An overview of methods in official statistics based on Bayesian networks

Mauro Scanu [1]

Abstract

Bayesian networks are a graphical formalisation of a joint multivariate distribution. They are efficiently exploited in many different applied settings. In these last years, some applications in official statistics have been defined. This paper illustrates at first the concept of Bayesian networks, and then focusses on applications in official statistics.

Keywords: graphical models, imputation of missing items, complex survey designs.

1 Italian National Institute of Statistics - Istat (scanu@istat.it).
The views and opinions expressed are those of the author and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

1. Introduction

"*...Bayesian networks are complex diagrams that organize the body of knowledge in any given area by mapping out cause-and-effect relationships among key variables and encoding them with numbers that represent the extent to which one variable is likely to affect another...*" The previous quotation is from the Los Angeles Times (Helm, 1996). In that article Bill Gates and other researchers at *Microsoft* explain how the usual computers were deaf, dumb, blind and clueless, and how "Bayesian stuff" could be used in order to make computers more interactive with human beings. In the following example, *Microsoft* applications with Bayesian networks are briefly reviewed.

Example - The first Bayesian network application in *Microsoft* programmes is the so called paperclip (or *Office assistant*, see Figure 1), firstly programmed by Horvitz, a researcher at *Microsoft*. The annoying features of the paperclip may suggest the reader to immediately stop understanding and using Bayesian networks! However, as stated in the following quotation from a newspaper article (The Economist, 2001), the original tool has been modified: "*...The paperclip in question, as even casual users of Microsoft's Office software will be aware, is a cheery character who pops up on the screen to offer advice on writing a letter or formatting a spreadsheet. That was the idea, anyway. But many people regard the paperclip as annoyingly over-enthusiastic, since it appears without warning and gets in the way. To be fair, that is not Dr Horvitz's fault. Originally, he programmed the paperclip to use Bayesian decision-making techniques both to determine when to pop up, and to decide what advice to offer….The paperclip's problem is that the algorithm (sequence of programming steps) that determined when it should appear was deemed too cautious. To make the feature more prominent, a cruder non-Bayesian algorithm was substituted in the final product, so the paperclip would pop up more often….*".

Figure 1: The paperclip (*Office Assistant*) implemented in *Microsoft Office*

This first attempt has been followed by many Bayesian networks based tools more respectful of Bayesian network theory (see for instance the following web page: http://www.microsoft.com/research/default.aspx). They include the selection of the items in the sometimes long context lists, user modelling and intelligent user interfaces (not only the already discussed *Office Assistant* implemented in *MS Office*, but models, theory and systems implemented in *Priorities*), diagnostics, trouble shooting and sensor fusion. All these tools use the *Windows*-based application for Bayesian belief network (Belief network is a synonym of Bayesian network) construction and inference called *Microsoft Belief Networks* (MSBN, see Kadie *et al*, 2001), available free for non-commercial purposes (http://research.microsoft.com/adapt/MSBNx/).

All the applications described in the previous example deal with the "decision making" problem. This is not the only problem that Bayesian networks tackle. Among the others, Bayesian networks have been proved to be useful for discovering causal relationships, prediction, assessment of risk, evolution in a simulated world, data mining, reliability analysis. The application fields are the most diverse, from biology (analysis of gene expression data) to medicine (diagnostics), psychology (cognitive psychology), artificial intelligence, speech recognition and weather forecasting (for a complete overview of Bayesian networks applications see Neapolitan, 2004, Chapter 12). The use of Bayesian networks in all these fields is justified by the interaction between an easily manageable set of multivariate statistical models and the existence of fast and efficient statistical algorithms for their estimation and use. This aspect is the motivation of a profitable use also in many different official

statistics problems. Applications in official statistics are yet in their infancy. Preliminary results date to the beginning of this century (Getoor *et al*, 2001a; Sebastiani et al, 2001b, Thibaudeau *et al*, 2002). The topics of imputation of missing items and of the multivariate structure of estimators in finite survey sampling has been studied to a certain level of detail in a number of papers, and show that the models offered by Bayesian networks in official statistics are an extremely promising tool.

This paper is organised as follows. At first (Section 2) Bayesian networks are defined and some theoretical aspects are highlighted. Note that this paper does not aim at aim at giving a complete and mathematically exhaustive explanation of Bayesian networks: just those elements that will be of interest in the applications to official statistics are described at a certain level of detail, leaving the rest to the relevant literature. This section is based on many references (mainly Cowell et al, 1999, and Neapolitan, 2004; but also Charniak, 1991, and the web page on Bayesian networks managed by Kevin P Murphy: http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html). Sections 2.1, 2.2 and 2.3 are mainly based on Neapolitan (2004). Bayesian networks applications in official statistics (Section 3) include the treatment of missing items (Section 3.1, based on the results in Di Zio *et al*, 2003, 2004a-c, 2005) and the use in sampling from finite populations (Section 3.2 based on the results in Ballin et al. 2005a-e). At the end of each of these two last sections, the role of Bayesian networks and the advantages in their use are highlighted in separate comments. Section 3.3 describes some other Bayesian networks applications. Finally, possible future developments are discussed in Section 4.

2. Bayesian Networks

Usually dependence relationship between variables are modelled with specific functions of their parameters, as in the generalised linear models or in the loglinear models. Bayesian networks are different. They are a class of models based on 2 elements:

i. the presence or absence of (any kind of) probabilistic relationship between the variables, and

ii. the possibility to represent these probabilistic relationships graphically in such a way that it is possible to associate the joint probability distribution to the graphical representation in a non-ambiguous way.

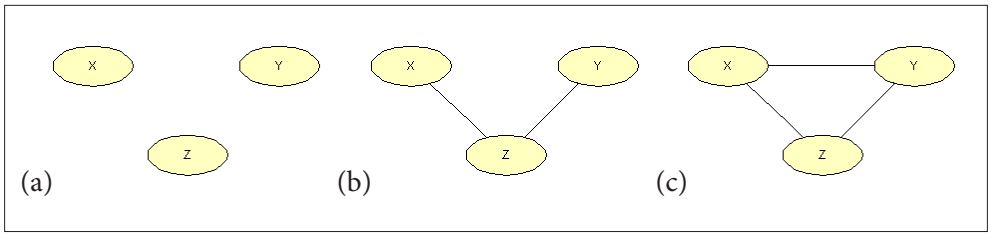
The first requirement makes this class quite general, and not dependent on specific functional definition of the dependence relationship of the variables. The second one is restrictive (for instance just some, but not all, of the loglinear models are Bayesian networks, see Warning 2 in Section 2.1).

In the following Bayesian networks are defined formally starting from the concept of Conditional Independence Graph (CIG) and Directed Acyclic Graph (DAG) as in Whittaker (1990). Finally the Bayesian network characteristics are shown with the help of some simplifying examples.

In general, a graphical representation of a multivariate variable $(X_1,...,X_K)$ is composed of a set of nodes V, each node representing one of the K variables, and a set of edges connecting pairs of nodes, E.

Conditional Independence Graphs (CIG) - A CIG is a graphical representation of the multivariate variable V composed of the pair (V,E), such that the edges in the set E are undirected and a pair of nodes is not connected by an undirected edge if and only if the two nodes are independent given all the other variables. Examples of CIG for three variables are in Figure 2.

Figure 2: Three CIGs for three variables *X*, *Y*, and *Z*



CIG (a) represents the situation of independence of the three variables, CIG (b) that *X* and *Y* are independent given *Z*, and CIG (c) that no conditional independencies characterize the three variables.

As a matter of fact, a CIG illustrates important features of the variables in *V*, in particular their dependence relationship. However, the joint probability distribution of *V* cannot be represented graphically, hence it is not yet useful for operative purposes.

Directed Acyclic Graphs (DAG) – In order to be operative, DAGs are appropriate. A DAG is a pair (*V,E*) of nodes and edges. Differently from CIGs, a DAG uses directed edges, henceforth arrows, for connecting pairs of nodes. The following elements characterize a DAG:

1. if there is an arrow from X to *Y* or from *Y* to *X*, *X* and *Y* are called *adjacent*
2. if there is an arrow from *X* to Y, *X* is a *parent* of *Y* and *Y* is a *child* of *X*;
3. the set of arrows connecting two nodes *X* and *Y* is called a *path*;
4. if there is a path from X to *Y*, *X* is an *ancestor* of *Y* and *Y* is a *descendent* of *X*;
5. if there is not a path from *X* to *Y*, *Y* is a *nondescendent* of *X*

The DAG has not associated any particular probabilistic feature of the variables in *V*, yet. One possibility that allows the operative use of the graphical representation linking the DAG with the probabilistic features of the variables is offered by the so called Markov condition.

> *Markov condition*. Let *P* be the joint probability distribution of the random variables represented by the nodes in *V*, and let the

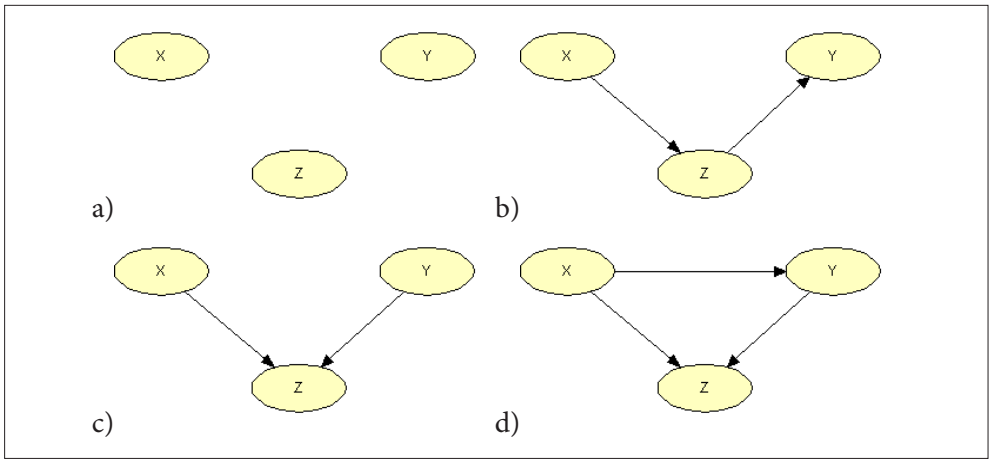
pair $G=(V,E)$ be a DAG associated to V. Then, the pair (P,G) satisfies the Markov condition if, for each variable X in V, X is independent of all its nondescendants given all its parents.

The pair (P,G) is a Bayesian network when it satisfies the Markov condition. Hence, the Bayesian network is an operative graphical representation of the joint probability distribution of the nodes in V. It is enough to associate each node X_j, $j=1,...K$, with the conditional distribution of X_j given its parents $pa(X_j)$ (when $pa(X_j)$ is the empty set, i.e. X_j is a root of the network, this conditional distribution is simply the marginal distribution of X_j). Then, the joint distribution function of the variables V is given by (chain rule):

$$P(X_1,...,X_k) = \prod_{j=1}^{K} P(X_j \mid \boldsymbol{p}\ (X_j)) \qquad (1)$$

Note that each multivariate variable can be factorised in the product of conditional distributions, but not all these decompositions correspond to a Bayesian network of the set of variables. As already said at the beginning of this section, a key issue is represented by the fact that the decomposition should be able to represent graphically the probabilistic relationship among the variables and describe it in a non-ambiguous way. Sometimes, this is not possible. For this reason, Bayesian networks are just a subclass of all the possible multivariate models: the Bayesian networks are the set of models for which it is possible to represent graphically the probabilistic relationship among the variables according to the Markov condition. Section 2.1 shows what this means in the case of three variables X, Y and Z.

Figure 3: Four possible Bayesian network structures for three variables



2.1 Meaning of different structures

Figure 3 shows some DAGs for three variables. According to the chain rule (1), these networks have the following interpretation (a thorough introduction on the concept of conditional independence is in Dawid, 1979).

Dag a) It has associated the following factorisation of the joint probability distribution: $P(X,Y,Z)=P(X)P(Y)P(Z)$. This case corresponds to the model of independence of the variables X, Y, Z.

Dag b) The joint probability distribution is $P(X,Y,Z)=P(X)P(Z|X)P(Y|Z)$. This is the case of conditional independence of X and Y given Z.

Dag c) The joint probability distribution is $P(X,Y,Z)=P(X)P(Y)P(Z|X,Y)$. This case corresponds to marginal independence of X and Y (just marginalize the joint probability with respect to Z) but conditional dependence of X and **Y** given Z. Note that it would not be possible to have at the same time X and Y marginal independent and conditional independent given Z in this network, unless (Z,X) is independent of Y or (Z,Y) is independent of X (as in the extreme case of DAG a) of complete independence; see also the following Warning 1).

Dag d) The joint probability distribution is factorised as $P(X,Y,Z)=P(X)P(Y|X)P(Z|X,Y)$. This is the *complete* model: all the dependencies between the variables are present. This model is also called *clique*.

When more than three variables are available, the possible dependence relationships are combination of the ones previously described. Two warnings are in order.

Warning 1 - As a matter of fact, the previous representations are not unique. In fact, for each CIG there can possibly be more than one Bayesian network, or better, given the same joint multivariate distribution P, more than one DAG. For instance, in Figure 2 conditional independence between X and Y given Z can be expressed uniquely by the CIG (b). On the contrary, different DAGs representing the situation of conditional independence of X and Y given Z can be defined via a suitable redirection of the arrows. These are shown in Figure 4. Their justification lies on the fact that, when X and Y are independent given Z, their joint probability distribution can be equivalently factorised as:

$$P(X,Y,Z)=P(X)P(Z|X)P(Y|Z)=P(Y)P(Z|Y)P(X|Z)=P(Z)P(X|Z)P(Y|Z).$$

Figure 4: Three equivalent Bayesian networks when *X* and *Y* are independent given *Z*

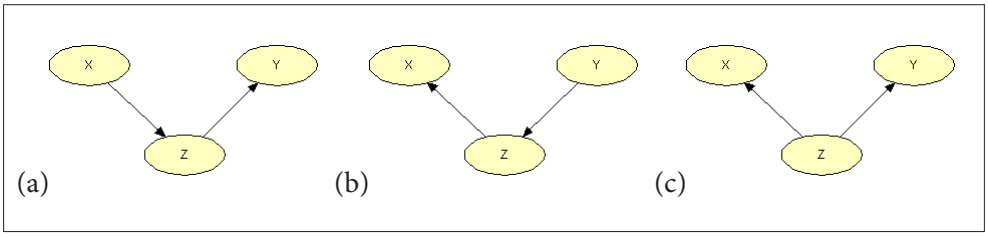


Figure 4 does not include the graph with the edges X -> Z and Y -> Z, i.e. Figure 3 c). In fact, this network has a complete different meaning. In order for Figure 3 c) to be consistent with the model represented by the equivalent networks of Figure 4, it is necessary to include an additional arrow linking X and Y. In other words, it is necessary to resort to a more complicated network than necessary (the clique, i.e. Figure 3 d). Hence, particular caution should be posed on the redirection of the arrows of a Bayesian network. The rules for arrows redirection and the definition of equivalent Bayesian networks are in Verma *et al* (1990).

Warning 2 - As already said, it is always possible to factorize a joint probability distribution, but it is not always possible to define a Bayesian network. An example is offered by loglinear models for categorical variables. It is easy to see that all the hierarchical loglinear models for three variables can be expressed as Bayesian networks but one: the one with the three way interaction set to zero. This loglinear model has a very peculiar aspect: the dependence relationship between the variables is not defined in terms of the joint probability distribution of all the variables, but by means of all the bivariate tables (distributions) of each couple of variables. In other words, it is true that each variable is connected with the others, although it is not the complete model (the saturated one). When factorizing the joint distribution of three variables X, Y, Z satisfying this model, the result is (no matter the order of the variables in the factorisation):

$$P(X,Y,Z)=P(X)P(Y|X)P(Z|X,Y).$$

Again, the result is the clique which is the appropriate factorisation for the saturated model (in other words, the factorisation of the joint distribution is the one associated to a more complicated dependence model). The appropriate representation of this loglinear model would actually involve three Bayesian networks, one for each minimal sufficient table for the model. Each model is a clique of respectively the pairs (X,Y), (X,Z), (Y,Z). As a matter of fact, there is not the possibility to describe this dependence relationship with a unique Bayesian network. Generally speaking, all those models that are defined via dependence relationship between subsets of variables in V and that cannot be expressed by an appropriate factorisation of the joint distribution function, do not have estimates of the parameters in closed form (e.g. the Iterative Proportional Fitting algorithm is used, calibrating successive estimates to the dependence relationship contained in each minimal sufficient table of the loglinear model). All these models are excluded by the set of models expressible as Bayesian networks.

Note that the previous problem does not apply to normal variables, i.e. multivariate normal variables can always be represented by Bayesian networks. This is due to the fact that multivariate normal variables are actually defined by the pairwise relationship of each couple of variables (subject to appropriate constraints on the variance matrix).

2.2 BN estimation

In the previous paragraph we have described a Bayesian network as a particular (graphical) model. Nothing "statistical" has been described. When just a sample of records where the variables V are observed is available, the Bayesian network should be estimated. There are many algorithms and methods for the estimation of a Bayesian network, some of them implemented in commercial or free software tools. A complete and updated reference is Neapolitan (2004). Here we review just the most important features on Bayesian network estimation.

The most important thing is that a Bayesian network is the pair (P,G), where P is the multivariate distribution of the variables V, and $G=(V,E)$ is the DAG. In this setting, only the set of nodes V is known in advance. The object of the inference is composed of two distinct elements:

1. the set of arrows E, or in other words the structure of a DAG
2. the conditional distribution of each node given its parents

In fact, the previous two elements define the Bayesian network and, by the chain rule (1), are able to define also the joint distribution of the variables V. It is worthwhile to mention three alternative approaches in estimating a Bayesian network.

The first one estimates at first the DAG structure, checking by appropriate independence and conditional independence tests whether undirected edges should be considered or not. Appropriate rules for the specification of the direction of the edges are defined in order to account for the relationship between variables (whether it is marginal or conditional independence). This estimation procedure of the structure is called *PC algorithm* (see Spirtes *et al*, 2000). Once the DAG structure is known, standard estimation methods (e.g. maximum likelihood estimation) can be applied in order to estimate the parameters of the conditional distribution of each node given its parents. This method is already implemented in commercial software tools, as Hugin (http://www.hugin.com). This approach is suitable when the data set is complete. Actually, some software tools allow to use this method also for incomplete data sets. In this last case, the PC algorithm is applied only on the subdata set of complete records, while the parameter estimation phase can be

performed on the overall data set. For instance, given the estimated structure, maximum likelihood estimation of the parameters can be performed with the EM algorithm.

The second approach is able to estimate with a unique procedure both the DAG structure and the parameters of the model given the structure via maximisation of the likelihood function (suitably penalised in order to avoid overspecification of the estimated model). This procedure has also been generalised to the case of partially observed data sets (Friedman, 1997). This approach is based on an extension of the *Expectation-Maximisation* (EM) algorithm for model selection problems that performs search for the best structure inside the EM procedure. Friedman proves the convergence properties of this algorithm, called *Model Selection EM*, and of one of its simplifications (in order to reduce the computational burden) *Alternating MS-EM*.

The third approach is just for incomplete data sets. It is a Bayesian approach developed by Sebastiani *et al* (2001a). This approach has the particular merit to highlight the different missingness mechanisms with the possibility to estimate the structure of a BN. Actually, the missingness mechanism can be considered as a set of additional dichotomous variables, showing whether each variables is actually observed or not. The multivariate structure of the variables of interest should take into account also their relationship with the indicators of missingness. This approach has not been implemented in any software tool, yet.

For a complete list of software codes and tools for using and estimating Bayesian networks and of their characteristics, see the webpage managed by Kevin P. Murphy (http://http.cs.berkeley.edu/~murphyk/Bayes/bnsoft.html) and the one of the gR project (graphical models in R: http://www.r-project.org/gR/).

2.3 Efficient use of the information in a BN

The Markov condition allows the identification of the relationship between a variable and its nondescendents. However it is still not clear the relation with all the other variables in V. The question is, given a variable X in V, which is the subsets of variables V' in V that makes X independent of all the

other variables in *V* given *V''*? *V'* is called the Markov blanket of *X*, henceforth *MB(X)*, and can be graphically determined in the Bayesian network structure via the following definition.

> *Markov blanket* – The Markov blanket *MB(X)* of a node *X* in *V* is composed by all the parents, children and parents of the children of *X*.

While it is evident the direct relationship of *X* with its parents and children, more attention should be given to its children's parents. The easiest example is offered by Network c) in Figure 3. In that case, *MB(X)* is composed by *Z* (its child) and *Y* (its child's parent). As already remarked, this network corresponds to considering marginal independence between *X* and *Y*, but conditional dependence of *X* and *Y* given *Z*. This last characteristic implies that *Y* should be included in *MB(X)* (*Z* alone is unable to make *X* independent of all the other variables given itself). Hence, in a multivariate setting the *MB(X)* is the subset of relevant variables for *X*: once *MB(X)* is known, all the other variables do not contain additional information on *X*.

3. Use in Official Statistics

Multivariate statistical models, as regression equations and loglinear models, are efficiently exploited in different official statistics problems: are Bayesian networks able to add something? The answer is yes, in many respects. First of all, Bayesian networks define models of interdependence between all the variables: variables relationship are easy to recognize. Secondly, this interdependence model allows a simplification of the joint distribution of the variables induced by the chain rule (1). Thirdly, each factor of the joint distribution can be easily estimated and used for operative purposes. Finally, when additional information is available (evidences, new distributions, additional records in the sample and so on) it can be easily used in order to update the joint distribution according to well established algorithms (see Cowell *et al*, 1999 and Cowell, 1998). All these elements suggest that some of the typical methodologies used up to now are just components of a larger family (see Ballin *et al*, 2005e for sampling and Di Zio *et al*, 2004a, for imputation). In the following a quick review of the use of Bayesian networks in official statistics is given. Note that most of the results have been obtained in the last 5 years. They should still be considered as research problems, and many issues have not yet been investigated. In the following, only categorical variables are studied. In fact, applications in this setting can be easily performed by means of the available software tools. The case of continuous variables still need to be further studied.

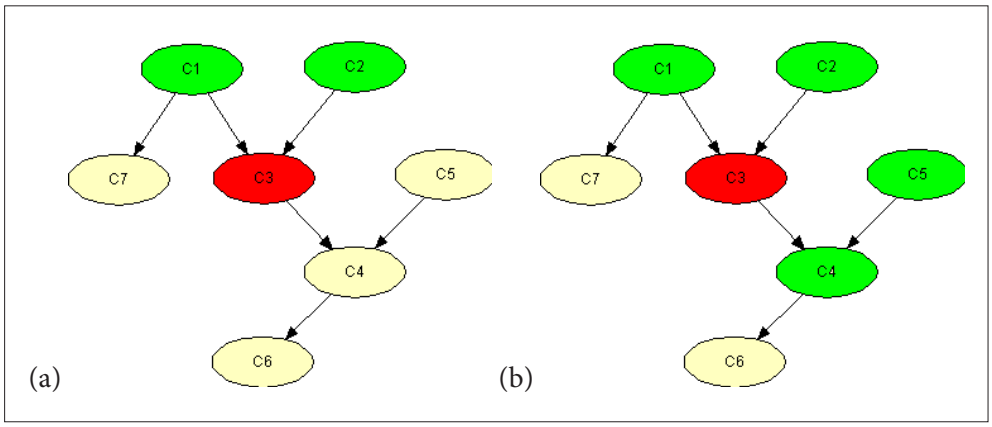
3.1 Imputation of missing items

This is maybe the most straightforward application of Bayesian networks, at least when missing data follow a Missing at Random mechanism (henceforth MAR; see Little *et al*, 1987, and references therein for a formal definition of MAR). Let $x=(x1_i,...,xk_i)$, $i=1,...n$, be a sample of n i.i.d. observations of the r.v. $X=(X_1,...,X_k)$, and assume that these records are just partially observed. Let $o(i)$ and $m(i)$ be subsets of $\{1,\ldots k\}$ such that $o(i) \grave{E}\ m(i) \cup \{1,\ldots,k\}$ and let $x_{o(i)}$ and $x_{m(i)}$ be respectively the observed and missing part of the record xi, i=1,…n. A usual practice for partially observed data set is imputation of missing values, i.e. generation of suitable values $\tilde{\mathbf{x}}_{m(i)}$ for the unobserved $x_{m(i)}$. Different imputation procedures have been defined. A "perfect" imputation

procedure would impute the missing part of the record with a random generation from the distribution of $X_{m(i)}$ given $X_{o(i)}$. This procedure can be considered as "perfect" because the new imputed data set would maintain the characteristic to be a random sample of n i.i.d. observations of X. Actually this procedure can be simplified in the sense that not all the conditional variables are necessary. A simplification that preserves the property to maintain the inferential characteristics of the imputed data set would consider a generation of imputed values from the distribution of $X_{m(i)}$ given $MB(X_{m(i)})$, where $MB(X_{m(i)})$ can possibly be a subset of $X_{o(i)}$. Hence, the identification of the Markov Blanket of $X_{m(i)}$ greatly simplifies the imputation procedure reducing the sets of conditionals and adapting the set of conditionals to the pattern of missing data in the record. For this reason, Bayesian networks are a useful tool for identifying which of the observed variables are necessary for imputation.

A preliminary formalisation of the use of the Bayesian network representation of the dependence relationship of the variables for imputation is in Thibaudeau *et al* (2002). In their paper, given a DAG structure, each missing variable is imputed drawing a value at random from its probability distribution given its parents. The imputation procedure starts from those nodes without parents. When all the missing items in these variables have been filled in, all the remaining variables whose parents are within the already imputed variables are imputed. When also these variables have been imputed, all the remaining variables whose parents are among those already imputed are imputed and so on.

Figure 5: Use of the dependence structure suggested by a Bayesian network. The alternative use of just the parents and Markov blanket of C3 is highlighted respectively in (a) and (b)



Their approach has been studied and generalised in some papers (Coppola *et al*, 2002a, 2002b, and Di Zio *et al* 2004a). In particular Di Zio et al (2004a) explains how logical constraints in terms of structural zeros can be easily considered in this setting. In fact, rules of compatibility between the observations on a unit can be defined as a fundamental aspect of the multivariate model for X. The possibility to specify Bayesian networks subject to logical rules, as the structural zeros, is a powerful approach that can be easily implemented during the Bayesian network estimation procedure. However, this approach actually does not exploit all the information in the data set: imputation of a missing variable is performed only by means of its parents, given an ordering among the variables (e.g. C_3 in Figure 5 (a) is imputed drawing randomly a value for its distribution given C_1 and C_2). Other papers (Di Zio *et al*, 2003, 2004b-c) have defined algorithms for the imputation of missing items with respect to the corresponding Markov blanket (e.g. C_3 is imputed conditioning on C_1, C_2, C_4 and C_5, see Figure 5 (b)). Manipulation of the Bayesian network in order to perform this operation is part of a software code in C++, described in Di Zio *et al* (2005).

An extension to the case of missing items in longitudinal surveys is in Righi (2005).

Comment: Bayesian networks appear as a device for exploiting most of the statistical information contained in the observed data set. Although the

use of random generation of imputations via conditional distributions is not new, Bayesian networks are a novel practice as far as the definition of the conditional variables is concerned. In a sense, the use of the Markov blanket of the unobserved variables makes the set of conditionals *adaptive* with respect to the pattern of missing values in each record. Adaptation is justified by the statistical relationship of the overall multivariate distribution. The variables not used as conditionals are independent of the missing variables given the conditional ones.

As a matter of fact, the multivariate distribution and the DAG structure should be estimated. The use of maximum likelihood estimators is particularly appropriate in this setting for their consistency. When the data set is large, the maximum likelihood estimate of the joint distribution function should be reasonably "near" to the true but unknown one. Hence, the imputed data set can be considered as "almost" generated by the true, and unknown, joint distribution function. Up to now, imputation by Bayesian networks has always been performed via estimation of the Bayesian network structure by the PC algorithm and, given the estimated structure, the conditional probability distributions are estimated via maximum likelihood. Other approaches in the estimation of Bayesian network structures for imputation are under study.

3.2 Estimation with completely observed samples drawn according to complex survey schemes

Also sampling methods from finite populations benefit of the multivariate relationship among the variables of interest (e.g. regression estimators). In general, special attention should be given to the sampling design. In fact, as stated in every modern textbook on sampling theory (e.g. Chambers *et al*, 2003), the sampling design is itself a variable and plays a very important role in the estimation process. Let $X_1,\ldots X_k$ be k variables of interest on a finite population of N units. Let a sample of n units be drawn from the population according to a complex survey scheme, with sample weights (defined by the pair design/estimator) w_i, i=1,…n. One of the most used estimators of the joint distribution function of the k variables is the ratio estimator:

$$\hat{F}(x_1,...,x_k) = \sum_{i=1}^{n} I_{x_1 \ldots x_k}(x_{1i},...,x_k) \frac{w_i}{\sum_{i=1}^{n} w_i} \qquad (2)$$

where $I(\ )$ is the indicator function, and $x_{1i},...x_{ki}$, i=1,…n, are the n observed records in the sample. The previous estimator can equivalently be rewritten via a Bayesian network model (preliminary results were obtained in Ballin *et al*, 2005a; advances are written in Ballin *et al* 2005b,c,d; and further extensions are in Ballin *et al*, 2005e). This new formalisation of estimator (2) is obtained via a new variable, S. This is the "design variable", with as many categories as the different inclusion probabilities, say $w_{(1)},\ldots,w_{(H)}$, and with marginal probability given by the fraction of the total weight of the units with the same sample weight:

$$P(S=h)=\frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}}$$

where n_h is the number of units with equal first inclusion probability $w_{(h)}$, h=1,…H. Given that S contains all the information on the sample design, conditioning on this variable produces estimators that are sample weights free. For instance, denoting with s_h the set of labels of the n_h units with weight $w_{(h)}$:

$$P(X_1=x_1 \mid S=h)=\frac{\sum_{i\in s_h} I_{x_1}(x_{1i})w_{(h)}}{n_h w_{(h)}}=\frac{\sum_{i\in s_h} I_{x_1}(x_{1i})}{n_h}$$

$$P(X_1=x_1 \mid X_2=x_2, S=h)=\frac{\sum_{i\in s_h} I_{x_1x_2}(x_{1i}x_{2i})w_{(h)}}{\sum_{i\in s_h} I_{x_2}(x_{2i})w_{(h)}}=\frac{\sum_{i\in s_h} I_{x_1x_2}(x_{1i}x_{2i})}{\sum_{i\in s_h} I_{x_2}(x_{2i})}$$

These definitions allow to rewrite (2) as the following:

$\hat{F}(x_1,...,x_k)=$

$$= \sum_{h=1}^{H} \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \sum_{i \in s_h} \frac{I_{x_1}(x_{1i})}{n_h} \sum_{i \in s_h} \frac{I_{x_1 x_2}(x_{1i} x_{2i})}{\sum_{i \in s_h} I_{x_1}(x_{1i})} \ldots \sum_{i \in s_h} \frac{I_{x_1 x_2 \ldots x_k}(x_{1i} x_{2i} \ldots x_k)}{\sum_{i \in s_h} I_{x_1 x_2 \ldots x_{k-1}}(x_{1i} x_{2i} \ldots x_{k-1i})}$$

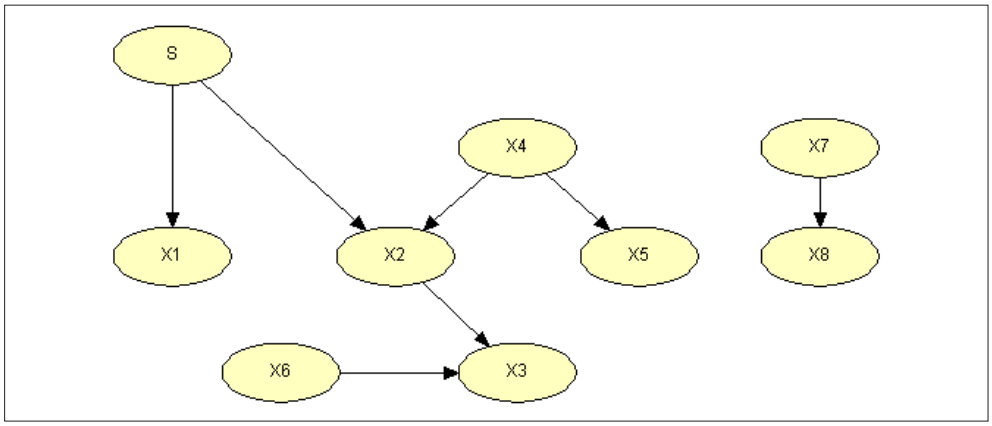
$$= \sum_{h=1}^{H} P(S=h) P(X_1 = x_1 \mid S=h) \prod_{j=2}^{k} P(X_j = x_j \mid X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, S=h) .$$

As a matter of fact, the usual ratio estimator $\hat{F}(x_1,...,x_k)$ implicitly assumes a particular model: the complete dependence model among ($S,X_1,\ldots,X_k$). In the Bayesian network terminology, the implicit model is the clique. If the dependency model for ($S,X_1,\ldots,X_k$) is simpler, the estimator (2) may result inefficient. An example taken from Ballin *et al* (2005e) is represented in Figure 6.

In order to define estimators that fulfil the dependence relationship between the variables and, at the same time, always use the sample weights, four different type of nodes have been defined.

- Type (a) nodes: these nodes admit S as a parent. In Figure 6, nodes X_1 and X_2 are type (a) nodes.
- Type (b) nodes: these nodes have at least a type (a) ancestor but S is not one of their parents. Node X_3 in Figure 6 is a type (b) node.
- Type (c) nodes: these are those nondescendants of type (a) and/or (b) nodes that do not admit S as a parent but that are (indirectly) linked to S. Figure 6 has two distinct groups of type (c) nodes: the first one is composed by the pair (X_4,X_5); the second one by X_6.
- Type (d) nodes: these are the nodes disconnected with S. In Figure 6, the couple (X_7,X_8) is a group of type (d) nodes.

Figure 6: Example of Bayesian networks for finite populations



The estimator of the joint distribution function will be of the following form:

$$\hat{F}(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) =$$

$$= \hat{F}(X_4, X_5)\hat{F}(X_6)\hat{F}(X_1, X_2 | X_4)\hat{F}(X_3 | X_2, X_6)\hat{F}(X_7, X_8)$$

where each component is estimated marginalizing their joint distribution with S with respect to S:

type (c)
$$\hat{F}(X_4, X_5) = \sum_{h=1}^{H} P(S = h)P(X_4, X_5 | S = h) \quad ;$$

$$\hat{F}(X_6) = \sum_{h=1}^{H} P(S = h)P(X_6 | S = h)$$

type (a)
$$\hat{F}(X_1, X_2 | X_4) = \sum_{h=1}^{H} P(S = h)P(X_1, X_2 | X_4, S = h)$$

type (b)
$$\hat{F}(X_3 | X_2 X_6) = \sum_{h=1}^{H} P(S = h)P(X_3 | X_2, X_6, S = h)$$

$$\hat{F}(X_7, X_8) = \sum_{h=1}^{H} P(S=h) P(X_7, X_8 \mid S=h)$$

type (d)

Note that type (b), (c) and (d) nodes may admit more than one subgroup (the two type (c) subgroups in Figure 6 are just an example). As shown in the previous example, each of these subgroups should be estimated distinctly. In general, if there are T, V and W distinct type (b), (c) and (d) nodes, with labels in the sets B_t, t=1,…T, C_v, v=1,…,V, D_w, w=1,…W, the general form of the Bayesian network (BN) based estimator is (Ballin et al, 2005e):

$$\hat{F}(X_1,...,X_k) = \left[\prod_{v=1}^{V} \hat{F}(\mathbf{X}_{C_v})\right] \hat{F}\left(\mathbf{X}_A \middle| X_{C_v}, v=1,...V\right) \left[\prod_{t=1}^{T} \hat{F}\left(\mathbf{X}_B \middle| \mathbf{X}_A, \mathbf{X}_{C_v}, v=1,...V\right)\right]\left[\prod_{w=1}^{W} \hat{F}\left(\mathbf{X}_{D_w}\right)\right]$$

A Monte Carlo experiment in Ballin et al 2005(c) shows that the BN based estimators can be much more efficient than the usual ratio estimators. The key idea is that the use of estimators linear in the weights introduce implicitly dependence induced by marginalisation with respect to S. For this reason, each type of node and each subgroup should be estimated distinctly with respect to S. As a result, if the interest is just on a few marginal tables instead of the complete joint distribution of the variables of interest, this approach gives results which are internally consistent (see Ballin *et al*, 2005d), i.e. if two tables contain the same variable, its marginal distribution is always the same. Ballin *et al* (2005e, Proposition 1) define a list of necessary and sufficient conditions that ensure that the dependence model of the set of variables is respected (and hence the disseminated tables are consistent). Finally, Ballin *et al* (2005b) and (2005e) show that the usual calibration estimators (that in case of categorical variables are poststratification estimators) can be equivalently defined as updating procedures in a BN, and this ensures the possibility to enlarge the set of possible poststratification procedures.

Comment: As a matter of fact, it seems that survey weights may have an unpleasant effect on the usual estimators computed as linear functions of the weights: the introduction of dependencies that actually do not hold true. The introduction of a wrong dependence relationship makes the estimator

structure more complex, and consequently less efficient. BN based estimators are non linear in the weights but still make use of the weights, without the unpleasant introduction of spurious dependencies.

All the previous results are obtained given the BN structure. Estimation of a structure of a BN in a finite population setting is still an unsolved problem. The possible translation of the PC algorithm through changes of the test statistics in order to take into account the complexity of the survey design is discussed in Ballin *et al* (2005a).

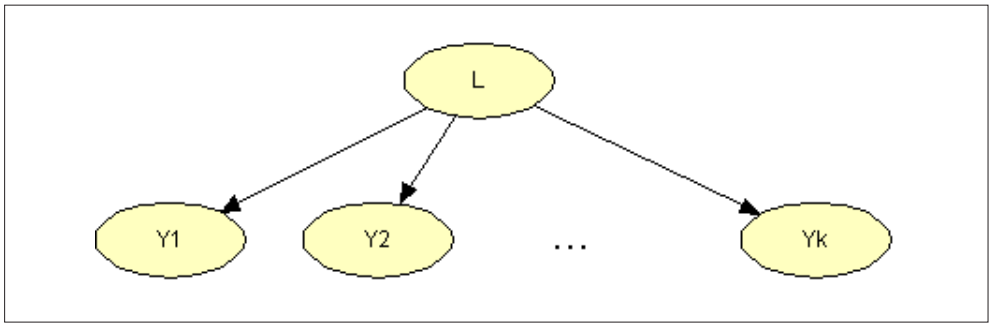
3.3 Other applications of Bayesian networks and possible extensions

Statistical Information Systems - One of the first Bayesian networks applications in official statistics is in Getoor et al. (2001a). They show how a very complex data base, as the one of the 1990 U.S. Census, can be easily and efficiently represented by a Bayesian network. The Bayesian network shows which contingency tables are necessary in order to describe the overall statistical information in the census, reducing the figures to store. They also show what numerical computations are necessary for any queries, i.e. how to join information from the different tables suggested by the Bayesian network. The example that the authors consider is relative to a data set which refers to just one kind of statistical unit. In general, the available tables may refer to different kinds of units: for instance some variables may refer to individuals, other to families, other to geographical (regions, counties,…) or institutional (hospitals, schools,…) entities. Getoor *et al* (2001b) show how to extend the concept of Bayesian network to the case of data referring to multiple kinds of units: they call this tool *Probabilistic Relational Model* (PRM). This tool is based on a knowledge representation language describe in Koller *et al* (1997). It seems particularly suitable for designing statistical information systems.

Record Linkage – When it is necessary to match records belonging to the same statistical unit in two data sets, but the record identifiers in the two data sets are subject to error, record linkage procedures are used (ISTAT, 2003, and references therein). Winkler (2002) shows which DAG structure is implicitly used for the naïve record linkage procedure. Assuming that the status of matched and unmatched pairs of records is represented by a (latent) variable

L, the DAG structure for the naïve record linkage procedure is represented in Figure 7. As a matter of fact, it corresponds to the so called conditional independence assumption.

Figure 7: Bayesian network for the naïve record linkage procedure, where *L* is the latent status of pair, and $Y_1,\ldots Y_k$ are the comparison of the two records in the pair with respect to the *k* matching variables



It is well discussed how the naïve record linkage procedure can lead to misleading results. It is important to investigate other approaches. For instance, Friedman (1997) shows how to estimate Bayesian networks in presence of latent variables. This approach can suggest alternative multivariate models able to link appropriately the record pairs.

Time series – Penny *et al* (2004) describe by means of BNs the multivariate dependence structures of time series. They apply this description to the quarterly gross national expenditure in New Zealand. Their objective is to identify which components of the gross national expenditure deserve to be improved in terms of timeliness.

4. Further developments

First of all, the applications discussed in Section 3 still need to be further explored and compared to the "traditional" ones. Nevertheless, it seems that Bayesian networks can be useful in many other different topics. Two of them appear particularly promising.

1. Integration of surveys – Following Ballin *et al* (2001), the different surveys can be designed as a *junction tree* (i.e. the tool used for the propagation of information in a Bayesian network, see Cowell, 1998, and Jensen, 1996). This network can perform as a tool for jointly analyzing variables only when strict model assumptions hold (this case corresponds to the statistical matching problem, see D'Orazio *et al*, 2005). Nevertheless it seems to be a formidable tool for updating survey results according to new information from archives or new surveys. In this case, it is necessary to understand the interaction between BN based estimators (Section 3.2) and calibration, poststratification, ratio raking (Harora *et al,* 1977a-b), and repeated weighting (Houbiers, 2003) estimators
2. Editing – The possibility to include logical rules in the estimation of the joint distribution of multiple variables, as well as to include the definition of "rare" events to be further investigated, suggest that editing procedures can be appropriately defined via Bayesian networks.

Acknowledgments

This is hopefully a comprehensive review article on the use of Bayesian networks in official statistics. I am indebted with all those who had the patience to introduce me to the different topics touched in this paper and with those I had the chance and fortune to discuss and work with: in alphabetical order Marco Ballin, Marco Di Zio, Orietta Luzi, Julia Mortera, Paola Vicard.

References

Ballin, M., M. Scanu, and P. Vicard. 2005a. "Bayesian Networks for finite populations". In *Atti del Convegno Metodi di Indagine e di Analisi per le Politiche Agricole (MIAPA)*: 95-106. Pisa, Italy, 21-22 October 2004.

Ballin, M., M. Scanu, and P. Vicard. 2005b. "Information propagation in finite survey sampling: Bayesian networks and poststratification". In Liseo, B., G.E. Montanari, e N. Torelli (*a cura di*). *Metodi Statistici per l'Integrazione di Dati da Fonti Diverse*. Milano: Franco Angeli.

Ballin, M., M. Scanu, e P. Vicard. 2005c. "Reti bayesiane per la costruzione di stimatori in popolazioni finite". In *Atti del Convegno Agristat, Verso un Nuovo Sistema di Statistiche Agricole*. Firenze 30-31 Maggio 2005.

Ballin, M., M. Scanu, and P. Vicard. 2005d. "Coherence of sample estimates for finite populations: some results based on Bayesian networks". In *Atti Convegno S.Co2005, Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*. Bressanone, Italy, 15-17 September 2005.

Ballin, M., M. Scanu, and P. Vicard. 2005e. "Bayesian networks and complex survey sampling from finite populations". In *FCSM Conference*. Arlington, Virginia, 14-17 November 2005.

Ballin, M., and P. Vicard. 2001. "A proposal for the use of graphical representation in official statistics". In *Atti del Convegno S.Co2001, Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*. Bressanone, Italy, 24-26 September 2001.

Chambers, R.L., and C.J. Skinner (*eds.*). 2003. *Analysis of Survey Data*. Chichester, UK: John Wiley & Sons.

Charniak, E. 1991. "Bayesian networks without tears". *AI Magazine*: 50-63.

Coppola, L., M. Di Zio, O. Luzi, A. Ponti, and M. Scanu. 2002. "Bayesian networks for imputation in official statistics: A case study". In *Proceedings of the Data Clean Conference*. Jyvaskyla, Finland, 29-31 May 2002.

Coppola L., M. Di Zio, O. Luzi, A. Ponti, and M. Scanu. 2002b. "On the use of Bayesian networks in official statistics". In *Atti della XLI Riunione Scientifica della Società Italiana di Statistica:* 237-240. Milano, Italy, 5-7 June 2002.

Cowell, R.G. 1998. "Introduction to inference for Bayesian networks". In Jordan, M.I. (*eds.*). *Learning in Graphical Models*. NATO ASI Series (Series D: Behavioural and Social Sciences), Volume 89. Dordrecht, The Netherlands: Springer.

Cowell, R.G., A.P. Dawid, S.L. Lauritzen, and D.J. Spiegelhalter. 1999. *Probabilistic Networks and Expert Systems*. New York, NY, U.S.: Springer-Verlag.

D'Orazio, M., M. Di Zio, and M. Scanu. 2006. *Statistical Matching: Theory and Practice*. Chichester, UK: John Wiley & Sons.

Dawid, A.P. 1979. "Conditional independence in statistical theory". *Journal of the Royal Statistical Society*, Series B (*Methodological*), Volume 41, N. 1: 1–31.

Di Zio, M., M. Scanu, and P. Vicard. 2003. "Open problems and new perspectives for imputation using Bayesian Networks". In *Atti del Convegno* S.Co2003, *Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*: 170-175. Treviso, Italy, 4-6 September 2003.

Di Zio, M., M. Scanu, L. Coppola, O. Luzi, and A. Ponti. 2004a. "Bayesian networks for imputation". *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Volume 167, Issue 2: 309-322.

Di Zio, M., G. Sacco, M. Scanu, and P. Vicard. 2004b. "Some approaches in imputing missing items with Bayesian networks". In *Atti della XLII Riunione Scientifica della Società Italiana di Statistica*. Bari, Italy, 9-11 June 2004.

Di Zio, M., G. Sacco, M. Scanu, and P. Vicard. 2004c. "Multivariate techniques for imputation based on Bayesian networks". In Antoch, J. (*ed.*). *Proceedings Compstat 2004*: 928-934. 16th Symposium of IASC, Prague, 23-27 August 2004. Heidelberg, Germany: Physica-Verlag, Springer.

Di Zio, M., G. Sacco, M. Scanu, and P. Vicard. 2005. "Methodology and software for imputation of missing values by Bayesian networks". In Liseo, B., G.E. Montanari, e N. Torelli (*a cura di*). *Metodi Statistici per l'Integrazione di Dati da Fonti Diverse*. Milano: Franco Angeli.

Friedman, N. 1997. "Learning belief networks in the presence of missing values and hidden variables". In *Fourteenth International Conference on Machine Learning (ICML97)*.

Getoor, L., B. Taskar, and D. Koller. 2001a. "Selectivity estimation using probabilistic models". In *ACM-Sigmod*. Santa Barbara, CA, U.S., 21-24 May 2001.

Getoor, L., N. Friedman, D. Koller, and A. Pfeffer. 2001b. "Learning probabilistic relational models". In Dzeroski, S., and N. Lavrac (*eds.*). *Relational Data Mining*. Heidelberg, Germany: Springer-Verlag.

Harora, H.R., and G.J. Brackstone. 1977a. "An investigation of the properties of raking ratio estimators I: with simple random sampling". *Survey Methodology*, Volume 3: 62-83.

Harora, H.R., and G.J. Brackstone. 1977b. "An investigation of the properties of raking ratio estimators II: with cluster sampling". Survey Methodology, Volume 3: 232-243.

Helm, L. 1996. "Improbable inspiration". *Los Angeles Times*, October 28th, 1996.

Houbiers, M., P. Knottnerus, A.H. Kroese, R.H. Renssen, and V. Snijders. 2003. "Estimating consistent table sets: position paper on repeated weighting". *Discussion paper* N. 03005, CBS - Statistics Netherlands.

Istituto Nazionale di Statistica – Istat. 2003. "Metodi Statistici per il Record Linkage". *Collana Metodi e Norme*, N. 16. Roma: Istat.

Jensen, F.V. 1996. *An introduction to Bayesian Networks*. New York, NY, U.S.; Springer-Verlag.

Kadie, C.M., D. Hovel, and E. Horvitz. 2001. "MSBNx: A component-centric toolkit for modeling and inference with Bayesian networks". *Microsoft Research Technical Report*, MSR-TR-2001-67, July 2001.

Koller, D, A. Levy, and A. Pfeffer. 1997. "P-classic: a tractable probabilistic description logic". In *Proceedings of the Fourteenth Conference on Artificial Intelligence* (AIII-97): 390-397. Providence, Rhode Island, August 1997.

Little, R.J.A., and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Chichester, UK: John Wiley & Sons.

Neapolitan, R.E. 2004. *Learning Bayesian Networks*. Upper Saddle River, NJ, U.S.: Prentice Hall.

Penny, R.N., and M. Reale. 2004. "Using graphical modelling in official statistics". *Quaderni di Statistica*, N. 6: 31-47.

Righi, P. 2005. "Trattamento delle mancate risposte nelle indagini longitudinali mediante modelli grafici ricorsivi". *Tesi di Dottorato in Metodi Statistici per l'Economia e l'Impresa*. Roma, Italia, Università degli Studi Roma Tre, XVI ciclo.

Sebastiani, P., and M. Ramoni. 2001a. "Bayesian Selection of Decomposable Models With Incomplete Data". *Journal of the American Statistical Association*, Volume 96, N. 456: 1375-1386.

Sebastiani, P., and M. Ramoni. 2001b. "On the use of Bayesian networks to analyse survey data". *Research in Official Statistics - ROS*, Volume 4, N. 1: 52-64.

Spirtes, P., C. Glymour, and R. Scheines. 2000. *Causation, Prediction and Search. Second edition*. Cambridge, MA, U.S.: MITCogNet, MIT Press.

The Economist. 2001. "Son of paperclip". *The Economist, print edition*, March 24th 2001.

Thibaudeau, Y., and W.E. Winkler. 2002. "Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints". Technical report, RRS2002/9. U.S. Bureau of the Census.

Verma, T.S., and J. Pearl. 1990. "Equivalence and synthesis of causal models". In *Proceedings of the Sixth Conference on Uncertainty in AI*: 220-227. Cambridge, MA, U.S., 27-29 July 1990.

Whittaker, J. 1990. *Graphical Models in Applied Multivariate Statistics*. Chichester, UK: John Wiley & Sons.

Winkler, W.E. 2002. "Methods for record linkage and Bayesian network". *Research Report*, Series n. 2002/5. U.S. Bureau of the Census.