Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

MODEL ASSISTED APPROACHES TO COMPLEX SURVEY SAMPLING FROM FINITE POPULATIONS USING BAYESIAN NETWORKS: A TOOL FOR INTEGRATION OF DIFFERENT SOURCES

Marco Ballin, Mauro Scanu, and Paola Vicard[1]

ABSTRACT

A class of estimators based on the dependency structure of a multivariate variable of interest and the survey design is defined. The dependency structure is the one described by the Bayesian networks. This class allows ratio type estimators as a subclass identified by a particular dependency structure. It will be shown by a MonteCarlo simulation how the adoption of the estimator corresponding to the population structure is more efficient than the others. It will also be underlined how this class adapts to the problem of integration of information from two surveys through the probability updating system of the Bayesian networks.

KEY WORDS: Graphical models; probability update.

1. INTRODUCTION

The whole information collected by one or more surveys is undoubtedly a complex system. For this reason, it is not only important to organize these surveys in an appropriate system (e.g. for sample coordination or coherence of definitions) but also to provide tools suitable to support representation, estimation and updating.

In this paper we will show how Bayesian Networks (BN) can help in achieving such goals taking into account the survey design. BNs have been successfully applied in several contexts as artificial intelligence, forensic statistics, genetics, computer troubleshooting and other fields where it is common to deal with a large number of variables linked by a complex dependence structure (Neapolitan 2004). BNs have already been used also in official statistics, e.g. for the description of some census results (Getoor et al. 2001) and for imputation of missing values (Thibaudeau and Winkler 2002, Di Zio et al. 2005 and references therein).

It should be noted that, while in the previous analyses the i.i.d. assumption is natural and well justified, in the context of complex survey designs this assumption does not hold anymore. Our proposal is to explicitly represent the sampling design in the BN and to highlight its statistical relationship with the variables of interest. This approach defines a class of estimators of the joint distribution of these variables that includes the usual ratio type estimators. Moreover information propagation properties of BNs (Cowell et al. 1999) will be used to update the estimates of a survey once an informative shock resulting from another survey occurs. In this sense BNs help in visualizing and understanding how information coming from different surveys interacts.

2. ESTIMATORS BASED ON BAYESIAN NETWORKS

Let $\mathcal{P}$ be a finite population of size n, and let $X_1,\ldots,X_k$ be k variables of interest. For the sake of simplicity, let these variables be categorical, with frequency distribution in $\mathcal{P}$

[1] Marco Ballin, Istat, via A. Ravà 150, 00100 Roma Italy; Mauro Scanu, Istat, via C. Balbo 16, 00184 Roma Italy (scanu@istat.it); Paola Vicard, Università Roma Tre, via Ostiense 139, 00154 Roma Italy (vicard@uniroma3.it).

$$F(x_1,...,x_k) = \sum_{i=1}^{N} \frac{I_{x_1...x_k}(x_{i1},...,x_{ik})}{N} \tag{1}$$

where $I(.)$ is the indicator function. Equation (1) is the parameter of interest. Let $\mathcal{S}$ be a sample drawn from $\mathcal{P}$ according to a complex survey design defined by the survey weights w_i, $i \in \mathcal{S}$. A natural and intuitive estimator of the distribution (1) is:

$$\hat{F}(x_1,...,x_k) = \sum_{i \in S} I_{x_1...x_k}(x_{i1},...,x_{ik}) \frac{w_i}{\sum_{i \in S} w_i} . \tag{2}$$

Estimator (2) is a ratio type estimator.

Let S be an additional (categorical) variable with as many categories as the different weight values, say $w_{(h)}$, $h=1,...H$, and marginal probability distribution

$$F(S=h) = \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}}, \quad h=1,...,H, \tag{3}$$

where n_h is the size of the subsample of units s_h of $\mathcal{S}$ with equal weight $w_{(h)}$, $h=1,...H$. Using S in (2), it is possible to hide the role of the weights conditioning on S, e.g.

$$\hat{F}(x_j \mid S=h) = \sum_{i \in s_h} \frac{I_{x_j}(x_{ij})}{n_h}, \quad \hat{F}(x_j \mid X_l = x_l, S=h) = \sum_{i \in s_h} \frac{I_{x_j x_l}(x_{ij}, x_{il})}{\sum_{i \in s_h} I_{x_l}(x_{il})} \tag{4}$$

for any $j=1,...K$, and $l \neq j$, and to highlight the role of S in the estimator (2) by using the chain rule (Cowell et al. 1999). This is possible by rewriting estimator (2) via a recursive factorization involving the factors in (3) and (4), i.e.

$$\begin{aligned} \hat{F}(x_1,...,x_k) &= \sum_{h=1}^{H} \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \sum_{i \in s_h} \frac{I_{x_1}(x_{i1})}{n_h} \sum_{i \in s_h} \frac{I_{x_1 x_2}(x_{i1}, x_{i2})}{\sum_{i \in s_h} I_{x_1}(x_{i1})} \cdots \sum_{i \in s_h} \frac{I_{x_1...x_K}(x_{i1},...,x_{iK})}{\sum_{i \in s_h} I_{x_1...x_{K-1}}(x_{i1},...,x_{iK-1})} \\ &= \sum_{h=1}^{H} F(S=h) \hat{F}(x_1 \mid S) \hat{F}(x_2 \mid X_1, S) \cdots \hat{F}(x_K \mid X_1, \cdots, X_{K-1}, S) . \end{aligned} \tag{5}$$

The factorization (5) shows that the usual ratio estimator (2) relies on a particular dependence model among the variables $X_1,...,X_K$, S. This dependence model is the situation of complete dependence between these variables, i.e. the *saturated* model. This can be represented graphically as a BN called *clique*. For instance, when there are three variables of interest X, Y, and Z under the design S, the clique corresponds to the network (a) of Figure 1. It should be noted that the order of the variables in factorization (5) is not unique and leads to alternative but equivalent graphical structures.

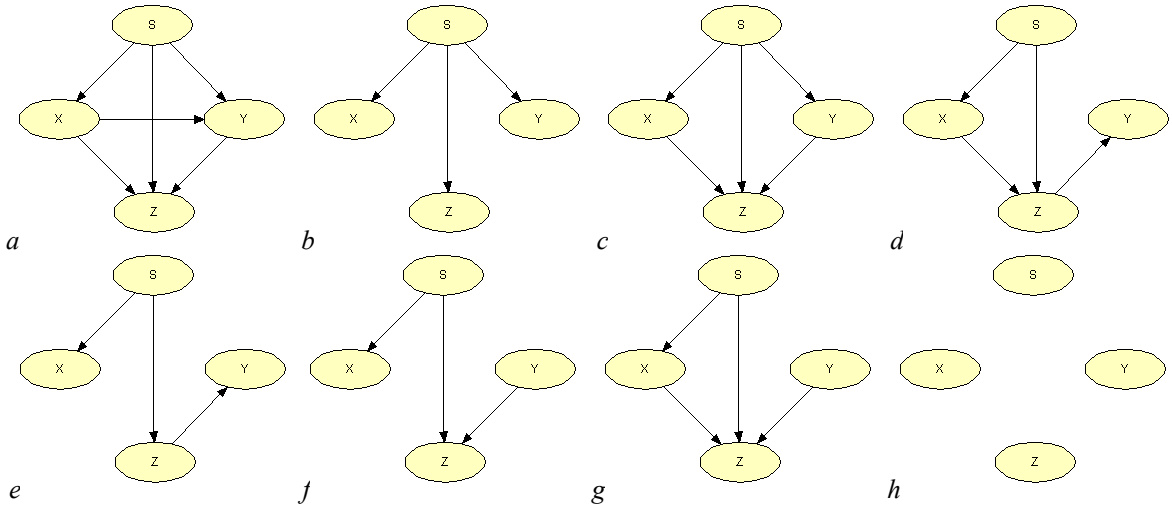

Figure 1 – Eight possible BNs representing the dependence structure between three variables of interest *X, Y, Z* and the survey design *S*.

It might happen that the clique is an overparameterized model because some variables are marginally or conditionally independent. For instance, networks (b)-(h) in Figure 1 show some simplified BNs. The dependence structure of a BN suggests an estimator based on the chain rule, whose general form is:

$$\hat{F}_{BN}(x_1,...,x_K) = \sum_{h=1}^{H} F(S) \prod_{j=1}^{K} \hat{F}(x_j \mid pa(X_j)) \tag{6}$$

For instance, the BN based estimator of the BN (f) of Figure 1:

$$\hat{F}_f = \left[\sum_{h=1}^{H} F(S)\hat{F}(x \mid S)\hat{F}(z \mid Y,S) \right] \hat{F}(y) = \left[\sum_{h=1}^{H} \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \sum_{i \in s_h} \frac{I_x(x_i)}{n_h} \sum_{i \in s_h} \frac{I_{yz}(y_i, z_i)}{\sum_{i \in s_h} I_y(y_i)} \right] \sum_{i \in S} \frac{I_y(y_i)}{n}$$

Note that $pa(X_j)$ does not necessarily include *S*, as in networks (d)-(h) of Figure 1. In that case, the estimator of the conditional distribution does not use sample weights. Furthermore, the class of estimators (6) based on the different BN structures is finite for a fixed survey design *S*, and always includes estimator (2).

Despite each estimated factor in (6) is design unbiased when the sum of the sample weights in each *S* category is constant with respect to sample variability (e.g. in the stratified sampling design), unbiasedness is ensured only for the estimator relative to the clique, i.e. estimator (2) (the only condition is that the sum of the sample weights equals *N*). The other estimators might be biased with respect to the multivariate joint frequency distribution of $X_1,...X_K$ in $\mathcal{P}$. Nevertheless there is empirical evidence that the exploitation of the dependency structure of the variables of interest and the survey design might lead to better estimators (Section 3).

A different class of BN based estimators is discussed in Ballin et al (2005). It consists in producing weights estimates also for those factors in Equation (6) that do not admit *S* in the conditional $pa(X_j)$. The comparison between these two alternative definitions will be provided elsewhere.

3. MONTE CARLO SIMULATION

A MonteCarlo simulation has been carried out in order to evaluate the performance of the estimators suggested by BNs and the consequences of possible model misspecifications. In particular, eight populations consisting of 10000 units have been generated according to the eight BNs of Figure 1. For each population 500 samples of size 1000 have been drawn according to the stratified sample design described in Table 1 (note that sampling fraction is not proportional to stratum size).

Stratum code h	Stratum size N_h	$F(S=h)$	Sample size n_h
$S=1$	5995	0,5995	100
$S=2$	2959	0,2959	200
$S=3$	1046	0,1046	700

Table 1: stratum and sample sizes

For each population and for each MonteCarlo replication the estimates of the joint distribution resulting from the estimator suggested by the corresponding BN have been compared with the true joint distribution. The performance of the estimator has been measured by the mean of MonteCarlo estimates of relative MSE of each element of the joint distribution:

$$MSE\left(\hat{F}_{BN}\right)=\frac{1}{V}\sum_{v=1}^{V}\sum_{x,y,z}\frac{\left[F(x,y,z)-\hat{F}_{BN,v}(x,y,z)\right]^2}{F(x,y,z)},$$

where V is the number of MonteCarlo replications and $\hat{F}_{BN,v}(x,y,z)$ is the estimate of $F(x,y,z)$ obtained with the v-th sample. In order to have an idea of the robustness of estimates against model misspecification, the joint distribution has been estimated using the estimators suggested by BNs associated to the other seven BN structures. The results are reported in Table 2, where each row refers to a population and each column refers to an estimator.

Pop	MSE($\hat{F}_a$)	MSE($\hat{F}_b$)	MSE($\hat{F}_c$)	MSE($\hat{F}_d$)	MSE($\hat{F}_e$)	MSE($\hat{F}_f$)	MSE($\hat{F}_g$)	MSE($\hat{F}_h$)
a	0.32	0.54	0.36	3.07	3.51	2.71	2.57	8.59
b	0.27	0.15	0.22	3.16	3.14	10.41	10.43	11.93
c	0.29	0.45	0.25	10.17	11.29	9.17	9.12	11.07
d	0.28	6.25	0.45	0.10	0.47	0.63	0.52	15.34
e	0.30	0.37	0.25	0.12	0.11	0.18	0.22	7.12
f	0.28	0.25	0.23	1.44	1.43	0.11	0.15	7.25
g	0.30	0.31	0.25	1.30	1.46	0.28	0.18	6.98
h	0.37	0.14	0.31	0.14	0.11	0.14	0.25	0.03

Table 2 – MonteCarlo estimates of relative MSE for the BN based estimators suggested by the BNs in Figure 1

In Table 2 the estimator is labelled with the same index of the corresponding population (the first column refers to the BN based estimator suggested by the BN (a) of Figure 1, the second to the BN based estimator suggested by the BN (b) of Figure 1, and so on.). This simulation shows that the estimators suggested by the corresponding BN perform always better than estimator (2) (compare values contained in the diagonal with those of the first column). Even if it is difficult to measure the "distance" between the data generating model and the model assumed by the usual ratio estimator $\hat{F}_a$, the comparison among values of the diagonal and values of the first column suggests that the gain of efficiency depends on such "distance". For example, in the case of population (b) (characterized by only three links) the gain of efficiency is higher (0.15 vs 0.27) than in the case of population (c) (0.25 vs 0.29) which differs from the usual model for only the link between X and Y.

The previous result is rather general. For population (h) the different estimators seem ordered according to the number of additional arrows. The worst estimator is actually $\hat{F}_a$, whose structure is the most distant. From the other rows it is possible to see that adding arrows seems to have a milder effect than deleting arrows. This is due to the

fact that a simplified structure is unable to take into account the actual dependences, while when data show independence this is almost preserved also when adding (a few) arrows.

Table 3 shows the estimated bias and variance contributions to the relative MSEs of the estimators. It can be noted that for the usual estimator $\hat{F}_a$ (that is unbiased) the contribution of estimated bias to MonteCarlo MSE is negligible. The contribution of bias is negligible along the diagonal too. This result suggests that estimators including a known dependence structure are approximately unbiased. Outside the diagonal the contribution of bias is usually higher.

Pop	$Bias_a$	Var_a	$Bias_b$	Var_b	$Bias_c$	Var_c	$Bias_d$	Var_d	$Bias_e$	Var_e	$Bias_f$	Var_f	$Bias_g$	Var_g	$Bias_h$	Var_h
a	2.0	98.0	74.0	26.0	25.9	74.1	57.8	42.2	70.4	29.6	69.3	30.7	48.1	51.9	85.1	14.9
b	1.9	98.1	3.0	97.0	1.9	98.1	57.8	42.2	59.1	40.9	88.7	11.3	85.3	14.7	94.9	5.1
c	1.6	98.4	71.5	28.5	2.6	97.4	82.9	17.1	85.5	14.5	79.2	20.8	69.4	30.6	91.2	8.8
d	0.4	99.6	97.4	2.6	41.2	58.8	1.6	98.4	68.5	31.5	61.4	38.6	31.1	68.9	97.7	2.3
e	3.8	96.2	62.5	37.5	1.9	98.1	5.2	94.8	7.1	92.9	17.8	82.2	12.8	87.2	90.7	9.3
f	3.9	96.1	49.9	50.1	1.7	98.3	28.3	71.7	30.5	69.5	4.6	95.4	1.8	98.2	83.3	16.7
g	3.1	96.9	56.9	43.1	2.1	97.9	12.1	87.9	44.9	55.1	44.2	55.8	1.8	98.2	51.0	49.0
h	4.1	95.9	6.2	93.8	1.7	98.3	2.5	97.5	3.9	96.1	4.0	96.0	1.7	98.3	6.3	93.7

Table 3. MonteCarlo estimates of bias and variance contribution to relative MSE

The same results hold also for the marginal estimated distributions.

4. INTEGRATION ISSUES

In a complex and integrated system of two or more surveys it is important to have tools to update estimates computed with a survey when results of other surveys or new archives become available. In this case we are in presence of an *informative shock* and we need to propagate the additional information to results of previous surveys in order to achieve consistency between surveys (external consistency). This problem can be faced in official statistics by means of calibration estimators allowing to estimate the parameters of interest under linear constraints. Here we propose the use of BNs since they are a natural tool for information propagation (Cowell et al. 1999).

BNs can be updated when an informative shock occurs. With new information here we mean a new frequency distribution for one or more variables of interest gained from an archive or a new survey. The relationship among the variables of a BN (*i.e.* the arrows connecting them) is the road for the propagation of this kind of information. For the sake of simplicity, consider a BN composed of just two nodes, X_1 and X_2, joined by the arrow $X_1 \rightarrow X_2$. Hence, the probability distributions $F(X_1 = x_1)$, $F(X_2 = x_2 \mid X_1 = x_1)$ can be associated to the BN. Let the marginal probability distribution for X_2 be changed into $F^*(X_2 = x_2)$. In order for the network to absorb the new distribution $F^*(X_2=x_2)$ leaving the relationship between the variables unchanged, *i.e.* the conditional distribution of X_2 given X_1, it is necessary to update the marginal distribution of X_1:

$$F^*(X_1 = x_1) = \sum_{x_2} F(X_1 = x_1 \mid X_2 = x_2) F^*(X_2 = x_2) = \sum_{x_2} F(X_1 = x_1, X_2 = x_2) \frac{F^*(X_2 = x_2)}{F(X_2 = x_2)} .$$

In other words, the old joint distribution $F(X_1 = x_1, X_2 = x_2)$ is updated via the ratio $F^*(X_2 = x_2)/F(X_2 = x_2)$ (called *update ratio*) between the new and old marginal distributions of X_2.

Here information propagation mechanism has been illustrated via a two-variables example. The results above can be generalized to the multivariate case when new information updates more than one variable distributions. To this purpose, different efficient algorithms based on the concept of *junction trees* (see Jensen, 1996) have been defined.
The propagation process just illustrated can be applied to perform the traditional poststratification procedure. Consider the ratio type estimator (2). For the sake of simplicity let the informative shock be relative to just variable

X_1 whose updated frequency distribution is N^*_{1q}, $q=1,\ldots,Q$. Propagating this shock through the BN corresponds to poststratify with respect to X_1. More precisely the old sample weights w_i are changed into:

$$w_i^* = w_i \frac{N^*_{1q}}{\sum_i w_i I_{x_{1i}}(q)} = w_i \frac{N^*_{1q}}{\hat{N}_{1q}}, \quad i:\ I_{x_{1i}}(q) = 1, \quad q = 1,\ldots, Q$$

where $\hat{N}_{1q}$ are the frequency estimates computed on the old survey weights. The shock produces a change in the node S, which modifies into a new node S^* in a way such that: (i) S^* categories are given by the Cartesian product of the S and X_1 categories, *i.e.* (h,q), $h=1,\ldots,H$, $q=1,\ldots,Q$; (ii) the units in the same category (h,q) have the same weight, $w^*_{(h,q)}$. Again, Bayes theorem allows the computation of the probability distribution of S given X_1:

$$F(S = h \mid X_1 = q) = \frac{F(S = h)F(X_1 = q \mid S = q)}{\sum_{h=1}^{H} F(S = h)F(X_1 = q \mid S = q)}, \quad q = 1, \ldots, Q;\ h = 1,\ldots, H.$$

Poststratification leaves unchanged the previous distribution, *i.e.* the statistical relationship between S and X_1 according to the initial survey design. Furthermore post-stratifying with respect to the new distribution of X_1, $N^*_1(q)$, $q=1,\ldots,Q$, or better to the relative frequency distribution: $F_1^*(q) = N^*_{1q} \big/ N, q = 1,\ldots,Q$, corresponds to consider the following new joint distribution:

$$F(S^* = (h,q)) = F(S = h, X_1 = q) = F(S = h \mid X_1 = q)F_1^*(q) = \frac{F(S = h)F(X_1 = q \mid S = h)}{\sum_{h=1}^{H} F(S = h)F(X_1 = q \mid S = h)} F_1^*(q) =$$

$$= \frac{n_h w_{(h)}}{\sum_{h=1}^{H} n_h w_{(h)}} \frac{n_{hq}}{n_h} \frac{F_1^*(q)}{\hat{F}_1(q)}, \quad q = 1,\ldots,Q; h = 1,\ldots,H.$$

S^* is characterized by constant weights, $w^*_{(h,q)}$, for all the units in the same category (h,q). *Let* n_{hq} be the size of this category. Hence

$$w^*_{(h,q)} = \frac{\sum_{h=1}^{H} n_h w_{(h)}}{n_{hq}} F(S^* = (h,q)) = w_{(h)} \frac{F_1^*(q)}{\hat{F}_1(q)} = w_{(h)} \frac{N_1^*(q)}{\hat{N}_1(q)} \frac{\hat{N}}{N}$$

with $\hat{N} = \sum_{i=1}^{N} w_i$.

The poststratification procedure described above can be directly applied when the informative shock involves the joint distribution of two or more linked variables. This procedure is straightforward for estimator (2). In the case of any other BN structure, the preservation of the shock on the joint distribution of a subset of variables is preserved only when that subset is a clique (this should be considered as a constraint for any poststratification).

In this section we have illustrated how to update estimates produced with a survey on the base of results from a new survey or an archive. This information propagation can be seen as part of a more general and extended system composed by more than two surveys and archives organized according to various criteria, among which their reliability and time sequence of performance. The use of BNs allows an efficient update of estimates due to the junction tree based propagation. In this sense it is necessary to organize the system of surveys so that new information available on a variable, say X_1, can reach, following the arrows, all the variables known to be statistically dependent on X_1. In technical terms the surveys must be connected in such a way that the so called junction property (Lauritzen, 1996) is fulfilled.

As a further development, we think that object-oriented Bayesian network (OOBN, see Koller and Pfeffer, 1997) can be a valid support to represent, manage and use a system of surveys. OOBN is a recent extension of BN technology. It allows hierarchical definition and construction of a BN, using simple modular building blocks. Additional complexity can easily be introduced by adding new modules or refining existing ones. In our case the OOBN would be composed by as many modules (instances) as the number of surveys. These modules would be connected by a system of output/input nodes - representing in our case identity relations between variables observed in different surveys - allowing updating, *i.e.* the output node would be the new estimate for say variable X to be propagated to previous surveys where X was observed as well. The information can then be extended to the other variables by means of the procedure illustrated above. It is interesting that with OOBNs macro coherence would be explicitly represented and the final user would have not to interpret a very complex structure. However the clear picture of any single survey is not lost since, whenever necessary, it is possible to focus and work on any of them. We think that this way is promising (OOBNs are implemented in the software Hugin version 6, http://www.hugin.com) although still to be explored so further research on it is needed.

REFERENCES

Ballin, M., Scanu, M., and Vicard, P. (2005), "Bayesian networks and complex survey sampling from finite populations". Proceedings of the 2005 FCSM Symposium, Arlington (Virginia), November 14-16, 2005.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999), *Probabilistic Networks and Expert Systems*, Heidelberg: Springer.

Di Zio, M., Sacco G., Scanu M., Vicard P. (2005), "Multivariate techniques for imputation based on Bayesian networks". *Neural Network World*, 2005/4, pp. 303-309.

Getoor L., Taskar B., Koller D. (2001), "Selectivity estimation using probabilistic models". *ACM-SIGMOD*, Santa Barbara, California, USA.

Jensen, F. V. (1996), *Introduction to Bayesian Networks*. Springer.

Koller, D., and Pfeffer, A. (1997), "Object-oriented Bayesian networks". *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, (ed. D. Geiger and P. Shenoy), Morgan Kaufmann Publishers, San Francisco. pp. 302-313.

Lauritzen, S. L. (1996), *Graphical Models*. Oxford University Press.

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

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