

Measurement Error Modelling using Object-Oriented Bayesian Networks

Daniela Marella and Paola Vicard

Abstract Variables are rarely, if ever, measured without error. In this paper we propose to use the Object-Oriented Bayesian Networks architecture to model measurement errors. A mixed measurement error model is introduced to model the respondent error. Then an Object-Oriented Bayesian network, implementing the model above, is developed to represent how the actually observed values are generated from the original ones. Furthermore, potentialities and possible extensions of such an approach are discussed.

Key words: Bayesian Networks, Measurement errors.

1 Introduction

With the term measurement errors we denote those errors that occur during the data collection stage. Formally, the measurement error can be characterized as the difference between the true and the observed value of the study variable for a given unit ([9]). Groves ([7]) distinguishes among four sources of measurement error: the questionnaire, the respondents, the interviewer and last but not least the data collection mode. Such errors can never be avoided in practice. The use of standard analyses in the presence of measurement errors usually gives misleading results. More specifically, measurement errors can lead to large bias effects on estimation and data analysis and will often reduce the precision of the estimates and the power of statistical tests. As a consequence, the evaluation of their effect on statistical analysis is an important goal of national statistical agencies and requires the introduction

Daniela Marella

Università Roma Tre, via del Castro Pretorio 20, 00185 Roma, Italy, e-mail: dmarella@uniroma3.it

Paola Vicard

Università Roma Tre, via Silvio D'Amico 77, 00145 Roma, Italy, e-mail: vicard@uniroma3.it

of a model describing their generation. Different way to model measurement errors and their effects on estimation and data analysis are discussed in [6], [2].

In this paper the measurement error process will be modelled using the Object-Oriented Bayesian Network (OOBN) architecture [8]. Bayesian networks (BN) have been already used to model and solve problems typical of official statistics. In particular they appeared to be very useful in missing item imputation ([11], [4] and [5]) and contingency table estimation ([1]).

OOBNs architecture represents an extension of BN which can be used over large and more complex domains described in terms of inter-related objects: the objects are instances of Bayesian networks. An Object-Oriented Network is a network that, in addition to the usual nodes, contains instance nodes. An instance node is a node representing an instance of another network. In other words, an instance node is given by a subnetwork.

OOBNs are particularly useful when they are used to model problems where specific graphical structures (subnetworks) are repeated identically in the overall model. For these reason OOBNs have been already successfully applied in DNA profiling problems typical of forensic statistics (for an example see [3]). Also in measurement error model specification there is repetition across individuals of model basic structure. Therefore OOBNs appear to be a natural and promising framework for combining the different measurement error sources (represented by local inter-dependent submodels) into a global analysis.

The insertion of the evidence (*e.g.* observed value) and the propagation of its effect through the network allows to estimate for each unit the probability distribution of the true value given the observed one. An appropriate synthesis of such a distribution (*e.g.* the mode or a random draw) can be used to predict the individual true value.

The paper is organized as follows. In Section 2 various measurement error models are described stressing the role of interclass transition probabilities. Such models describe the error generating mechanism when the respondent provides consciously or unconsciously the incorrect answer. In Section 3 an Object-Oriented Bayesian Network formalizing the measurement process is introduced and explained. Finally, in Section 4 we discuss possible extensions of such an approach.

2 Error Model

Suppose to consider a variable X characterized by K classes of response whose frequencies p_k , $k = 1, \dots, K$ are supposed to be known. When a measurement error takes place the observed class for a given unit is different from the true class. Formally, the measurement model expresses the relationship between the observed class and the true class, *i.e.* it describes the measurement errors generating mechanism. Let $q_{i \rightarrow j}$ be the interclass transition probability from class i to class j . Then, the transition probabilities matrix (TPM) can be defined as follows

$$\mathbf{TPM} = \begin{pmatrix} q_{1 \rightarrow 1} & q_{1 \rightarrow 2} & \cdots & q_{1 \rightarrow j} & \cdots & q_{1 \rightarrow k} \\ q_{2 \rightarrow 1} & q_{2 \rightarrow 2} & \cdots & q_{2 \rightarrow j} & \cdots & q_{2 \rightarrow k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ q_{k \rightarrow 1} & q_{k \rightarrow 2} & \cdots & q_{k \rightarrow j} & \cdots & q_{k \rightarrow k} \end{pmatrix}$$

where the following constraint $\sum_j q_{i \rightarrow j} = 1$ holds.

In order to estimate the $K(K-1)$ model probabilities, we could refer to previous surveys or carry out a pilot survey. An alternative way to proceed when K is large is to express the transition probabilities $q_{i \rightarrow j}$ by means of models characterized by a smaller number of parameters to be estimated. Formally, the transition probabilities can be parameterized as follows

$$q_{i \rightarrow j} = \lambda s_{i \rightarrow j} \quad (1)$$

where λ is the unknown measurement parameter to be estimated,[12]. Different measurement models can be identified by specifying the quantities $s_{i \rightarrow j}$. For instance, in the *scalar mutation model* the nonnegative quantities $s_{i \rightarrow j}$, with $i \neq j$, are supposed to be known. In the *proportional measurement model*

$$s_{i \rightarrow j}^{prop} = p_j, i \neq j \quad (2)$$

the larger is the frequency p_j of the response class j , the more likely is that class i mutates into class j . This model can be useful to represent situations when the respondent wants to hide himself/herself giving an answer consistent with the population average behaviour. This model is very simple but unrealistic by itself.

In real cases most measurement errors are towards a neighboring class. As a consequence an alternative model is the *one-two step measurement model*, where

$$s_{i \rightarrow j}^{MM2} = \begin{cases} 1/2 & \text{if } |i-j| = 1, i \neq 1, 2, \text{ or } i \neq K-1, K \\ 1/2 & \text{if } |i-j| = 2, i \neq 1, 2, \text{ or } i \neq K-1, K \\ 1 & \text{if } |i-j| = 1, i = 1 \text{ or } i = K \\ 1 & \text{if } |i-j| = 2, i = 1 \text{ or } i = K \\ 1/2 & \text{if } |i-j| = 1, i = 2 \text{ or } i = K-1 \\ 1 & \text{if } |i-j| = 2, i = 2 \text{ or } i = K-1 \\ 0 & \text{(otherwise)} \end{cases} \quad (3)$$

In order to get a more realistic and plausible representation of the measurement error generating process, we consider a mixed measurement model given by a mixture of the *one- two step model* (3) and the *proportional model* (2). In this way it

possible to model both the willingness of respondents to provide the average answer and the natural tendency to provide a plausible answer, *i.e.* not too far from the true class. Formally we have

$$s_{i \rightarrow j}^{mix} = h s_{i \rightarrow j}^{SMM2} + (1 - h) s_{i \rightarrow j}^{prop} \quad (4)$$

where the mixture parameter h , reflecting the relative importance of each model, is supposed to be known. This model is completely general and can be particularized to various contexts by an appropriate choice of h , whose value will depend on the nature of the variable at study.

The model (4) will be considered in the following analysis in Section 3. Clearly, the measurement error model cannot be transferred from one survey to another without any adjustment. In fact, the measurement model can be differently specified for different analyses (in this sense it could be sufficient to set an appropriate h given the variable of interest). For instance, the measurement error for an income survey will differ from the measurement model for other survey topics. The selection of a model needs careful analysis of survey social context, study variable nature as well as characteristics of the target population. The researcher experience is important in understanding data and in formalizing appropriate models.

3 Analysis with Object-Oriented Bayesian Network

In this paper we focus attention on one source of measurement error: the respondent error. Sudman and Bradburn ([10]) state that the motivation of the respondent plays an important role in the quality of the data provided. The respondent may consciously or unconsciously provide incorrect answers. Deliberate reporting of false values often occurs with personal questions such as income, drug use or abortion. Unconscious response error are those due to telescoping and memory error. As previously stressed, the OOBN architecture is particularly useful for the measurement model specification. The Bayesian network software system Hugin supports OOBNs. We use bold face to indicate a network class, and teletype face to indicate an instance or regular node. The main network (top-level) describing the problem is represented in Fig. 1.

This graph represents the structure of a classical measurement model whereby the observed measurement (OC-observed class) is the result of the true value (TC-true class) and an error term. Note that, our measurement model assumes independence between measurement errors: the error in an observation do not affect other observations. Such a hypothesis is more realistic for mail survey or for data collection modes in which there is not a common interviewer. In fact, several studies have shown that responses obtained by the same interviewer from two different respondents tend to be correlated because of factors such as interviewer experience and interviewer's personal characteristics (age, race, sex, etc). For instance, the interviewer may affect

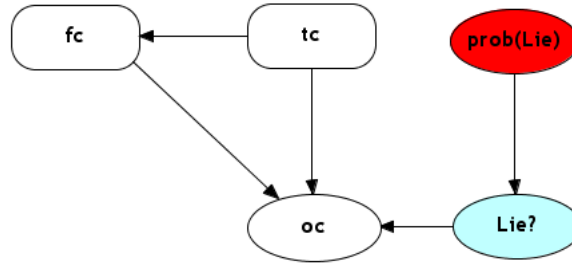


Fig. 1 Top-level network for the measurement error model for the respondent.

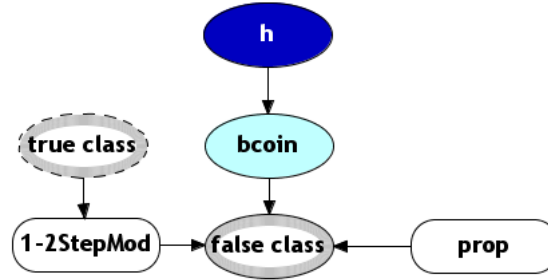


Fig. 2 Subnetwork for the false class.

the accuracy of the response by misreading questions, influencing the respondent to provide an incorrect response or recording the response incorrectly.

Clearly, the observed class coincides with the true class if no measurement errors occurs, otherwise the true value changes according to the mixed measurement model encoded in the subnetwork of class **model** in Fig.2. Going into details, in the respondent measurement model in Fig. 1 the true class is represented by an instance of class **tc** associated with the probability distribution of the variable of interest; the observed class is given by a standard random node **oc** that is defined by means of the following logical expression: `if (Lie? == 1, fc, oc)`. This means that if the respondent is a *liar* then the observed class is different from the original one and it is modelled by means of the mixed measurement model (4) implemented in the instance **fc** of class **model** network. If the respondent is *honest* then the observed class coincides with the original one. In Fig. 1 the fact that the respondent may be a *liar* or not is represented by means of node **Lie?** associated with a Bernoulli distribution of parameter **prob(Lie)**. We have represented the Bernoulli parameter (**prob(Lie)**) with a separated node in order to have the possibility to study the sensitivity of our model to the probability of getting a false answer.

The network in Fig. 2 of class **model** encodes the mixed model (4), *i.e.* it describes and models as the true value changes giving rise to the false one according to (4). The node **false class** is associated with the following logical expres-

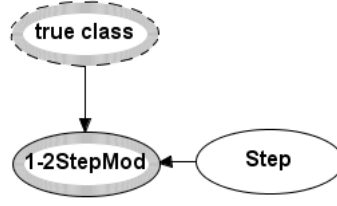


Fig. 3 Subnetwork for the one-two step model.

sion `if (bcoin == 1, 1-2StepMod, prop)`, that is the essence of (4). In particular the false class is chosen between a class (`prop` node) generated by the proportional model (2) and a class (`1-2StepMod` node) generated by the one-two step model (3) according to a biased coin toss. The biased coin toss is represented by node `bcoin` that is associated with a Bernoulli variable of parameter h (node `h`). Note that the choice to model the mixing mechanism by means of the Bernoulli distribution has been done in order not only to perform future sensitivity analysis with respect to h , but also to have the possibility to specify different (and reasonable) values of h for different variables.

The modularity property of BNs and the hierarchical structure of OOBNs allow to model the one-two step measurement model (3) in a separate subnetwork (Fig. 3) in order not to complicate readability of **model** class. Node `1-2StepMod` in Fig. 2 is an instance of class **one-two-step** represented in Fig. 3. Here the node `Step` is associated with a probability distribution on the possible steps, *i.e.* $-2, -1, +1, +2$. The probabilities of the different jumps can be appropriately specified for the given variable of interest.

4 Discussion and open problems

The variable of interest is generally assumed to be measured without error. Since, in most situations such an assumption is unrealistic, it is important to specify an appropriate statistical model in order to quantify the impact that measurement errors can have on statistical analysis. This paper proposed the Object-Oriented Bayesian Network architecture to model the measurement error due to the respondent.

In order to investigate the characteristics of the proposed model, a simulation experiment will be performed. The study will be based on data sets of true and observed values generated according to the Object-Oriented network described in Fig. 1. Given the model parameters `prob(Lie)` (the probability of getting a false answer) and h (mixture parameter), the insertion of the evidence (*e.g.* observed value) and the propagation of its effect through the network would allow to estimate the individual probability distribution of the true value given the observed value. As a consequence, for each unit the true value could be predicted by a random draw

from such a distribution. Furthermore, in order to investigate the sensitivity of measurement model to the probability of getting a false answer the simulation could be repeated for different values of `prob(Lie)`.

As a matter of fact, the measurement error is the outcome of an interactive process involving various error sources (the questionnaire, the respondents, the interviewer, the data collection mode) then further developments could include the construction of a more complex Object-Oriented Network. Under this circumstance, the OOBN architecture could be particularly useful since it would allow to define submodels (one for each measurement error source) as building blocks in order to represent the entire problem domain. Clearly, such submodels are inter-related instances of Bayesian networks that, suitably combining the different error sources, allow to easily visualize, understand and solve the overall measurement error problem.

Acknowledgements We thank Flaminia Musella for useful comments. The research was supported by MIUR/PRIN 2007.

References

1. Ballin, M., Scanu, M., Vicard, P.: Estimation of contingency tables in complex survey sampling using probabilistic expert systems. *J.Statist.Plann.Inference* **140**, 1501–1512 (2010)
2. Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S.,(1991): *Measurement Errors in Surveys*. New York: Wiley.
3. Dawid, A.P., Mortera, J., Vicard P.: Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International*, **169**, 195–205 (2007).
4. Di Zio, M., Scanu, M., Coppola, L., Luzi, O., Ponti, A.: Bayesian Networks for Imputation. *J. Roy. Statist. Soc./A* **167**,2, 309–322 (2004)
5. Di Zio, M., Sacco, G., Scanu, M., Vicard, P.: Multivariate techniques for imputation based on Bayesian networks. *Neural Network World* **4**, 303–309 (2005)
6. Fuller, W.A.,(1987): *Measurement Error Models*. New York:Wiley.
7. Groves, R.M.,(1987): Research on Survey Data Quality. *Public Opinion Quarterly*, **51**, S156-S172.
8. Koller D., Pfeffer A.,(1997). *Object-Oriented Bayesian Networks*. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence*, 302-313.
9. Lessler J.T.,Kalsbeek W.D.,(1992): *Nonsampling error in surveys*. New York:Wiley.
10. Sudman S.,Bradburn N.M.,(1974): *Response effects in surveys: A review and synthesis* . Chicago:Aldine.
11. Thibaudeau, Y., Winkler, W.E. : Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints. In: Research Report RRS2002/9 - 2002. U.S. Bureau of the Census (2002) Available via DIALOG. www.census.gov/srd/papers/pdf/rrs2002-09.pdf.
12. Vicard, P., Dawid, A.P., Mortera, J., Lauritzen, S. L.: Estimation of mutation rates from paternity casework. *Forensic Science International Genetics*, **2**, 1, 9–18 (2008).