Incoherence correction strategies in statistical matching

Andrea Capotorti
Dip. di Matematica e Informatica,
University of Perugia, Italy
capot@dmi.unipg.it

Barbara Vantaggi
Dip. Scienze di Base e Applicate per l'Ingegneria,
University La Sapienza, Roma, Italy
vantaggi@dmmm.uniroma1.it

Abstract

We deal with the statistical matching problem and in particular we study the problem related to the managing of inconsistencies. In fact, when logical relations among the variables are present incoherence can arise in the probability evaluations. The aim of this paper is to remove such incoherences by using different methods. Specific precise distances minimization or least committal imprecise probability extensions are adopted. We compare these methods using a practical example that brings to light the peculiarities of the statistical matching problem.

Keywords. Statistical matching, incoherence, inference, specialized discrepancy measure.

1 Introduction

In several economic applications there is a need to consider different data sources and to integrate the information coming from them [3, 13, 23, 25, 26]. In particular, we deal with the so called statistical matching problem, that can be represented by the following simple situation: there are two different sources, A and B, with some overlapping variables and some variables collected only in one source. Let X represent the common variables, Y denotes the variables collected only in A, and Z those only in B. Thus, the data consist of a first sample (X, Y) and a second sample on (X, Z). In this context data are missing by design since they have been already collected separately, and to get jointly data on Y and Z would be expensive and time-consuming.

Traditionally, to cope with these problems the available data are combined with assumptions strong enough to point-identify the joint probability distribution (see references in [26]): we recall, for example, those based on a conditional independence assumption, i.e. the variables Y and Z are independent conditional on X.

However, in several situations the independence assumption is not adequate, as first raised by Sims [31] (see also [25, 28, 29, 32]). Other methods aim at incorporating auxiliary information about relationships between Y and Z to avoid or to relax conditional independence assumption (see, e.g. [32]). Although this is an important case, it is not always feasible because the required external knowledge may not be available.

Actually, since there are many distributions on (X, Y, Z) compatible with the available partial information on (X, Y) and (X, Z), it is too restrictive to consider just one of the compatible distributions, obtained perhaps by taking a specific assumption (as already noted in [14, 17, 30] and for the missing data problem [11, 22, 34]).

This problem has been faced in a coherent conditional probability setting in [35, 36]: coherence allows us to check the compatibility of partial (conditional) assessments, to manage further available knowledge, for example coming from field experts; moreover it allows us to draw inferences by considering all the compatible distributions.

A further remarkable advantage of using this approach is that we are able to consider multiple integration, that is important for real applications (for instance, see [33] for some economic Hungarian applications based on the combination of three different surveys).

Moreover, this approach [36] allows to manage logical constraints characterizing the relevant links among variables describing the phenomenon. In particular, in [36] it is proved that when there is no logical constraint among the variables, coherence is always satisfied by also requiring conditional independence, then this hypothesis is legitimate from a syntactical point of view (even if it is useful to look for all compatible coherent extensions). On the other hand, when logical constraints are present it is necessary to check global coherence of the relevant partial assessments drawn from the different sources and if coherence is not satisfied we need to remove incoherences. In [35] this is done by looking for the "minimal" incoherent assessments and to remove them in order to restore coherence by using the $L1$ norm.

The aim of this paper is to deal with incoherences and to look for the coherent assessment "closest" to the given one with respect to different distances ($L1, L2$, Kulback-Leibler divergence, discrepancy). Then, when coherence is restored we can draw inference: for each (conditional) event we can directly build the interval of all coherent probability values solely on the base of a partial assessment, i.e. it is not needed to artificially fulfill the missing values of the data base. It is important to remark that the interval bounds are computed analytically.

Actually, our aim is in the same line of those based on multiple imputation [30] and its extension [26], which aims at approximating the lower and upper bounds for the quantities of interest in the multinormal setting. A similar approximation for these bounds is carried out in [14] on the base of maximum likelihood approach.

To let this paper be as much as possible self-contained, in Section 2 we introduce the basic notions and characteristic of coherent conditional partial assessments, either based on precise $\mathbf{p}$ or on imprecise **lub** evaluations given on a finite domain $\mathcal{E}$. Coherence of an assessment is required to perform a sound inference that for partial assessments coincide with a coherent extension. Hence also basic extension notions, both for the precise and imprecise context, are given. Afterwards in Section 3 the main (pseudo)distances between conditional assessments are introduced. It is in fact thanks to their minimization that consistent correction of incoherent assessments will be possible. Such (pseudo)distances can be based on geometrical properties, e.g. $L1$ and $L2$ norms, or on information theoretic foundation, e.g. KL divergence, or can derive from proper scoring rules, e.g. discrepancy Δ, suitably tailored for partial conditional assessments. In Subsection 3.1 it is sketched an alternative way of restoring consistency: whenever it is possible to identify a coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$, it can be coherently extended to the rest of the initial domain $\mathcal{F} = \mathcal{E} \setminus \mathcal{G}$. This inevitably produces an imprecise conditional probability assessment. Subsequently, in Section 4, the statistical matching problem is reformulated inside a conditional probability assessment framework and conditions are given that guarantee the coherence of the whole assessment. On the contrary, whenever there are logical constraints among the variables under investigation, even starting from separately coherent sources of information, the whole assessment could result incoherent. In Section 5 this is well described by a simple example. This section is the core of our contribution, where the previous concepts are merged together and the two main approaches for inconsistency correction, the minimization of (pseudo)distances or the extension of a coherent sub-assessment, are specialized to the statistical matching problem. It is also shown how the peculiarity of the statistical matching suggests a specialization of the general discrepancy Δ into a peculiar one Δ_{mix}. This new discrepancy is a mixture of the original one applied to the different scenarios and the consequent inconsistency correction, obtained by its minimization, leaves unchanged the marginal distribution of the common variables X. To better show the advantages and drawbacks of the proposed methods, in Section 6 we introduce an example built from data taken from [14]. The final short concluding Section 7 sums up the proposed methodologies.

2 Preliminaries about coherent conditional probability

Whenever several sources of information, that could represent expert's opinions and/or knowledge bases, are merged together, we can generally start to deal with an overall domain $\mathcal{E} = [E_1|H_1, \dots, E_n|H_n]$.

The events E_i's represent the situations under consideration, while the H_i's usually represent the different contexts, or scenarios, under which the E_i's are evaluated.

The basic events $E_1, \dots, E_n, H_1, \dots, H_n$ can be endowed with logical constraints, that represent dependencies among particular configurations of them (e.g. incompatibilities, implications, partial or total coincidences, etc.).

In the following E_iH_i will denote the logical connection "E_i and H_i" ($E_i \wedge H_i$), E_i^c will indicate "not E_i", the contrary of E_i, and the event $H^0 = \bigvee_{i=1}^n H_i$ will represent the whole set of contexts.

Starting with the basic events $E_1, \dots, E_n, H_1, \dots, H_n$, it is possible to span a sample space $\Omega = \{\omega_1, \dots, \omega_k\}$, where ω_j represents a generic atom that is the elementary element in the algebra generated by the E_i and H_i. Note that the sample space Ω, together with H^0, are not part of the assessment but only auxiliary tools.

Every probability mass function $\alpha : \mathcal{P}(\Omega) \to \mathbb{R}$ corresponds to a non-negative vector $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_k]$, with $\alpha_j = \alpha(\omega_j)$, then for every event E it results $\alpha(E) = \sum_{\omega_j \subseteq E} \alpha_j$.
We need to introduce a nested hierarchy among probability distributions sets:

- $\mathcal{A} = \left\{ \boldsymbol{\alpha}, \sum_1^k \alpha_i = 1, \alpha_j \geq 0, j = 1, \ldots, k \right\}$;
- $\mathcal{A}_0 = \left\{ \boldsymbol{\alpha} \in \mathcal{A} | \alpha(H^0) = 1 \right\}$;
- $\mathcal{A}_1 = \{ \boldsymbol{\alpha} \in \mathcal{A}_0 | \alpha(H_i) > 0, i = 1, \ldots, n \}$.

It is easy to see that the set $\mathcal{A}_1$ is a convex set and $\mathcal{A}_0$ is the closure of $\mathcal{A}_1$ in the usual topology.

We focus our attention on coherent (conditional) probability assessments $\mathbf{p}$, that can be reduced to the compatibility with a conditional probabilities, as introduced by Dubins [15] and De Finetti [12] (see also Krauss [18] and Rényi [27]).

Definition 1 *Let $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$ be an arbitrary set of conditional events, an assessment $\mathbf{p}$ on $\mathcal{E}$ is said to be a coherent conditional probability if there exists a conditional probability $P(\cdot|\cdot)$ defined on $\mathcal{B} \times (\mathcal{B} \setminus \emptyset)$ (with $\mathcal{B}$ the algebra spanned by $E_1, H_1, ..., E_n, H_n$) which restriction to $\mathcal{E}$ coincides with $\mathbf{p}$.*

Every probability distribution $\boldsymbol{\alpha} \in \mathcal{A}_1$ generates a coherent conditional probability assessment $\mathbf{q}_{\boldsymbol{\alpha}}$ on $\mathcal{E}$ through the usual formula

$$q_{\alpha_i} = \sum_{\omega_j \subseteq E_i H_i} \alpha_j / \sum_{\omega_j \subseteq H_i} \alpha_j \quad \text{for all } i = 1, \ldots, n. \quad (1)$$

Note that $\mathbf{q}_{\boldsymbol{\alpha}}$ is a continuous function of $\boldsymbol{\alpha}$ when $\boldsymbol{\alpha} \in \mathcal{A}_1$. When $\boldsymbol{\alpha} \in \mathcal{A}_0$, the previous formula (1) defines $\mathbf{q}_{\boldsymbol{\alpha}}$ only on

$$\mathcal{E}_{\boldsymbol{\alpha}} := \{ E_i|H_i \in \mathcal{E}, \alpha(H_i) > 0 \}. \quad (2)$$

To cover the case of conditioning events with null probability, in fact we need to resort to a suitable class of probability distributions $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ agreeing with $P(\cdot|\cdot)$ (for more details refer to the characterization theorem reported e.g. in [9, 10]).

Coherence is crucial since it is a prerequisite for a sound inference, that means extension of the given assessment to any new conditional event. In fact the following theorem, essentially due to [12], holds:

Theorem 1 *Let $\mathbf{p}$ be an assessment on an arbitrary family $\mathcal{E}$; then there exists a (possibly not unique) coherent extension of $\mathbf{p}$ to any family $\mathcal{K} \supset \mathcal{E}$* if and only if $\mathbf{p}$ *is a coherent conditional probability on $\mathcal{E}$.*

Moreover, if $\mathbf{p}$ is a coherent conditional probability on $\mathcal{E}$, then the coherent probability values for any conditional event $F|K \in \mathcal{K} \setminus \mathcal{E}$ belong to a closed interval $[\underline{p_{F|K}}, \overline{p_{F|K}}]$.

The aforementioned coherent interval $[\underline{p_{F|K}}, \overline{p_{F|K}}]$ can be obtained by solving specific linear optimization problems (for details refer again to [10]) based on suitable classes of probability distributions $\{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l\}$ agreeing with $\mathbf{p}$.

The notion of coherence also apply to imprecise conditional assessments, i.e. whenever the numerical part of the assessment is elicited through interval values

$$\mathbf{lub} = ([lb_1, ub_1], \ldots, [lb_n, ub_n]). \quad (3)$$

Of course, some of the intervals $[lb_i, ub_i]$ could degenerate to a precise value p_i.

For assessments such as $(\mathcal{E}, \mathbf{lub})$, although defined on finite spaces, there could be different kinds of consistency requirements (for a detailed exposition, among others, refer to [24]). The basic consistency notion is the so called *avoiding of partial loss*, while in this paper we focus on the most stringent one: *(strong) coherence*. By taking into account a Bayesian sensitivity analysis interpretation, coherent lower-upper conditional probability assessments $(\mathcal{E}, \mathbf{lub})$ are such that intervals' lower (upper) extremes lb_i (ub_i) can be obtained as lower (upper) envelopes of sets of coherent precise conditional probability assessments on $\mathcal{E}$. It follows that to have a coherent lower-upper assessment $(\mathcal{E}, \mathbf{lub})$, there should exist a *set* of probability *classes* $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_l$ such that they induce probabilities for the $E_i|H_i$ inside the ranges $[lb_i, ub_i]$, and moreover each lower (lbi_s) and upper (ubi_s) bound on a conditional event is attained in at least one distribution.

Also, starting from a coherent lower-upper assessment $(\mathcal{E}, \mathbf{lub})$, it is possible to infer coherent bounds $[\underline{p_{F|K}}, \overline{p_{F|K}}]$ for the probability of any target conditional event $F|K$ through specific sequences of linear optimization problems and/or satisfiability of logical configurations (for details refer to [4]).

3 Coherent adjustments

Given an incoherent conditional probability assessment, for example, on a domain arising from the merging of separately coherent partial probability assessments, we need to restore coherence in a way to preserve, as much as possible, the information on the initial assessments, without introducing exogenous information. This goal is obtained generally by minimizing some kind of distance among partial conditional assessments.

(Pseudo)distances among probability distributions are usually defined through divergencies (e.g. Euclidean distance, Kulback-Leibler divergence, Csiszár f-divergences, etc.). Some of them can be applied only among unconditional probability distributions; others could be applied in our context of partial conditional assessments, but could have no probabilistic justification, being purely geometrical tools.

Given two conditional assessments $\mathbf{p} = [p_1, \ldots, p_n]$ and $\mathbf{q} = [q_1, \ldots, q_n]$ on the same set of conditional events $\mathcal{E}$, the most widely adopted divergencies among them are:

1. $L1(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |q_i - p_i|$;
2. $L2(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i - p_i)^2$;
3. $KL(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} (q_i \ln(q_i/p_i) - q_i + p_i)$.

$L1$ and $L2$ are usual metric distances, endowed with all their geometric properties, but until now remain without an intuitive probabilistic interpretation for conditional assessments. Moreover, their use in conditional context could lead to numerical troubles due to non-convexity of coherent assessments, as the following simple example borrowed from [2] shows:

Example 1 *Consider $\mathcal{E} = [A|H, B|AH, AB|H]$ with A, B, H logically independent. Hence the sample space is composed by 8 atoms, 4 of them inside $H^0 \equiv H$. The set of coherent assessments $\mathcal{Q}_\mathcal{E}$ is formed by the triples $[q_1, q_2, q_3] \in [0,1]^3$ with $q_3 = q_1 q_2$.*

Then, the set $\mathcal{Q}_\mathcal{E}$ is evidently non-convex.

KL is the so called logarithmic Bregman divergence. In the unconditional framework, such divergence is the most frequently adopted, because of its information theoretic properties. In fact, it generalizes the well known Kulback-Leibler divergence [19] to partial assessments. Anyhow, it is known that this Bregman divergence is generated by a logarithmic scoring rule that has a peculiarity that in some cases it is better to avoid: it evaluates only the events that occur, without considering those that turn out to be false.

To overcome this characteristic and to encompass the need of considering the conditional framework where the assessment is given, recently in [6, 8] for partial conditional assessments $\mathbf{v} = [v_1, \ldots, v_n] \in (0,1)^n$ over $\mathcal{E} = [E_1|H_1, \ldots, E_n|H_n]$, the following random variable has been proposed as scoring rule:

$$S(\mathbf{v}) := \sum_{i=1}^{n} |E_i H_i| \ln v_i + \sum_{i=1}^{n} |E_i^c H_i| \ln(1 - v_i) \quad (4)$$

with $|\cdot|$ the indicator function of unconditional events.

The motivation of such a score is that the assessor "loses less" the higher the probabilities are of occurring events, and at the same time, the lower the probabilities of events are, which do not occur. The values assessed on events that turn out to be undetermined do not influence the score. Such a score $S(\mathbf{v})$ is an extension to partial and conditional probability assessments of the "total-log proper scoring rule" for probability distributions proposed by Lad in [20, pag. 355].

By considering the difference between the expected penalties suffered by the two evaluations $\mathbf{p}$ and $\mathbf{q}_{\boldsymbol{\alpha}}$ as distance criterion, it is possible to define the "discrepancy" $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ between a partial conditional assessment $\mathbf{p}$ over $\mathcal{E}$ and a distribution $\boldsymbol{\alpha} \in \mathcal{A}_0$ through the expression

$$\sum_{i|\alpha(H_i)>0} \alpha(H_i) \left(q_i \ln \frac{q_i}{p_i} + (1 - q_i) \ln \frac{(1 - q_i)}{(1 - p_i)} \right) \quad (5)$$

taking the convention $0 \ln(0) = 0$. Note that in $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ each term is weighted by $\alpha(H_i)$, which reflects the "relevance" of each context H_i with respect to all the assessments.

The main idea is to take as coherent correction of $\mathbf{p}$ the assessment $\mathbf{q}_\mathbf{p} \equiv \mathbf{q}_{\widetilde{\boldsymbol{\alpha}}}$ generated by the distribution $\widetilde{\boldsymbol{\alpha}}$ solution of the nonlinear optimization program

$$\min_{\boldsymbol{\alpha} \in \mathcal{A}^0} \Delta(\mathbf{p}, \boldsymbol{\alpha}). \quad (6)$$

The motivation for this choice is that (intuitively) the assessor of $\mathbf{p}$ would expect to suffer the penalty $S(\mathbf{p})$, hence we select the coherent assessment $\mathbf{q}_\mathbf{p}$ that has a (probabilistic) expected score as close as possible.

In [8] it is formally proved that $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ is a non negative function on $\mathcal{A}_0$ and that $\Delta(\mathbf{p}, \boldsymbol{\alpha}) = 0$ if and only if $\mathbf{p} = \mathbf{q}_{\boldsymbol{\alpha}}$; moreover $\Delta(\mathbf{p}, \cdot)$ admits a minimum on $\mathcal{A}_0$. Finally if $\boldsymbol{\alpha}, \boldsymbol{\alpha}^0 \in \mathcal{A}_0$ are distributions that minimize $\Delta(\mathbf{p}, \cdot)$, then for all $i \in \{1, \ldots, n\}$ such that $\alpha(H_i) > 0$ and $\alpha^0(H_i) > 0$ we have $(\mathbf{q}_\alpha)_i = (\mathbf{q}_{\alpha^0})_i$; in particular if $\Delta(\mathbf{p}, \cdot)$ attains its minimum value on $\mathcal{A}_1$ then there is a unique coherent assessment $\mathbf{q}_{\underline{\alpha}}$ such that $\Delta(\mathbf{p}, \underline{\alpha})$ is minimum. On the contrary, if the minimum is attained in $\mathcal{A}_0 \setminus \mathcal{A}_1$, i.e. there exists some conditioning event forced to have null probability, the optimization program (6) can be iterated by restricting the assessment only on the different "zero layers" (for details refer again to [8]). Moreover, the discrepancy measure $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ can be used to correct incoherent assessments and to aggregate expert opinions [6, 8]. $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ can even be applied to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities [7] but this feature will not be used here since the statistical matching analysis will be based on a precise initial assessment $\mathbf{p}$. Imprecise probabilities can appear whenever a consistent sub-assessment is selected and it is coherently extended to the rest of the domain, as it is shown in the next sub-section.

3.1 Coherent Extension

Another possibility to adjust the initially incoherent assessment $(\mathcal{E}, \mathbf{p})$ could be to determine a coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ and coherently extend it to the rest $\mathcal{F} = \mathcal{E} \setminus \mathcal{G}$ as prescribed by the generalized Bayesian updating scheme (see e.g. [9, 10, 36] among others). Since, in general, coherent extension produces intervals of plausible values, with this approach the whole assessment turns out to be imprecise due to the interval values $((\mathcal{F}, [\underline{\mathbf{p}_{\mathcal{F}}}, \overline{\mathbf{p}_{\mathcal{F}}}]))$. Also in such a situation, inference can pe performed again through the generalized Bayesian updating scheme but applied to imprecise evaluations (see e.g. [1, 4] among others). Whenever such inferences are too vague, i.e. when the intervals are very wide (close to [0,1]), they can be eventually reduced by a procedure proposed in [5] that enucleates coherent cores, i.e. surely coherent subintervals with highest degree of support. This is motivated by the fact that, in general, not all the subintervals of the extensions are coherent, whereas this is guaranteed by the choice of such coherent cores since they are *total* coherent (for this stronger consistency notion refer e.g. to [16]).

The choice of the coherent sub-assessment $(\mathcal{G}, \mathbf{p}_{|\mathcal{G}})$ should follow some criterion, since it could not be uniquely determined. Anyhow, for the specific application to statistical matching that is the scope of the present paper, such a choice comes quite naturally since in [35] it has been shown that it is possible to detect the incoherent sub-assessment $(\mathcal{F}, \mathbf{p}_{|\mathcal{F}})$ with minimal cardinality.

4 Integration of sources in a coherent setting

We briefly describe how the problem of integration of sources, named statistical matching, can be formalized in the coherent conditional probability setting. In particular here we refer to the case of two sources as already described in [35], while the case of more sources has been studied in [36].

Let us denote by $(X_1, Y_1), ..., (X_{n_A}, Y_{n_A})$ and by $(X_{n_A+1}, Z_{n_A+1}), ..., (X_{n_A+n_B}, Z_{n_A+n_B})$ two random samples (with a finite range) related to two sources A and B. We suppose that the two samples both concern the same population of interest and are drawn according to the same sampling scheme. We can regard, under the above conditions, $(X_1, Y_1), ..., (X_{n_A}, Y_{n_A})$ (analogously $(X_{n_A+1}, Z_{n_A+1}), ..., (X_{n_A+n_B}, Z_{n_A+n_B})$) exchangeable, as well as the sequence $X_1, ..., X_{n_A}$, $X_{n_A+1}, ..., X_{n_A+n_B}$.

We can elicit from the two files the relevant probability values: from file A the conditional probabilities

$$\mathcal{y}_{j|i} = P_{Y|(X=x_i)}(Y = y_j), \tag{7}$$

that the next unit has $Y = y_j$ on the hypothesis that $(X = x_i)$ (for any x_i taken by X), and analogously from file B the conditional probability values

$$\mathcal{z}_{k|i} = P_{Z|(X=x_i)}(Z = z_k). \tag{8}$$

Moreover, from data on both files we can evaluate

$$\mathcal{x}_i = P_X(X = x_i). \tag{9}$$

Given $\mathcal{y}_{j|i}, \mathcal{z}_{k|i}, \mathcal{x}_i$, for any i, j, k, one needs to check coherence of the whole assessment $(\mathcal{E}, \mathbf{p})$, that is

$$\begin{array}{l}\mathcal{E} = \\ \left\{\begin{array}{r}(X = x_i),\ (Y = y_j)|(X = x_i),\ (Z = z_k)|(X = x_i) \\ \text{for any } x_i, y_j, z_k\end{array}\right\}, \\ \mathbf{p} = \{\mathcal{x}_i, \mathcal{y}_{j|i}, \mathcal{z}_{k|i}\}_{i,j,k} \quad .\end{array} \tag{10}$$

Now we recall the result proved in [36], that claims that when the partitions $\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$ associated to the variables are logically independent (i.e. for any $A \in \mathcal{E}_X,\ B \in \mathcal{E}_Y,\ C \in \mathcal{E}_Z,\ A \wedge B \wedge C \neq \emptyset$) coherence is assured.

Theorem 2 *Let X, Y, Z be three finite random variables and consider the following three coherent assessments $\{P_X(X = x_i)\}_i$, $\{P_{Y|X=x_i}(Y = y_j)\}_j$ and $\{P_{Z|X=x_i}(Z = z_k)\}_k$.*

Then the assessment

$$\{P_X(X = x_i)\,,\ P_{Y|X=x_i}(Y = y_j) : \text{ for any } x_i, y_j\}$$

(analogously $\{P_X(X = x_i), P_{Z|X=x_i}(Z = z_k)$: for any $x_i, z_k\}$) is coherent.

Moreover, if the partitions $\mathcal{E}_Y, \mathcal{E}_Z$ are logically independent with respect to $\mathcal{E}_X$ (i.e. $(X = x_i, Y = y_j, Z = z_k)$ is possible for any value x_i of X, y_j of Y, z_k of Z s.t. the events $(X = x_i, Y = y_j)$ and $(X = x_i, Z = z_k)$ are possible), then the whole assessment (10) is coherent.

On the other hand, when there are some logical constraints among the variables Y and Z, the coherence of the whole assessment (10) is not assured by coherence of the single assessments (7-9) (see [35]). Notice that the need of managing logical constraints arises from practical applications [14].

5 Removing inconsistencies in statistical matching

We have now all the elements to specialize the general approaches for inconsistencies correction described in

Section 3 for the specific setting of the statistical matching as depicted in the previous Section 4.

The starting point is that the whole assessment (10) is not coherent, then inconsistencies must be detected in order to restore coherence. This kind of problem has already been studied (e.g. see [21]) in combining assessments given by different experts: the approach to the identification and reconciliation of incoherence uses an external observer equipped with a prior distribution and likelihood function. Actually, this approach does not seem suitable in the context of statistical matching because of the lack of information on the variables not jointly observed, so that the prior distribution cannot be updated and the likelihood function has a flat ridge (as already noted in [30]). Hence we propose a different method: to restore coherence we can easily find the minimal restriction of the whole assessment which is not coherent (as proposed in [36]) and adjust it by a specialization of the techniques presented in Section 3. Let us see it into details.

As claimed by Theorem 2, in statistical matching incoherences are related to conditional events with the same conditioning event $(X = x_i)$. Hence the check of coherence of the whole assessments (10) can be reduced to the check of coherence for the sub-assessments

$$\left\{ \mathcal{Y}_{j|i}, \, \mathcal{Z}_{k|i} : \text{ for fixed } i \text{ and any } j, k \right\}. \tag{11}$$

Once not coherent sub-assessments of type (11) have been disclosed, they can be adjusted by finding coherent values that minimize some of the (pseudo)distances presented in Section 3.

Whereas classical distances - $L1$, $L2$ and KL - can be directly applied to such minimal incoherent restriction since their arguments are directly the conditional probabilities, for the discrepancy $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ a "reformulation" is required. In fact, we require that its expression (5) specifically acts on values for any conditioning events $(X = x_i)$. This is possible by considering the following mixture of discrepancies $\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_i\}_i)$:

$$\sum_i \mathcal{X}_i \left[\sum_j \left(q_{j|i}^{\alpha_i} \ln \frac{q_{j|i}^{\alpha_i}}{\mathcal{Y}_{j|i}} + (1 - q_{j|i}^{\alpha_i}) \ln \frac{(1 - q_{j|i}^{\alpha_i})}{(1 - \mathcal{Y}_{j|i})} \right) + \right.$$
$$\left. + \sum_k \left(q_{k|i}^{\alpha_i} \ln \frac{q_{k|i}^{\alpha_i}}{\mathcal{Z}_{k|i}} + (1 - q_{k|i}^{\alpha_i}) \ln \frac{(1 - q_{k|i}^{\alpha_i})}{(1 - \mathcal{Z}_{k|i})} \right) \right] \tag{12}$$

where each distribution α_i works just on the sample space spanned by the conditional events $\{(Y = y_j)|(X = x_i), (Z = z_k)|(X = x_i)\}$, it is constrained to fulfill the normalizing condition

$$\alpha_i(X = x_i) = \mathcal{X}_i, \tag{13}$$

and generates the conditional probabilities

$$q_{j|i}^{\alpha_i} = \frac{\alpha_i(Y = y_j)}{\alpha_i(X = x_i)} \quad q_{k|i}^{\alpha_i} = \frac{\alpha_i(Z = z_k)}{\alpha_i(X = x_i)}. \tag{14}$$

As already mentioned, coherence of the overall assessment $(\mathcal{E}, \mathbf{q})$, with

$$\mathbf{q} = \{\mathcal{X}_i, q_{j|i}^{\alpha_i}, q_{k|i}^{\alpha_i}\}_{i,j,k}$$

is guaranteed by Theorem 2.

Since the specialized discrepancy defined in equation (12) is a mixture of discrepancies, each one working on a specific scenario $(X = x_i)$, its use in an optimization program like (6) allows to adjust only the values inside sub-domains of $\mathcal{E}$ conditioned to scenarios $(X = x_i)$ where some incoherence appear, without changing the other values. This characteristic differentiates the specialized discrepancy (12) from the original discrepancy (5), as the following simple example shows:

Example 2 *Let $\{x_1, x_2\}$, $\{y_1, y_2, y_3\}$, $\{z_1, z_2, z_3\}$ be the sample space of three r.v. X, Y, Z with constraints*

$$(Z = z_1) \wedge ((Y = y_1) \vee (Y = y_2)) = \emptyset$$

and

$$(Z = z_2) \wedge (Y = y_1) = \emptyset.$$

Consider the following conditional assessment $\mathbf{p}$*:*

$$\begin{array}{lll} \mathcal{X}_1 = \frac{1}{3} & \mathcal{X}_2 = \frac{2}{3} & \\ \mathcal{Y}_{1|1} = \frac{387}{1111} & \mathcal{Y}_{2|1} = \frac{102}{1111} & \mathcal{Y}_{3|1} = \frac{622}{1111} \\ \mathcal{Y}_{1|2} = \frac{2}{3} & \mathcal{Y}_{2|2} = 0 & \mathcal{Y}_{3|2} = \frac{1}{3} \\ \mathcal{Z}_{1|1} = \frac{179}{1108} & \mathcal{Z}_{2|1} = \frac{443}{1108} & \mathcal{Z}_{3|1} = \frac{486}{1108} \\ \mathcal{Z}_{1|2} = \frac{2}{3} & \mathcal{Z}_{2|2} = \frac{1}{9} & \mathcal{Z}_{3|2} = \frac{2}{9} \end{array}.$$

It is easy to check that $\mathbf{p}$ *on events $(X = x_i)$ is coherent, as well as $\mathcal{Y}_{j|i} = P(Y = y_j|X = x_i)$ (and analogously $\mathcal{Z}_{k|i} = P(Z = z_k|X = x_i)$) for any $(X = x_i)$. However, the whole assessment is not coherent, and incoherence is localized on events conditioned to $(X = x_2)$.*

By applying either $\Delta(\mathbf{p}, \boldsymbol{\alpha})$ or $\Delta_{mix}(\mathbf{p}, \{\boldsymbol{\alpha}_i\}_i)$ the same correction on those values is induced (see Table 1), whereas with the former also the unconditional values for $P(X = x_i)$ are modified, even if they are coherent.

Note that, with such specialized discrepancy, the subdomains, where incoherence must be removed, are implicitly detected, without the need of a preliminary

$\mathcal{E}$	P	Δ	Δ_{mix}
$X = x_1$	0.3333	**0.3726**	-
$X = x_2$	0.6667	**0.6274**	-
$Y = y_1 \vert X = x_1$	0.3483	0.3483	0.3483
$Y = y_2 \vert X = x_1$	0.0918	0.0918	0.0918
$Y = y_3 \vert X = x_1$	0.5599	0.5599	0.5599
$Z = z_1 \vert X = x_1$	0.1616	0.1616	0.1616
$Z = z_2 \vert X = x_1$	0.3998	0.3998	0.3998
$Z = z_3 \vert X = x_1$	0.4386	0.4386	0.4386
$Y = y_1 \vert X = x_2$	0.6667	*0.4156*	*0.4156*
$Y = y_2 \vert X = x_2$	0	*0.0996*	*0.0996*
$Y = y_3 \vert X = x_2$	0.3333	*0.4848*	*0.4848*
$Z = z_1 \vert X = x_2$	0.6667	*0.4848*	*0.4848*
$Z = z_2 \vert X = x_2$	0.1111	*0.0996*	*0.0996*
$Z = z_3 \vert X = x_2$	0.2222	*0.4156*	*0.4156*

Table 1: Correction comparison for Example 2. In boldface changes associated to unconditional events, while in italic changes associated to conditional ones

inspection of the assessment $(\mathcal{E}, \mathbf{p})$. Moreover the adjustments are weighted by the relevance of the scenarios expressed through the $\mathcal{x}_i$'s in (12).

From these data we can also get a comparison between Δ and Δ_{mix}: actually Δ also changes the probability distribution on $(X = x_i)$'s in order to reduce the minimum value taken from Δ even if Theorem 2 assures the separate coherence of the probability assessments $(\mathcal{x}_i, \mathcal{y}_{j|i})$ and $(\mathcal{x}_i, \mathcal{z}_{j|i})$, for any $i = 1, 2$ and $j = 1, 2, 3$. Then, we can stress that for the statistical matching problem Δ_{mix} seems to be more appropriate than Δ.

Another criterion (further than the quoted ones based on $L1, L2, KL$ minimizations) for restoring coherence could be based on the maximum likelihood criterion: when the evaluations are obtained through the maximum likelihood criterion, we can maximize the "partial likelihood function" on the set of events generating incoherence. Also in this situation we have an optimization problem with a non-linear objective function and a set of linear constraints.

Note that if we apply this criterion to data in Example 2 the marginal distribution of X does not change and the adjustment is localized on the assessment over $(X = x_2)$, analogously to what happens with Δ_{mix}. We have not reported these values on Table 1 because the aim of the example is just to stress the difference between Δ and Δ_{mix}. Explicit results of the maximum likelihood criterion will appear in the next section.

6 A practical example

In order to show our proposal we develop an example with data taken from [14] and studied also in [36]. The data are a subset of 2313 employees (people at least 15 years old) extracted from 2000 pilot survey of the Italian Population and Household Census. Three categorical variables have been analyzed: Age, Educational Level and Professional Status. In file A, containing 1148 units, the variables Age and Professional Status are observed, while file B, consisting of 1165 observations, the variables Age and Educational Level are considered. The variables are grouped in homogeneous response categories as follows: A_1=15-17 years old, A_2=18-22 years old, A_3=23-64 years old, A_4=more than 65 ; E_1=None or compulsory school, E_2=Vocational school, E_3=Secondary school, E_4=Degree; S_1=Manager, S_2=Clerk, S_3=Worker.

Logical constraints between the variables Age and Educational level (Age and Professional Status) are denoted by the symbol "–" (to be distinguished from the zero frequencies) in Table 2 (Table 3): for example, in Italy a 17 years old person cannot have a University degree. Tables 2 and 3 show, respectively, the distribution of Age and Professional Status in file A, and in file B that related to Age and Educational Level.

	Prof. Status			
Age	S_1	S_2	S_3	Tot.
A_1	–	–	9	9
A_2	–	5	17	22
A_3	179	443	486	1108
A_4	6	1	2	9
Tot.	185	449	514	1148

Table 2: Distribution of Age and Professional Status in file A.

	Educ. level				
Age	E_1	E_2	E_3	E_4	Tot.
A_1	6	0	–	–	6
A_2	14	6	13	–	33
A_3	387	102	464	158	1111
A_4	10	0	3	2	15
Tot.	417	108	480	160	1165

Table 3: Distribution of Age and Educational level in file B.

Additional logical constraints involving both the variables Professional Status and Educational level are the following ones:

$$S_1 \wedge (E_1 \vee E_2) = \emptyset \text{ and } S_2 \wedge E_1 = \emptyset.$$

By considering the frequencies (that, whenever coherent, correspond also to the maximum likelihood estimations) as evaluation of the relevant conditional probabilities, we get the assessment reported in Table 4. Such conditional probability assessment is not

	A_1	A_2	A_3	A_4
$P(\cdot)$	0.0065	0.0238	0.9594	0.0104
$P(S_1\|\cdot)$	$--$	$--$	0.1616	0.6667
$P(S_2\|\cdot)$	$--$	0.2273	0.3913	0.1111
$P(S_3\|\cdot)$	1	0.7727	0.4293	0.2222
$P(E_1\|\cdot)$	1	0.4242	0.3419	0.6667
$P(E_2\|\cdot)$	0	0.1818	0.0918	0
$P(E_3\|\cdot)$	$--$	0.3940	0.4176	0.2
$P(E_4\|\cdot)$	$--$	$--$	0.1422	0.1333

Table 4: Conditional probability assessment elicited from frequencies of Tab.2 and Tab.3.

coherent as shown in [36]. The incoherencies need to be identified and removed. It comes out that $P(\cdot|A_4)$ is not coherent since from logical constraints between Educational Level and Professional Status it follows $E_1 \wedge S_1 = \emptyset$ and $E_1 \subseteq S_3$, respectively, while from Table 4 result $P(E_1|A_4) + P(S_1|A_4) + P(S_3|A_4) > 1$ and $P(E_1|A_4) > P(S_1|A_4)$.

Then, we could either identify, as proposed in [36], the minimal set of conditional events involved in incoherencies that is $\mathcal{F} = \{E_1|A_4, S_1|A_4, S_3|A_4\}$, or adjust, with respect to a suitable distance, the whole distribution on Professional Status and Educational Level conditioned to A_4.

Different corrections are considered and the results are shown in Table 5, where

- $L1_{|\mathcal{F}}$ gives the solution proposed in [36] by minimizing $L1$ distance only among $\mathcal{F}$, the minimal incoherent subset of $\mathcal{E}$;
- $L1_{|A4}$, $L2_{|A4}$, $KL_{|A4}$ gives the solutions obtained by minimizing usual distances discussed in Section 3 only among events conditioned to A_4;
- Δ_{MIX} gives the solution obtained by minimizing the specific discrepancy (12);
- ML gives the maximum likelihood estimation;
- $IP_{\mathcal{E}\setminus\mathcal{F}}$ gives the coherent lower-upper extension induced by the given assessment on $\mathcal{E} \setminus \mathcal{F}$;
- $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$ gives the coherent lower-upper extension induced by the given assessment on $\mathcal{E} \setminus \{S_i|A_4, E_j|A_4 \,:\, i = 1, 2, 3\,;\, j = 1, ..., 4\}$;
- the last column gives the extensions of the respective corrections on the inference target $S_3|E_4$ with the respective "core" rows showing the total coherent sub-interval extension with maximum support in line with [5].

Note that only the values conditioned to A_4 are reported, those involved in the incoherence (the other 18 values remaining the same as the given assessment $\mathbf{p}$).

Firstly, we compare the rows related to remove the minimal set of incoherence, and it seems that $L1_{|\mathcal{F}}$ and $IP_{\mathcal{E}\setminus\mathcal{F}}$ perform similarly. Even though we can observe a drastic change on the probability values, mainly induced by removing not all the set of conditioning events with conditioning A_4 but just a subset (a minimal subset), they induce quite reasonable inference bounds. In particular, the imprecise adjustment $IP_{\mathcal{E}\setminus\mathcal{F}}$ performs quite well. In fact it induces inference bounds for $S_3|E_4$ similar to the precise corrections with the advantage of having the possibility to enucleate the "core" sub-interval. This sub-interval, even though it remains quite vague, has the positive aspect of bounding away from zero the lower probability, and this is seen very often as a positive aspect.

Note that $L1_{|A4}$ and ML give similar results and in particular they leave to 0 the probability of $E_2|A_4$ since the absence of observations in the original data. And the impossibility to change null values is one of the peculiarities of maximum likelihood criterion.

On the other hand, we observe that precise adjustments on the whole assessment conditioned to A_4 have all quite similar behaviors for the other distances taken into consideration, and in particular they also modify the assessment related to $E_2|A_4$, where there is no observation.

The advantage of Δ_{mix} correction is its automatic localization of the scenarios (in this specific example A_4) where the adjustment can be performed and their relative importance expressed by the unconditional probabilities x_i. Note that we apply Δ_{mix}, instead of Δ, in order to avoid any change on the probability distribution of X, that is coherent with any conditional probability on $Y|(X = x)$ (or equivalently $Z|(X = x)$), for any x, as shown in Theorem 2. In fact, Δ tighten to change also the distribution of X (through the weights) in order to reduce the inconsistencies, as shown in Example 2.

On the other hand, the wider imprecise correction $IP_{\mathcal{E}\setminus\{\cdot|A_4\}}$, being the one with less assumption requirement, surely performs worst. Its vagueness on the values conditioned on A_4 is due to freedom induced by the coherence characterization, and this re-

	$S_1\|A_4$	$S_2\|A_4$	$S_3\|A_4$	$E_1\|A_4$	$E_2\|A_4$	$E_3\|A_4$	$E_4\|A_4$	$S_3\|E_4$
p	0.6667	0.1111	0.2222	0.6667	0	0.2000	0.1333	$\emptyset$
$L1_{\|\mathcal{F}}$	0.2222	-	0.6667	0.6667	-	-	-	[0,0.6285]
$L1_{\|A4}$	0.5266	0	0.4734	0.4734	0	0.2836	0.2431	[0,0.6234]
$L2_{\|A4}$	0.5333	0.0389	0.4278	0.4278	0.0389	0.3	0.2333	[0,0.6238]
$KL_{\|A4}$	0.4856	0.1179	0.3965	0.3965	0.1179	0.2914	0.1942	[0,0.6257]
Δ_{mix}	0.4985	0.0939	0.4077	0.4077	0.0939	0.2943	0.2042	[0,0.6252]
ML	0.4286	0.0714	0.5000	0.5000	0	0.3000	0.2000	[0,0.6254]
$IP_{\mathcal{E}\setminus\mathcal{F}}$	[0 , 0.2222]	-	[0.6667 0.8889]	-	-	-	-	[0,0.6386]
core								[0.0017,0.6286]
$IP_{\mathcal{E}\setminus\{\cdot\|A_4\}}$	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0,0.6607]
core								[0,0.6349]

Table 5: Several incoherence correction with associated inference results for the target $S3|E4$

flects also on the inference performances.

Note that we report, just as an example, only the extension values for a conditional event, however we could compute all the values of the (conditional) events of interest, as for example on the partition generated by the three random variables.

7 Conclusion

Checking coherence and removing incoherences in the data is a long debated problem in literature, we have studied it by focusing on statistical matching applications. In fact, in this kind of application the incoherence can arise when the variables are linked by logical relations.

We have applied several incoherence adjustment procedures in this specific ambit. From this study some differences among these adjustments come out. Due to peculiarities of source integration and lack of information on the variables not jointly observed, usual divergences techniques can be specialized. In particular, a specific adjustment of a discrepancy, originally introduced for general conditional probability assessment, shows the advantage of an automatic and weighted localization of the sub-domains where incoherence must be removed.

We have analyzed also a very simple practical application and we have shown that better results are obtained not simply focusing on the minimal number of incoherent values, but involving all the elements conditioned to the same scenarios, where incoherence arises. On the other hand, coherent imprecise adjustment performs better focusing on the minimal number of incoherent values. This entail a minimal number of changes with respect the original assessment, but has as counterpart obvious vaguer inference conclusions. Vagueness that can however be reduced by the aforementioned "maximally supported" sub-intervals detection.

References

[1] V. Biazzo and A. Gilio: A generalization of the fundamental theorem of de Finetti for imprecise conditional probability assessments. *Int. J. of Approximate Reasoning*, 24: 251-272, 2000.

[2] V. Biazzo, A. Gilio. Some Theoretical Properties of Conditional Probability Assessments. In *LNAI (Lecture Notes in Computer Science)* 3571: 775–787, 2005.

[3] E.C. Budd. The creation of a microdata file for estimating the distribution of income. *Review of Income and Wealth*, 17: 317–333, 1971.

[4] A. Capotorti, L. Galli and B. Vantaggi. How to use locally strong coherence in an inferential process based on upper-lower probabilities. *Soft Computing*, 7(5): 280–287, 2003.

[5] A. Capotorti and M. Zagoraiou. Implicit Degree of Support for Finite Lower-Upper Conditional Probabilities Extensions, in *Proceed. of Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU'06*, EDK Paris - France, vol III: 2331– 2338, 2006.

[6] A. Capotorti and G. Regoli. Coherent correction of inconsistent conditional probability assessments. Proc. of IPMU'08 - Malaga (Es), 2008.

[7] A. Capotorti, G. Regoli and F. Vattari. On the use of a new discrepancy measure to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities. Proc. of ISIPTA'09 - Durham (UK), 2009.

[8] A. Capotorti, G. Regoli and F. Vattari. Correction of incoherent conditional probability assessments. *International Journal of Approximate Reasoning*, 51(6): 718–727, 2010.

[9] G. Coletti. Coherent numerical and Ordinal probabilistic assessments. *IEEE Transaction on Systems, Man, and Cybernetics*, 24: 1747-1754, 1994.

[10] G. Coletti, R. Scozzafava. *Probabilistic Logic in a Coherent Setting*, Dordrecht: Kluwer, Series "Trends in Logic", 2002.

[11] G. De Cooman, and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159: 75-125, 2004.

[12] B. de Finetti. Sull'impostazione assiomatica del calcolo delle probabilità. *Annali Univ. Trieste*, 19:3-55, 1949. - Engl. transl. in: Ch.5 of *Probability, Induction, Statistics*. Wiley, London, 1972.

[13] M. D'Orazio, M. Di Zio and M. Scanu. Statistical Matching for Categorical Data: displaying uncertainty and using logical constraints. *Journal of Official Statistics*, in press.

[14] M. D'Orazio, M. Di Zio and M. Scanu. *Statistical Matching: Theory and Practice*. Wiley, 2006.

[15] L.E. Dubins. Finitely additive conditional probabilities, conglomerability and disintegration. *The Annals of Probability*, 3:89–99, 1975.

[16] A. Gilio and S. Ingrassia. Totally coherent set-valued probability assessments. *Kybernetika*, 34(1): 3–15, 1998.

[17] J. B. Kadane. Some statistical problems in merging data files. *Journal of Official Statistics*, 17: 423–433, 2001.

[18] P.H. Krauss. Representation of Conditional Probability Measures on Boolean Algebras. *Acta Math. Acad. Scient. Hungar.*, 19:229–241, 1968.

[19] S. Kulback,*Information Theory and Statistics*, New York: John Wiley, 1957.

[20] F. Lad, *Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction*, New York: John Wiley, 1996.

[21] D.V. Lindley, A. Tversky, R.V. Brown. On the reconciliation of probability assessments. *Journal of Royal Statistical Society* Ser. A 142(2): 146–180, 1979.

[22] C.F. Manski, *Identification problems in the social sciences*. Cambridge, MA: Harvard University Press, 1995.

[23] B. A. Okner. Data matching and merging: an overview. *Annals of Economic and Social Measurement*, 3(2): 347–352, 1974.

[24] E. Miranda, Updating coherent previsions on finite spaces, *Fuzzy Sets and Systems*, 160(9): 1286–1307, 2009.

[25] G. Paass. Statistical match: evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*, (Eds. G.H. Orcutt and H. Quinke) Elsevier Science, Amsterdam, 401–422, 1986.

[26] S. Rässler. Statistical matching: a frequentist theory, practical applications and alternative Bayesian approaches. *Lecture Notes in Statistics*, Springer Verlag, 2002.

[27] A. Rényi. On a new axiomatic theory of probability. *Acta mathematica Academiæ Scientiarium Hungaricæ*, 6:285–335. 1955.

[28] R. H. Renssen. Use of statistical matching techniques in calibration estimation. *Survey Methodology*, 24: 171–183, 1998.

[29] W. L. Rodgers. An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2(1): 91–102, 1984.

[30] D.B. Rubin. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 2(1): 87–94, 1986.

[31] C. A. Sims. Comments (on Okner). *Annals of Economic and Social Measurement*, 1(3): 343–345, 1972.

[32] A. C. Singh, H. J.Mantel, M. D. Kinack and G. Rowe. Statistical matching: use of auxiliary information as an alternative to conditional independence assumption. *Survey Methodology*, 19(1): 59–79, 1993.

[33] P. Szivós, T. Rudas, I.G. Tóth. A tax-benefit microsimulation model for Hungary. *Workshop on Microsimulation in the New Millennium: Challenges and Innovations*, Cambridge, 1998.

[34] J.L. Schafer, *Analysis of incomplete multivariate data*. Chapman & Hall, London, 1997.

[35] B. Vantaggi. The role of coherence for the integration of different sources. Proc. 4th International Symposium on Imprecise Probabilities and their Applications ISIPTA'05 (Pittsburgh, USA), 369–378, 2005.

[36] B. Vantaggi. Statistical matching of multiple sources: A look through coherence, *International Journal of Approximate Reasoning*, 49(3):701–711, 2008.