Statistical Journal of the IAOS 32 (2016) 537–544
DOI 10.3233/SJI-150951
IOS Press

Estimation from contaminated multi-source data based on latent class models

Ugo Guarnera* and Roberta Varriale
Italian National Statistical Institute ISTAT, via Cesare Balbo 16, Roma

Abstract. Recently, many statistical institutes have been moving from traditional estimation approaches based on sample survey data to new approaches that try to exploit the increased availability of administrative data, due to the need of reducing the response burden and providing users with more reliable statistical information. In this context, problems concerning the use of multiple sources for estimation purposes have been receiving an increasing attention in Official Statistics. A commonly adopted strategy is to rely on a "hierarchy" of the sources, based on preliminary analyses of the data quality of each source. In this work, we propose an alternative approach based on the concept of latent variables, where one takes advantage of the simultaneous availability of information from different sources. The true values of the target variable are viewed as realizations from a latent (unobserved) variable and the distinct (possibly coinciding) observed values from different sources are considered as imperfect measurements of this latent variable. According to this approach, all the available information is used and "weighted" according to its reliability, and a prediction of "true" values of some numeric variable of interest is obtained conditional on *all* the available information.

Keywords: Multi-source data, data integration, contamination models, latent variables

1. Introduction

In recent years, statistical analysis based on different data sources has become an active area of research in both theoretical and applied statistics. In particular, due to the increasing availability of administrative data, problems concerning the use of multiple sources for estimation purposes have been receiving an increasing attention in Official Statistics. Frequently, National Statistical Institutes (NSIs) try to combine data from available sources in order to build "statistical" archives to be used in different phases of the statistical production process. Massive use of "external" data is being considered by NSIs as an important alternative to the traditional approaches based on survey data. In fact, this approach allows NSIs to move resources previously allocated in conducting surveys to other activities, reducing at the same time the response burden on respondents. Moreover, statistical analysis based on large datasets may result in more accurate estimates than the ones that can be obtained through sample surveys. On the other hand, combining data to build a statistical information system is a complex task. In fact, administrative data are typically collected by different institutions for specific purposes (for instance, data on enterprises provided by the tax agency have "fiscal nature") and may not be usable in their original form for statistical purposes. Thus, a lot of "preprocessing" work has to be done in activities, such as harmonization of definitions, variable standardization, etc., aiming at providing users with data that satisfy their informative requirements. Another important issue is related to the possibility of partial (or total) overlapping among informative contents from different sources. This is of course of no concern when differences among values of corresponding items are negligible, but problems can arise when, due to "measurement errors", strong discrepancies are observed. In the latter case some decision strategy is necessary. A possible approach is to rely on a "hierarchy" based on preliminary analyses of the data quality of each source: in presence of discordant values, the source with the "highest" score according to the established hierarchy is chosen. Getting information from a single source for each statistical unit has the advantage of preserving coherence among different items. This approach has been used for instance at the Italian Institute of Statistics (Istat), to build a statistical information system for an-

*Corresponding author: Ugo Guarnera, Via Tuscolana 1788, 00173 Roma. Tel.: +39 0646736637; E-mail: guarnera@istat.it.

nual Structural Business Statistics (SBS) on small and medium enterprises [1]. The problem with the hierarchical approach is that it is not obvious how to define the hierarchy among sources. Moreover, in some situations, information from sources with low score in the hierarchy could be used when implausible values or missing values are observed in the highest quality source. These observations suggest an alternative approach where one takes advantage of the simultaneous availability of information from different sources. With this approach, all the available information is used and "weighted" according to its reliability. In this paper, a model for the prediction of "true" values of some numeric variable of interest conditional on *all* the available information is presented. The true values of the target variable are viewed as realizations from a latent (unobserved) variable and the distinct (possibly coinciding) observed values from different sources are considered as imperfect measurements of this latent variable. Given a model for the true data and a measurement error model for each available source through the specification of a conditional distribution of the data observed in the source given the true unobserved data, one can easily derive, via Bayes formula, the distribution of the true data given the observed data.

Other authors have used latent variable models to estimate the validity of administrative variables. For example, in [2] a structural equation model is used to assess and compare the quality of administrative sources for statistical use, and in [3] a simulation study is performed to test the robustness of this method to different amounts of measurement error, to misspecification of the measurement model, and to small sample size. In [4] the measurement error of a categorical target variable is determined by matching the information obtained by the longitudinal part of a survey with unique register data, taking into account that also register data are not error-free and that measurement error is likely to be correlated over time. In particular, the authors propose the estimation of the measurement error in the two sources using an extended hidden Markov model with two categorical observed indicators.

The approach proposed in the present paper deals with continuous data variables, and takes into account the measurement error in the different data sources. The proposed latent variable model can be used for different purposes. First, the error mechanism parameters can be used to assess the quality of the available sources. Second, individual predictions, obtained taking expectations of the true data distribution conditional on the observed data, can be used for editing activities. In this context, the proposed model can be considered as a natural extension of the contamination model used in [5] to identify influential errors in data from a single data source. Finally, predictions can be directly used to build a micro-data file for estimation purposes in particular circumstances, e.g. when the micro-data file has not to be disseminated to external users.

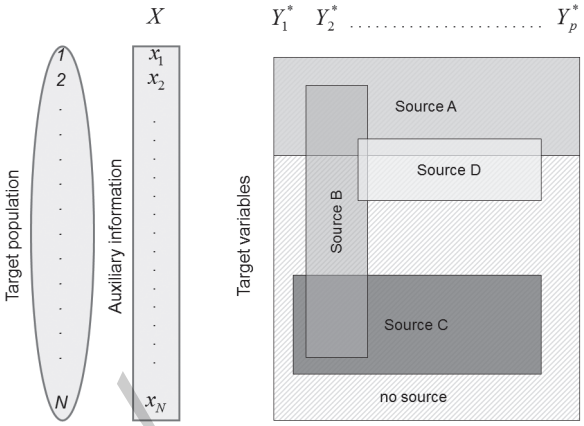

Fig. 1. Informative context with 4 data sources.

The paper is organized as follows: Section 2 describes the model focusing on the true data model (2.1), the error model (2.2) and the estimation process (2.3). Section 3 presents a simulation study using both simulated and true data. Section 4 concludes the paper.

2. The model

In this section we illustrate the model used to handle multisource data. The general informative context is represented in Fig. 1: p target variables are observed in G data sources, but not all the variables are available in each source, and the sources cover only subsets of the target population. Note that in some cases more than one data source is available, while in other cases there is no information at all.

In the paper, we focus on the univariate case, with only one target variable measured in G data sources. As in the general case, the different sources may cover only subsets of the whole target population.

2.1. The true data model

Let us assume that the true unobserved data are realizations from n *iid* random Gaussian variables Y_i^*, with mean μ_i and common variance $\sigma^2 (i = 1, ..n)$.

We also allow for the possibility of a linear dependence of the means μ_i on some set of q covariates $x_i = (x_{i0}, x_{i1} \ldots, x_{iq})'$ observed without error, i.e., we assume the relation $\mu_i = \beta' x_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}$, where $\beta_j (j = 0, \ldots, q)$ are unknown coefficients to be estimated and, as usually, we set $x_{i0} \equiv 1$. Thus, true data are modeled via the ordinary linear regression model:

$$Y_i^* = \beta' x_i + U_i, \quad i = 1, \ldots, n, \tag{1}$$

where U_i are *iid* Gaussian random variables with zero mean and variance σ^2.

In real applications on economic data, logarithms of data instead of data in their original scale are often assumed to be normally distributed. This does not imply substantial changes in the proposed methodology.

2.2. The error model

Assume that the variable of interest is observed with error in some (possible all) of the G sources $S^1, ..., S^G$ (for instance administrative archives) and let Y_i^g be the variable corresponding to the value observed in the source S^g for the unit i ($i = 1, \ldots, n$).

In order to complete the modeling, we have to specify the measurement error model for each source, that is, the conditional distribution of Y_i^g given the true value y_i^*. An essential feature of the error mechanism to be taken into account is its "intermittent" nature, where "intermittence" refers to the fact that in the present context, differently from common situations in experimental sciences, it is assumed that only a fraction of the available data are affected by errors, or, in other words, that data are only partially contaminated. This assumption naturally leads to the adoption of contamination models for the observed data. These models have been largely used to detect outliers or influential errors in statistical data available from a single data source [5,6].

In detail, we model the intermittent nature of the error on the different data sources via independent Bernoullian variables Z_i^g with parameters π_g, i.e., $Z_i^g = 1$ if an error occurs for the unit i in the source S^g, or in other words, if $Y_i^g \neq Y_i^*$, and zero otherwise. Also, given the event $\{Z_i^g = 1\}$, we assume that $Y_i^g = Y_i^* + \varepsilon_i^g$ where ε_i^g are mutually independent Gaussian variables with zero mean and variance $\alpha_g \sigma^2$, where α_g is a positive constant ($g = 1, \ldots, G$).

In short, the measurement error model can be described through the equation:

$$Y_i^g = Y_i^* + Z_i^g \varepsilon_i^g \quad g = 1, \ldots, G; i = 1, \ldots, n. \tag{2}$$

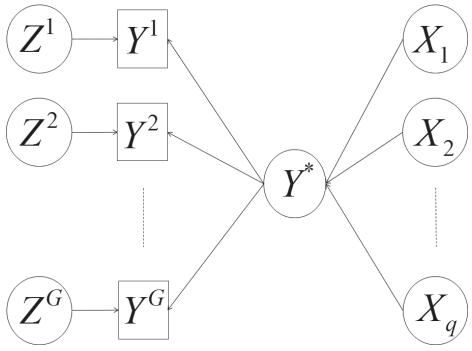

Fig. 2. Linear regression and measurement error model.

Equations (1) and (2) completely specify the model. We note that the parameters (π_g, α_g) can be thought of as quality indicators for the source S^g, representing, respectively, the (a priori) error probability and the effect of the error (variance inflation).

In Fig. 2 the model is illustrated. Following the conventions, the circle and the rectangles represent the latent and observed variables respectively, arrows connecting latent and/or observed variables represent direct effects, which do not need to be linear.

Some issues on the model assumptions need to be discussed. First, the assumption $E(\varepsilon_i^g) = 0$ implies that the errors in all the sources are "random", so that the model would be not appropriate in presence of systematic errors. It follows that the approach should be applied after some appropriate editing of the data to start with. Alternatively, one could introduce additional terms in the model to account for possible systematic errors such as definitional bias. For instance, the model $Y_i^g = a^g + b^g Y_i^* + Z_i^g \varepsilon_i^g$ could be used in place of model 2). However, the latter option implies the need to introduce some distributional constraints in order to identify the additional parameters a^g and b^g. This is what is usually done in the context of factor analysis, where the latent variables are hypothetical constructs to be interpreted. In the present context, where the non-observed variable Y_i^* has its own definition being related to some quantity of the real world, we define the Y^gs as variables coinciding with Y^* with probability greater than zero, and we treat the other relevant variables not having this property as *covariates*.

A second important issue is related to the fact that in the present model the ε_i^g variables, as well as the Z_i^g variables, are supposed to be mutually independent. This corresponds to assuming independence of the different measurement processes, or, more precisely, conditional independence of the available measures (Y_i^g variables) given the true data (Y_i^* variables). The as-

sumption of independence among the measurement errors in the different sources may be unrealistic in some circumstances. Specifically, local dependence effects may occur when some non-observed characteristics of the target population (e.g., specific of some subpopulations) have similar effects on the measures from the different sources. The inclusion of these effects in the modelling by relaxing the conditional independence assumption makes the correlation structure of the observed data (and thus of the estimation process) much more complex and will not be taken into account in the present work.

2.3. Estimation

In order to estimate the parameters of the model specified via Eqs (1) and (2), we need to derive the observed data distribution. Since we treat the case of partially overlapping sources, where for some units less than G sources may be available (see Fig. 1), the *observed data* for each unit i are the measures $y_i^{j_1}, \ldots, y_i^{j_m}$ corresponding to the m available sources $S^{j_1}, \ldots, S^{j_m}$, where $(j_1, \ldots, j_m) \subset (1, \ldots, G)$.

From the above model assumptions it follows that the distribution $f(y_i) = f(y_i^{j_1}, \ldots, y_i^{j_m})$ of the random vector $Y_i = (Y_i^{j_1}, \ldots, Y_i^{j_m})$ associated with the measures from $S^{j_1}, \ldots, S^{j_m}$ available for the ith unit is a mixture of probability distributions corresponding to the different error patterns across the sources. Formally:

$$f(y_i) = \sum_{k=1}^{2^m} w_k h_k(y_i; \beta, \sigma^2, \alpha), \qquad \alpha \equiv \alpha_{j_1}, \ldots, \alpha_{j_m}, \beta \equiv \beta_0, \ldots, \beta_q, \tag{3}$$

where the sum is over the 2^m error patterns across $S^{j_1}, \ldots, S^{j_m}$, and for the kth pattern, the "mixing weight" w_k is the product of m factors of the form π_g or $(1 - \pi_g)$ depending on whether the pattern k corresponds to an erroneous or correct value in the source S^g. The densities h_k in 3) are suitable products of Gaussian distributions possibly degenerated in mass points.

As an example, let us consider the case of three sources where the values of a target variable Y are reported. The pattern corresponding to correct (coinciding) values in the first two sources and to an error in the third source will be associated to the weight $(1 - \pi_1)(1 - \pi_2)\pi_3$ and to the density:

$$h(y_i^1, y_i^2, y_i^3) = N(y_i^1; \beta' x_i, \sigma^2)\delta(y_i^2 - y_i^1) N(y_i^3; y_i^1, \alpha_3 \sigma^2),$$

where $\delta(\cdot)$ is the Dirac's delta-function with mass at zero. In this example the first two sources give the same information (this is the reason why the delta-function appears), implying that the common reported value is correct (according to the assumed model, the probability that two erroneous values are equal is zero). On the other hand, the third source gives information only on the error mechanism since in this case the error $y_i^3 - y_i^* = y_i^3 - y_i^1 = y_i^3 - y_i^2$ is *actually observed*. From the example we argue that the problem of classifying the observations according to the different error patterns is partially supervised, in that the assignment is without uncertainty whenever at least two values from different data sources coincide.

The log-likelihood function based on the observed data distribution is obtained by taking logarithm of (3) and summing over the units ($i = 1, \ldots, n$). We implemented an appropriate Expectation Maximization (EM) algorithm for the maximum likelihood estimation (MLE) of the model parameter $\theta \equiv \beta_j, \sigma^2, \pi_g, \alpha_g$ ($j = 0, \ldots, q; g = 1, \ldots, G$). Programming codes for R, developed by the authors, are available on request.

As described in Section 2.2, the estimates of the error model parameters π_g, α_g can be used to assess the quality of the source S^g, being related to the proportion and magnitude of errors in S^g, respectively. Moreover, the estimation procedure provides, for each unit i and source S^g, the estimated posterior probabilities $\hat{\tau}_{ig}$ of presence of error defined as:

$$\hat{\tau}_{ig} = \sum_{k \in I_g} \frac{\hat{w}_k \hat{h}_{k,i}}{\sum_{l=1}^{2^m} \hat{w}_l \hat{h}_{l,i}},$$

where $\hat{w}_k$ and $\hat{h}_{k,i}$ are the estimates of the corresponding quantities in Eq. (3) and I_g is the set of indices associated with the error patterns where y_i^g is not correct. These posterior probabilities could be used to select, for each unit, the best source to take the information from. Finally, the single predictions for each unit can be used to build a micro-data file to be used for different estimation purposes. Moreover, the predictions can also be compared with the values reported in the different data sources in the context of data editing activities.

It is worthwhile noting that, because of the identifiability of finite mixture of Gaussian distributions [7], the intermittent nature of the error mechanism makes the multi-source model indentifiable also in situations

where the corresponding model with "genuinely continuous" random noise is not identifiable. For instance, in case $G = 1$, if U_i and ε_i are zero mean Gaussian variables with variances σ^2 and $\alpha\sigma^2$ respectively, the model specified through the equations a) $Y_i^* = U_i$ and b) $Y_i = Y_i^* + \varepsilon_i$ is not identified, while it becomes identified if equation b) is replaced by the equation b') $Y_i = Y_i^* + Z_i\varepsilon_i$, where Z_i is a Bernoullian variable with parameter π. In fact, in the latter case the observed data distribution $f(y_i)$ is a mixture of two Gaussian distributions with common (zero) mean and different variances:

$$f(y_i) = (1-\pi)N(y_i; 0, \sigma^2) + \pi N(y_i; 0, \alpha\sigma^2).$$

Another important issue that has to be considered in the estimation context is the number of possible error patterns involved in the (log-)likelihood function. This is equal to 2^m when m ($m \leqslant G$) sources are available, and becomes very large if there are many data sources. However, in most practical applications is not likely to be very large and, although the total number of possible error patterns increases exponentially, the number of model parameters increases only linearly as a function of G.

3. Simulation study

In this section, an evaluation of the proposed methodology based on two Monte Carlo (MC) studies is presented. The first study is based on completely simulated data, while in the second study only the error mechanism is simulated by randomly perturbing real economic data that are taken from a subset of the Istat statistical information system for annual SBS. In both studies a single covariate X is considered.

3.1. Study I: Completely simulated data

At each MC iteration, a sample of $n = 2556$ "error-free data" *($i = 1, \ldots, n$)* has been generated in logarithmic scale according to the regression model 1) with one X covariate. The sample size n is the same as the size of the real dataset used in study II.

In order to compare the results with the ones from the real data experiment, we have used, as true model parameters, the (robust) estimates obtained by regressing the Y variable on the X variable in the real economic (log)data. The corresponding values are $\beta_0 = 9.7$, $\beta_1 = 1.2$, $\sigma^2 = 0.32$. Three measurement processes ($G = 3$) have also been simulated according to (2) where different set of values have been used for parameters π_g and α_g. Finally, missing values have been randomly introduced in the three sources with missing rate 0.50 for S^1, 0.10 for S^2, and 0.20 for S^3, reproducing the missing rates observed in the real dataset used in study II.

For each set-up of the experiment and for each MC run, the EM-estimates of the parameters have been used to estimate the posterior probabilities τ_{ig} for each unit i. Moreover, predictions of true values Y_i^* conditional on the available information have also been computed for each unit. Three methods have been compared for building a single set of predicted micro-data. For all the methods, the "obvious" cases corresponding to at least two coinciding sources have been preliminary treated by considering the repeated values as true. The remaining values have been determined as follows. With the first method the values are chosen based on the source hierarchy (*hier*) that is defined, whenever possible, according to the "quality parameters" π_g and α_g. Specifically, when the "quality rank" of the sources is the same with respect to the (estimates of the) π_g and α_g parameters (i.e., when the order of the π_gs agrees with the order of the α_gs), the available source with highest rank is chosen. For example, if $\pi_1 < \pi_2 < \pi_3$ and $\alpha_1 < \alpha_2 < \alpha_3$, the first source is always chosen whenever it is available, while, in cases where it is missing, the second source is preferred, and the third source is used only if it is the only available source. When the "quality rank" of the sources is not the same with respect to the estimates of the π_g and α_g parameters, the source hierarchy is not defined. With the second method (*pps*), for each unit i the source S^g is chosen where the reported value has the "highest" (estimated) posterior probability τ_{ig} of being error-free. Finally, the third method (*pred*) is based on the model predictions, i.e., the micro-data are the expectation of true data conditional on data observed from the different sources. In all approaches, predictions for units where no source is available have been obtained by simply regressing the Y variable on the (always observed) X covariate, using the regression parameters from the EM algorithm.

Assuming that the target quantity is the population mean of the variable Y, we compute this quantity on the basis of the micro-data files obtained with methods *hier*, *pps* and *pred*. Then, for each method, the relative bias (RB) and the relative root mean square error (RRMSE) is estimated by averaging the squares of the estimation errors over 500 MC iterations. Re-

Table 1
Study I: RRMSE for methods based on source hierarchy (*hier*), posterior probabilities (*pps*), and model predictions (*pred*)

π	α	RB			RRMSE		
π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$	*hier*	*pps*	*pred*	*hier*	*pps*	*pred*
0.1, 0.2, 0.3	2, 4, 6	11.85	5.33	−0.28	12.74	6.61	0.86
0.1, 0.2, 0.3	6, 4, 2	Not defined	3.23	0.09	Not defined	3.77	0.86
0.1, 0.2, 0.3	2, 4, 8	13.49	7.82	0.00	15.05	10.12	0.71
0.2, 0.4, 0.6	2, 4, 6	23.69	12.16	0.07	24.53	13.14	1.23
0.2, 0.2, 0.2	2, 4, 6	11.41	3.62	0.72	12.08	4.26	0.84
0.2, 0.4, 0.8	2, 4, 6	26.87	15.75	0.15	27.67	16.61	1.32

Table 2
Study I: RB and RRMSE for parameter estimates

π	α	RB*			RRMSE*		
π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$	$\beta_0, \beta_1, \sigma^2$	π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$	$\beta_0, \beta_1, \sigma^2$	π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$
0.1	2	−0.12	0	−0.99	0.50	0.01	16.16
0.2	4	0.28	0	3.21	1.35	0.01	9.24
0.3	6	−0.63	0	0.12	3.04	0.01	7.59
0.1	6	0.01	0	1.06	0.50	0.01	14.60
0.2	4	0.03	0	1.16	1.38	0.01	9.88
0.3	2	−0.09	0	0.69	3.29	0.01	7.81
0.1	2	−0.07	0	−1.97	0.51	0.01	13.92
0.2	4	0.11	0	1.49	1.30	0.01	9.08
0.3	8	0.66	0	0.11	3.42	0.01	7.87
0.2	2	−0.07	0	3.33	0.58	0.02	12.50
0.4	4	0.07	0	0.72	1.51	0.01	7.62
0.6	6	−0.46	0	0.24	3.96	0.02	5.93
0.2	2	−0.01	0	−0.70	0.56	0.01	9.70
0.2	4	0.05	0	−0.30	1.45	0.01	8.91
0.2	6	−0.17	0	0.54	3.23	0.01	8.87
0.2	2	−0.06	0.01	−1.50	0.58	0.02	13.55
0.4	4	0.14	0	−0.83	1.56	0.02	6.41
0.8	6	0.28	0	0.24	3.73	0.01	6.54

*For the π parameters bias instead of RB and mean square error instead of RRMSE are used.

Table 3
Study II: RRMSE for methods based on source hierarchy (*hier*), posterior probabilities (*pps*), and model predictions (*pred*)

π	α	RB			RRMSE		
π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$	*hier*	*pps*	*pred*	*hier*	*pps*	*pred*
0.1, 0.2, 0.3	2, 4, 6	11.83	6.06	0.77	12.53	7.06	0.88
0.1, 0.2, 0.3	6, 4, 2	Not defined	3.96	0.86	Not defined	4.33	0.94
0.1, 0.2, 0.3	2, 4, 8	13.40	7.99	0.93	14.04	8.56	1.03
0.2, 0.4, 0.6	2, 4, 6	24.15	13.64	1.36	24.71	14.32	1.50
0.2, 0.2, 0.2	2, 4, 6	11.96	5.17	0.81	12.46	5.85	0.90
0.2, 0.4, 0.8	2, 4, 6	26.72	17.32	1.62	27.31	18.01	1.76

sults are reported in Table 1. RB and RRMSE are also reported for the model parameter estimates in Table 2; for "scale" reasons, bias instead of RB and mean square error instead of RRMSE are used for the π parameters.

3.2. Study II: Real data

The only difference with the previous study is that in the present case only the error mechanism is simulated. The number of employees and the corresponding Labour Cost have been considered as X and Y variables respectively. The set of error model parameters are the same as in the study on simulated data. The original unperturbed dataset is composed of the enterprises with more than 10 and less than 100 employees belonging to the Manufacture of textiles division (NACE code 13). A preliminary data analysis performed on the SBS data has shown that data can hardly be considered realizations from a Gaussian distribu-

Table 4
Study II: RB and RRMSE for parameter estimates

w	α	RB*			RRMSE*		
π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$	$\beta_0, \beta_1, \sigma^2$	π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$	$\beta_0, \beta_1, \sigma^2$	π_1, π_2, π_3	$\alpha_1, \alpha_2, \alpha_3$
0.1	2	0.36	0	179.23	0.40	0.01	184.33
0.2	4	−1.57	0	169.14	1.64	0.01	170.55
0.3	6	−62.45	0	165.29	62.46	0.01	166.49
0.1	6	0.35	0	166.34	0.39	0.01	170.27
0.2	4	−1.52	0	171.26	1.59	0.01	172.56
0.3	2	−62.56	0	172.02	62.56	0.01	173.08
0.1	2	0.37	0	179.65	0.41	0.01	183.66
0.2	4	−1.55	0	174.78	1.63	0.01	176.52
0.3	8	−62.62	0	170.23	62.63	0.01	171.41
0.2	2	0.38	0	183.64	0.47	0.02	186.56
0.4	4	−1.56	0	175.81	1.72	0.01	176.78
0.6	6	−62.91	0	172.13	62.93	0.01	173.06
0.2	2	0.36	0	174.2	0.39	0.01	175.78
0.2	4	−1.57	0	172.65	1.61	0.01	173.96
0.2	6	−62.53	0	171.11	62.53	0.01	172.66
0.2	2	0.40	0	184.94	0.47	0.02	188.39
0.4	4	−1.53	0	177.41	1.67	0.02	178.24
0.8	6	−63.61	0	176.37	63.63	0.01	177.02

*For the π parameters bias instead of RB and mean square error instead of RRMSE are used.

tion. Thus, this second study serves as "test of robustness" of the method with respect to departures from the normality assumption. Results on Y-mean estimation and on parameter estimation are reported in Tables 3 and 4 respectively.

3.3. Results

The results reported in Tables 1 and 3 show that, generally, micro data-files based on predictions of true values conditioned on all the available information (*pred*) allow one to obtain the best estimates of the finite population quantity. Moreover, the hierarchical approach based on the *a priori* choice of the best source provides the worst performances. However, if a statistical archive is to be built using only data that are actually observed, a possible option could be that of taking values from different sources for different units, choosing the source corresponding to the lowest error probability (*pps*). An interesting finding is that the dominating component of the mean square error is the bias for the methods *hier* and *pps*, and the variance for *pred*. In particular, also in case of real data (Study II) the estimates via *pred* of the population total for the target variable are only moderately biased.

The comparison between real data and simulated data experiments show that even when the (log)normality assumption for the true data distribution is not appropriate, the proposed modeling can be useful. In fact, although the estimates of the error model parameter (see Tables 2 and 4) are much less accurate for the real data experiment (especially for the α parameters) than for simulated data, the comparison of Tables 1 and 3 shows that the performances of the methods in terms of predictive accuracy differ only slightly. Moreover, as verified, the ordering of the π and α parameters is always correctly estimated for both studies. Thus, valid conclusions on the accuracy of the different data sources are obtained in both cases.

4. Conclusions and future research

In the paper we have presented a new approach to deal with multiple source data for different purposes, such as statistical estimation, quality assessment, editing. The advantage of this approach is that it exploits the information of all data sources, instead of relying on a predetermined hierarchy among sources.

At the moment, algorithm and software have been fully developed in the univariate case. Multivariate extensions imply more complex procedures for the likelihood maximization and are under investigation. In fact, when several target variables are simultaneously considered, the number of possible error patterns "explodes". However, if mutual independence among measurement errors on different variables is assumed, the number of corresponding parameters in the error model increases only linearly with the number of target variables. The latter independence assumption could

be not realistic in some practical applications and need to be evaluated case by case.

A further direction for future research is the extension of the proposed approach to more complex models that can account for non-normal true data distributions and more general error mechanisms. For example, it could be of interest to relax the conditional independence assumption among the measurement processes or to extend the model to situations where the a priori probabilities π_g and the variance inflation parameters α_g depend on some set of covariates.

References

[1] O. Luzi, M. Di Zio, F. Oropallo, A. Puggioni and R. Sanzo, Integrating administrative and survey data in the new Italian system for SBS: quality issues. The 3rd European Establishment Statistics Workshop, Nuremberg, Germany [Internet], 2013 [cited 2015 Nov 16]. Available from: http://enbes.wikispaces.com/EESW13.

[2] B.F.M. Bakker, Estimating the validity of administrative Variables, *IStatistica Neerlandica* **66**(1) (2012), 8–17.

[3] S. Scholtus and B.F.M. Bakker, Estimating the validity of administrative and survey variables through structural equation modeling. A simulation study on robustness. Discussion paper (201302), Statistics Netherlands, The Hague/Heerlen [Internet]. 2013 [cited 2015 Nov 16]. Available from: http://www.cbs.nl.

[4] D. Pavlopoulos and J.K. Vermunt, Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology* **41**(1) (2015), 197–214.

[5] M. Di Zio and U. Guarnera, A contamination model for Selective Editing, *Journal of Official Statistics* **29**(4) (2013), 539–555.

[6] B. Ghosh-Dastidar and J.L. Schafer], Outlier Detection and Editing Procedures for Continuous Multivariate Data, *Journal of Official Statistics* **22**(3) (2006), 487–506.

[7] H. Teicher, Identifiability of finite mixtures, *Ann Math Statist* **34** (1963), 1265–1269.