Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 137–157

Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints

Marcello D'Orazio[1], *Marco Di Zio*[1], *and Mauro Scanu*[1]

Statistical matching is a technique for combining information from different sources. It can be used in situations when variables of interest are not jointly observed and conclusions must be drawn on the basis of partial knowledge of the phenomenon. Uncertainty regarding conclusions arises naturally unless strong and nontestable hypotheses are assumed. Hence, the main goal of statistical matching can be reinterpreted as the study of the key aspects of uncertainty, and what conclusions can be drawn. In this article we give a formalization of the concept of uncertainty in statistical matching when the variables are categorical, and formalize the key elements to be investigated. A consistent maximum likelihood estimator of the elements characterizing uncertainty is suggested. Furthermore, the introduction of logical constraints and their effect on uncertainty are studied. All the analyses have been performed according to the likelihood principle. An example with real data is presented and a comparison with other approaches already defined is performed.

Key words: Data fusion; synthetical matching; constrained maximum likelihood estimation; EM algorithm.

1. Introduction

Information plays a key role with regard to understanding phenomena. In some cases it may be obtained by combining two or more data sources: some examples are database marketing (Kamakura and Wedel, 1997 and 2003) and economic research via microsimulation (e.g., the Social Policy Simulation Database created at Statistics Canada, Singh et al., 1993, and references therein). Another particularly suitable field of application is Official Statistics, this because of the large number of files maintained in National Statistical Institutes (NSIs) (D'Orazio et al., 2001).

Integration of data from different sources can be performed by means of three different methodologies: merging, record linkage and statistical matching. The first two are designed to link the same units from two or more different files: merging needs error-free matching variables, while record linkage is a statistical decision procedure that can be used when matching variables are affected by errors. Both these techniques require that the sets of observed units in the two sources overlap. Statistical matching faces the problem of integration when the files do not contain the same units. The main target of statistical matching is to give joint information on variables observed in different sources. This

[1] Italian National Statistical Institute (ISTAT), via Cesare Balbo 16, 00184 Rome, Italy. Emails: madorazi@istat.it, dizio@istat.it, and scanu@istat.it
Acknowledgments: We are very grateful to the referees and an Associate Editor for their helpful comments.

integration problem may be represented by the following situation. There are two different sources, A and B, two groups of variables never jointly observed, Y in A and Z in B, and one group of variables available in both data sources, X (see Figure 1).

Different statistical matching techniques have been developed since the 1970s (see references in Rässler 2002) and may be broadly divided into three large groups. The first one contains those techniques (implicitly) based on a specific model: Y and Z are probabilistically independent conditionally on X (Conditional Independence Assumption, CIA henceforth). When this model is not adequate, the integrated synthetical dataset may be significantly different from the truth, and the application of the usual parameter estimators may result in highly misleading estimates (Rodgers 1984, Paass 1986, Barry 1988, Goel and Ramalingam 1989, Singh et al. 1993, Renssen 1998). The second group of techniques faces this problem using auxiliary information on (Y, Z) (e.g., an additional outdated, proxy or confidential file C on (Y, Z) or (X, Y, Z), see Singh et al., 1993, and references therein). In particular, Singh et al. (1993) show by simulation studies how the accuracy of results of the matching procedure can be improved in this setting. Although this is an important special case, it is not always feasible (Ingram et al., 2000) because the required external information on the parameters regarding the statistical relationships between Y and Z or on the (Y, Z) distribution is rarely available. Both the previous two groups of techniques are constrained to just a single *world*: in the first we assume that the world is that described by the CIA, in the second we describe the closest world (with respect to the Kullback-Leibler distance; see Csizár 1975) to that of the auxiliary information (e.g., previous year) and coherent with data currently observed. Actually many distributions on (X, Y, Z) are compatible with the available partial information, i.e., many worlds may have generated the observed data, and those worlds are indistinguishable. This problem leads to the third group of techniques that addresses the so-called *identification problem*. This approach consists in assessing all the possible worlds, i.e., all the parameter values consistent with the available information. This problem was defined by Manski (1995) for the missing data problem and addressed by Kadane (1978), Rubin (1986), Moriarity and Scheuren (2001, 2003) and Rässler (2002) in the matching problem for continuous variables.

In this article we describe an approach to statistical matching in the identification problem framework for categorical data. First (Section 2) we analyze the case of marginal complete information on (X, Y) and (X, Z) and discuss what we mean by uncertainty. Then (Section 3) we consider the case when marginal information on (X, Y) and (X, Z) is provided by two independent samples. In this case, uncertainty is estimated following the Maximum Likelihood (ML) approach, i.e., all the possible worlds maximizing the likelihood function are regarded as equally informative and are taken into consideration. We also suggest the use of the elements characterizing uncertainty to some conclusions

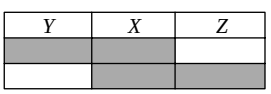

Fig. 1. Typical situation for statistical matching in a unit (row) by variable (column) matrix. Spaces in grey correspond to observed data, while white spaces are missing data. The first block of data corresponds to source A, and the second to source B.

(decisions) regarding parameter values. In order to exclude some impossible worlds, it is important to introduce logical constraints, i.e., constraints characterizing the phenomenon. In Section 4 we will consider structural zeros and inequality constraints between pairs of distribution parameters. Their introduction implies, as expected, a decrease of the overall parameter uncertainty. Finally, in Section 5, we introduce an example with NSI data to better show advantages and drawbacks of the proposed method. In the last section, concluding remarks and some directions for further research are presented.

All the considerations in the next sections have been developed when X, Y and Z are univariate variables. The extension to the multivariate context is straightforward, provided the blocks of observed data are as in Figure 1.

2. Uncertainty in a Statistical Matching Context

The statistical matching context is inevitably characterized by uncertainty, i.e., even in the optimal case of complete knowledge on the (X, Y) and (X, Z) distributions, it is not possible to draw unique and certain conclusions regarding the overall distribution (X, Y, Z) unless Y or Z can be exactly predicted by X due to a deterministic relationship between X and Y or between X and Z. Moreover in a real context, just two samples from respectively (X, Y) and (X, Z) are available, introducing an additional source of imprecision: sampling variability. The statistical analysis for this latter context is discussed in Section 3.

Let Δ define the cells of the table of the triplet (X, Y, Z) with respectively I, J, and K categories,

$$\Delta = \{(i,j,k) : i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K\}$$

The (X, Y, Z) joint distribution is multinomial with parameters:

$$\theta^*_{ijk} = P(X = i, Y = j, Z = k) \quad i,j,k \in \Delta \tag{1}$$

If the overall distribution (1) of (X, Y, Z) is unknown, the "true" but unknown parameter vector $\boldsymbol{\theta}^* = \{\theta^*_{ijk}\}$ may take values in the following set:

$$\Theta = \left\{ \boldsymbol{\theta} : \theta_{ijk} \geq 0; \sum_{(ijk)\in\Delta} \theta_{ijk} = 1 \right\} \tag{2}$$

Actually the vector $\boldsymbol{\theta}^*$ is *totally uncertain*.

Let us now assume that the marginal distributions for the two pairs (X, Y) and (X, Z) are perfectly known:

$$\theta^*_{ij} \quad i = 1, \ldots, I, j = 1, \ldots, J \tag{3}$$

$$\theta^*_{i.k} \quad i = 1, \ldots, I, k = 1, \ldots, K \tag{4}$$

Information in (3) and (4) restricts the set of possible parameters (2) to the following

subset:

$$\Theta = \begin{cases} \sum_k \theta_{ijk} = \theta^*_{ij.} & i = 1, \ldots, I, j = 1, \ldots, J \\ \sum_j \theta_{ijk} = \theta^*_{i.k} & i = 1, \ldots, I, k = 1, \ldots, K \\ \theta_{ijk} \geq 0; \sum_{(ijk)\in\Delta} \theta_{ijk} = 1 & \end{cases} \quad (5)$$

Each set (2) and (5) represents uncertainty, i.e., multiplicity of plausible solutions given the available information. In particular (5) may be considered as a description of the uncertainty connected to the statistical matching problem when complete knowledge regarding the marginal distributions is available.

Both (2) and (5) have the following characteristics:

- the true, but unknown, parameter θ^*_{ijk} lies in an interval $\theta^L_{ijk} \leq \theta^*_{ijk} \leq \theta^U_{ijk}$; in particular in set (2) $\theta^L_{ijk} = 0$ and $\theta^U_{ijk} = 1$ for all (i, j, k), while in set (5) $\theta^L_{ijk} > 0$ and $\theta^U_{ijk} < 1$ possibly for some (i, j, k);
- for each $(i, j, k) \in \Delta$, the frequency of all the plausible values for θ_{ijk} forms a distribution, $m_{ijk}(\theta)$, $\theta^L_{ijk} \leq \theta \leq \theta^U_{ijk}$. Generally speaking, this distribution "counts" all the parameter vectors $\boldsymbol{\theta} \in \Theta$ such that $\theta_{ijk} = \theta$. More formally, given that θ_{ijk} lives in a continuous space, $m_{ijk}(\theta)$ is a density function. An example of the shape of the distribution $m_{ijk}(\theta)$ is outlined in the Appendix.

These characteristics represent the uncertainty regarding each single parameter. In particular a key role is assumed by the distribution $m_{ijk}(\theta)$ and its dispersion: the less it is dispersed, the less we are uncertain about the parameter value, or, in other words, the more the imposed constraints are informative. We remark that this distribution suggests an additional dimension to consider in the sensitivity analysis for statistical matching: the distribution $m_{ijk}(\theta)$ is not just characterized by its range, but also by its variability. The range remains the only important information when $m_{ijk}(\theta)$ is flat for any $(i, j, k) \in \Delta$. For instance, this happens when $I = J = K = 2$ and the constraints in (5) hold.

The dispersion of $m_{ijk}(\theta)$ may be studied by traditional descriptive measures, among others the standard deviation, the coefficient of variation or the interquartile range. Furthermore, the dispersion of $m_{ijk}(\theta)$ can suggest also a punctual approximation of the true parameter θ^*_{ijk}. For instance, a reasonable approximation for θ^*_{ijk} can be the average with respect to $m_{ijk}(\theta)$. Let $\bar{\theta}_{ijk}$ be this average value:

$$\bar{\theta}_{ijk} = \int_{\theta^L_{ijk}}^{\theta^U_{ijk}} \theta m_{ijk}(\theta) d\theta \quad (6)$$

It is easy to see that $\boldsymbol{\theta} = \{\bar{\theta}_{ijk}\}$ describes a distribution (each parameter is nonnegative and their sum is 1). As a limit case, when the dispersion is null, the marginal distributions are sufficient for determining θ^*_{ijk}.

An overall measure of dispersion in Θ is the following: take the set of $\boldsymbol{\theta}$ compatible with the imposed constraints $\theta^*_{ij.}$ and $\theta^*_{i.k}$, $(i, j, k) \in \Delta$ and determine its volume, $V = \int_\Theta d\boldsymbol{\theta}$

(computed just on the uncertain parameters, i.e., those parameters with the corresponding uncertainty distribution $m_{ijk}(\theta)$ not concentrated on one value). This measure is particularly important when logical constraints (common in real cases) are available: for example in Official Statistics logical constraints are frequently used in the error localization phase (Fellegi and Holt 1976). These additional constraints are useful in order to exclude impossible distributions in (5), thus shrinking Θ, reducing V and, hence, the overall uncertainty. This aspect is studied in Section 4.

Note that similar definitions of uncertainty are not just for categorical variables, but hold also for continuous ones. For instance, let (X, Y, Z) be a trivariate multinormal random variable. In this case, uncertainty is defined fixing all the parameters suggested by the (X, Y) and (X, Z) distributions (means, variances, correlations for (X, Y) and (X, Z)) as the set of all the correlations for (Y, Z) compatible with the fixed parameters. It can be proved that the uncertainty distribution for ρ_{YZ} is a uniform distribution between a minimum and maximum value (these extrema are well studied and discussed in: Kadane 1978, Moriarity and Scheuren 2003, Rässler 2002). However, when either Y or Z is not univariate the uncertainty distributions for the uncertain parameters might be different, as suggested by Figure 3 in Rässler (2004). We defer discussion on uncertainty for continuous variables to a future work.

3. The Statistical Model

Let us consider n i.i.d. realizations of (X, Y, Z). Dealing with discrete variables from the multinomial distribution in (1) and denoting the vector of observed frequencies

$$\mathbf{n} = \{n_{ijk}, (i,j,k) \in \Delta\}$$

where n_{ijk} is the number of units in the sample with $(X = i, Y = j, Z = k)$ the *complete* likelihood is:

$$L(\boldsymbol{\theta}|\mathbf{n}) = \prod_{i,j,k} \theta_{ijk}^{n_{ijk}} \quad \boldsymbol{\theta} \in \Theta \tag{7}$$

The statistical matching context in Figure 1 is present when the n statistical units are divided in two subgroups A and B (two independent subsamples) of respectively n_A and n_B units, $n_A + n_B = n$. Let us assume that Z is not observed in A and Y is not observed in B. A usual assumption (see for instance: Kamakura and Wedel 1997, Rässler 2002, pp. 75–76) is that this missing data mechanism is ignorable. Under this assumption, in the situation of Figure 1, marginalization of the complete data likelihood (7) gives the *observed* data likelihood (Little and Rubin 1983):

$$L(\boldsymbol{\theta}|\mathbf{n}_A, \mathbf{n}_B) = \prod_{i,j} (\theta_{j|i}\theta_{i\cdot\cdot})^{n^A_{ij\cdot}} \prod_{i,k} (\theta_{k|i}\theta_{i\cdot\cdot})^{n^B_{i\cdot k}} \quad \boldsymbol{\theta} \in \Theta \tag{8}$$

where $\theta_{j|i} = \theta_{ij\cdot}/\theta_{i\cdot\cdot}$, $\theta_{k|i} = \theta_{i\cdot k}/\theta_{i\cdot\cdot}$, and $\mathbf{n}_A$ and $\mathbf{n}_B$ have the same meaning as $\mathbf{n}$. Note that the factorization in (8) is a straightforward application of the Factorization Lemma in Rubin (1974). Although (8) is a function of the overall vector $\boldsymbol{\theta} \in \Theta$, the right-hand side depends explicitly only on some marginal parameters. The maximum of (8) with respect to

these parameters is uniquely determined by:

$$\hat{\theta}_{j|i} = \frac{n^A_{ij.}}{n^A_{i..}}, \hat{\theta}_{k|i} = \frac{n^B_{i.k}}{n^B_{i..}}, \hat{\theta}_{i..} = \frac{n^A_{i..} + n^B_{i..}}{n} \quad (i,j,k) \in \Delta \tag{9}$$

The previous statements allow us to derive the final estimates for the following parameters:

$$\hat{\theta}_{ij.} = \frac{n^A_{ij.}}{n^A_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n}, \hat{\theta}_{i.k} = \frac{n^B_{i.k}}{n^B_{i..}} \frac{n^A_{i..} + n^B_{i..}}{n}, \hat{\theta}_{i..} = \frac{n^A_{i..} + n^B_{i..}}{n} \quad (i,j,k) \in \Delta$$

However, we are interested in estimating the overall vector $\boldsymbol{\theta}^*$. The maximum of the observed likelihood function (8) in θ_{ijk} is not unique. Every vector $\boldsymbol{\theta} = \{\theta_{ijk}\}$ that satisfies the following set of equations:

$$\begin{cases} \sum_k \theta_{ijk} = \hat{\theta}_{ij.} = \frac{n^A_{ij.}}{n^A_{i..}} \left(\frac{n^A_{i..} + n^B_{i..}}{n} \right) \\ \sum_j \theta_{ijk} = \hat{\theta}_{i.k} = \frac{n^B_{i..k}}{n^B_{i..}} \left(\frac{n^A_{i..} + n^B_{i..}}{n} \right) \\ \theta_{ijk} \geq 0, \sum_{i,j,k} \theta_{ijk} = 1 \end{cases} \tag{10}$$

is an ML estimate. The set constituted by all the ML estimates forms a region called the *likelihood ridge*. It is easy to see that the likelihood ridge is the ML estimator of the set (5), and consequently may be used for estimating the uncertainty of the statistical matching process. Given that the likelihood ridge is composed of ML estimates, all the distributions in it are equally informative, given the data. A consequence of the properties of ML estimators is that uncertainty is estimated according to the likelihood principle.

One of the most important features of (10) is that it is dependent on the samples $\mathbf{n}_A$ and $\mathbf{n}_B$ through the ML estimates $\hat{\theta}_{ij.}$ and $\hat{\theta}_{i.k}$, $(i,j,k) \in \Delta$. The sample variability of the likelihood ridge (10) decreases when n_A and n_B diverge to $+\infty$, due to the consistency of the ML estimators of the marginal distributions $\hat{\theta}_{ij.}$ and $\hat{\theta}_{i.k}$, $(i,j,k) \in \Delta$. In other words, the likelihood ridge converges (almost surely) to the set of distributions in (5) that describes the uncertainty connected with the statistical matching context when *complete* knowledge regarding (X, Y) and (X, Z) is available. Another consequence is that we can use the MLE counterpart of θ^U_{ijk}, θ^L_{ijk} and $m_{ijk}(\theta)$ in the following $\hat{\theta}^U_{ijk}$, $\hat{\theta}^L_{ijk}$ and $\hat{m}_{ijk}(\theta)$. All these estimators are consistent, and can be usefully considered for the computation of (6). Since uncertainty is a factor strongly characterizing statistical matching, the most important thing is to reduce uncertainty, i.e., reduce the dispersion of the distributions $m_{ijk}(\theta)$, $(i,j,k) \in \Delta$. One possibility is offered by logical constraints (Section 4).

We also underline that the likelihood ridge (10) contains the solutions under the CIA. In fact, the parameter vector $\boldsymbol{\theta}$ assumes the form:

$$\theta_{ijk} = \frac{\theta_{ij.}\theta_{i.k}}{\theta_{i..}} \quad \forall (i,j,k) \in \Delta \tag{11}$$

Consequently, the (unique) ML estimates are

$$\hat{\theta}_{ijk} = \frac{n^A_{ij.}}{n^A_{i..}} \left(\frac{n^B_{i.k}}{n^B_{i..}} \right) \left(\frac{n^A_{i..} + n^B_{i..}}{n} \right) \quad \forall (i,j,k) \in \Delta \tag{12}$$

and these estimated parameters are clearly inside the likelihood ridge (10). Similar results hold also in the continuous Gaussian case (e.g., Moriarity and Scheuren 2001).

4. Logical Constraints

There are situations when it is possible to introduce logical constraints. We intend for logical constraints those rules that make some of the parameter vectors in Θ illogical for the investigated phenomenon. Thus their introduction is needed in order to eliminate impossible worlds. There are various examples of logical constraints. Two frequent cases, which we will use in the next paragraphs, are:

- *existence of some quantities*: e.g., it cannot be accepted that a unit in the population is both ten years old and married;
- *inequality constraints*: e.g., the probability of being a worker with a diploma is higher than the probability of being a manager with a degree.

For the statistical model described in Section 3, they can be expressed as:

$$\theta_{ijk} = 0 \qquad \text{for some } (i,j,k) \tag{13}$$

$$\theta_{ijk} \leq \theta_{i'j'k'} \quad \text{for some } (i,j,k), (i',j',k') \tag{14}$$

Constraint (13) is usually called *structural zero* (see, e.g., Agresti 1990). This constraint occurs when: (a) (i, j, k) contains at least one pair of incompatible categories or (b) each pair in (i, j, k) is plausible but the triplet is incompatible.

Note that great caution should be exercised when it comes to the definition of the set of logical constraints. In fact, if the constraints are not compatible with each other, Θ results in an empty set (see Bergsma and Rudas 2002, for more details and references therein). From now on, we suppose that the chosen logical constraints are compatible.

A further remark is that not all the imposed constraints are of the same nature: some of them are certainties (this happens for the structural zeros) while others are validated by subject matter experts (it is very unlikely that they do not happen). In the following, we will not distinguish between them, and we will assume that they are all strictly valid. Further studies should be devoted to the introduction of a suitable probabilistic device to deal with the nonstrictly logical constraints.

The main effect of these constraints is the possible reduction of the dispersion of the plausible parameters in the likelihood ridge. It is clear that the size of the reduction is dependent on the amount of information introduced. In some circumstances, information carried by logical constraints can be so informative that, using them in addition to the observed marginal distributions (X, Y) and (X, Z), it is possible to reduce the likelihood ridge to a unique distribution. This happens, for instance, when $(J - 1)$ $(K - 1)$ independent structural zero constraints are set for each $X = i$, for $i = 1, \ldots, I$ (i.e., maximum dependence between Y and Z conditional on X). Structural zeros are also very effective because, with the exception of limit cases, parameter vectors $\{\theta_{ijk}\} \in \Theta$ based

on the CIA become illogical. In fact, when θ_{ijk} is set to 0 for some (i, j, k), the CIA (i.e., parameters as in (11)) holds only when $\hat{\theta}_{ij.} = 0$ and/or $\hat{\theta}_{i.k} = 0$, otherwise that distribution is outside the restricted parameter space and cannot be considered in the estimation phase (D'Orazio et al. 2002).

Let us suppose that the imposed logical constraints restrict Θ to a subspace $\Omega \subset \Theta$ which is closed and convex (any combination of structural zeros and inequality constraints leads to such a restriction). The problem of the likelihood function maximization when constraints are imposed may be solved following two different strategies. These strategies refer to these situations: 1) Ω has a nonempty intersection with the unconstrained likelihood ridge (10); 2) Ω has an empty intersection with the unconstrained likelihood ridge (10).

In the first case the likelihood ridge reduces to the set of solutions of:

$$\begin{cases} \sum_k \theta_{ijk} = \hat{\theta}_{ij.} = \dfrac{n^A_{ij.}}{n^A_{i..}} \dfrac{n^A_{i..} + n^B_{i..}}{n} \\ \sum_j \theta_{ijk} = \hat{\theta}_{i.k} = \dfrac{n^B_{i.k}}{n^B_{i..}} \dfrac{n^A_{i..} + n^B_{i..}}{n} \\ \boldsymbol{\theta} \in \Omega \end{cases} \tag{15}$$

In the second case, the set of equations in (15) has no solutions, i.e., Ω does not contain a vector $\boldsymbol{\theta}$ such that the likelihood function derivative with respect to some $\theta_{ij.}$ and $\theta_{i.k}$ is equal to zero. In other words, the maximum(s) of the likelihood (8) may be located only on the border of the subspace Ω (border determined by the admissible marginal distributions $\{\theta_{ij.}\}$ and $\{\theta_{i.k}\}$ that are in Ω). In this case, we suggest using an iterative algorithm in order to find the maximum in $\boldsymbol{\theta}$ of (8) constrained to $\boldsymbol{\theta} \in \Omega$. An example of such a situation will be given in Section 5.2.

The likelihood maximization problem in a proper closed and convex subset has been studied by many authors by means of different approaches (see e.g., Judge et al. 1980, Chapter 17). We adopt a version of the "projection method" described in Winkler (1993) that makes use of the EM algorithm (Dempster et al. 1979). It consists of the following steps:

- initialize the algorithm with a $\hat{\boldsymbol{\theta}}^0 \in \Omega$;
- if at iteration t, $t \geq 1$, the EM unconstrained estimate $\hat{\boldsymbol{\theta}}^t$ does not satisfy the constraints, such solution is "projected" to the boundary of the closed and convex subspace Ω; otherwise it is left unchanged.

Such an approach is convenient in our context because the likelihood in (7) is a mixture of multinomial distributions. In this case, a theorem by Haberman (Theorem 4, 1977; see also Winkler 1993) suggests the following: if $\hat{\boldsymbol{\theta}}^{t-1}$ and $\hat{\boldsymbol{\theta}}^t$, $t \geq 1$, are successive estimates, and $\hat{\boldsymbol{\theta}}^{t-1} \in \Omega$ while $\hat{\boldsymbol{\theta}}^t \notin \Omega$, then $\hat{\boldsymbol{\theta}}^t$ should be replaced by the linear combination of $\hat{\boldsymbol{\theta}}^{t-1}$ and $\hat{\boldsymbol{\theta}}^t$ so that $\alpha\hat{\boldsymbol{\theta}}^{t-1} + (1-\alpha)\hat{\boldsymbol{\theta}}^t$ lies on the boundary of Ω $(0 \leq \alpha \leq 1)$. Given that Ω is closed and convex, such α exists and is unique. Additionally, the theorem by Haberman states that the likelihood of the successive M step solutions of this modified EM algorithm (Winkler calls this method EMH) is nondecreasing. Winkler (1993) warns that this algorithm may stick at a relative maximum in Ω. In order to get only the global maxima, it

is worth starting the EMH algorithm from different points and analyzing the likelihood computed on the results. For instance, in the example in Section 5 we start from 100,000 different initial values and only the results with the (same) highest likelihood are retained. Another warning (Winkler 1993) relates to the computation of α that may require additional iterative algorithms, but in the present setting α may be computed directly. Structural zero constraints (13) may be easily fulfilled by setting to zero the corresponding $\hat{\theta}^0_{ijk}$ in the initialization step of the EM algorithm (for details see Schafer 1997, pp. 52–53). Also inequality constraints are easily fulfilled. In fact, Inequality (14) is satisfied when

$$\alpha = \frac{\hat{\theta}^t_{i'j'k'} - \hat{\theta}^t_{ijk}}{\hat{\theta}^{t-1}_{ijk} - \hat{\theta}^{t-1}_{i'j'k'} - \hat{\theta}^t_{ijk} + \hat{\theta}^t_{i'j'k'}}$$

If more than one inequality constraint is imposed, the smallest α should be considered.

We remark that the multinomial model in (7) is saturated, and consequently each *M* step in the EM algorithm gives solutions in closed form. However, if a different loglinear model is assumed, the ECM algorithm can be adopted instead of the EM algorithm, as in Winkler (1993).

We finally remark also that, as stated by Equation (15), the overall uncertainty represented by the volume of the set of parameters compatible with the estimable ones (the volume *V* of the likelihood ridge) is always reduced by the introduction of constraints. This is a straightforward result when constraints are compatible with maximum likelihood estimates of the estimable parameters. When such compatibility does not hold, the new likelihood ridge is the projection of the unrestricted likelihood ridge on the restricted space. Hence, also in this case there is an overall reduction of the uncertainty. An example of this situation is outlined in the next paragraph.

5. An Example

In order to show how to introduce logical constraints and the corresponding advantages and drawbacks, we have developed an example with Official Statistics data where logical constraints are frequently used. A subset of 2,313 employees (people at least 15 years old) has been extracted from the 2000 pilot survey of the Italian Population and Household Census. Only three variables have been analyzed: Age (AGE), Educational Level (EDU) and Professional Status (PRO). For the sake of simplicity and without loss of information in respect of our aim, the original variables have been transformed by grouping homogeneous response categories. The results of this grouping are shown in Table 1.

Table 1. Response categories for the variables considered in the example

Variables	Transformed response categories
Age (AGE)	"1" = 15–17 years old; "2" = 18–22; "3" = 23–64; "4" = 65 and above
Educational Level (EDU)	"C" = None or compulsory school; "V" = Vocational school; "S" = Secondary school; "D" = Degree
Professional Status (PRO)	"M" = Manager; "E" = Clerk; "W" = Worker

To reproduce the situation in Figure 1, the original file has been randomly split into two almost equal subsets. The variable Educational Level has been removed from the first subset (file A), containing 1,148 units, and the variable Professional Status has been removed from the second subset (file B), consisting of the remaining 1,165 observations.

Table 2 shows the true relative frequencies of the original dataset for each cell. Structural zeros are represented by "–". For instance, in Italy a 17-year-old person cannot have a university degree. Tables 3 and 4 show, respectively, the distribution of Age vs Professional Status in file A, and Age vs Educational Level in file B, after the original dataset has been split. Note that each structural zero in a marginal table implies a set of structural zeros on the joint distribution. The joint distribution has some additional structural zeros that cannot be inferred from the marginals in Table 3 and Table 4 because they refer to structural zeros of the variables (PRO, EDU). This happens, for instance, in Cells 3 and 4 that correspond to managers (PRO = "M") but with at maximum a compulsory school educational level (EDU = "C").

5.1. Matching Results

If the two files are matched by means of a technique based only on the common variable Age, without considering any auxiliary information about the relationship existing between the two variables Educational Level and Professional Status, the final output will give estimates under the CIA. In our case, results under the CIA are reported in the last column of Table 2. As can be observed, the CIA yields unrealistic estimates for some cells. In particular, it gives nonzero estimated probabilities for certain events that cannot happen in real life, i.e., structural zeros for (PRO, EDU). On the other hand, as expected, structural zeros are preserved when observed in the marginals in Tables 3 and 4, e.g., Cells 9, 25 and 41 corresponding to the structural zero (AGE = "1"), (EDU = "S").

In order to explore the likelihood ridge, we have decided to run the EM algorithm with different random starting points:

- **S0** starting point in the full space Θ;
- **S1** starting point in the space Ω restricted by structural zeros;
- **S2** starting point in the space Ω restricted by structural zeros and the following inequality constraint: P(AGE = "3", EDU = "D", PRO = "M") $\geq$ P(AGE = "3", EDU = "D", PRO = "E").

The last inequality states that a person with Age in class "3" and with Educational Level in class "D" has a higher probability of being a Manager (PRO = "M") (Cell 15) than of being a Clerk (PRO = "E") (Cell 31). Although this inequality is not true for general populations, it is consistent with our dataset, and we have used it to show the effect of this type of constraint on the results. In order to satisfy constraints in S2, the EMH has been applied instead of the EM algorithm.

It is worthwhile to underline that different starting points produce different EM results and, given that the EM algorithm may stick at suboptimal points, we have considered only the maximum likelihood ones.

Table 5 reports the simulation extremes of the likelihood ridge found by running EM 100,000 times for each of the above-mentioned starting vectors $\hat{\boldsymbol{\theta}}^0$ (note that each simulation gave rise to a global maximum likelihood result). As expected, when no

Table 2. True cell counts (n_{ijk}) and relative frequencies (θ_{ijk}), and corresponding CIA estimates ($\hat{\theta}_{ijk}$)

Cell	AGE	EDU	PRO	n_{ijk}	θ_{ijk}	$\hat{\theta}_{ijk}$
1	1	C	M	–	–	–
2	2	C	M	–	–	–
3	3	C	M	–	–	0.0540
4	4	C	M	–	–	0.0048
5	1	V	M	–	–	–
6	2	V	M	–	–	–
7	3	V	M	–	–	0.0143
8	4	V	M	–	–	–
9	1	S	M	–	–	–
10	2	S	M	–	–	–
11	3	S	M	142	0.0614	0.0649
12	4	S	M	4	0.0017	0.0013
13	1	D	M	–	–	–
14	2	D	M	–	–	–
15	3	D	M	220	0.0951	0.0220
16	4	D	M	5	0.0022	0.0009
17	1	C	E	–	–	–
18	2	C	E	–	–	0.0022
19	3	C	E	–	–	0.1336
20	4	C	E	–	–	0.0009
21	1	V	E	–	–	–
22	2	V	E	1	0.0004	0.0009
23	3	V	E	123	0.0532	0.0350
24	4	V	E	0	0	0
25	1	S	E	–	–	–
26	2	S	E	8	0.0035	0.0022
27	3	S	E	653	0.2823	0.1604
28	4	S	E	3	0.0013	0.0004
29	1	D	E	–	–	–
30	2	D	E	–	–	–
31	3	D	E	87	0.0376	0.0545
32	4	D	E	0	0	0
33	1	C	W	15	0.0065	0.0065
34	2	C	W	27	0.0117	0.0078
35	3	C	W	759	0.3281	0.1466
36	4	C	W	12	0.0052	0.0017
37	1	V	W	0	0	0
38	2	V	W	7	0.0030	0.0035
39	3	V	W	90	0.0389	0.0385
40	4	V	W	0	0	0
41	1	S	W	–	–	–
42	2	S	W	12	0.0052	0.0073
43	3	S	W	143	0.0618	0.1755
44	4	S	W	0	0	0.0004
45	1	D	W	–	–	–
46	2	D	W	–	–	–
47	3	D	W	2	0.0009	0.0597
48	4	D	W	0	0	0.0004
Total				2313	1.0000	1.0000

Table 3. Distribution of Professional Status vs Age in file A

Age	Professional Status			Total
	M	E	W	
1	–	–	9	9
2	–	5	17	22
3	179	443	486	1108
4	6	1	2	9
Tot.	185	449	514	1148

restrictions are imposed on the starting point (S0), EM produces nonnull estimates in correspondence with the structural zeros as under the CIA. In this case the CIA solution is always included in the range of values found through EM. On the other hand, when structural zeros are introduced in the starting point (S1), EM produces zero estimated probabilities in correspondence with structural zeros. Moreover, for nonnull probabilities it can be observed how the introduction of this kind of auxiliary information results in a general reduction of the ranges of estimated cell probabilities. When, in addition to structural zeros, the inequality constraint involving Cells 15 and 31 is introduced (S2), the results change quite markedly: it makes the likelihood ridge shrink (see e.g., Figures 2 and 3).

In general, in comparison with the initial situation of absence of auxiliary information about the phenomena under study (S0), an overall reduction of ranges for most of the estimated probabilities can be observed. Note that sometimes the maximum of $S1$ is larger than the maximum of $S0$. This is caused by the exploration of the likelihood ridge with a finite number of points. When there is compatibility between the unrestricted maximum likelihood estimates and the structural zeros, the maximum of $S1$ is never greater than the maximum of $S0$.

When the final ranges of estimated probabilities (S2) are compared with those of S1, it emerges that about half of them remain unchanged while a decrease occurs for the others. This reduction is really marked in case of Cell 31 where the upper bound for this estimated probability reduces from 0.1364 to 0.0678. On the other hand, in the case of Cell 15 the upper bound remains unchanged while the lower bound increases from 0 to 0.0260. In the case of Cells 33–36 it can be observed that the introduction of structural zeros is so

Table 4. Distribution of Educational Level vs Age in file B

Age	Educational Level				Total
	C	V	S	D	
1	6	0	–	–	6
2	14	6	13	–	33
3	387	102	464	158	1111
4	10	0	3	2	15
Tot.	417	108	480	160	1165

Table 5. Range of probability estimates in 100,000 runs of EM with three different starting settings compared with the true counts (n_{ijk}) and frequencies (θ^*_{ijk})

Cell	AGE	EDU	PRO	n_{ijk}	θ^*_{ijk}	S0		S1		S2	
						Min	Max	Min	Max	Min	Max
1	1	C	M	–	–						
2	2	C	M	–	–						
3	3	C	M	–	–	0.0000	0.1549				
4	4	C	M	–	–	0.0035	0.0067				
5	1	V	M	–	–						
6	2	V	M	–	–						
7	3	V	M	–	–	0.0000	0.0839				
8	4	V	M	–	–						
9	1	S	M	–	–						
10	2	S	M	–	–						
11	3	S	M	142	0.061	0.0000	0.1549	0.0186	0.1550	0.0186	0.1290
12	4	S	M	4	0.002	0.0000	0.0021	0.0024	0.0031	0.0024	0.0031
13	1	D	M	–	–						
14	2	D	M	–	–						
15	3	D	M	220	0.095	0.0000	0.1260	0.0000	0.1363	0.0260	0.1364
16	4	D	M	5	0.002	0.0000	0.0014	0.0013	0.0021	0.0013	0.0021
17	1	C	E	–	–						
18	2	C	E	–	–	0.0000	0.0054				
19	3	C	E	–	–	0.0000	0.3261				
20	4	C	E	–	–	0.0000	0.0012				
21	1	V	E	–	–						
22	2	V	E	1	0.000	0.0000	0.0043	0.0000	0.0043	0.0000	0.0043
23	3	V	E	123	0.053	0.0000	0.0880	0.0014	0.0881	0.0015	0.0881
24	4	V	E	0	0	0	0	0	0	0	0
25	1	S	E	–	–						
26	2	S	E	8	0.004	0.0000	0.0054	0.0011	0.0054	0.0011	0.0054
27	3	S	E	653	0.282	0.0000	0.3776	0.1591	0.3776	0.2279	0.3780
28	4	S	E	3	0.001	0.0000	0.0012	0.0000	0.0007	0.0000	0.0007
29	1	D	E	–	–						
30	2	D	E	–	–						
31	3	D	E	87	0.038	0.0000	0.1362	0.0000	0.1364	0.0000	0.0678
32	4	D	E	0	0	0.0000	0.0011	0.0000	0.0007	0.0000	0.0007
33	1	C	W	15	0.006	0.0065	0.0065	0.0065	0.0065	0.0065	0.0065
34	2	C	W	27	0.012	0.0047	0.0101	0.0101	0.0101	0.0101	0.0101
35	3	C	W	759	0.328	0.0000	0.3278	0.3342	0.3342	0.3342	0.3342
36	4	C	W	12	0.005	0.0000	0.0023	0.0052	0.0052	0.0052	0.0052
37	1	V	W	0	0	0	0	0	0	0	0
38	2	V	W	7	0.003	0.0000	0.0043	0.0000	0.0043	0.0000	0.0043
39	3	V	W	90	0.039	0.0000	0.0880	0.0000	0.0866	0.0000	0.0865
40	4	V	W	0	0			0	0	0	0
41	1	S	W	–	–						
42	2	S	W	12	0.005	0.0040	0.0094	0.0040	0.0083	0.0040	0.0083
43	3	S	W	143	0.062	0.0000	0.3926	0.0000	0.0866	0.0000	0.0866
44	4	S	W	0	0	0.0000	0.0021	0	0	0	0
45	1	D	W	–	–						
46	2	D	W	–	–						
47	3	D	W	2	0.001	0.0000	0.1361	0.0000	0.0855	0.0000	0.0859
48	4	D	W	0	0	0.0000	0.0014	0	0	0	0

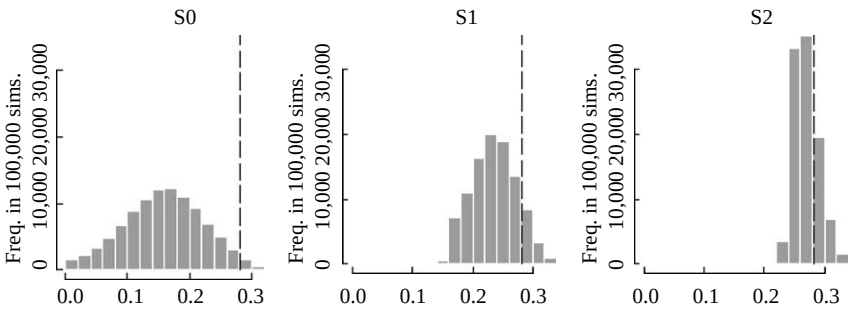

Fig. 2. Likelihood ridge for Cell 27 when no constraints (left), structural zeros (center) and the additional inequality constraint (right) are imposed. The vertical bar is the true probability.

informative (in terms of degrees of freedom) that it makes the EM converge to a unique value, in all cases close to the true one.

The width is not the only element to consider as an evaluation measure of the uncertainty on the parameters. The density $m_{ijk}(\theta)$ of each single parameter θ_{ijk} in the likelihood ridge is another important aspect (Section 2). We have approximated such density with the frequency distribution relative to the 100,000 simulations. In general, the dispersion reduces, also for those cells where the width of the interval does not change from S1 to S2.

In Figures 2 and 3, we represent the evolution of this density in the three simulation contexts here considered for Cells 15 and 27.

The joint analysis of the range and dispersion is essential to understand uncertainty. Figure 2 shows how the final distribution (S2) of the parameter in the likelihood ridge is totally different from the initial one (S0), and in particular it is more concentrated. In other words, there is at the same time a shrinkage of the interval of values for the parameter and a higher concentration in its distribution. In Figure 3, the dispersion and width of the interval again decreases from S1 to S2. This does not happen when S2 is compared with S0, although the effect of the constraints is still important. In fact, in this case the S2 distribution is more concentrated near the true value than S0.

In Table 6 we have reported the average value ($\bar{\theta}_{ijk}$) of the 100,000 estimates respectively for S0, S1 and S2, as representative parameter values among those in the ridge. In addition, as a synthetic measure of the dispersion and closeness of estimates with

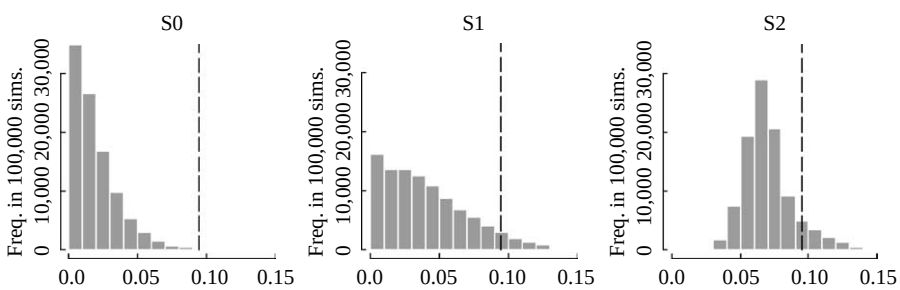

Fig. 3. Likelihood ridge for Cell 15 when no constraints (left), structural zeros (center) and the additional inequality constraint (right) are imposed. The vertical bar is the true probability.

Table 6. True probabilities (θ^*_{ijk}), average values according to the density $m_{ijk}(\theta)$ in the likelihood ridge ($\bar{\theta}_{ijk}$) and corresponding RRMSE

Cell	AGE	EDU	PRO	θ^*_{ijk}	$\bar{\theta}_{ijk}$			RRMSE		
					S0	S1	S2	S0	S1	S2
11	3	S	M	0.0614	0.0683	0.1146	0.0850	0.5661	0.9925	0.4812
12	4	S	M	0.0017	0.0014	0.0027	0.0027	0.3375	0.5648	0.5645
15	3	D	M	0.0951	0.0196	0.0404	0.0700	0.8136	0.6548	0.3236
16	4	D	M	0.0022	0.0010	0.0018	0.0018	0.5662	0.1982	0.1979
22	2	V	E	0.0004	0.0009	0.0017	0.0017	2.4742	4.1619	4.1506
23	3	V	E	0.0532	0.0361	0.0700	0.0730	0.5023	0.4363	0.4679
24	4	V	E	0	0	0	0			
26	2	S	E	0.0035	0.0022	0.0037	0.0037	0.5330	0.3135	0.3138
27	3	S	E	0.2823	0.1587	0.2353	0.2698	0.4933	0.2125	0.0852
28	4	S	E	0.0013	0.0002	0.0004	0.0004	0.8432	0.6739	0.6735
31	3	D	E	0.0376	0.0553	0.0782	0.0408	0.9327	1.3897	0.3823
32	4	D	E	0	0.0001	0.0003	0.0003			
33	1	C	W	0.0065	0.0065	0.0065	0.0065	0	0	0
34	2	C	W	0.0117	0.0078	0.0101	0.0101	0.3527	0.1355	0.1355
35	3	C	W	0.3281	0.1453	0.3342	0.3342	0.5875	0.0184	0.0184
36	4	C	W	0.0052	0.0016	0.0052	0.0052	0.7077	0	0
37	1	V	W	0	0	0	0			
38	2	V	W	0.0030	0.0034	0.0026	0.0026	0.3088	0.3823	0.3815
39	3	V	W	0.0389	0.0403	0.0181	0.0150	0.5409	0.6751	0.7248
40	4	V	W	0	0	0	0			
42	2	S	W	0.0052	0.0072	0.0057	0.0057	0.4662	0.2259	0.2254
43	3	S	W	0.0618	0.1737	0.0508	0.0459	2.0993	0.3882	0.4178
44	4	S	W	0	0.0005	0	0			
47	3	D	W	0.0009	0.0615	0.0178	0.0257	75.6145	25.2834	33.7404
48	4	D	W	0	0.0003	0	0			

respect to the true parameter values, we have computed the Relative Root Mean Squared Error (RRMSE in Table 6). In particular, for each cell the MSE is obtained as:

$$\frac{1}{100,000}\sum_{s=1}^{100,000}\left(\hat{\theta}^{(s)}_{ijk}-\theta^*_{ijk}\right)^2 \tag{16}$$

RRMSE is then computed by dividing the squared root of (16) by the true value. If, for example, Cell 15 is considered, we see how the inequality constraint on it yields a marked decrease of the corresponding RRMSE from 0.6548 (S1) to 0.3236 (S2). The same happens for Cell 31. In this case, however, the introduction of structural zeros (S1) results in an increase of RRMSE with respect to S0. Generally speaking, there can be cases when the reduction of the overall uncertainty (i.e., the reduction of the number of vectors $\boldsymbol{\theta}$ compatible with the imposed constraints) does not lead to a reduction of the marginal uncertainty of some parameters (see the Appendix for an extreme case, when a structural zero induces a flat marginal uncertainty distribution in one parameter, and certainty in the others). Further studies should be devoted to the relationship between the reduction of the overall uncertainty (i.e., the shrinkage of Θ) and of the marginal uncertainty distributions.

5.2. *Empty Intersection Between the Unconstrained Maximum Likelihood Ridge and the Imposed Logical Constraints*

As a last remark, logical constraints not only help in reducing uncertainty for inestimable parameters, but also may improve the ML estimators for estimable parameters (i.e., the ones with a unique ML estimate, in our example the ones for the marginal distributions (AGE, PRO) and (AGE, EDU)). In particular, the ML estimator for estimable parameters has the following behavior:

- for those samples whose unconstrained likelihood ridge is compatible with the constraints (i.e., some distributions in the ridge satisfy the constraints), the ML estimate of the estimable parameters remains unchanged in the constrained and the unconstrained case;
- for those samples whose likelihood ridge is not compatible with the constraints, the ML estimate of the estimable parameters is forced to respect the constraints.

Hopefully, in the second case each constrained ML estimate is moved towards the true parameter.

We have shown this last situation in our example, where some structural zeros (imposed in S1) are not compatible with the unconstrained likelihood ridge (10) when the samples in Tables 3 and 4 are observed. For instance, let us consider

$$\theta_{4CM} = P(AGE = \text{“4”},\ \ EDU = \text{“}C\text{”},\ \ PRO = \text{“}M\text{”})$$

The unconstrained ML estimates are

$$\hat{\theta}_{4.M} = \frac{6}{9} \times \frac{9+15}{2313}, \quad \hat{\theta}_{4C.} = \frac{10}{15} \times \frac{9+15}{2313}, \quad \hat{\theta}_{4..} = \frac{9+15}{2313}$$

From the standard inequality

$$\max\{0, \theta_{ij.} + \theta_{i.k} - \theta_{i..}\} \leq \theta_{ijk} \leq \min\{\theta_{ij.}, \theta_{i.k}\}$$

it holds that

$$\hat{\theta}_{4CM} \geq \frac{6}{9} \times \frac{9+15}{2313} + \frac{10}{15} \times \frac{9+15}{2313} - \frac{9+15}{2313} > 0$$

which is not compatible with the constraint so far introduced that (AGE = “4”, EDU = “*C*”, PRO = “*M*”) is a structural zero. As a consequence, this structural zero restricts Θ to a set Ω where $\theta_{4C.}$ cannot be equal to its unconstrained ML estimate $\hat{\theta}_{4C.}$.98 In fact, the constrained ML estimate of this parameter is moved towards the true value. Table 7 shows the effect of all the imposed structural zeros (S1) on some parameter

Table 7. Comparison among some probability estimates and the corresponding marginal true probabilities θ^*

	θ^*	MLE (S0)	MLE (S1)
P(PRO = *M*, AGE = 4)	0.0039	0.0069	0.0044
P(EDU = *S*, AGE = 4)	0.0030	0.0021	0.0031
P(EDU = *D*, AGE = 4)	0.0022	0.0014	0.0021
P(EDU = *C*, AGE = 4)	0.0050	0.0069	0.0052

estimates. Note that the constrained ML estimates for these marginal parameters are unique, and may be obtained by marginalizing the corresponding estimates of $\bar{\theta}_{ijk}$.

Furthermore the derivative of $L(\boldsymbol{\theta})$ with respect to $\theta_{4CM} = 0$ cannot be zero because it is not an unconstrained ML estimate. Given that the restricted space Ω is closed and convex, it cannot happen that the maximum likelihood function is inside Ω, otherwise the derivative would have been zero for each parameter. Thus the maximum must lay on the boundary of Ω. In this simple case, the boundary is the zero value for θ_{4CM}, given that only that value of θ_{4CM} is admissible.

6. Conclusions

In recent years, the main goal of the statistical matching procedures has been reinterpreted, it may be said, as the efficient use and combination of all available and relevant information. However, in the statistical matching context, the available information is in terms of partial knowledge of the phenomenon (e.g., two independent samples of some marginal distributions). Therefore, we can just draw conclusions under uncertainty. Whenever it is possible to also use auxiliary information, we can draw conclusions regarding the parameters with a lower degree of uncertainty. Extreme cases are those where particular models can be assumed, such as the CIA, or external information like that in Singh et al. (1993). In these cases it is possible to provide a unique conclusion regarding the joint distribution.

When neither the CIA nor external auxiliary information is usable, uncertainty regarding conclusions can rarely be avoided. Following ideas already applied for continuous variables, we analyze uncertainty through the description of all the plausible solutions, i.e., all those distributions coherent with the observed data according to the likelihood principle. In this context we have focused on some aspects of uncertainty and we have proposed some statistics in order to draw conclusions concerning the phenomenon at different levels, for instance regarding either single parameters of the distribution or the entire distributions.

We also describe the situation where particular auxiliary information is available: logical constraints. Logical constraints can be very useful for reducing uncertainty as to parameter values. Their usage is not immediate, and an algorithm for introducing them in the statistical matching procedure has been proposed.

Finally we remark that all the analyses of uncertainty due to partial knowledge of the phenomenon investigated are made with respect to the likelihood principle.

This article considers only categorical variables. The case of continuous variables has been frequently studied in statistical matching. Future research will be devoted to the use of logical constraints also for continuous variables.

Since the concept of uncertainty and partial knowledge has been deeply investigated in other contexts than statistical matching, for instance in artificial intelligence, we feel it is very important to analyze the common aspects and the solutions proposed there. The contamination with other frameworks is worth studying not only concerning uncertainty (as with the imprecise probabilities; see Walley 1991), but also for the use of logical constraints (e.g., Coletti and Scozzafava 2002 and Vantaggi 2003). Further research will be devoted to the inspection of these other scientific contexts.

Appendix

As in Section 2, let us assume that the marginal distribution $\{\theta_{ij.}\}$ of (X, Y) and $\{\theta_{i.k}\}$ of (X, Z) are perfectly known. One of the key elements in assessing uncertainty of the parameter values is the description of the set of vectors $\boldsymbol{\theta} \in \Theta$ that satisfy the set of conditions in (2) or (5). For this reason, in Section 2 we introduced the distribution $m_{ijk}(\theta)$, $\theta_{ijk}^L \leq \theta \leq \theta_{ijk}^U$ for each parameter θ_{ijk}, $(i,j,k) \in \Delta$. The distribution $m_{ijk}(\theta)$ describes, i.e., counts, the set of vectors $\boldsymbol{\theta}$ satisfying the constraints in (2) or the additional marginal constraint (5) for the single parameter θ_{ijk} and, usually, is not uniform. To illustrate this we will use the case where X has two categories ($I = 2$), Y has two categories ($J = 2$) and Z has three categories ($K = 3$). Given that the parameters of the multinomial distribution lie in a continuous interval, the distribution $m_{ijk}(\theta)$ will be described by its density. For the sake of simplicity, we will study only the parameter θ_{111}.

When the only (natural) constraint is that the parameters of the multinomial distribution are nonnegative and their sum is one, i.e., Θ is defined by (2), $m_{111}(\theta)$ assumes the form:

$$m_{111}(\theta) = c\int_0^{1-\theta} dx_1 \int_0^{1-\theta-x_1} dx_2 \ldots \int_0^{1-\theta-\sum_{v=1}^{9} x_v} dx_{10} = c\frac{(1-\theta)^{10}}{10!} \quad 0 \leq \theta \leq 1$$

where c is a normalizing constant. Note that once θ_{111} has been fixed to θ, there are just 10 parameters that are free to take values in a nondegenerate interval, due to the constraints in (2). Generally speaking, if the multinomial distribution has $W + 2$ categories, the previous distribution will have been $c(1-\theta)^W/W!, \theta \in (0,1)$. Note also that the density is decreasing in θ. This is due to the fact that there are more distributions with a small θ than with a large one. Additionally, the density function is a polynomial function in θ.

When the marginal distributions of (X, Y) and (X, Z) are known, i.e., Θ is reduced to (5), the aspect of $m_{111}(\theta)$ changes. First of all, once θ_{111} has been fixed to θ, just three other parameters are allowed to take values in nondegenerate intervals, say θ_{112}, θ_{211}, and θ_{212}. Analyzing the constraints for all the parameters θ_{ijk}, the new range of $m_{111}(\theta)$ is

$$\theta_{111}^L = \max\{\theta_{1.1} - \theta_{12.}; 0\} \leq \theta \leq \min\{\theta_{1.1}; \theta_{11.}\} = \theta_{111}^U$$

and the integrals to consider for determining the shape of $m_{111}(\theta)$ will have the following bounds:

$$b_{112}^L = \max\{0; \theta_{11.} - \theta - \theta_{1.3}\}; b_{112}^U = \min\{\theta_{1.2}; \theta_{11.} - \theta\}$$
$$b_{211}^L = \max\{\theta_{2.1} - \theta_{22.}; 0\}; b_{211}^U = \min\{\theta_{21.}; \theta_{2.1}\}$$
$$b_{212}^L = \max\{0; \theta_{21.} - \theta_{211} - \theta_{2.3}\}; b_{212}^U = \min\{\theta_{2.2}; \theta_{21.} - \theta_{211}\}$$

where b_{ijk}^L and b_{ijk}^U are the bounds for θ_{ijk} conditional on $\theta_{111} = \theta$ (and they may differ from the unconstrained bounds θ_{ijk}^L and θ_{ijk}^U). Note that the bounds for θ_{211} and θ_{212} contribute with a constant in θ to the computation of $m_{111}(\theta)$. As an example,

Table 8. Marginal distributions of (X, Y) and (X, Z)

	$Y=1$	$Y=2$	$Z=1$	$Z=2$	$Z=3$
$X=1$	0.43	0.16	0.16	0.25	0.18
$X=2$	0.17	0.24	0.15	0.18	0.08

consider the marginal distributions in Table 8. In this case, the possible θ_{111} values are between 0 and 0.16, and its density is:

$$m_{111}(\theta) = c\int_{b_{112}^L}^{b_{112}^U} d\theta_{112}\int_{b_{211}^L}^{b_{211}^U} d\theta_{211}\int_{b_{212}^L}^{b_{212}^U} d\theta_{212} = \tilde{c}\int_{0.25-\theta}^{0.25} d\theta_{112} = \tilde{c}\theta \tag{17}$$

where c is a normalizing constant and $\tilde{c}$ is c times the result of the integrals with respect to θ_{211} and θ_{212}.

Additional constraints change the form of $m_{111}(\theta)$. If there is the structural zero $\theta_{121} = 0$, then $m_{111}(\theta)$ is concentrated in just one point, i.e., $\theta_{111} = \theta_{1.1} = 0.16$ with certainty. If there is the structural zero $\theta_{122} = 0$, then $m_{111}(\theta)$ is uniform in (0, 0.16). Actually it seems there is more variability, contradicting our view. But in this case, each time θ_{111} is held fixed in θ, the parameter θ_{112} is constrained to assume just one value; consequently the joint density is less variable. Finally, an inequality constraint, as $\theta_{111} > \theta_{121}$ reduces (17) to assume values only in $(\theta_{1.1}/2 = 0.08; 0.16)$.

When the marginal distributions of (X, Y) and (X, Z) are estimated from the two samples at hand, the distribution $m_{ijk}(\theta)$ is estimated from the subset of $\boldsymbol{\theta}$ that lies on the likelihood ridge. All the other $\boldsymbol{\theta}$ are not considered.

7. References

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley.

Barry, J.T. (1988). An Investigation of Statistical Matching. Journal of Applied Statistics, 15, 275–283.

Bergsma, W.P. and Rudas, T. (2002). Marginal Models for Categorical Data. The Annals of Statistics, 30, 140–159.

Coletti, G. and Scozzafava, R. (2002). Probabilistic Logic in a Coherent Setting. Trends in Logic 15, Dordrecht: Kluwer Academic Publishers.

Csizár, I. (1975). I-divergence Geometry of Probability Distributions and Minimization Problems. The Annals of Probability, 3, 146–158.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39, 1–38.

D'Orazio, M., Di Zio, M., and Scanu, M. (2001). Statistical Matching: A Tool for Integrating Data in National Statistical Institutes. Proceedings NTTS – ETK 2001 (New Techniques and Technologies for Statistics and Exchange of Technology and Know-how) Hersonissos (Greece), 18 – 22 June, 433–441.

D'Orazio, M., Di Zio, M., and Scanu, M. (2002). Statistical Matching and Official Statistics. Rivista di Statistica Ufficiale, 1/2002, 5–24.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

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

Haberman, S. (1977). Product Models for Frequency Tables Involving Indirect Observation. Annals of Statistics, 5, 1124–1147.

Ingram, D.D., O'Hare, J., Scheuren, F., and Turek, J. (2000). Statistical Matching: A New Validation Case Study. Proceedings of the American Statistical Association, Section on Survey Research Methods,746–751.

Judge, G.G., Griffiths, W.E., Hill, R.C., and Lee, T.C. (1980). The Theory and Practice of Econometrics. New York: John Wiley.

Kadane, J.B. (1978). Some Statistical Problems in Merging Data Files. In Compendium of Tax Research, Department of Treasury, U.S. Government Printing Office, Washington D.C., 159–179 (Reprinted in 2001, Journal of Official Statistics, 17, 423–433).

Kamakura, W.A. and Wedel, M. (1997). Statistical Data Fusion. Journal of Marketing Research, 34, 485–498.

Kamakura, W.A. and Wedel, M. (2003). List Augmentation with Model Based Multiple Imputation: A Case Study Using a Mixed-Outcome Factor Model. Statistica Neerlandica, 57, 46–57.

Little, R.J.A. and Rubin, D.B. (1983). On Jointly Estimating Parameters and Missing Data by Maximising the Complete-Data Likelihood. The American Statistician, 37, 218–220.

Manski, C.F. (1995). Identification Problems in the Social Sciences. Cambridge, Massachusetts: Harvard University Press.

Moriarity, C. and Scheuren, F. (2001). Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure. Journal of Official Statistics, 17, 407–422.

Moriarity, C. and Scheuren, F. (2003). A Note on Rubin's Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation. Journal of Business and Economic Statistics, 21(1), 65–73.

Paass, G. (1986). Statistical Match: Evaluation of Existing Procedures and Improvements by Using Additional Information. In Microanalytic Simulation Models to Support Social and Financial Policy, G.H. Orcutt, and H. Quinke (eds), Amsterdam: Elsevier Science, 401–422.

Rässler, S. (2002). Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches. Lecture Notes in Statistics, New York: Springer Verlag.

Rässler, S. (2004). Data Fusion: Identification Problems, Validity, and Multiple Imputation. Austrian Journal of Statistics, 33, 153–171.

Renssen, R.H. (1998). Use of Statistical Matching Techniques in Calibration Estimation. Survey Methodology, 24, 171–183.

Rodgers, W.L. (1984). An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91–102.

Rubin, D.B. (1974). Characterizing the Estimation of Parameters in Incomplete-Data Problems. Journal of the American Statistical Association, 69, 467–474.

Rubin, D.B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics, 4, 87–94.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Singh, A.C., Mantel, H., Kinack, M., and Rowe, G. (1993). Statistical Matching: Use of Auxiliary Information as an Alternative to the Conditional Independence Assumption. Survey Methodology, 19, 59–79.

Vantaggi, B. (2003). Conditional Independence Structures and Graphical Models. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 11(5), 545–571.

Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. London: Chapman & Hall.

Winkler, W.E. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the American Statistical Association, Section on Survey Research Methods, 274–279.

Received September 2003
Revised October 2005