

Applications of Bayesian networks in official statistics¹

Paola Vicard

Dipartimento di Economia, Università Roma Tre
Via Silvio D'Amico 77, Roma, e-mail: vicard@uniroma3.it

Abstract: this paper is devoted to the use of Bayesian networks in topics of central interest in official statistics. In particular we study problems related to: missing item imputation; estimation of contingency tables in complex survey sampling; quality of data production process.

Keywords: contingency table, imputation, influence diagrams, Markov blanket, paradata, stratified sampling design.

1. Introduction

Statistical analyses can be particularly complex when are referred to surveys and databases produced by a National Institute of Statistics. The complexity is mainly due to: high number of surveys carried out by the institute, sampling design complexity, high number of variables and huge sample size. In this context it can be useful to analyse and exploit the dependence structures.

Bayesian networks (Cowell *et al.*, 1999), from now on BNs, are a multivariate statistical model able to represent and manage complex dependence structures. The theoretical setting of BNs and of graphical models is the basis for developing methods for efficiently representing and managing survey systems. A known (or previously estimated) dependence structure can help: in model based and model assisted estimators computation; when coherence constraints among different surveys must be fulfilled; in missing data imputation. Furthermore, since BNs can be extended to embody decision and utility nodes giving rise to BNs for decisions (Jensen, 2001, and Lauritzen and Nielson, 2001), they can be used for data collection monitoring that is carried out on the basis of various factors determining the final quality of the survey.

Therefore BNs can be thought of as a unique framework by which it is possible to manage a survey from planning, passing through imputation and estimation to integration with partially overlapping surveys.

In the next sections we will survey recent results on the application of BNs in official statistics contexts focusing on categorical variables.

2. Missing data imputation

The treatment of missing values in survey data is one of the most important aspect to be taken into account in the data production process (Little and Rubin, 2002). A common approach is to impute missing items with artificial plausible values, under the assumption that the missing data mechanism is missing at random. A wide variety of

¹ The present paper is financially supported by MIUR-PRIN 2007 research project.

imputation techniques has been developed. While most of them have desirable properties for univariate characteristics, the preservation of relationships among variables is still a problem. BNs are a well-known tool to study the relationships among variables therefore they can be usefully applied for imputation.

A first approach was presented by Thibaut and Winkler (2002). Further research has been developed by Di Zio *et al* (2004, 2005, 2006).

Let us consider a sample where k variables, $X_1, \dots, X_k$, are observed on n units. Suppose item x_{ij} is missing and we want to impute it according to information carried by the other variables. BNs help in efficiently collecting and propagating information about variable X_j . The proposed methodologies differ in the way information collection is performed. Given a network for the dependence structure among $X_1, \dots, X_k$ and given an ordering of the variables in terms of their reliability, in Di Zio *et al.* (2004) a missing value of X_j is imputed on the bases of information carried by all the variables directly influencing X_j , *i.e.* linked to X_j by an arrow pointing to it. These variables are named *parents* of X_j (written $pa(X_j)$). Information is then automatically propagated through the network giving rise to the updated conditional distribution of X_j given its parents. Finally, the imputed value is drawn at random from this conditional distribution.

As a matter of fact, given a network, the information on a node, say variable X_j , is carried not only by its parents but by its *Markov blanket* constituted by the parents, the children and the parents of the children of X_j . In Di Zio *et al.* (2005) an imputation algorithm based on Markov blanket is presented. The reliability ordering of the interest variables is still important. Imputation is now performed iteratively on each variable at a time; starting from the most reliable. Its conditional distribution is updated according to its Markov blanket and the missing values are replaced by a random draw from the updated conditional distribution. Then the initial data set is replaced by a new one where the column of the first variable is complete. We next pass to consider the second reliable variable and impute its missing items as above using the updated data set. The algorithm end once the missing values of all the variables have been imputed.

A different algorithm based on the Markov blanket has been proposed in Di Zio *et al.* (2006). For each record, the overall vector of missing variables is imputed simultaneously by a random draw from its probability distribution conditional on its Markov blanket. Simulation and experiments have been carried out and it was shown that these algorithms have good performance when compared with other imputation practices. Furthermore a software, BNimput, has been developed to perform the above algorithms.

3. Estimation of contingency tables in complex survey sampling

BNs are particularly useful to integrate two (or more) surveys. In fact BNs are characterized, among the others, by two positive aspects (Ballin and Vicard, 2001): i) an easy-to-interpret, concise and, above all, informative way to represent a set of surveys with their dependence structure; ii) a method to update the information produced by a system of surveys, *i.e.* to pass information between surveys. The first is a crucial aspect and requires the development of BN-based methods to derive estimators when the sampling design is complex. Having BN-based estimators, it becomes possible to apply poststratification or calibrate estimates by means of the propagation algorithm based on *junction tree*. Ballin *et al.* (2005) have proposed BN-based contingency table estimators

under a stratified sampling design both in a model based and in a model assisted framework.

In a model based approach the design variables are modelled together with the variables of interest therefore the BN includes, in addition to the k nodes associated to the k interest variables, a node, say D , representing the sampling design, with as many states as the number, H , of strata. Moreover we assume that node D has no parents (*i.e.* is a root of the network). By the chain rule proper of BNs the joint distribution of $D, X_1, \dots, X_k$ can be written as

$$\theta_{hx_1 \dots x_k} = \theta_h \prod_{j=1}^k \theta_{x_j | pa(x_j)}$$

$\theta_{hx_1 \dots x_k}$ being the frequency distribution of $D, X_1, \dots, X_k$. The BN-based estimator is

$$\hat{\theta}_{x_1 \dots x_k}^{BN} = \sum_{h=1}^H \theta_h \prod_{j=1}^k \hat{\theta}_{x_j | pa(x_j)} \cdot \quad (1)$$

In the model assisted framework the design variable is not directly modelled with $X_1, \dots, X_k$. Information on D is incorporated in survey weights. Assuming a BN for $X_1, \dots, X_k$, whose joint probability distribution is $\theta_{x_1 \dots x_k}$, its BN-assisted estimator is

$$\hat{\theta}_{x_1 \dots x_k}^{BN} = \prod_{j=1}^k \hat{\theta}_{x_j | pa(x_j)} \quad (2)$$

where each factor is a weighted estimator of the (conditional) distributions. It is easy to show that when the network is complete, *i.e.* all the nodes are directly connected between each other, (1) and (2) coincide with the Horvitz-Thompson estimator.

Simulation studies have been performed to compare these estimators (1) and (2) with Horvitz-Thompson estimator and among themselves. It has been shown that both model based and model assisted BN estimator are better than Horvitz-Thompson estimator when the network is not complete. Moreover estimator (2) is more robust than (1) in case of model misspecification.

4. Monitoring data production process

Data collection monitoring is carried out on the basis of various factors determining the final quality of the survey. The indices used to monitor the survey are called *paradata* while a survey process that can be modified on the bases of paradata is called *responsive design* (Heeringa and Grooves, 2004 and Hansen and Maher, 2005). In a responsive design the survey production is treated as a decision process where every decision (with associated cost/benefit) is taken by minimizing a cost function – the expected cost is updated on the basis of paradata. In this context it is important to first formalize the decision process.

Ballin *et al.* (2006) have proposed to use BNs for decisions (usually named *influence diagrams*) to model and solve this problem.

Influence diagrams are made up of: chance nodes representing random variables; decision nodes representing decisions; utility nodes representing utility functions. When applied to monitoring data production process, the different kinds of nodes play the following roles: chance nodes represent paradata; decision nodes represent the eventual intervention needed to improve the survey quality and utility nodes represent survey costs. In this way a monitoring and decision supporting system is set up. Once evidence on quality indicators (paradata) arrives, it is inserted and both propagation and optimization algorithms allow to efficiently propagate and solve a decision problem about the necessity to eventually modify the data collection process.

Aknowledgements

Thanks are due to Marco Ballin, Marco Di Zio, Giuseppe Sacco and Mauro Scanu.

References

- Ballin M., Vicard P. (2001) A proposal for the use of graphical representation in official statistics, in: *Proceedings of SCO2001*, September 2001, Bressanone/Brixen.
- Ballin M., Scanu M., Vicard P. (2005) Model assisted approaches to complex survey sampling from finite populations using Bayesian networks. Working paper n. 54, Università Roma Tre. URL: <http://tinyurl.com/dxufm9>.
- Cowell R. G., Dawid A. P., Lauritzen S. L., Spiegelhalter D. J. (1999) *Probabilistic Networks and Expert Systems*, Springer, Heidelberg.
- Di Zio M., Scanu M., Coppola L., Luzi O., Ponti A. (2004) Bayesian Networks for Imputation, *Jour. Roy. Statist. Soc./A*, 167 (2), 309-322.
- Di Zio M., Sacco G., Scanu M., Vicard P. (2006) Metodologia e software per l'imputazione di dati mancanti tramite le reti bayesiane, in: *Metodi statistici per l'integrazione di dati da fonti diverse*, Liseo, B., Montanari, G. E. and Torelli, N., (Eds.), Franco Angeli, 307-321.
- Di Zio M., Sacco G., Scanu M., Vicard P. (2005) Multivariate Techniques for Imputation based on Bayesian Networks *Neural Network World*, 15, n. 4, 303-309.
- Hansen, S. E., Maher, P. (2005) Using Process Data for Responsive Design, in *Proceedings of the XXII International Methodology Symposium*, Ottawa, Canada.
- Heeringa S. G., Groves R. M. (2004) Responsive Design for Household Surveys, in: *Proceedings of the Survey Research Methods Section, ASA*.
- Jensen F. V. (2001) *Bayesian Networks and Decision Graphs*, Springer, New York.
- Lauritzen S. L., Nilsson D. (2001) Representing and Solving Decision Problems with Limited Information, *Management Science*, 47, 1235-1251
- Little J., Rubin D., (2002) *Statistical Analysis with Missing data*, Wiley, New York.
- Thibaudeau Y., Winkler W.E. (2002) Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints. Technical report RRS2002/9, U.S. Bureau of the Census. URL: www.census.gov/srd/papers/pdf/rrs2002-09.pdf