

On the use of Bayesian Networks in Official Statistics

Sull'uso delle Reti Bayesiane nella Statistica Ufficiale

Lucia Coppola, Marco Di Zio, Orietta Luzi,

Alessandra Ponti, Mauro Scanu

ISTAT, via Cesare Balbo 14, 00184 Roma, e-mail: scanu@istat.it

Riassunto: Le reti bayesiane forniscono una rappresentazione sintetica delle relazioni di dipendenza fra un certo gruppo di variabili e della corrispondente distribuzione congiunta. Tale informazione è utile in diversi campi della statistica ufficiale. In questo lavoro si propone l'impiego delle reti bayesiane per imputare valori mancanti di un'indagine. L'applicazione è stata svolta simulando diverse percentuali di dati mancanti a partire da un data-set completo e reale. I risultati fino ad ora ottenuti confermano le potenzialità del metodo.

Keywords: Graphical structure, multivariate analysis, imputation techniques.

1. Bayesian networks and official statistics

Official Statistics often has to deal with multivariate distributions. They may be considered as an output of the surveys but they are also crucial as an input for many Official Statistics purposes. A synthetic way to describe multivariate distributions is offered by Bayesian Networks (Cowell et al., 1999), in the following denoted BN. Getoor et al. (2001) use BN as compact graphical representations for high-dimensional joint distributions. Their reasoning is that only a few statistical relationships among variables are important for defining the structure of the graphical representation, and they try to both identify the structure and to estimate the distribution under the learned structure. They consequently extend such reasoning to Probabilistic Relational Models, when dealing with different kinds of units. As an example, they show an application to the 1993 Current Population Survey of the U.S. Census Bureau. The multivariate distribution given in a BN may be applied in many different contexts. Thibaudau et al. (2001) suggest the use of such multivariate distribution for the imputation of missing values. Other contexts are analysed in Sebastiani et al. (2001) and Winkler (2000). This work is devoted to the problem of imputation of missing values by using BN.

2. Imputation by BN

Item non-responses correspond to failures in obtaining information on some questionnaire items. In Official Statistics, this kind of non-response is usually dealt with *imputation*, i.e. methods whose aim is the prediction of missing items. Kalton et al. (1982) give a review of many different imputation methods pointing out their advantages and drawbacks. Imputation methods generally make use of all the available information (from the sampling frame, reported in questionnaires, from administrative sources, etc.). It is clear that whether there is no relationship between such information

and the items to be imputed, or the information is poorly modelled by the imputation method, imputations could be an additional source of bias. Furthermore, imputations are generally performed under the assumption that missing values are Missing At Random (MAR) (Little et al., 1987), roughly non-respondents and respondents are assumed to have similar behaviours and characteristics.

The imputation technique we want to investigate has been formulated by Thibaudau et al. (2001) and it exploits the characteristics of a BN. A BN is characterised by a structure S , that describes the presence or absence of statistical relationship between pairs of variables, the directions of such relationships and a joint distribution of the variables conditional to the structure S . The structure S can be estimated by the PC algorithm (references in Cowell et al., 1999). It iteratively tests the effectiveness of the relationship between pairs of variables, conditional to appropriate sets of variables, by means of “chi-square type” tests. It also defines the edges direction, even if sometimes it is possible to redirect some of them according to common sense, under some rules (e.g. cycles are forbidden). Given the structure S , the joint distribution of the variables can be determined by maximum likelihood estimation. When missing values are present, the EM algorithm can be usefully applied. The structure S helps in partitioning the set of variables in k mutually exclusive subsets (P_i $i=0,\dots,k-1$), where k is the number of variables in the longest chain. The first subset P_0 contains variables with no parents. The second subset P_1 contains the remaining variables whose parents are only in P_0 . The third subset P_2 contains the remaining variables whose parents are in $P_0 \cup P_1$, and so on. Once the variables have been partitioned, the estimated joint distribution is considered. For each unit a , $a=1,\dots,n$, in the sample, we check the presence of a missing value starting from the variables in P_0 . If a variable X in P_0 is missing, X is imputed by randomly generating a value from the marginal distribution of X . If the variables in P_{i-1} ($i>1$) are either present or imputed and a variable X in P_i is missing, X is imputed by randomly generating a value from the distribution of X conditional to its parents. Such procedure is iterated for all the variables in the k groups and all the units $a=1,\dots,n$.

3. Data description and error simulation

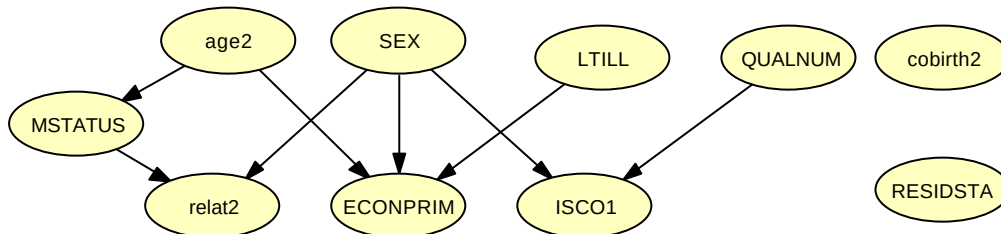
The application presented in the paper has been performed by using only categorical test data. In particular, a subset of anonymous individual records and variables surveyed in 1991 U.K Population Census (SARS) have been used. For each sampled household member, only 10 demographic variables were analysed: age, sex, marital status (*mstatus*), relationship to household head (*relat2*), primary economic position (*econprim*), long term illness (*ltill*), occupation code (*isco2*), number of higher education qualifications (*qualnum*), country of birth (*cobirth*), whereabouts on census night (*residsta*). The experiment consists of two steps: 1) random generation of artificial missing items in the set of complete and coherent SARS data; 2) missing items imputation by means of the BN-based method. In this first experiment we generate only Missing Completely At Random items (MCAR). SARS data were contaminated by different percentages of missing items (1%, 5%, 10%) in order to evaluate the effects on results of different levels of contamination. For each contamination level, the structure S of the dependencies among the variables has been estimated calibrating the significance level of the tests at 1% and 5%. Both the structure and the distribution were estimated using respectively the complete data-set (optimal case) and the “observed” data-sets, i.e.

the ones with the generated missing values. Finally the imputation procedure described in section 2 has been applied. A software code in JAVA has been written for the purpose. The estimated structure and results are shown in the next section.

4. Results and considerations

The estimated structures in the different simulation contexts contain just minor differences. In general, variables without ancestors are *sex*, *age*, *ltill* and *qualnum*, while *cobirth2* and *residsta* are independent (disconnected) from the other variables. In all the estimated structures, the PC algorithm suggests a link from *mstatus* towards *age2*. According to common sense, we have inverted the direction of the arrow. Instead of such a posterior action, it would be desirable to investigate the possibility of introducing available information before the learning process starts (as in Sebastiani et al., 2001). Figure 1 shows the structure we estimated from the complete data case.

Figure 1: Estimated structure for the complete data file



As far as the imputation process is concerned, the previous graph suggests, for example, that *relat2* should be imputed taking care only of its parents *sex* and *mstatus* (whenever *mstatus* is missing the imputation should check the ancestor *age2*), but the other variables are unnecessary. Another interesting result is that imputation of *residsta* and *cobirth2* is performed independently from the other variables. As a result, the complexity of the problem is greatly reduced.

Generally, the quality of an imputation method is evaluated with respect to three criteria (Chambers, 2000): 1) the individual predictive accuracy (preservation of micro data); 2) the distributional accuracy (preservation of marginal and joint distributions); 3) the final data plausibility (coherence among individual items with respect to logical constraints).

Table 1: Expected frequency (EF) of correct imputations for the variable *residsta*

Percentage of missing items for <i>residsta</i>	1%	5%	10%
<i>EF</i>	79,9%	73,7%	73,2%

Regarding the first criterion, descriptive indices are usually computed. Instead this approach allows the computation of an expected frequency (*EF* in the following) of correctly recovering original values. This can be obtained by computing for each observation the probability of correctly imputing the missing items with respect to the probabilities estimated by BN and averaging with respect to the imputed units. Table 1 shows the expected number of correct imputations for the variable *relat2* for three

different contamination situations (the networks are estimated from the data-sets with missing values; the significance level of the tests is set to 1%). The *EF* for the data-set with an average of 10% missing items per variable needed a double order of expectations because some of the missing *residsta* items have at least one missing parent. In some cases the expectation with respect to a missing parent may have a positive effect on *EF*. As far as the second criterion is concerned, the proposed imputation strategy is expected to preserve joint distributions. The capability of preserving joint distributions has been evaluated by the expected value of a distance between the true and the imputed values distributions within the set of units with missing items. This index is zero when the distributions are identical and one when the distributions are concentrated in two different categories. The indices for the different simulated data-sets are close to zero for all the partial distributions (results are available upon request). The last criterion is generally the most difficult to be fulfilled by imputation techniques. In addition to this problem, techniques allowing coherent imputations need a subject matter expert to define logical edit rules. On the contrary, if our observed data satisfy the logical edit rules (e.g. data obtained through a follow up) BN imputations will be expected to satisfy logical constraints. This can be inferred by the fact that the probabilities of the cells with no observations in the complete true data-set, are set to zero in the learned BN model. When the distribution is estimated from data with missing items, such probabilities usually assume positive but near zero values. In order to avoid, according to edit rules, the possibility to impute a forbidden value it would be desirable to constrain the learning process to logical and structural zeros. In our example, most of the selected variables are related through logical rules (edits) used in the imputation strategy adopted by the U.K. National Statistical Office and expressing conditions for acceptable or coherent information. All these edit rules are fulfilled by the imputed data-set.

Further investigations are also needed for evaluating the performance of the proposed approach in the case of MAR mechanism and with errors in data.

References

- Chambers R. (2000) Evaluation Criteria for Statistical Editing and Imputation, *T001.05, EUREDIT report* (www.cs.york.ac.uk/euredit/).
- Cowell R.G., Dawid A.B., Lauritzen S.L. and Spiegelhalter D.J. (1999) *Probabilistic Networks and Expert Systems*, Springer Verlag, New York.
- Getoor L, Taskar B. and Koller D. (2001) Selectivity estimation using probabilistic models, *ACM-SIGMOD*, Santa Barbara, California, USA.
- Kalton G., Kasprzyk D. (1982) Imputing for missing survey responses, *Proceedings of the section of Survey Research Methods*, American Statistical Association, 22-31.
- Little, J., Rubin, D. (1987) *Statistical Analysis with Missing data*, J. Wiley, New York.
- Sebastiani P., Ramoni, M. (2001). "On the use of Bayesian networks to analyse survey data". *Research in Official Statistics*, **4**, 52-64.
- Thibaudreau Y., Winkler W.E. (2001) "Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints". Technical report, U.S. Bureau of the Census.
- Winkler W.E. (2000) Machine learning, information retrieval and record linkage, *2001 Proceedings of the Section on Survey Research Methods, ASA*, 20-29.