The role of coherence for the integration of different sources

Barbara Vantaggi
Dip. Me.Mo.Mat, Univ. "La Sapienza", Roma
vantaggi@dmmm.uniroma1.it

Abstract

In several economic applications (e.g. marketing research, microsimulation models) there is the need to consider different data sources and to integrate the information coming from them. In this paper we show how integration problems can be managed by means of coherence for partial conditional probabilistic assessments. Coherence allows us to combine the knowledge coming from the different sources, included those (possibly) given from field experts, without necessarily assuming further hypothesis (as conditional independence). Moreover, inferences and decisions can be drawn taking in consideration also logical constraints among the variables. An example showing advantages and drawbacks of the proposed method is given.

Keywords. Coherent conditional probability, data fusion, statistical matching, inference.

1 Introduction

The integration problem of knowledge coming from several separate micro data bases, which have some variables in common as well as some variables recorded only in one data base, occurs in several economic applications, some examples are marketing research [19] and microsimulation modeling (e.g. from economic and administrative sources for public policy and social research) [21, 22, 23, 34] (see also Section 2). This problem may be represented by the following simple situation: there are two different sources, A and B, the first one contains data from one sample on the variables (X, Y) and the second one collects data from another sample on (X, Z), so data on X are available in both sources. 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.

There are analytic techniques for combining data from different sources, which have been developed since 1970s (see references in [24]): we recall, for example, those based on conditional independence assumption, i.e. the variables Y and Z are independent conditionally on X. However, in several situations the independence assumption is not adequate, as first raised by Sims [29] (see also [23, 25, 26, 30]). Other methods aim at incorporating auxiliary information about relationships between Y and Z to avoid or to relax conditional independence assumption (see, e.g. [30]). 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 [10, 28] and for missing data problem [11, 18, 20, 32]). This aspect is here faced by considering a coherence notion for partial conditional assessments (first introduced by de Finetti [13] and studied by several authors, see, e.g. [8]). In fact, coherence allows to check the compatibility of partial (conditional) assessments and to manage further available knowledge, for example coming from field experts (see also [6]).

We describe how the data contained in the different sources can be used to evaluate conditional probability assessments and how to manage logical constraints (called also structural zeros) characterizing the relevant links among variables describing the phenomenon. We prove that when there is no logical constraint among the variables, coherence is always satisfied and thus conditional independence assumption is legitimate from a syntactical point of view (even if it is useful to look for all compatible coherent extensions). Further we show that when logical constraints are present it is necessary to check the global coherence of the relevant partial assessments drawn from the different sources (see Section 6). When coherence

is not satisfied we need to detect where incoherencies are localized by looking for the "minimal" incoherent assessments and to remove them in order to restore coherence (see Section 4). This aspect is faced by means of the algorithms proposed in [4, 5, 16], which allows also to draw inferences in this setting and to take into account logical constraints. For each (conditional) event this approach aims at building directly the interval of all coherent values without to achieve a preliminary complete (artificial) data base containing all variable of interest (as e.g. in [21, 24, 28, 36]).

Finally in Section 6 we introduce an example built from data taken from [10] to better show advantages and drawbacks of the proposed method.

2 Some relevant applications in economics

As already remarked in the Introduction, the main aim of integration processes is to give joint (or conditional) knowledge on the variables related to data available in different sources. This problem arises from several applications in different areas.

Among the applications we recall those related to integrated analysis of economical variables as consumer's expenditures and income. For example, in Italy (as noted in [9]) there is no source describing these variables jointly, in fact income is related to the Household Balance survey managed by the Bank of Italy, while different sources can be used for expenditures: among the others, the Household Expenditure survey and the Household Multipurpose survey, both managed by National Institute of Statistics. Examples of fusion of such data sets are the following ones [9]: construction of social accounting matrices including economic indicators (e.g. per capita income and economic growth); analysis between income and health expenditures; microsimulation models for analysis of public policies (to predict the impact of policy changes and aggregate characteristics of tax, social security and welfare benefit programs).

Early work in the area of tax and benefit modeling has been described in [3, 21]: in particular in the first paper Tax Model with data from the 1965 Current Population Survey has been matched with the Survey of Financial Characteristics of Consumers, which provides extra knowledge on income from rent and interest. The merged file helps to improve the evaluations of the income size distribution made by the Office of Business Economics, which previously relied on relating cross-tabulations from several different sources. A similar statistical match has been performed in the second quoted paper, where matching data from the 1967 Survey of Economic Opportunity with the 1966 Tax File are involved and provide demographic information, data on non-taxable income, and income data for families who do not file a tax return. This latter group consists mainly of families with low incomes. However, Survey of Economic Opportunity does not collect knowledge on capital gains, and underestimates higher incomes. The Tax File filled these gaps, providing more complete (and probably more accurate) knowledge on taxable income. The merged database provides comprehensive information on the distribution of income. This has been used to study the distribution of federal state and local taxes.

We recall also the economic Hungarian applications based on the combination of knowledge from three different surveys [31]: income and demographic variables from the Household Panel Survey, consumption variables from the Household Budget Survey of the Hungarian Central Statistic Office and tax variables from administrative tax records. The analysis have been implemented for producing outputs showing the gains or losses due to policy changes, thus the distribution impact of a policy measure on different types of families or income levels is derived.

Integration of sources regards also problems related to market and social research: advertisers and media planners rely on measurement of media usage (e.g. television ratings and magazine and newspaper readership), which are usually collected in separate sources (see, e.g. [19]).

3 Coherent conditional probability

The notion of coherence for an uncertainty measure is necessary for managing functions defined on sets not necessarily closed with respect to the usual operations (more precisely it is not required that the function is defined on a Boolean algebra or on the product of a Boolean algebra and an additive set). In this paper we consider coherence for conditional probability in the sense of de Finetti [13] and Dubins [14]:

Definition 1 *Given a Boolean algebra $\mathcal{B}$ and an additive set $\mathcal{H}$ (closed under finite unions) such that $\mathcal{H} \subset \mathcal{B}$ and $\emptyset \notin \mathcal{H}$, a conditional probability on $\mathcal{B} \times \mathcal{H}$ is a function $P(\cdot|\cdot)$ into $[0,1]$, which satisfies the following conditions:*

(i) $P(H|H) = 1$ for every $H \in \mathcal{H}$,

(ii) $P(\cdot|H)$ is a finitely additive probability on $\mathcal{B}$ for any $H \in \mathcal{H}$,

(iii) $P(E \wedge A|H) = P(E|H)P(A|E \wedge H)$, whenever $E, A \in \mathcal{B}$ and $H, E \wedge H \in \mathcal{H}$

Actually, coherence gives a condition allowing to regard a partial assessment as restriction of a (conditional) probability.

Definition 2 *Given an arbitrary set of conditional events $\mathcal{F}$, a real function P on $\mathcal{F}$ is a coherent conditional probability assessment if there exists $\mathcal{E} \supseteq \mathcal{F}$ with $\mathcal{E} = \mathcal{B} \times \mathcal{H}$ (where $\mathcal{B}$ is an algebra, and $\mathcal{H}$ is an additive set with $\mathcal{H} \subset \mathcal{B}$) such that there exists a conditional probability $P'(\cdot|\cdot)$ on $\mathcal{E}$ extending P.*

A characterization of coherence for coherent conditional probabilities has been given in [8], we recall this result only in the finite case, for the general case we refer to the quoted paper.

Theorem 1 *Let $\mathcal{F} = \{E_1|H_1, ..., E_n|H_n\}$ be an arbitrary finite family of conditional events. Denote by $\mathcal{B}$ and $\mathcal{C}$ the algebra and the set of atoms generated by $\mathcal{U}_\mathcal{F} = \{E_1, H_1, ..., E_n, H_n\}$. For a real function P on $\mathcal{F}$ the following statements are equivalent:*

(i) P is a coherent conditional probability on $\mathcal{F}$;

(ii) there exists (at least) a family of probabilities $\mathcal{P} = \{P_0, ..., P_k\}$, each probability being defined on a suitable subset $\mathcal{A}_\alpha \subseteq \mathcal{B}$ (with $\mathcal{A}_0 = \mathcal{B}$ and, for $\alpha = 1, ..., k$, $\mathcal{A}_\alpha = \{E \in \mathcal{A}_{\alpha-1} : P_{\alpha-1}(E) = 0\}$), such that for any $E_i|H_i \in \mathcal{F}$ there exists a unique P_α with $P_\alpha(H_i) > 0$ and

$$P(E_i|H_i) = \frac{P_\alpha(E_i \wedge H_i)}{P_\alpha(H_i)}.$$

(iii) there exists a sequence of compatible systems $(\mathcal{S}_\alpha)$ with unknowns $x_r^\alpha \geq 0$ associated to the atoms $C_r \in \mathcal{C}$,

$$\begin{cases} \displaystyle\sum_{C_r \subseteq E_i \wedge H_i} x_r^\alpha = P(E_i|H_i) \sum_{C_r \subseteq H_i} x_r^\alpha, \quad \text{if } \sum_{C_r \subseteq H_i} \mathbf{x}_r^{\alpha-1} = 0 \\ \displaystyle\sum_{C_r \subseteq H_0^\alpha} x_r^\alpha = 1 \\ x_r^\alpha \geq 0 \end{cases}$$

where $\mathbf{x}^\alpha$ (with r-th component $\mathbf{x}_r^\alpha$) is the solution of the system $\mathcal{S}_\alpha$, and $\mathbf{x}_r^{-1} = 0$ for any C_r; moreover, H_0^α denotes the union of the conditioning event H_i such that $\sum_{C_r \subseteq H_i} \mathbf{x}_r^{\alpha-1} = 0$.

Condition *(ii)* gives a characterization in terms of a class of unconditional probabilities $\{P_0, \ldots, P_k\}$ and it allows a "local" representation of a conditional probability as a "ratio" of suitable unconditional probabilities of the above class. While condition *(iii)* gives an operative tool to check coherence by solving a sequence of linear systems where unknowns are probabilities of atoms (i.e. the possible events of the form $E_i^* \wedge H_1^* \wedge \ldots \wedge E_n^* \wedge H_n^*$, where E_i^* - analogously H_i^* - stands for either E_i or E_i^c).

The problem of checking coherence presents computational difficulties related to the number of atoms and so to the construction of matrix of atoms (the problem is NP-complete, see [1]), hence some strategies have been given to circumvent these computational difficulties: in [4] conditions allowing to split the coherence problem in subproblems have been given to avoid to build the whole matrix, while in [16, 17] a generation column technique, which considers only suitable sub-matrices, is proposed.

Another important feature is the possibility of extending a coherent conditional probability assessment on $\mathcal{F}$ to new (conditional) events [8].

Theorem 2 *If P is an assessment on a family of conditional events $\mathcal{F}$, then there exists a (possibly not unique) coherent extension of P to an arbitrary family $\mathcal{F}'$ of conditional events, with $\mathcal{F}' \supset \mathcal{F}$, if and only if P is coherent on $\mathcal{F}$.*

In particular, supposed that P on $\mathcal{F}$ is coherent, if $\mathcal{F}' = \{E|H\} \cup \mathcal{F}$, the coherent values $p = P(E|H)$ are all the values of a suitable closed interval $[\underline{p}, \overline{p}] \subseteq [0, 1]$, with $\underline{p} \leq \overline{p}$ (see e.g. [8]).

If the events $E \wedge H$ and H are logically dependent on $\mathcal{U}_\mathcal{F}$ (i.e. they are union of some atoms generated by $\mathcal{U}_\mathcal{F}$), the problem is to find the minimum and maximum value of

$$P(E|H) = \frac{P_\alpha(E \wedge H)}{P_\alpha(H)}$$

with α such that $P_\alpha(H) > 0$ for every class $\mathcal{P}$ agreeing with P (in the sense of condition *(ii)* of Theorem 1).

Actually, the problem can be solved by adding to systems $\mathcal{S}_\alpha$ (with $\alpha \geq 0$) the constraint $\sum_{C_r \subseteq H} x_r^\alpha = 0$ till the system is compatible. If for $\overline{\alpha}$ the system $\mathcal{S}_{\overline{\alpha}}$ with the above constraint has no solution, then all possible solutions of system $\mathcal{S}_{\overline{\alpha}}$ give positive probability to H. Then, the minimum and maximum coherent value for $P(E|H)$ coincides with

$$\min / \max \sum_{C_r \subseteq E \wedge H} y_r^{\overline{\alpha}}$$

under $\mathcal{S}'_{\overline{\alpha}}$ that is

$$\begin{cases} \displaystyle\sum_{C_r \subseteq E_i \wedge H_i} y_r^{\overline{\alpha}} = P(E_i|H_i) \sum_{C_r \subseteq H_i} y_r^{\overline{\alpha}} \quad \text{if } P_{\overline{\alpha}-1}(H_i) = 0 \\ \displaystyle\sum_{C_r \subseteq H} y_r^{\overline{\alpha}} = 1 \\ y_r^{\overline{\alpha}} \geq 0 \qquad\qquad\qquad\qquad C_r \in \mathcal{C}_\mathcal{F} \cap \mathcal{A}_\alpha \end{cases}$$

Note that the unknowns $x_r^{\overline{\alpha}}$ and $y_r^{\overline{\alpha}}$ are linked by a normalization constant, $x_r^{\overline{\alpha}} = \dfrac{y_r^{\overline{\alpha}}}{\displaystyle\sum_{C_r \in \mathcal{C}_\mathcal{F} \cap \mathcal{A}_\alpha} y_r^{\overline{\alpha}}}$.

Actually, when any solution $\mathbf{x}^o$ of $\mathcal{S}_0$ is such that $\sum_{C_r \subseteq H} \mathbf{x}^o_r > 0$, the coherent interval $[\underline{p}, \overline{p}]$ for $E|H$ coincides with the so-called natural extension [35].

When $E \wedge H$ or H are not logically dependent on $\mathcal{U}_\mathcal{F}$, the bound $\underline{p}$ [or $\overline{p}$] is related to the conditional probability of the maximum [minimum] (with respect to $\subseteq^*$) event logically dependent on $\mathcal{U}_\mathcal{F}$ contained [containing] $E|H$, where inclusion operation $\subseteq^*$ between conditional events is defined as follows (see, e.g. [8]),

$$A|H \subseteq^* B|K \Longleftrightarrow AH \subseteq BK \text{ and } B^cK \subseteq A^cH.$$

Therefore, the maximum event (with respect to $\subseteq^*$) contained in $E|H$ is

$$(E|H)_* = (E \wedge H)'|(E \wedge H)' \vee (E^c \wedge H)''$$

and the minimal event containing $E|H$ is

$$(E|H)^* = (E \wedge H)''|(E \wedge H)'' \vee (E^c \wedge H)',$$

with $(G)' = \bigvee_{C_r \subseteq G} C_r$ and $(G)'' = \bigvee_{C_r \wedge G \neq \emptyset} C_r$.

Algorithms to make inference on the base of coherence have been studied in [2, 4, 5, 7]. One of the main features of these algorithms is that both logical constraints among events and numerical assessments can be managed.

Notice that the extension values are obtained simply from the probabilistic partial assessment without requiring other assumptions. Moreover, one can include other judgements coming from a field expert or other sources: for example, conditional independence judgements among some events [33] or preferences expressing the idea "not more probable than" [6].

4 Integration of sources in a coherent setting

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 (whose variables have finite range). We suppose that the two samples are related to 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 $X_1, ..., X_{n_A}$, $X_{n_A+1}, ..., X_{n_A+n_B}$ (actually the above variables could be seen as partially exchangeable [12]).

We can evaluate from the two files the relevant probability values: from file A the conditional probabilities $P_{Y|(X=x_i)}(y_j)$, that the next unit has $Y = y_j$ on the hypothesis that $(X = x_i)$ (for any value x_i taken by X), and from file B the conditional probability values $P_{Z|(X=x_i)}(z_k)$ (that the next unit has $(Z = z_k)$ if $(X = x_i)$). In the following we denote, if no misunderstanding occurs, the above conditional probabilities by $p_{j|i}$ and $p_{k|i}$, respectively. Moreover, from data on both files we can evaluate $p_i = P(X = x_i)$.

By evaluating the relevant probabilities from the frequencies it follows $p_{j|i} = \frac{n_A^{ij\cdot}}{n_A^{i\cdot\cdot}}$ (if $n_A^{i\cdot\cdot} > 0$), analogously $p_{k|i} = \frac{n_B^{i\cdot k}}{n_B^{i\cdot\cdot}}$ (if $n_B^{i\cdot\cdot} > 0$), and $p_i = \frac{n_A^{i\cdot\cdot} + n_B^{i\cdot\cdot}}{n_A + n_B}$, where $n_A, n_A^{ij\cdot}, n_A^{i\cdot\cdot}$ denote, respectively, the number of observations in file A, those taking the value (x_i, y_j) and those having x_i (the quantities related to file B are defined analogously). Note that the above relevant quantities $p_{j|i}$, $p_{k|i}$ are well-defined if $n_A^{i\cdot\cdot}$ and $n_B^{i\cdot\cdot}$, respectively, are greater than 0, while, when e.g. $n_A^{i\cdot\cdot} = 0$, that means there is no observation in file A taking value x_i, then the value of $p_{j|i}$ cannot be assessed through frequencies, but it could be given by a field expert or deduced from auxiliary knowledge (even it is not required to assess $p_{j|i}$ for any x_i).

Notice that the values $p_{j|i}$, $p_{k|i}$, p_i can be evaluated according to different paradigms, but we do not want to stress which evaluation is better to be considered, while we focus on the relevant aspects arising after this choice. For example, the above assessment can be derived by assuming that the variables (X, Y, Z) have multinomial distribution with parameters $p_{ijk} = P(X = x_i, Y = y_j, Z = z_k)$ and by taking the maximum likelihood estimations. Another choice consists into assuming that (X, Y, Z) given $\Theta = \theta_{ijk}$ has multinomial distribution and Θ Dirichlet distribution with suitable parameters, then one might take expected values of posterior distribution or the maximum posterior values as evaluations of the relevant conditional probabilities.

Given $p_{j|i}, p_{k|i}, p_i$, for any i, j, k obtained, one needs to check coherence of the whole assessment. First of all, note that coherence of each single assessment e.g. $\{p_{j|i}\}_{j,i}$ (analogously $\{p_{k|i}\}_{k,i}$ or $\{p_i\}_i$) is actually assured, as shown by the following result, by the following condition: for any x_i, $\sum_j p_{j|i} = 1$ (and $p_{j|i} = 0$ for any y_j such that $(X = x_i) \wedge (Y = y_j) = \emptyset$).

Proposition 1 *Let X, Y be two random variables. If for any value x_i of X the assessment $\{P_{Y|X=x_i}(y_j)\}_{y_j}$ is coherent, then the whole assessment*

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

is coherent iff $\sum_{x_i} P_X(x_i) = 1$ and $P_X(x_i) \geq 0$ for any x_i.

Proof: If $\{P_{Y|X=x_i}(y_j)\}_{y_j}$ is coherent for a given x_i, then the system $\mathcal{S}_0$ admits a solution $\mathbf{x}^i$ with components $\mathbf{x}_{ij}$. Since the events $(X = x_i)$, with x_i in the range of X, are pairwise incompatible, coherence of the assessment $\{P_{Y|X=x_i}(y_j) : \text{ for any } x_i\}$

follows easily from Theorem 1: in fact, the relevant system $\mathcal{S}_0$ related to the above assessment has groups of equations (each group related to a given x_i) independent from the others, in the sense that the unknowns present in the equations related to x_i are not in those related to x_j, for $j \neq i$. The vector $\mathbf{y}_i = \frac{\mathbf{x}^i}{p_i}$, with $p_i = P_X(x_i)$, is solution of the group of equation related to x_i. Then, a solution of system $\mathcal{S}_0$ is given from the vector $\mathbf{y}$ with components the subvectors $\mathbf{y}_i$; in fact also the last equality of system is satisfied if and only if the sum of p_i is equal to 1. □

Now, one needs to check coherence of the global assessment

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

Actually, 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 as proved by the following result:

Theorem 3 *Let X, Y, Z be three finite random variables. Given the following three coherent assessments $\{P_X(x_i)\}_{x_i}$, $\{P_{Y|X=x_i}(y_j)\}_{y_j}$ and $\{P_{Z|X=x_i}(z_k)\}_{z_k}$, if the partitions $\mathcal{E}_Y, \mathcal{E}_Z$ are logically independent with respect to $\mathcal{E}_X$ (i.e. $A \wedge B \wedge C \neq \emptyset$ for any $A \in \mathcal{E}_X$, $B \in \mathcal{E}_Y$, $C \in \mathcal{E}_Z$ s.t. $A \wedge B \neq \emptyset \neq A \wedge C$), then the whole assessment*

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

is coherent.

Proof: From Theorem 1 one has that an assessment is coherent iff there exists a sequence of compatible systems $\mathcal{S}_\alpha$. From the fact that the three assessments are coherent one has that, for any y_j, $P_{Y|X=x_i}(y_j) \geq 0$ and $\sum_{y_j} P_{Y|X=x_i}(y_j) = 1$. Then, by denoting the atoms $C_{ijk} = (X = x_i) \wedge (Y = y_j) \wedge (Z = z_k)$ and the associated unknown (i.e. probabilities of atoms) by x_{ijk}, the system $\mathcal{S}_0$ is

$$\begin{cases} \sum_k x_{ijk} = P_{Y|X=x_i}(y_j) \sum_{k,j} x_{ijk} & \text{for any } y_j \\ \sum_j x_{ijk} = P_{Z|X=x_i}(z_k) \sum_{k,j} x_{ijk} & \text{for any } z_k \\ \sum_{j,k} x_{ijk} = P_X(x_i) \sum_{i,k,j} x_{ijk} & \text{for any } x_i \\ \sum_{i,k,j} x_{ijk} = 1 \\ x_{ijk} \geq 0 \end{cases}$$

For any atom C_{ijk} one can define e.g.

$$\mathbf{x}_{ijk} = P_{Y|X=x_i}(y_j) P_{Z|X=x_i}(z_k) P_X(x_i),$$

in fact the first block (analogously the second one) of equations, related to y_j (related to z_k), is satisfied being $\sum_k x_{ijk} = P_{Y|X=x_i}(y_j) P_X(x_i)$ and moreover $\sum_{k,j} x_{ijk} = P_X(x_i)$. The last equation holds since the assessment $P_X(x_i)$ is coherent. Some equations of the first system could be satisfied trivially, it means that the solution associated to the conditioning events is 0, so in this case one should check the compatibility of the system $\mathcal{S}_1$, which again has solution, that is of the same form of that given for $\mathcal{S}_0$. □

Notice that logical independence of $\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$ implies logical independence of $\mathcal{E}_Y$ and $\mathcal{E}_Z$ with respect to $\mathcal{E}_X$, then Theorem 3 shows that coherence is assured also when each logical constraint links either $\mathcal{E}_X$ and $\mathcal{E}_Y$ or $\mathcal{E}_X$ and $\mathcal{E}_Z$.

Remark 1 The solution found in the proof of Theorem 3 is such that Y and Z are independent conditionally on X (according to the classical independence definition), so when there is no constraint involving the variables Y and Z the assessment is coherent and the independence assumption, used by many authors in the integration problem (as recalled in the Introduction), is legitimate from coherence in the sense that it is compatible from a syntactical point of view. However, as noted in several papers [23, 25, 26, 29, 30] this assumption could be not adequate and we must look for all the solutions.

On the other hand, when there are some logical constraints among the variables Y and Z, the whole coherence is not assured by coherence of the single assessments (see next section). Notice that the need of managing logical constraints arises from practical applications, as shown by the following example.

Example 1 A population of N persons, has two municipalities. A file A contains the variable municipality (M_1 and M_2 denote respectively that a person lives in the first or second municipality) and the variable age X, which has two categories: < 18 and ≥ 18. While data in file B are related to possession of a driving licence (event D) and to municipality variable. Since in some countries (as e.g. in Italy) one cannot have driving licence if his/her age is less than 18, then, a logical constraint between the variables collected in different files is present: $(X < 18) \wedge D = \emptyset$ (or equivalently $D \subseteq (X \geq 18)$).

Consider the assessments: that evaluated from file A

$$P((X < 18)|M_1) = \frac{7}{22}, P((X < 18)|M_2) = \frac{9}{29},$$

that from file B

$$P(D|M_1) = \frac{3}{4}, P(D|M_2) = \frac{11}{16}.$$

The assessment computed from file A (analogously that obtained from file B) is coherent, but it is easy to check that the whole assessment with $p_1 = P(M_1)$ is not coherent: the atoms are the following omes $C_{i12} = M_i \wedge (X < 18) \wedge D^c$, $C_{i21} = M_i \wedge (X \geq 18) \wedge D$, $C_{i22} = M_i \wedge (X \geq 18) \wedge D^c$, with $i = 1, 2$ and $\mathcal{S}_0$

admits no solution since the sub-system (related to $P((X<18)|M_1)$, $P(D|M_1)$ and $p_1 = P(M_1)$)

$$\begin{cases} x_{112} = \frac{7}{22}(x_{112}+x_{121}+x_{122}) \\ x_{121} = \frac{3}{4}(x_{112}+x_{121}+x_{122}) \\ x_{112}+x_{121}+x_{122} = p_1 \\ x_{ijk} \geq 0 \end{cases}$$

has no solution. This shows that when there are logical constraints among variables related to different assessments it is necessary to check coherence. □

When the global assessment is not coherent, then incoherence must be localized, in the sense we need to find the minimal restriction of the whole assessment, which is not coherent. Then, starting from the whole assessment on the set of conditional events $\mathcal{F}$, i.e.

$$\{(X=x_i),(Y=y_j)|(X=x_i),(Z=z_k)|(X=x_i)\}_{i,j,k}$$

we need to remove the minimal set $\mathcal{E}$ of events from $\mathcal{F}$ such that a restriction on $\mathcal{G} \subseteq \mathcal{F} \setminus \mathcal{E}$ is coherent. The set $\mathcal{G}$ can be not univocally defined since the values generating incoherence can be induced also from other values. While the set $\mathcal{E}$ is uniquely determined and must be found among the conditional events

$$\{(Y=y_j)|(X=x_i)\,,\,(Z=z_k)|(X=x_i)\}$$

involved in the logical constraints, as follows from the result below.

Theorem 4 *Let X, Y, Z be three finite random variables. Given three coherent assessments $\{P_X(x_i)\}_{x_i}$, $\{P_{Y|X=x_i}(y_j)\}_{y_j}$ and $\{P_{Z|X=x_i}(z_k)\}_{z_k}$, if for each x_i the assessment $\{P_{Y|X=x_i}(y_j), P_{Z|X=x_i}(z_k)\}_{\{y_j,z_k\}}$ is coherent, then the whole assessment*

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

is coherent.

Proof: If for any given x_i the assessment $\{P_{Y|X=x_i}(y_j), P_{Z|X=x_i}(z_k)$ for any y_j and $z_k\}$ is coherent, then it means that for any x_i the system

$$\begin{cases} \sum_k x_{ijk} = P_{Y|X=x_i}(y_j)\sum_{k,j} x_{ijk} & \text{for any } y_j \\ \sum_j x_{ijk} = P_{Z|X=x_i}(z_k)\sum_{k,j} x_{ijk} & \text{for any } z_k \\ \sum_{k,j} x_{ijk} = 1 \\ x_{ijk} \geq 0 \end{cases}$$

admits a solution $\mathbf{x}^i$ with components $\mathbf{x}_{ijk}$. Then, let $\mathbf{y}_{ijk} = \frac{\mathbf{x}_{ijk}}{p_i}$ where $p_i = P(X=x_i) > 0$, it is easy to check that the vector $\mathbf{y} = (\mathbf{y}_{ijk})_{ijk}$ is a solution of the relevant system $\mathcal{S}_0$. If there are some x_i such that $p_i = 0$, then, analogously to previous step, $\mathbf{x}_{ijk}$ is solution of $\mathcal{S}_\alpha$. □

Remark 2 Theorem 4 shows how to localize the incoherencies, i.e. to find the set $\mathcal{E}$ and to determine a "maximal" coherent restriction of the initial assessment on $\mathcal{F}$. Moreover, it allows to split the problem of coherence in subproblems, in fact it shows that in the above case it is enough to check coherence "locally" for any given x_i, so the number of atoms to compute each time decreases.

Given a coherent restriction on $\mathcal{G}$ then, by means of the inference procedure shown in Section 3, we can find the interval of coherent values for the removed conditional events. Inside these intervals we would like to choose the values (on the set of removed conditional events) by looking for the minimal changes from the evaluations obtained from data, according to any given norm. This criterion corresponds to the minimal change from data. However, other criteria could be proposed: for example if the evaluations are obtained by means of maximum likelihood principle, then we could look for the coherent evaluations that maximize the likelihood function by choosing the likelihood as objective function. Analogously, when the evaluations are obtained by means of maximum posterior values, then we could look for the coherent evaluations that maximize the posterior.

When the minimal change according to L_1 norm is chosen the values can be found by solving a linear programming problem (while in the other above cited cases we can get optimization problems with nonlinear objective function).

Suppose that the set of removed conditional events is $\mathcal{E} = \{(Y=y_j)|(X=i),(Z=z_k)|(X=i)\}$ (for some $j \in J$, $k \in K$), we need to find the solution $\mathbf{x}$ of $\mathcal{S}_0$, which minimizes

$$\sum_i \left(\sum_{j\in J} \mid \frac{\sum_k x_{ijk}}{p_X(x_i)} - p_{Y|X=x_i}(y_j) \mid + \right.$$

$$\left. + \sum_{k\in K} \mid \frac{\sum_j x_{ijk}}{p_X(x_i)} - p_{Z|X=x_i}(z_k) \mid \right).$$

Notice that if in the family $\mathcal{E}$ there are some assessments $p_{Y|X=x_i}(y_j), p_{Z|X=x_i}(z_k)$ such that $p_X(x_i) = 0$, these evaluations do not make a constraint in the minimization problem under $\mathcal{S}_0$ (see Section 3), and the optimization problem can be split in subproblems (due to locally strong coherence, see [4]), and for any i with $p_X(x_i) = 0$ we need to minimize

$$\sum_{j\in J} \mid \sum_k y^1_{ijk} - p_{Y|X=x_i}(y_j) \mid + \sum_{k\in K} \mid \sum_j y^1_{ijk} - p_{Z|X=x_i}(z_k) \mid$$

under the system $\mathcal{S}_1$ and the constraint $\sum_{jk} y^1_{ijk} = 1$.

Then, if the assessment is globally coherent (or coherence is restored by the above procedure) we make inference on any (conditional) event of interest (as

shown in Section 3) by solving suitable linear programming problems, for example to find the coherent values for $(X = x_i, Y = y_j, Z = z_k)$ we should to compute the minimum and the maximum of x_{ijk} under the system $\mathcal{S}_0$. Analogously, for $(Y = y_j)|(Z = z_k)$ we need to add the sequence of systems $\mathcal{S}_\alpha$ until the constraint $\sum_{i,j} x_{ijk} = 0$ is satisfied. Then, when $\mathcal{S}_\alpha$ with the additional constraint is not compatible, we add to the system $\mathcal{S}_\alpha$ the constraint $\sum_{i,j} y_{ijk} = 1$ and we find the minimum and maximum values of $\sum_i y_{ijk}$.

The approach introduced in this section will be applied in Section 6.

4.1 A comparison with Bayesian methods

Bayesian approach is particularly useful to handle assessments given by different experts and auxiliary information [15], however in the context of statistical matching some problems come out (see, e.g. [18, 27]): in fact, the posterior distribution of association between Y and Z given X is equal to the prior distribution, due to the lack of joint knowledge on (Y, Z) given X. Multiple imputation [28] aims ar carrying out a sensitivity analysis with respect to different assumption parameters in the multinormal setting, in particular $\rho_{Y,Z|X}$. In [24] an extension of multiple imputation is given in order to find the lower and upper bounds for $\rho_{Y,Z}$. This approach is performed by completing m times the concatenated data sets and by generating m independent values from predictive distribution (obtained by assuming a uniform prior on $\rho_{Y,Z|X}$ in the hypercube). Hence, a random inspection of values assumed by $\rho_{Y,Z|X}$ is carried out.

A direct comparison of our method with those studied in [18, 28, 24] is not feasible since they are related to multinormal variables, while we study here only the case for finite variables. However, our aim is in the same line of those in the quoted papers, that is to find lower and upper bounds for quantities related to Y and Z given X: multiple imputation yields to a random inspection, while in our setting we detect all coherent probability values taking into account also logical constraints.

5 Integration by means of different sampling schemes

In Section 4 we study integration problem when samples are drawn from the same population by means of the same sampling scheme. Now we suppose that the two samples are still drawn from the same population, but according to different sampling schemes: let S_s (with $s = 1, 2$) be the event "the unit is drawn according to the s-th sampling scheme". If data in file A are drawn according to first sampling scheme, while those in B are drawn according to the second one, we evaluate from the first file the probability $p^A_{j|i}$ that the next unit, drawn according to the first sampling scheme, has $Y = y_j$ supposing that $X = x_i$ (i.e. $P(Y = y_j|(X = x_i) \wedge S_1)$), analogously we can get $p^B_{k|i}$. By supposing that the probability that a unit is selected by both the sampling schemes is 0, we get

$$p_i = p_i^{S_1} P(S_1) + p_i^{S_2} P(S_2) \qquad (1)$$

where $p_i^{S_s} = P(X = x_i|S_s)$ (obtained as the previous evaluations) and $P(S_s)$ represents the probability that a unit is sampled from a population according to s-th sampling scheme. Notice that in this way we reinterpret some of the results given in [28].

Obviously, the above hypothesis can be removed and so a unit could be selected according to both the sampling schemes with positive probability; then p_i is not generally univocally defined as in (1), but we can assess the evaluations $p_i^{S_s} = P(X = x_i|S_s)$ and $P(S_s)$ as above, while p_i can be computed by looking for not just one value but by considering all coherent values (see in Section 3).

6 Example

In order to show our proposal we develop an example with data taken from [10]. 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 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 1 (Table 2): for example, in Italy a 17 years old person cannot have a University degree. Table 1 and 2 show, respectively, the distribution of Age and Professional Status in file A, and in file B that related to Age and Educational Level. Additional logical constraints involving both the variables Professional Status and Educational level are the following ones:

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

	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 1: 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 2: Distribution of Age and Educational level in file B.

By considering the frequencies as evaluation of the relevant conditional probabilities, we get the assessment for the variable Age

$$p(A_1) = \frac{15}{2313}, p(A_2) = \frac{55}{2313},$$
$$p(A_3) = \frac{2219}{2313}, p(A_4) = \frac{24}{2313};$$

for the Professional Status given the Age

$$p(S_2|A_2) = \frac{5}{22}, p(S_3|A_2) = \frac{17}{22},$$
$$p(S_1|A_3) = \frac{179}{1108}, p(S_2|A_3) = \frac{443}{1108}, p(S_3|A_3) = \frac{486}{1108},$$
$$p(S_1|A_4) = \frac{2}{3}, p(S_2|A_4) = \frac{1}{9}, p(S_3|A_4) = \frac{2}{9};$$

for the Educational level given the Age

$$p(E_1|A_1) = 1, p(E_2|A_1) = 0, p(E_1|A_2) = \frac{14}{33},$$
$$p(E_2|A_2) = \frac{6}{33}, p(E_3|A_2) = \frac{13}{33}, p(E_1|A_3) = \frac{387}{1111},$$
$$p(E_2|A_3) = \frac{102}{1111}, p(E_3|A_3) = \frac{464}{1111}, p(E_4|A_3) = \frac{158}{1111},$$
$$p(E_1|A_4) = \frac{2}{3}, p(E_2|A_4) = 0,$$
$$p(E_3|A_4) = \frac{1}{5}, p(E_4|A_4) = \frac{2}{15}.$$

The above assessment is not coherent, in fact the system $\mathcal{S}_0$ admits no solution, so the incoherencies need to be localized as described in previous section (by algorithm given in [4]). It comes out that the following restriction of the previous assessment

Atom	Probab.
C_{113}	[0.0065, 0.0065]
C_{123}	[0, 0]
C_{213}	[0.0101, 0.0101]
C_{222}	[0, 0.0043]
C_{223}	[0, 0.0043]
C_{232}	[0.0011, 0.0054]
C_{233}	[0.0040, 0.0083]
C_{313}	[0.3342, 0.3342]
C_{322}	[0.0014, 0.0881]
C_{323}	[0, 0.0866]
C_{331}	[0.0186, 0.1550]
C_{332}	[0.1591, 0.3821]
C_{333}	[0, 0.0866]

Atom	Probab.
C_{341}	[0, 0.1363]
C_{342}	[0, 0.1364]
C_{343}	[0, 0.0866]
C_{413}	[0.0069, 0.0069]
C_{422}	[0, 0]
C_{423}	[0, 0]
C_{431}	[0.0009, 0.0011]
C_{432}	[0, 0.0011]
C_{433}	[0, 0]
C_{441}	[0.0002, 0.0014]
C_{442}	[0, 0.0011]
C_{443}	[0, 0]

Table 3: Atoms and their coherent values.

$$p(E_1|A_4) = \frac{2}{3}, p(S_1|A_4) = \frac{2}{3}, p(S_3|A_4) = \frac{2}{9}$$

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.

Then, identified the minimal set of conditional events

$$\mathcal{E} = \{E_1|A_4, S_1|A_4, S_3|A_4\}$$

involved in incoherencies, we remove this set from the initial one $\mathcal{F}$, and the restriction of p to $\mathcal{G} = \mathcal{F} \setminus \mathcal{E}$ is coherent.

Now, we need to restore coherence on the conditional events in $\mathcal{E}$, so first of all we find the coherent extensions for the conditional events in $\mathcal{E}$. Actually, for $E_1|A_4$ there is only one value coherent (implied from the restriction of $p(E_i|A_4)$ for $i = 2, 3, 4$), which is obviously $\frac{2}{3}$. While the interval of coherent values for $S_1|A_4$ is $\left[0, \frac{2}{9}\right]$ and that of $S_3|A_4$ is $\left[\frac{2}{3}, \frac{8}{9}\right]$.

Then, by looking for the extensions closer to the evaluations arising from the frequencies (since the aim is to change the starting assessment as little as possible), we get that

$$p(E_1|A_4) = \frac{2}{3},\ p(S_1|A_4) = \frac{2}{9},\ p(S_3|A_4) = \frac{2}{3}.$$

Now, coherence has been restored on the family $\mathcal{F}$, we are able to make inferences on the relevant (unconditional or conditional) events as shown in Section 3. Note that the number of atoms is 25, this reduction is due to the above logical constraints. The atoms $C_{ijk} = A_i \wedge E_j \wedge S_k$, generated by the variables Age, Educational level and Professional Status, and their coherent values are given in Table 3.

Prof. Status	Educ. level E_1	E_2
S_1	–	–
S_2	–	[0.00145, 0.09240]
S_3	0.35767	[0, 0.09095]

Prof. Status	Educ. level E_3	E_4
S_1	[0.01947, 0.15706]	[0.00023, 0.13782]
S_2	[0.16014, 0.38867]	[0, 0.13759]
S_3	[0.00396, 0.09491]	[0, 0.08662]

Table 4: Coherent values for the joint distribution of Educational level and Professional Status.

While the coherent values of the joint distribution of Educational level and Professional Status are represented in Table 4.

Since the unconditional values yield on intervals it is not possible to get conditional probabilities directly from them, but we get coherent conditional probabilities by means of the procedure shown in Section 3 and deeply described in [4].

For example, the probability values of the age given that a person is a manager and his/her educational level is secondary school are the following: $P(A_i|S_1 \wedge E_3) = 0$, for $i = 1, 2$, comes out from logical constraints, while$P(A_3|S_1 \wedge E_3) \in [0.89939, 0.99408]$, $P(A_4|S_1 \wedge E_3) \in [0.00591, 0.10060]$.

Moreover, the coherent probability that a person is a manager having a degree is $P(S_1|E_4) \in [0.00167, 1]$, and $P(S_2|E_4) \in [0, 0.99832]$, $P(S_3|E_4) \in [0, 0.62854]$; note that the length of the above intervals is close to 1, while this does not happen for the coherent values of $(S_i \wedge E_4)$ (see Table 4).

7 Conclusions

Integration of different sources has been studied in a conditional probability setting: we have shown that is necessary to check coherence when there are logical constraints linking the variables Y and Z observed in different files, moreover logical constraints must be managed since they arise naturally from applications. On the other hand, when no logical constraint is present the assessments, which arise (as shown in Section 4) from different files, are always coherent. However, the problem of restoring coherence has been faced by solving optimization problems. Moreover, inferences are drawn in a general setting by means of coherence.

A future inspection on the integration problem of n files by means of this approach is needed, and possibly considering also variables not necessarily finite.

Moreover, future research will be devoted to the general cases where samples are drawn from different populations and according to different sampling schemes, a first hint in this direction is given in Section 5.

References

[1] M. Baioletti, A. Capotorti, S. Tulipani and B. Vantaggi. Elimination of Boolean variables for probabilistic coherence. *Soft Computing*, 4(2): 81-88, 2000.

[2] V. Biazzo and A. Gilio. On the linear structure of betting criterion and the checking of coherence. *Annals of Mathematics and Artificial Intelligence*, 35: 83–106, 2002.

[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 and B. Vantaggi. Locally strong coherence in inferential processes. *Annals of Mathematics and Artificial Intelligence*, 35: 125–149, 2002.

[5] A. Capotorti, L. Galli and B. Vantaggi. Locally strong coherence and inference with lower-upper probabilities. *Soft Computing*, 7(5): 280–287, 2003.

[6] G. Coletti. Coherence principles for handling qualitative and quantitative partial probabilistic assessments. *Mathware & Soft Computing*, 3: 159-172, 1994.

[7] G. Coletti and R. Scozzafava. Exploiting zero probabilities. *Proc. of EUFIT97* (ELITE Foundation, Aachen, Germany) 1499-1503, 1997.

[8] G. Coletti and R. Scozzafava. *Probabilistic logic in a coherent setting*. Trends in logic n.15, Dordrecht/Boston/London: Kluwer, 2002.

[9] M. D'Orazio, M. Di Zio and M. Scanu. Statistical matching and official statistics. *Rivita di Statistica Ufficiale*, 1/2002: 5–24, 2002.

[10] M. D'Orazio, M. Di Zio and M. Scanu. Statistical matching and the likelihood principle:uncertainty and logical constraints. Technical Report *Contributi* 2004/1, Italian Statistical Ististute, Roma.

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

[12] B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7: 1–68, 1937.

[13] B. de Finetti. *Teoria della probabilitá*. Torino: Einaudi, 1970 (Engl. Transl. (1974) Theory of probability, London: Wiley & Sons).

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

[15] C. Genest and J. V. Zidek. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science*, 1(1): 114–148, 1986.

[16] B. Jaumard, P. Hansen and M. Poggi de Aragão. Column generation methods for probabilistic logic. *ORSA Journal on Computing*, 3: 135–148, 1991.

[17] P. Hansen, B. Jaumard and M. Poggi de Aragão. Mixed-integer column generation algorithms and the probabilistic maximum satistability problem. *European Journal of Operational Research*, 108: 671–683, 1998.

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

[19] W. A. Kamakura and M. Wedel. Statistical data fusion for cross-tabulation. *Journal of Marketing Research*, 34: 485–498, 1997.

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

[21] B.A. Okner. Constructing a new data base from existing microdata sets: the 1966 merge file. *Annals of Economics and Social Measurement*, 1(3): 325–342, 1972.

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

[23] 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 p. 401–422, 1986.

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

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

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

[27] D.B. Rubin. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American StatisticalAssociation*, 69: 467-474, 1974.

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

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

[30] 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.

[31] 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.

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

[33] B. Vantaggi. The role of coherence for handling probabilistic evaluations and independence. *Soft Computing*, 2005 (published online october 2004).

[34] M. Wolfson, S. Gribble, M. Bordt, B. Murphy and G. Rowe. The social policy simulation database and model: an example of survey and administrative data integration. *Survey of Current Business*, 69: 36–41, 1989.

[35] P. Walley. Statistical Reasoning with Imprecise Probabilities. London: Chapman & Hall, 1991.

[36] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105: 5-21, 2002.