2	1.13	1.16	-.50	-.007	.054	.58	.68	.72
3	1.13	1.16	-.39	.025	.065	.58	.67	.71
4	1.15	1.15	-.28	.025	.072	.59	.67	.70
5	1.10	1.13	.04	.043	.081	.62	.61	.65

Notes: 1. Data set numbers for file A correspond respectively to five values chosen for Cov(Y,Z).

2. For the particular 4x3x3x4 partition used in LLI^x, the measures $I_A(X,Y)$, $I_A(X,Z)$ and η turned out to be around .021, .045, and 43,000 respectively for all the five data sets.

6. CONCLUDING REMARKS

In this paper we have described two types of log-linear imputation for the problem of statistical matching: one for the Y ignorable case (or equivalently, a single combined file situation) denoted by LLI-S; and the other for the Y non-ignorable case (or equivalently a multiple file situation) denoted by LLI-M. In practice, the choice between

LLI-S and LLI-M would depend on the validity of the untestable assumption of conditional independence of Y and Z given X. Generally this assumption would be unrealistic and hence LLI-M would be preferable as it uses Y information. Among the five steps of LLI, namely,

(i) choice of X predictors,
(ii) choice of optimal partition and hence optimal imputation classes,
(iii) smoothing via log-linear modelling,
(iv) estimating the conditional distribution for proportional allocation, and
(v) completing the missing Z values;

the method LLI-M uses Y information in all the steps except in (iii).

The above observation suggests an important direction for further research in which LLI-M could be generalized, see Singh (1988). Specifically, in the smoothing step via log-linear modelling, the two-factor $(Y_\star, Z_\star)$ and the three-factor $(X_\star, Y_\star, Z_\star)$ effects are necessarily dropped because there are no data containing (Y,Z) and (X,Y,Z) values. It is interesting to note that under the categorical conditional independence assumption, these are precisely the parameters that are omitted. Now, it is conceivable that reasonable estimates of some or all of these parameters could be obtained from some other source of information. This can then be incorporated to obtain modified estimates of expected proportions and in turn a new version of LLI. As regards evaluation of LLI, the preliminary results reported in this paper are promising. More extensive work on evaluation of LLI is planned using synthetic data as well as a data file created by exact matching of information from Revenue Canada and the Survey of Consumer Finance.

ACKNOWLEDGEMENTS

The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. A version of this paper was presented in the annual meeting of the American Statistical Association, New Orleans, August 22-25, 1988. The authors would like to thank D. Drew, Section Chief, Social Survey Methods Division for helpful discussion and to Christine Larabie and Jackie Walker for their efficient manuscript processing.

REFERENCES

KADANE, J.B. (1978). Some statistical problems in merging data files. In 1978 Compendium of Tax Research, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 159-171.

LITTLE, R.J.A. and RUBIN, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley.

RODGERS, W.L. (1984). An evaluation of statistical matching. Journal of Business and Economic Statistics, 2, 91-102.

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

SIMS, C.A. (1978). Comments on Kadane's work on matching to create synthetic data. In 1978 Compendium of Tax Research, Office of Tax Analysis, U.S. Department of the Treasury, Washington, D.C.: U.S. Government Printing Office, 172-177.

SINGH, A.C. (1988). Log-linear Imputation. Methodology Branch Working Paper, Statistics Canada, SSMD 88-029E.

U.S. DEPARTMENT OF COMMERCE (1980). Report on exact and statistical matching techniques. Statistical Policy Working Paper 5, Federal Committee on Statistical Methodology.

WOLFSON, M., GRIBBLE, S., BORDT, M., MURPHY, B. AND ROWE, G. (1988). The Social Policy Simulation Database: an example of survey and administrative data integration. Proceedings of the Symposium on Statistical Uses of Administrative Data, Statistics Canada Ottawa November 23-25, 1987 (J.W. Coombs and M.P. Singh, eds), 201-229.