ROMA TRE
UNIVERSITÀ DEGLI STUDI

COLLANA DEL
DIPARTIMENTO DI ECONOMIA

OBJECT-ORIENTED BAYESIAN NETWORKS FOR MODELLING THE RESPONDENT MEASUREMENT ERROR

Daniela Marella - Paola Vicard

ISSN 2279-6916 Working papers
(Dipartimento di Economia Università degli studi Roma Tre) (online)

Working Paper n° 167, 2012

I Working Papers del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all'approvazione del Comitato Scientifico.

Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall'art. l del D.L.L. 31.8.1945, n. 660 e successive modifiche.

Copie della presente pubblicazione possono essere richieste alla Redazione.

esemplare fuori commercio
ai sensi della legge 14 aprile 2004 n.106

REDAZIONE:
Dipartimento di Economia
Università degli Studi Roma Tre
Via Silvio D'Amico, 77 - 00145 Roma
Tel. 0039-06-57335655 fax 0039-06-57335771
E-mail: dip_eco@uniroma3.it

ROMA TRE
UNIVERSITÀ DEGLI STUDI

DIPARTIMENTO DI ECONOMIA

OBJECT-ORIENTED BAYESIAN NETWORKS FOR MODELLING THE RESPONDENT MEASUREMENT ERROR

Daniela Marella - Paola Vicard

Comitato Scientifico:

Fabrizio De Filippis
Anna Giunta
Paolo Lazzara
Loretta Mastroeni
Silvia Terzi

Object-Oriented Bayesian Networks for Modelling the Respondent Measurement Error

Daniela Marella
Dipartimento di Scienze dell'Educazione
Università Roma Tre, Italy

Paola Vicard
Dipartimento di Economia
Università Roma Tre, Italy

Abstract

In this paper Object-Oriented Bayesian networks are proposed as a tool to model measurement errors in a categorical variable due to respondent. A mixed measurement error model is presented and an Object-Oriented Bayesian network implementing such a model is introduced. The insertion of evidence represented by the observed value and its propagation throughout the network yields for each unit the probability distribution of the true value given the observed. Two methods are used to predict the individual true value and their performance is evaluated via simulation.

JEL Classification: C110, C180, C800, C830.

Keywords: Bayesian networks, Measurement errors, Respondent Error.

1 Introduction

Variables are hardly ever measured without error. If the study variable suffers from measurement error, then the values observed during the data collection stage differ from the true values we are interested in. The use of standard analyses in the presence of measurement errors usually gives misleading results leading to erroneous conclusions of various severity. More specifically, measurement errors can lead to large bias effects on estimation and data analysis and will often reduce the estimates precision. The evaluation of these effects requires the introduction of a statistical model describing their generating mechanism. Different ways to model measurement errors and their effects on estimation and data analysis are discussed in Fuller (1987), Biemer et al. (1991).

In this paper the measurement error generating mechanism is described by a mixed measurement model that is implemented using the Object-Oriented Bayesian network (OOBN) architecture (Koller and Pfeffer, 1997). OOBNs represent an extension of Bayesian networks (BNs), (Cowell et al., 1999), which can be used over large and complex domains described in terms of inter-related objects. An Object-Oriented network is a network that, in addition to the standard random nodes (each representing a random variable), contains nodes that are instances of other networks. OOBNs allow hierarchical definition and construction of a BN by means of building blocks (network classes). Thanks to the property of *modularity*, complex problems can easily be represented by introducing new building blocks. OOBNs are also particularly useful to model problems where specific graphical structures (network classes) are repeated identically in the global model.

For these reasons OOBNs have already been successfully applied in forensic genetics (Dawid et al., 2007). They are a promising tool to formulate and represent models for complex phenomena whose complexity can be broken down into a series of networks for subproblems; see Jordan (2004), Best et al. (2010).

In addition to measurement errors, survey data are typically affected by selection bias due to sampling design and nonresponse. Such problems have already been modelled using BNs. In particular, BNs appear to be very useful in missing item imputation (Thibaudeau and Winkler, 2002; Di Zio et al., 2004; Di Zio et al., 2005) and contingency table estimation for complex survey sampling (Ballin et al., 2010). The OOBN architecture is a natural framework for combining different error sources into a global analysis. Fig. 1 shows the OOBN representation of a general survey process starting from the sampling design to the analysis of the survey variables. Round-shaped rectangles represent instances of network classes (instance nodes). Instance nodes are linked by arrows representing identity relationships between the output node (full line) and the input node (dashed line). In Fig. 1 only interface (input and output) nodes are represented. The standard random nodes can be visualized by expanding the specific network class containing these nodes. This means that the user can decide the level of precision and the specific part of the general model he/she needs to look at. In Fig. 1 let (X, Y, Z) be the random variables of interest and assume that (X, Y) are affected by measurement errors.

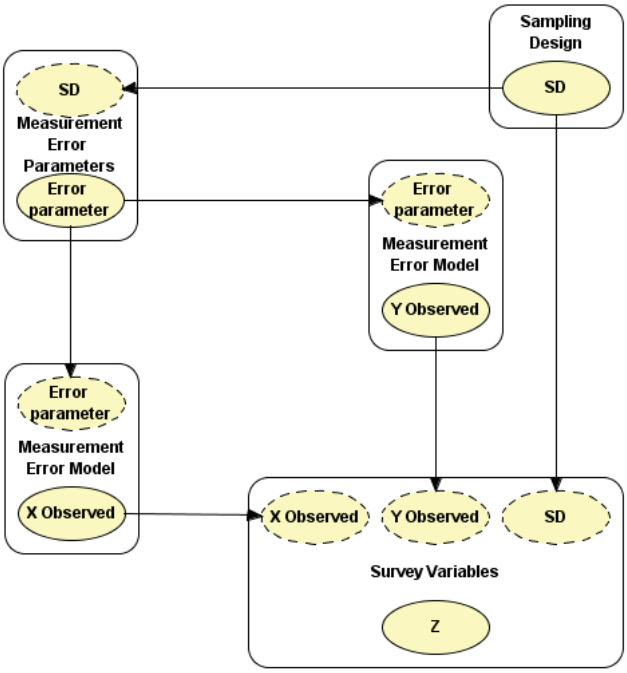

Figure 1: OOBN representation of a survey process

The observed values of X and Y are termed X *Observed* and Y *Observed*, respectively. The class *Sampling Design* takes into account the sampling design complexity by an appropriate use of sampling weights. The class *Measurement Error Parameters* imports information on the sampling design and exports the parameter (Error parameter) to the *Measurement Error Model* class network. This last class is repeated twice in order to model the measurement error of X and Y giving rise to X *Observed* and Y *Observed*, respectively. The *Survey Variables* network contains the variables of interest together with the sampling weights.

Groves (1987) distinguishes among four sources of measurement error: the questionnaire, the respondent, the interviewer and last but not least the data collection mode. Such errors cannot be avoided in practice. The measurement error is the outcome of an interactive process involving various error sources. The OOBN architecture could be particularly useful to represent the whole measurement error process by defining a submodel for each measurement error source. These submodels are inter-related instances of Bayesian networks that, suitably combined, allow to easily visualize, understand and solve the overall measurement error problem.

In this paper a mixed measurement model for the measurement error due to respondent is proposed and implemented in an Object-Oriented Bayesian network. Evidence (*i.e.* observed values) is inserted and propagated throughout the network to estimate, for each unit, the probability distribution of the true value given the observed. The individual true value can be predicted by an appropriate synthesis of such a distribution. Two prediction techniques are considered here: the mode and the random draw method.

The paper is organized as follows. In Section 2 various measurement error models are described stressing the role of intercategory transition probabilities. These models stem from the genetic mutation models (see Vicard and Dawid, 2004; Dawid et al., 2007; Vicard et al.,2008) and are explained in the measurement error context. In Section 2 a mixed measurement error model is proposed and in Section 3 an Object-Oriented Bayesian network formalizing the respondent measurement process is given. Estimates of measurement model parameters are proposed in Section 4. In Section 5 a simulation experiment is implemented so as to evaluate the performance of the two alternative prediction techniques: the mode and the random draw method.

2 The Measurement Model

Sudman and Bradburn (1974) wrote that the motivation of the respondent plays an important role in the quality of the data provided. The respondent may either consciously or unconsciously provide incorrect answers. Deliberate reporting of false values often occurs with personal questions about income, drug use or abortion. Unconscious response errors are those due to telescoping and memory errors. In this paper, we focus on classification error in an ordered categorical variable due to respondent providing a false answer.

Let X be an ordered categorical variable with K response categories whose frequencies p_k, $k = 1, \ldots, K$, are assumed known from administrative archives and/or census data. When a measurement error takes place the observed category for a given unit is different from the true category. The measurement error model describes the relationship between the observed category and the true category, *i.e.* how the true category mutates into the observed category. Let $q_{i->j}$ be the intercategory transition probability from the true category i to the observed category j, where $\sum_{j=1}^{K} q_{i->j} = 1$.

In order to estimate the $K(K-1)$ probabilities $q_{i->j}$, we could carry out an interview-reinterview study. Formally, a subsample of respondents are revisited after the original survey and a second measurement is obtained. Then the transition probability $q_{i->j}$ is estimated by the proportion of units in the subsample that mutate from the true category i to the observed category j. This method relies on knowledge of gold standard measurements (error-free measurements). We assume that for each unit the classification obtained in the reinterview is the true classification. Gold standard measurements are very difficult to obtain in practice, since the reconciled interviews data, as the original measurements, can be erroneous; see for instance Biemer and Forsman (1992). Attempts have been made to obtain them from: reconciled reinterviews surveys (see Forsman and Schreiner, 1991), in-depth probing reinterviews, record check studies (Biemer, 1988). The interview-reinterview method requires an explicit commitment of the survey programme, because it is expensive and time-consuming. Furthermore, when K is large the size of the subsample must be large enough to guarantee an accurate estimation with a substantial increase of survey costs.

An alternative way to proceed is to express the transition probabilities $q_{i->j}$ by means

of models characterized by a smaller number of parameters to be estimated. Here we use the scalar mutation model (Vicard and Dawid, 2004) in the measurement error context and call it *scalar measurement model*. This gives a simple but flexible description of the measurement error process, and it is given by

$$q_{i->j} = \lambda s_{i->j} \tag{2.1}$$

where $s_{i->j}$, $j \neq i$, and λ are the unknown measurement parameters to be estimated. The parameter λ will be called *Error parameter*. Since $\lambda \sum_{j:j\neq i} s_{i->j}$ is the probability of a measurement error when i is the true category, the overall measurement error rate μ can be expressed in terms of λ as follows

$$\begin{aligned} \mu &= \sum_i p_i \lambda \sum_{j:j\neq i} s_{i->j} \\ &= \beta\lambda \end{aligned} \tag{2.2}$$

where

$$\beta = \sum_i \sum_{j\neq i} p_i s_{i->j}. \tag{2.3}$$

Different measurement models can be identified by specifying the nonnegative quantities $s_{i->j}$. In the *proportional measurement model*

$$s^{prop}_{i->j} = p_j, i \neq j \tag{2.4}$$

the larger the frequency p_j of category j, the more likely category i mutates into category j. The model (2.4) reflects the assumption that, whenever a measurement error occurs, the observed value is generated at random from the population frequency distribution. Note that for this model $\beta^{prop} = 1 - \sum_i p_i^2$. The proportional model is a stationary model, which implies that if the true category has frequency p_i and the observed category has frequency p_i^*, then $p_i = p_i^*$. Thus the measurement error model is stationary if the population frequency distribution is unaffected by measurement error, *i.e* the population frequency distribution is constant over time. The proportional measurement model (2.4) can be useful to represent situations where respondents do not provide the true response and give answers consistent with the population average behaviour. Furthermore, the model (2.4) can also be used to model the measurement error when the categorical variable is not ordered.

The model (2.4) is very simple but unrealistic. In real cases most measurement errors occur in one of the L nearest neighbour categories. An alternative model for $L = 2$, say, is the *one-two step measurement model* with

$$s_{i->j}^{MM2} = \begin{cases} \alpha_1^- & \text{if} \quad (i-j) = 1,\ i \neq 1 \\ \alpha_1^+ & \text{if} \quad (i-j) = -1,\ i \neq K \\ \alpha_2^- & \text{if} \quad (i-j) = 2,\ i \neq 1, 2 \\ \alpha_2^+ & \text{if} \quad (i-j) = -2, \quad i \neq K-1, K \\ \alpha_1 & \text{if} \quad |i-j| = 1,\ i = 1 \quad \text{or} \quad i = K \\ \alpha_2 & \text{if} \quad |i-j| = 2,\ i = 1, 2 \quad \text{or} \quad i = K-1, K \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where $\alpha_1 = \alpha_1^- + \alpha_1^+$, $\alpha_2 = \alpha_2^- + \alpha_2^+$ and $\alpha_1 + \alpha_2 = 1$. The model (2.5) implies that the observed category can only be a neighbouring category or a category two steps away. For each category i a probability distribution is specified over the neighbouring categories. For instance, when $\alpha_1^- = \alpha_1^+ = \alpha_2^- = \alpha_2^+ = 1/4$ a uniform probability distribution is associated with the possible steps $(-2, -1, +1, +2)$. It is straightforward to show that for the one-two step measurement model $\beta^{MM2} = 1$. This is a non stationary model.

A more realistic and plausible representation of the measurement error generating process is the *mixed measurement model*, given by a mixture of the one-two step model (2.5) and the proportional model (2.4). In this way both the willingness of a respondent to provide the average answer and the natural tendency to provide a plausible answer, *i.e.* not too far from the true category, are modelled. Formally

$$s_{i->j}^{mix} = (1-h)s_{i->j}^{prop} + hs_{i->j}^{MM2} \tag{2.6}$$

where the mixture parameter h (for h between 0 and 1) reflects the relative importance of each model component. For instance, a value $h = 0.75$ emphasizes one-two step errors, retaining a nonnegligible probability for measurement errors towards further categories.

It is straightforward to show that for the mixed measurement model (2.6)

$$\beta^{mix} = (1-h)\left(1 - \sum_i p_i^2\right) + h. \tag{2.7}$$

The model (2.6) is completely general, and can be applied to various contexts thanks to the measurement model parameters $(\mu, \alpha_2^-, \alpha_1^-, \alpha_1^+, \alpha_2^+, h)$. Their values depend on the nature of the study variable, as well as on the features of the target population. For instance, information on income is generally affected by measurement errors. The main

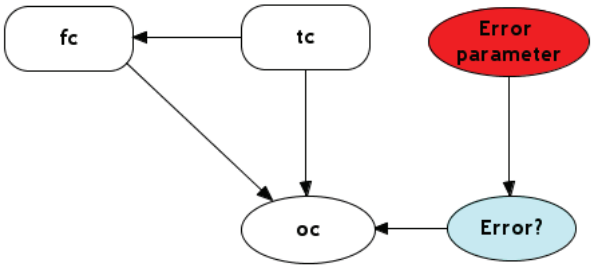

Figure 2: Top-level network for the measurement error model for the respondent.

reason is that questions about income are considered to be rather sensitive. If the target population is composed by high-income earners, the respondents will probably tend to under-report their income; the distribution probability over the neighbouring categories $(\alpha_2^-, \alpha_1^-, \alpha_1^+, \alpha_2^+)$ will be skewed to the right.

The mixed model (2.6) is non stationary, so that after any survey the frequency distribution of the categories is slightly modified. Our approach is based on assuming a model for the measurement error, with a limited number of parameters to be estimated from the data. This model represents a compromise between mathematical tractability, statistical feasibility and realism. Alternative models can be specified. For instance, it could be reasonable to assume a model where the probability of measurement error from category i to category j decreases as a function of the difference in length between the categories i and j, see Dawid et al. (2002). This model is stationary, and it requires the estimation of all measurement error rates $q_{i->j}$; a database containing all possible measurement errors must be available (Egeland and Mostad, 2002). However, a database may not contain all possible measurement errors and/or the estimation procedure could be extremely expensive. Alternatively, the non stationary decreasing model $q_{i->j} = \gamma_i r^{|i-j|}$ (Egeland and Mostad, 2002) where $0 < r < 1$ is a constant and γ_i is chosen so that $\sum_{j=1}^{K} q_{i->j} = 1$, can be considered. This model requires the estimation of $K + 1$ parameters, so that it could represent a valid alternative when K is small. Nevertheless, the model (2.6) is particularly convenient to be represented by BNs.

3 Object-Oriented Bayesian Network for the measurement model

We now build the Object-Oriented Bayesian network for the scalar measurement model (2.6). Our networks are implemented by using the program HUGIN version 7 (`www.hugin.com`). We use bold face to indicate a network class, and teletype face to indicate an instance or regular node. Fig. 2 shows the main network (top-level) representing the overall problem.

In Fig. 2 the true category is represented by an instance of class **tc** associated with the probability distribution of the variable of interest; the observed category is given by a standard random node `oc`, that is defined by the logical expression: `if (Error? ==`

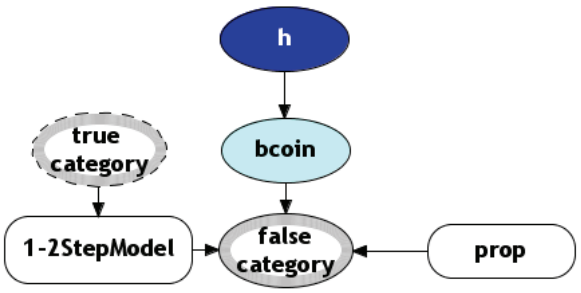

Figure 3: Subnetwork for the false category.

`1, fc, oc)`. This means that if the respondent is a *lier*, coded as `Error?`=1, then the observed category is different from the original one. It is then modelled by means of the mixed measurement model (2.6) implemented in the instance `fc` of class **model** network in Fig. 3. If the respondent is *honest*, coded as `Error?`=0, then the observed category coincides with the original one. In Fig. 2 the fact that a respondent may be a lier or not, is represented by means of node `Error?` associated with a Bernoulli distribution of parameter λ, where from (2.2) $\lambda = \mu/\beta$, *i.e.* it is a linear transformation of the overall measurement error rate μ.

The network in Fig. 3 of class **model** encodes the mixed measurement model (2.6), *i.e.* it describes and models how the true value changes giving rise to the false one according to (2.6). The node `false category` is determined by the logical expression `if (bcoin == 1, 1-2StepModel, prop)`, that is the essence of (2.6). Thus, the false category is chosen as either a category (`prop` node) generated by the proportional model (2.4) or a category (`1-2StepModel` node) generated by the one-two step model (2.5) 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`), *i.e* the mixture parameter into (2.6).

The mixture mechanism is modelled as a Bernoulli distribution in order to specify different values for h for different variables.

The modularity property of BNs and the hierarchical structure of OOBNs allow one to model the one-two step measurement model (2.5) in a separate subnetwork (Fig. 4). This helps readability of **model** class. Node `1-2StepModel` in Fig. 3 is an instance of class **one-two-step** represented in Fig. 4. The node `Step` is associated with a probability distribution on the possible steps, $-2, -1, +1, +2$. The probabilities of the different jumps can be appropriately specified for the given variable of interest.

4 Measurement Model Parameters Estimation

The network in Fig. 2 can be used to estimate for each unit the probability distribution of the true value given the observed value. In order to accomplish this, the measurement model parameters $(\mu, \alpha_2^-, \alpha_1^-, \alpha_1^+, \alpha_2^+, h)$ need to be estimated from data. Let $i \rightarrow j$ be

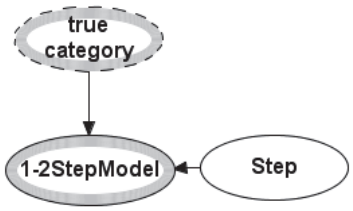

Figure 4: Subnetwork for the one-two step model.

the event *the true category i mutates into the observed category j*; it can easily be shown that

$$\alpha_1^- = \sum_{i=2}^{K} q_{i->(i-1)} \tag{4.1}$$

$$\alpha_1^+ = \sum_{i=1}^{K-1} q_{i->(i+1)} \tag{4.2}$$

$$\alpha_2^- = \sum_{i=3}^{K} q_{i->(i-2)} \tag{4.3}$$

$$\alpha_2^+ = \sum_{i=1}^{K-2} q_{i->(i+2)} \tag{4.4}$$

where $\alpha_1^-(\alpha_1^+)$ represents the minus one disagreement rate (plus one disagreement rate) while $\alpha_2^-(\alpha_2^+)$ represents the minus two disagreement rate (plus two disagreement rate). The parameters (4.1)-(4.4) can be estimated through an interview-reinterview study as follows. Suppose that a subsample of m respondents are revisited after the original survey and asked the same question. The basic assumption assuring unbiased estimation is that the survey conditions are identical in the two occasions. The model (2.5) specifies the same probability distribution for each category i, hence the estimation process does not rely on the knowledge of the true values for the m subsample units. As a result, we have two measurements on the same variable X for each subsample unit and the interview-reinterview data can be summarized as shown in Table 1

Table 1: *Interview-Reinterview table*

Interview A	*Reinterview B*
0	m_0
-1	m_1^-
$+1$	m_1^+
-2	m_2^-
$+2$	m_2^+
$\|step\| > 2$	m_s
	m

where

1. m_0 denotes the number of subsample units classified in the same category on both occasions;

2. m_1^- is the number of subsample units classified in the category i in the interview and in the category $i-1$ in the reinterview, with $i = 2\ldots, K$;

3. m_1^+ is the number of subsample units classified in the category i in the interview and in the category $i+1$ in the reinterview, with $i = 1\ldots, K-1$;

4. m_2^- is the number of subsample units classified in the category i in the interview and in the category $i-2$ in the reinterview, with $i = 3\ldots, K$;

5. m_2^+ is the number of subsample units classified in the category i in the interview and in the category $i+2$ in the reinterview, with $i = 1\ldots, K-2$;

6. m_s is the number of subsample units classified in the category i in the interview and in the category j in the reinterview with $|i-j| > 2$.

Let $m_1^- + m_1^+ + m_2^- + m_2^+ = m - m_0 - m_s$ be the number of subsample units inconsistently classified in the two interviews with $|i-j| \leqslant 2$. The measurement model parameters (4.1)-(4.4) can be estimated as follows

$$\widehat{\alpha}_1^- = \frac{m_1^-}{m_1^- + m_1^+ + m_2^- + m_2^+} \tag{4.5}$$

$$\widehat{\alpha}_1^+ = \frac{m_1^+}{m_1^- + m_1^+ + m_2^- + m_2^+} \tag{4.6}$$

$$\widehat{\alpha}_2^- = \frac{m_2^-}{m_1^- + m_1^+ + m_2^- + m_2^+} \tag{4.7}$$

$$\widehat{\alpha}_2^+ = \frac{m_2^+}{m_1^- + m_1^+ + m_2^- + m_2^+} \tag{4.8}$$

with $\widehat{\alpha}_1^- + \widehat{\alpha}_1^+ + \widehat{\alpha}_2^- + \widehat{\alpha}_2^+ = 1$. From Table 1, the estimated overall measurement error rate μ is

$$\widehat{\mu} = \frac{m_1^- + m_1^+ + m_2^- + m_2^+ + m_s}{m} \tag{4.9}$$

and, by (2.2), the estimate of Error parameter λ is

$$\widehat{\lambda} = \frac{\widehat{\mu}}{\beta}. \tag{4.10}$$

Analogously the estimate of the mixture parameter h is

$$\hat{h} = \frac{m_1^- + m_1^+ + m_2^- + m_2^+}{m - m_0}. \tag{4.11}$$

As previously stated, the basic assumption in the estimation process is that the two replicate measurements must be parallel. As emphasized in Biemer (2009), this assumption is seldom satisfied in practice since the general conditions that existed during the interview are likely to have changed by the time of reinterview. In addition, respondents may have been conditioned by the first interview and their reinterview responses may reflect this conditioning.

As a consequence, the measurement model parameters estimators in (4.5)-(4.8),(4.9) and (4.11) could be affected by bias due to hidden measurement errors since the probability of observing an incompatibility differs from the probability that a measurement error takes place. For instance, in the reinterview the respondents may recall their interview responses and they could repeat them so that

$$E(\widehat{\mu}) < \mu$$

and the estimator (4.9) is a biased downwards as an estimator of μ.

5 Simulation Study

Given the OOBN in Fig. 1, the true value is predicted for each unit using two alternative techniques: the first one is based on the mode, and the other on a random draw from the distribution of the true value conditionally on the observed one. Their performance is evaluated through a simulation experiment.

The simulation study is conducted by using RHugin (Konis, 2009), a R software package. Let X be an ordered categorical variable with $K = 5$ categories and let the node `Step` in Fig.4 be associated with a uniform probability distribution on the possible steps $(-2, -1, +1, +2)$. Then $\alpha_1^- = \alpha_1^+ = \alpha_2^- = \alpha_2^+ = 1/4$. This situation may represent unconscious measurement error, *e.g.* the respondent gives the wrong answer not deliberately, due for example to a memory or telescoping error. In this case the observed category may be one of the 2 nearest neighbour categories with uniform probability. The simulation analysis involves the following steps.

1. Assign a value to the measurement model parameters (μ, h).

2. Generate a dataset S of size $n = 5000$ of true and observed values of X according to the OOBN in Fig. 2.

3. For each unit $i \in S$ such that the true value differs from the observed value estimate the individual probability distribution of the true value given the observed value. Furthermore, predict the unknown true value by (a) the mode and (b) a random draw from the distribution.

4. Repeat steps 2 to 4 for different values of the measurement model parameters (μ, h).

Evaluation criteria have been introduced to investigate the performance of the imputation techniques. Let S^* be the subset of S of size n^* ($n^* \leqslant n$) composed of units affected by measurement errors. Let further f_j be the true relative frequency of category j of X of the units in S^* and f_j^* be the corresponding frequency after imputation. A first evaluation criterion to analyse the univariate distribution preservation is based on the following indicator

$$\Delta = \frac{1}{2}\sum_{j=1}^{K} |f_j - f_j^*| \tag{5.1}$$

where the sum is over the categories of X. Δ takes values between zero and one. The preservation of individual data is evaluated by

$$\varphi = \frac{1}{n^*}\sum_{i \in S^*} |x_i - x_i^*| \tag{5.2}$$

and

$$\xi = \frac{1}{n^*}\sum_{i \in S^*} I_{x_i}(x_i^*), \tag{5.3}$$

where x_i is the true value for unit i, x_i^* is the corresponding imputed value and I_{x_i} is the indicator function, equal to 1 if $x_i^* = x_i$ and 0 otherwise. Note that the indicator (5.2) evaluates the average distance between the true and the imputed values in S^* while (5.3) provides the percentage of correct imputations.

5.1 Simulation Results

We begin by analysing the performance of the proposed imputation methods as the shape of population distribution varies. We denote by $(\Delta^m, \varphi^m, \xi^m)$ and $(\Delta^r, \varphi^r, \xi^r)$ the indicators (5.1), (5.2), (5.3) for the mode and for the random prediction method, respectively. The values $h = 0.25, 0.75$ and $\mu = 0.2$ have been considered. Table 2 shows the results for $h = 0.25, \mu = 0.2$. The value $h = 0.25$ weights the proportional model more, but retains a nonnegligible probability for the one-two step measurement model. The value $\mu = 0.2$ implies that 20% of observed data is affected by error. Parameters β and λ have been obtained from (2.7) and (2.2), respectively.

Regardless of the population distribution shape, the mode method performs better than the random method in terms of distribution preservation Δ and of individual data reconstruction φ. The two methods are equivalent when the population distribution shape

Table 2: Δ, φ, ξ ***for mode and random draw method as the shape of population distribution varies*** $(\mu = 0.2, h = 0.25)$**.**

Population	β	λ	Δ^m	Δ^r	φ^m	φ^r	ξ^m	ξ^r
uniform	0.85	0.24	0.03	0.02	1.60	1.66	0.14	0.19
simmetric	0.84	0.24	0.05	0.12	1.44	1.58	0.15	0.20
bimodal	0.82	0.24	0.06	0.21	1.36	1.46	0.16	0.20
asymmetric	0.75	0.27	0.14	0.34	1.13	1.41	0.24	0.19

is uniform. Δ^m increases from 0.05 for symmetric distributions to 0.14 for asymmetric distributions but it is always smaller than Δ^r. Furthermore, the difference in terms of distribution preservation $(\Delta^r - \Delta^m)$ increases with the complexity of population distribution shape. Analogously, φ^m is always smaller than φ^r. With regard to the indicator ξ^r, for uniform, symmetric and bimodal distribution its value is larger than ξ^m. The closer ξ is to one, the better the performance of the imputation technique. Such a circumstance does not alter the previous conclusion. For instance, if the population is bimodal then $\varphi^m = 1.36 < \varphi^r = 1.46$ and $\xi^r = 0.20 > \xi^m = 0.16$. This means that, on one hand the random mode is able to correctly reconstruct 20% of data affected by measurement errors, and on the other hand the mode method provides an imputed value closer to the true value in 80% of the remaining cases. The same results are obtained with $h = 0.75$ and $\mu = 0.2$, as shown in Table 3.

The sensitivity of the model to the probability of getting a false answer is evaluated by analyzing the performance of the two imputation methods as the parameter μ changes. Here, we assume that μ is smaller than 0.5, *i.e.* at most 50% of observed data is affected by measurement error. Table 4 shows the results obtained for an asymmetric population distribution and $h = 0.25$ (where $\beta = 0.75$).

Table 3: Δ, φ, ξ ***for mode and random draw method as the shape of population distribution varies*** $(\mu = 0.2, h = 0.75)$**.**

Population	β	λ	Δ^m	Δ^r	φ^m	φ^r	ξ^m	ξ^r
uniform	0.95	0.21	0.09	0.02	1.48	1.59	0.05	0.21
symmetric	0.95	0.21	0.05	0.10	1.41	1.52	0.05	0.21
bimodal	0.94	0.21	0.16	0.20	1.36	1.53	0.05	0.19
asymmetric	0.92	0.22	0.09	0.34	1.01	1.37	0.28	0.21

As the value of parameter μ increases, the mode method performs better than the random method. Furthermore, when $\mu \geqslant 0.3$ the method imputes the true value in 35% of the records affected by measurement errors. On the other hand, the random technique is less sensitive to changes in the probability of getting a false answer than the mode method

Table 4: Δ, φ, ξ ***for mode and random draw method as*** μ ***changes*** $h = 0.25$***.***

λ	μ	Δ^m	Δ^r	φ^m	φ^r	ξ^m	ξ^r
0.13	0.1	0.15	0.34	1.11	1.38	0.25	0.20
0.27	0.2	0.14	0.34	1.13	1.41	0.24	0.19
0.40	0.3	0.04	0.34	0.97	1.38	0.35	0.22
0.54	0.4	0.04	0.33	0.97	1.39	0.35	0.20
0.67	0.5	0.04	0.36	0.99	1.44	0.34	0.18

both for Δ and for (φ, ξ). The same results hold for $h = 0.75$ as shown in Table 5 where $\beta = 0.92$.

Table 5: Δ, φ, ξ ***for mode and random draw method as*** μ ***changes with*** $h = 0.75$***.***

λ	μ	Δ^m	Δ^r	φ^m	φ^r	ξ^m	ξ^r
0.11	0.1	0.31	0.34	1.26	1.45	0.08	0.19
0.22	0.2	0.09	0.34	1.01	1.37	0.28	0.21
0.33	0.3	0.09	0.37	1.00	1.41	0.29	0.20
0.44	0.4	0.10	0.34	1.01	1.39	0.29	0.21
0.55	0.5	0.15	0.35	0.89	1.40	0.34	0.20

In order to investigate the sensitivity of the proposed measurement model to changes in the mixture parameter h, we proceed by analysing the performance of the two imputation methods as h varies between 0 and 1. Table 6 shows the results obtained for an asymmetric population distribution and $\mu = 0.2$.

Table 6: Δ, φ, ξ ***for mode and random draw method as*** h ***changes with*** $\mu = 0.2$

h	β	λ	Δ^m	Δ^r	φ^m	φ^r	ξ^m	ξ^r
0	0.66	0.3	0.03	0.34	1.06	1.34	0.33	0.22
0.25	0.75	0.27	0.14	0.34	1.13	1.41	0.24	0.19
0.5	0.83	0.25	0.23	0.32	1.21	1.35	0.16	0.20
0.75	0.92	0.22	0.09	0.34	1.01	1.37	0.28	0.21
1	1	0.2	0.12	0.33	1.00	1.43	0.26	0.19

The best performance of the mode method is obtained when $h = 0$, *i.e.* when the mixed measurement model (2.6) coincides with the proportional model. In this case, the estimate of the individual probability distribution of the true value given the observed takes advantage from information regarding the known population distribution. On the

other hand, the random technique is less sensitive to changes in the mixture parameter h than the mode method for both Δ and (φ, ξ).

6 Conclusions

In this paper the OOBN architecture has been proposed as a tool to represent by a single model the entire survey process. In fact, if on one hand the modularity of OOBN allows to easily represent the model building process, on the other hand the package $RHugin$ allows to reduce the compilation time in a big network. The latter is particularly useful both in measurement errors and nonresponse imputation (Thibaudeau and Winkler, 2002; Di Zio et al., 2004; Di Zio et al., 2005) where specific graphical structures are repeated identically in the overall model for each sample unit. Furthermore, the overall network can be partitioned into as many independent subnetworks as the error sources. Hence, each subnetwork can be planned and modified separately. Modularity of the OOBN allows easy extensions to similar but different situations.

In an income survey, for instance, the rich tend to under-report their income while the poor tend to over-report it (mean-reversion in measurement errors). A simple modification to the mixed measurement model (2.6) yields different measurement models in different sample groups. More specifically, the model (2.1) does not incorporate differences between units with respect to their response behaviour, that is each unit is assumed to have the same error probability. If auxiliary information is available, a more realistic model can be obtained by partitioning the original sample s into $G(s)$ groups s_g $(g = 1, \ldots, G(s))$ based on this information. The basic assumption is that all elements within the same group have the same error probability. One easy way to incorporate such an extension is to add a group indicator in the model (2.1). Correspondingly, the OOBN in Fig. 1 is modified by simply adding a node representing the group indicator in the modules *Measurement Model Parameters* and *Measurement Error Model* for X and Y, respectively. In this way we can account for the auxiliary information without modifying the general model structure. As a consequence, in a income survey it could be possible to assign a right skew probability distribution on the steps $(-2, -1, +1, +2)$ for high income earners and left skew probability distribution for low income earners without modifying the main structure of the model.

Acknowledgement

The research was supported by MIUR/PRIN 2007. We thank Julia Mortera for helpful suggestions and the anonymous referee for useful comments.

References

Ballin, M., Scanu, M., Vicard, P. (2010). Estimation of contingency tables in complex survey sampling using probabilistic expert systems. Journal of Statistical Planning and Inference 140:1501–1512 .

Best, N., Jackson, C., Richardson, S., (2010). Modelling complexity in health and social sciences: Bayesian graphical models as a tool for combining multiple sources of information. In: *Proceedings of the 3rd ASC International Conference on Survey Research Methods*. Association for Survey Computing . ISBN 0954674804

Biemer, P.P. (1988). Measuring data quality. In: Groves, R., Biemer, P.P., Lyberg, L., Massey, J., Nicholls,W., Waksberg, J. (Eds.), *Telephone Survey Methodology*. New York: Wiley.

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., Sudman, S.(1991). *Measurement Errors in Surveys*. New York: Wiley.

Biemer, P.P. (2009). Measurement errors in sample survey. In: Pfeffermann, D., Rao, C.R., *Handbook of Statistics, Vol. 29A, Sample Surveys:Design, Methods and Applications*. Amsterdam: North-Holland.

Biemer, P.P., Forsman, G. (1992). On the quality of reinterview data with applications to the current population survey. Journal of the American Statistical Association 87:915–923.

Cowell, R.G., Dawid, P., Lauritzen, S.L., Spiegelhalter, D.J. (1999).*Probabilistic networks and expert systems*. New-York: Springer.

Dawid, A.P., Mortera, J., Vicard, P. (2007). Object-oriented Bayesian networks for complex forensic DNA profiling problems. Forensic Science International 169:195–205.

Dawid, A.P., Mortera, J., Pascali, V., Van Boxel, D. (2002). Probabilistic expert systems from forensic inference from genetic markers. Scandinavian Journal of Statistics 29:577–595.

Di Zio, M., Scanu, M., Coppola, L., Luzi, O., Ponti, A. (2004). Bayesian Networks for Imputation. Journal of the Royal Statistical Society Series A 167:309–322.

Di Zio, M., Sacco, G., Scanu, M., Vicard, P. (2005). Multivariate techniques for imputation based on Bayesian networks. Neural Network World 4:303–309.

Egeland, T., Mostad, P.F. (2002). Statistical genetics and genetical statistics:a forensic perspective. Scandinavian Journal of Statistics 29:297–307.

Forsman, G., Schreiner, I. (1991). The design and analysis of reinterview: an overview. In: Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (Eds.), *Measurement Errors in Surveys*. New-York: Wiley.

Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley.

Groves, R.M.(1987). Research on Survey Data Quality. Public Opinion Quarterly 51: S156–S172.

Jordan, M.I. (2004). Graphical Models. Statistical Science 19: 140–155.

Koller, D., Pfeffer, A. (1997). Object-Oriented Bayesian Networks. In: *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence*.

Konis, K. (2009). The RHugin Package Homepage. URL http://rhugin.r-forge.r-project.org/.

Sudman, S., Bradburn, N.M.(1974).*Response effects in surveys:A review and synthesis*. Chicago: Aldine.

Thibaudeau, Y., Winkler, W.E. (2002). Bayesian networks representations, generalized imputation, and synthetic micro-data satisfying analytic constraints. In: *Research Report RRS2002/9 - 2002. U.S. Bureau of the Census*. Available via DIALOG. www.census.gov/srd/papers/pdf/rrs2002-09.pdf.

Vicard, P., Dawid, A.P. (2004). A statistical treatment of biases affecting the estimation of mutation rates. Mutation Research 547:19–33.

Vicard, P., Dawid, A.P., Mortera, J., Lauritzen, S. L. (2008). Estimation of mutation rates from paternity casework. Forensic Science International Genetics 2:9–18.