Working paper 32

**UNITED NATIONS**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Budapest, Hungary, 14-16 September 2015)

Topic (v): Emerging methods and data revolution

# Estimation and editing for data from different sources. An approach based on latent class model

Prepared by U. Guarnera and R. Varriale, Istat, Italy

## I. INTRODUCTION

1. Over the last few years, due to both the increased availability of different administrative sources of information and the need for reducing the response burden, many statistical institutes have been changing some statistical production processes from essentially "survey data based" to "administrative data based". Specifically, in some circumstances, administrative data are viewed as primary source of information instead of merely auxiliary information. The new role of the external sources implies the need for dealing with new operational and methodological problems. In fact, administrative data are typically collected by different institutions for specific purposes and may not be usable in their original form for statistical purposes. Thus, a preliminary work is necessary to harmonize definitions and transform variables in order to meet user needs. Even if many efforts are devoted to these pre-processing activities, some statistical methodology has to be applied to treat "errors" in data that cannot be explained in terms of systematic discrepancies between administrative and statistical concepts. The need for statistical editing of data from external sources is particularly evident in cases where several partially overlapping data sources are available (multi-source data), and discrepancies among corresponding values are observed. In these cases, if a statistical archive is to be built, the problem arises of which source has to be chosen. 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, one can choose the source with the "highest" score according to the established hierarchy. However, in many circumstances, it is not obvious how to establish the hierarchy. Moreover, for some units, values from sources with lower overall "quality rank" may result to be more plausible than the corresponding values from higher quality sources.

2. In this work, we suggest an alternative approach for numerical variables, where one takes advantage of the simultaneous availability of information from the different sources. With this approach, all the available information is used and "weighted" according to its reliability. This is done by modeling both the true data and the measurement processes for the different sources. Specifically, 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 the specification of a conditional distribution of the data observed in each source given the true unobserved data, one can easily derive, via Bayes formula, the distribution of the true data given all the observed data. The latter can be used to obtain predictions of true values to be used for (selective) editing of the administrative sources, or to be directly placed in a statistical archive.

3. The paper is structured as follows. In Section II the models for the true data and measurement process are introduced. In Section III an application to real survey data is described. Some concluding comments are provided in Section C.

## II. Modeling contaminated multi-source data through latent variables

4. In this section the general approach for modeling numerical data from different sources is illustrated. According to the usual notation, capital letters will denote random variables while small letters will be associated with actual realizations of random variables. We assume that the estimates of interest are based on a set of $p$ continuous variables $Y^* = Y_1^*, \ldots, Y_p^*$ (*target variables*) on a population of $n$ units, and measures of these variables are available from $G$ sources of information $S_1, \ldots, S_G$. We assume that in each source $g = 1, \ldots, G$ data can contain errors, and we denote the corresponding observed variables with $Y^g \equiv Y_1^g, \ldots, Y_p^g$. Moreover, information may be partial in the sense that in the different data sources some variables can be not available, and it can happen that only subsets of the target population are covered. Finally, a set of covariates $X$ can also be available for all the population units. Fig. 1 shows a representation of an informative context with $G = 4$ data sources, $p$ target variables and a single covariate $X$.

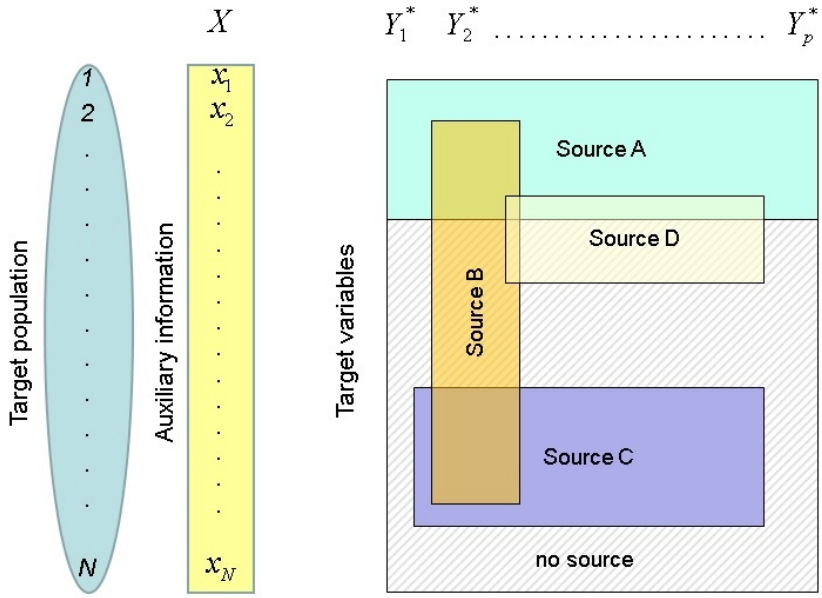


FIGURE 1. Informative context with 4 data sources, $p$ target variables and a single covariate $X$ available for all the population units

In our framework we adopt a practical criterion to distinguish $X$ variables (*covariates*) from $Y$ variables (*target variables* measured with error). We say that a variable $Y$ is a target variable ($Y^*$ variable) measured with error if its conditional distribution (density) given $Y^*$, denoted by $f(y|y^*)$, can be expressed as

$$f(y|y^*) = (1-\pi)\delta(y-y^*) + \pi h(y|y^*), \tag{1}$$

where $0 < \pi < 1$, $\delta(y-y^*)$ denotes the Dirac delta function with mass on the point $y^*$, and $h(y|y^*)$ is an arbitrary density function. Equation (1) together with the strict inequality $\pi < 1$ express a measurement model based on the assumption of an *intermittent* error mechanism, implying that only a proportion of the data are affected by errors. This assumption seems to be reasonable in the context of

official statistics, being the basic implicit assumption for many common data cleaning approaches based on *error localization* techniques. Note that we have excluded the trivial case of perfect measurement ($\pi = 0$). Differently from the $Y$ variables, the $X$ variables are such that, intuitively speaking, the events $X = Y^*$ have probability zero (we could express more formally this condition in terms of absolute continuity of the relevant probability measures with respect to each another). Furthermore, the probability distribution of the $X$ variables is not of interest, so that the $X$ variables can be considered as genuine *auxiliary information*.

5. According to the above terminology, the model is specified through the conditional distribution of the $Y^*$ variables given the covariates ("true data model"), and of the contaminated target variables $Y^g$ given the $Y^*$ variables (error models for each source $S_g$). Note that, in this framework, the $Y^*$ variables are latent variables in the sense that they are not directly observed. However, due to the intermittent nature of the error mechanisms characterizing the measurement processes associated with $S_1, \ldots S_G$, there is a strictly positive probability that measures from the different data sources are error-free. Moreover, since the relevant probability distributions are supposed to be continuous, true values are *surely* observed in the special case of coincidence of corresponding observations from different data sources.

6. The characterization of the $Y$ variables via equation (1) makes the present approach different from many common approaches for the quality assessment of the administrative sources. For instance, when dealing with continuous variables, in Bakker [2012] a structural equation model is used to assess and compare the quality of administrative sources for statistical use and in Scholtus and Bakker [2013] 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 these works the random variables representing the true data and the measurement processes are typically continuous, and special contraints on the model parameters are necessary in order to identify the model. In our approach, differently from common settings in the Factor Analysis and Item Response Theory literature, the latent variables represent *true* (although not directly observed) quantities rather than hypothetical constructs to be interpreted, so that the corresponding parameters cannot be subject to identification constraints (see Skrondal and Rabe-Hesketh [2004] for a discussion on variables measured with error vs hypothetical constructs).
In the context of categorical data, an approach similar to the one used in the present paper is adopted in Pavlopoulos and Vermunt [2015], where 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.

7. In this paper we illustrate the approach focusing on the case of only one $Y^*$ variable (univariate case) and a simple additive Gaussian error model. Specifically, we assume that the true unobserved data (possibly in log scale) are realizations from $n$ independent Gaussian random variables $Y_i^*$ $(i = 1, \ldots, n)$ whose expected values $\mu_i$ are linear combinations of a set of $q$ covariates: $\mu_i = \beta' x_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \ldots + \beta_1 x_{iq}$, where $x_{i0} = 1$ and $\beta_0, \ldots \beta_q$ are parameters to be estimated. Thus, we assume the regression model:

$$Y_i^* = \beta' x_i + U_i, \qquad U_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n. \tag{2}$$

The error model for each source $S_g$ is specified trough the conditional distribution of $Y_i^g$ given $Y_i^*$. This is defined by the equation:

$$Y_i^g = Y_i^* + Z_i^g \epsilon_i^g, \quad i = 1, \ldots, n \quad g = 1, \ldots, G \tag{3}$$

where $Z_i^g$ are independent Bernoullian random variables with parameter $\pi_g$ and $\epsilon_i^g$ are independent Gaussian random variables with zero mean and variance $\sigma_g^2$. The Bernoullian variables $Z_i^g$ are introduced in order to characterize the measurement error process as an intermittent mechanism (see Di Zio and Guarnera [2013] for the single data source case). Note that due to the independence of $Z$ and $\epsilon$ variables, the measurement variables $Y^g$ are conditionally independent given the true variable $Y^*$. In Fig. 2 the model is illustrated: 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. In the above setup, the parameters $\pi_g$ and $\sigma_g^2$ can be viewed as "quality parameters" for the source $S_g$, being associated with the "error frequency" and the "error magnitude", respectively.

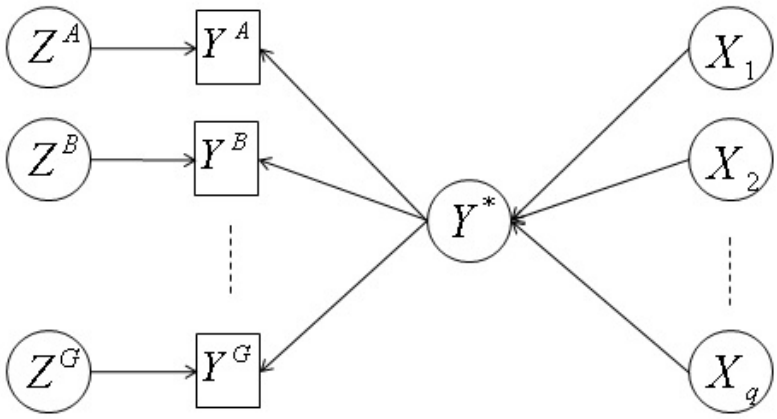


FIGURE 2. Model for the true data distribution and the error mechanism.

8. If the model parameters $\theta \doteq (\beta, \sigma^2, \pi_g, \sigma_g^2)$ were known, one could easily obtain predictions $\hat{y}_i$ of the true values $y_i^*$ conditional on the observed values $y_i^1, \ldots, y_i^G$ from the available sources. In fact, using the Bayes formula, the conditional distribution of $Y_i^*$ given $y_i^1, \ldots, y_i^G$ can be easily derived from equations (2) and (3), and expectation from this distribution can be used as $\hat{y}_i$. Comparing the predictions $\hat{y}_i$ with the observed values $Y_i^g$, $(i = 1, \ldots, n; \quad g = 1, \ldots, G)$, one can try to identify gross errors in the different data sources, possibly prioritizing units to be interactively revised (selective editing). On the other hand, from an estimation perspective, "robust" estimates can be obtained by replacing the observed values with the predictions in the corresponding estimators. Moreover, using again the Bayes formula, the probabilities of the events $\{Y_i^* = Y_i^g\} \equiv \{Z_i^g = 0\}$ conditional on the observed values $y_i^g$ can also be derived. These probabilities could be used to choose, for each unit, the most reliable source. Different sources will be possibly chosen for different units. Since the model parameters are unknown, they have to be estimated based on the set of available observations. A short description of the estimation procedure is reported in the following section.

## A. Parameter estimation

Starting from the model specification (Equations (2) and (3)), we can derive the probability distribution of observed data that is a $G - variate$ mixture of $2^G - 1$ degenerate and one non-degenerate Gaussian distributions. Specifically, each mixture component corresponds to a particular error pattern across the data sources. As an example, for three sources, the density $h_{001}(y_i^1, y_i^2, y_i^3)$ associated with the

error pattern $\{Z_i^1 = 0, Z_i^2 = 0, Z_i^3 = 1\}$ (error only on the third source) is

$$h_{001}(y_i^1, y_i^2, y_i^3) = N(y_i^1; \beta' x_i, \sigma^2)\delta(y_i^2 - y_i^1)N(y_i^3; y_i^1, \sigma_3^2),$$

where $\delta(\cdot)$ denotes the Dirac's delta function with mass in zero and $N(\cdot, \mu, \xi^2)$ is the Gaussian density with mean $\mu$ and variance $\xi^2$. The corresponding mixing weight is $(1 - \pi_1)(1 - \pi_2)\pi_3$. Note that the coincidence of $y_i^1$ with $y_i^2$ implies that in this case the true value $y_i^*$ (and thus the realized error on the source $S_3$) is actually observed (with probability 1). This example shows that the intermittence assumption of the error mechanism makes the estimation problem partially supervisioned, in that the class membership is known whenever the same value is observed in two or more data sources.

9. Based on a set of $G$-dimensional (possibly incomplete) observations, maximum likelihood estimates of the model parameters can be obtained via an appropriate EM algorithm. An R-code version has been implemented for the case $G = 3$. Using this algorithm, we have applied the methodology to both simulated and real data. Some applications are described in the following section.

## III. Application

10. In this section, we present an application of the methodology described in Section II. First, we give an evaluation of the methodology based on a Monte Carlo (MC) study where the error mechanism for three data sources is simulated and the true data $Y^*$ are taken from a subset of the Istat statistical information system for annual SBS [Luzi et al., 2013]. Second, we apply the methodology to real $Y$ variables. In both applications a single covariate $X$ is used.

### A. Simulated data

11. In the MC study, $Y^*$ and $X$ variables are Earnings and Number of employees, respectively, extracted from a subset of the Istat statistical information system for annual SBS, namely the set of the enterprises with more than 10 and less than 100 employees belonging to the Manufacture of textiles division (NACE code 13). We simulated three measurement processes ($G = 3$) according to Equation (3), with different set of values for parameters $\pi_g$ and $\alpha_g$, and we randomly introduced missing values in the three sources according to the missing rates $0.50, 0.10, 0.20$, respectively. For each set-up of the experiment and for each MC run, the posterior probabilities $\tau_{ig}$ for each unit $i$ have been estimated by using the EM-estimates of the parameters. We also computed, for each unit, the predictions of true values $Y_i^*$ conditional on the available information.
To evaluate the model performance, the relative bias (RB) and root mean square error (RRMSE) have been computed for the model parameters: $\alpha_g$ $(g = 1, \ldots, G)$, $\beta_0$, $\beta_1$ and $\sigma^2$. For the $\pi$ parameters bias instead of RB and mean square error instead of RRMSE are used. Results on $Y$-mean estimation and on parameter estimation are reported in Tables 1 and 2, respectively.
As previously mentioned, the proposed methodology can be used to build a complete statistical archive for general estimation purposes. In order to extract a single micro-data file from the different data sources, we compared three methods. 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.

(1) Method *hier*: the values are chosen based on the source hierarchy that is defined according to the "quality parameters" $\pi_g$ and $\alpha_g$. Specifically, when the "quality rank" of the sources is the same with respect to the (estimates of the) $\pi_g$ and $\alpha_g$ parameters (i.e., when the order of the $\pi_g$ agrees with the order of the $\alpha_g$), 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 $\pi_g$ and $\alpha_g$ parameters, the source hierarchy is not defined;

(2) Method *pps*: for each unit $i$ the source $g$ is chosen where the reported value has the "highest" (estimated) posterior probability $\tau_{ig}$ of being error-free;

(3) Method *pred*: the micro-data are the expectation (model predictions) 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 root mean square error (RRMSE) is estimated by averaging the squares of the estimation errors over 500 MC iterations.

12. Table 1 and 2 show the results for the second MC study. 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, while the hierarchical approach (*hier*) based on the a priori choice of the best source provides the worst performances. The method based on the choice of values from different sources for different units based on the the source corresponding to the lowest error probability (*pps*), even if is not the best performing, can be used when a statistical archive is to be built using only data that are actually observed. Table 2 shows that the bias is the predominant component of the RRMSE. This is probably due to the misspecification of the model assumed for the true data. However, even when the (log)normality assumption for the true data distribution is not appropriate, the proposed modeling can still be useful. Indeed, although the estimates of the error model parameters are not very accurate (especially for the $\alpha$ parameters), the performance in terms of predictive accuracy is quite satisfactory. Moreover, as verified, the ordering of the $\pi$ and $\alpha$ parameters is almost always correctly estimated (the maximum error in the ordering of $\alpha$ parameters is 5 out of 500 iterations). Thus, valid conclusions on the accuracy of the different data sources can be obtained.

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

| $\pi$ | $\alpha$ | RRMSE | | |
|---|---|---|---|---|
| $\pi_1, \pi_2, \pi_3$ | $\alpha_1, \alpha_2, \alpha_3$ | *hier* | *pps* | *pred* |
| $0.1, 0.2, 0.3$ | $2, 4, 6$ | 51.98 | 29.73 | 1.92 |
| $0.1, 0.2, 0.3$ | $6, 4, 2$ | $Not\ defined$ | 11.46 | 1.81 |
| $0.1, 0.2, 0.3$ | $2, 4, 8$ | 54.02 | 32.29 | 1.91 |
| $0.2, 0.4, 0.6$ | $2, 4, 6$ | 87.30 | 53.64 | 3.06 |
| $0.2, 0.2, 0.2$ | $2, 4, 6$ | 78.66 | 65.36 | 1.88 |

## B. **Real data**

13. In the application on real data, we used as $Y$ variable the Earnings measured in three administrative sources, namely *Financial Statements*, *Studi di Settore* and *Unico*. The units of Financial statements (FS) are the companies, mainly corporate firms, liable to fill in the financial statement. The Studi di settore (SDS) is a Fiscal Authority survey; the units compiling the SDS form are the enterprises with a turnover less than $7,500,000$ euros belonging to many activity sectors. Unico (UNI) is an archive with fiscal data available for almost all enterprises. The $X$ variable is the number of employees from the Istat register ASIA. Figure 3 shows the $Y$s and $X$ variables. While for the bulk

TABLE 2. Study II: RB and RRMSE for parameter estimates

| $\pi$ | $\alpha$ | RB* | | | RRMSE* | | |
|---|---|---|---|---|---|---|---|
| $\pi_1, \pi_2, \pi_3$ | $\alpha_1, \alpha_2, \alpha_3$ | $\beta_0, \beta_1, \sigma^2$ | $\pi_1, \pi_2, \pi_3$ | $\alpha_1, \alpha_2, \alpha_3$ | $\beta_0, \beta_1, \sigma^2$ | $\pi_1, \pi_2, \pi_3$ | $\alpha_1, \alpha_2, \alpha_3$ |
| 0.1 | 2 | 3.79 | 0.00 | 0.35 | 3.82 | 0.01 | 20.00 |
| 0.2 | 4 | -7.39 | 0.00 | -3.32 | 7.48 | 0.01 | 10.97 |
| 0.3 | 6 | 8.12 | 0.00 | -5.86 | 9.60 | 0.01 | 11.18 |
| 0.1 | 6 | 3.72 | 0.00 | -5.22 | 3.75 | 0.01 | 17.82 |
| 0.2 | 4 | -7.25 | 0.00 | -4.26 | 7.33 | 0.01 | 10.98 |
| 0.3 | 2 | 8.60 | 0.00 | -6.31 | 9.65 | 0.02 | 11.45 |
| 0.1 | 2 | 3.90 | 0.00 | 0.00 | 3.93 | 0.01 | 19.02 |
| 0.2 | 4 | -7.65 | 0.00 | -2.24 | 7.75 | 0.01 | 11.23 |
| 0.3 | 8 | 7.70 | 0.00 | -4.93 | 8.82 | 0.01 | 9.38 |
| 0.2 | 2 | 3.89 | 0.01 | 4.50 | 3.95 | 0.02 | 18.38 |
| 0.4 | 4 | -7.63 | -0.01 | -1.55 | 7.84 | 0.02 | 8.93 |
| 0.6 | 6 | 5.97 | 0.00 | -3.78 | 7.88 | 0.02 | 8.42 |
| 0.2 | 2 | 3.90 | 0.00 | -5.65 | 3.92 | 0.02 | 13.85 |
| 0.2 | 4 | -7.66 | 0.00 | -4.52 | 7.74 | 0.01 | 11.51 |
| 0.2 | 6 | 8.47 | 0.00 | -4.42 | 9.45 | 0.01 | 11.32 |
| 0.2 | 2 | 3.86 | -0.01 | 10.95 | 3.93 | 0.03 | 26.99 |
| 0.4 | 4 | -7.50 | 0.00 | -2.44 | 7.75 | 0.02 | 10.82 |
| 0.8 | 6 | 5.14 | 0.00 | -3.08 | 7.65 | 0.02 | 8.15 |

of data the information on Earnings provided by the administrative sources is the almost the same, in some situations strong discrepancies are observed.

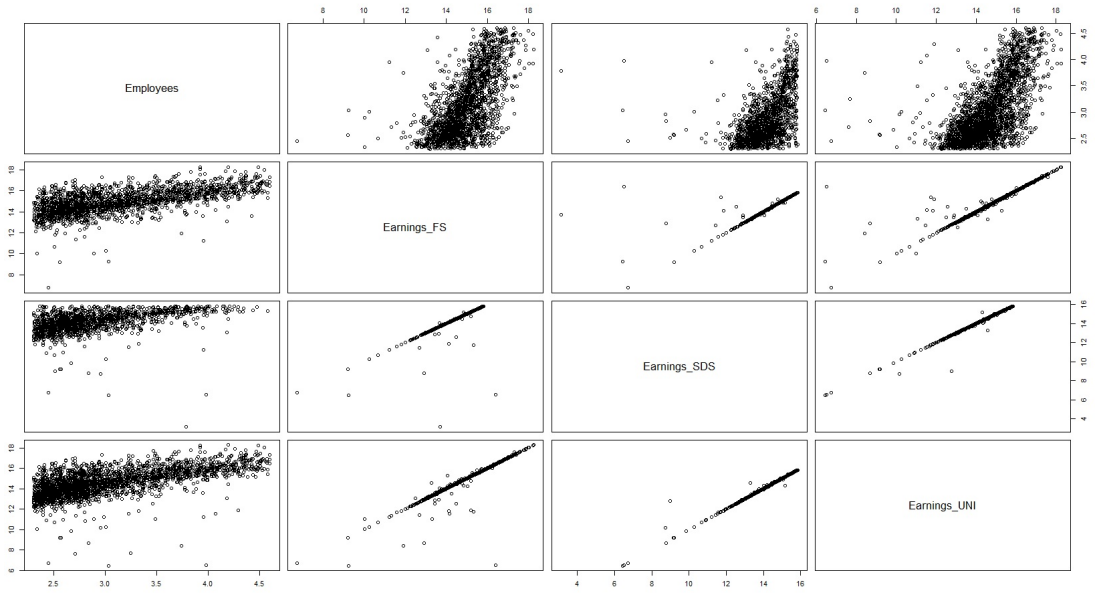


FIGURE 3. Earnings and Number of employees. Source: Istat statistical information system for annual SBS, year 2012, NACE code 13.

14. Note that when working with real economic data, one has to fix a tolerance value under which values from different sources have to be considered "coinciding" in order to distinguish strong

discrepancies from small differences due, for example, to rounding errors. The tolerance value is determined in terms of relative difference and is set to 0.01 in the present application. Running the EM algorithm provides the following estimates $\hat{\alpha}$ and $\hat{\pi}$ for the error parameters:

TABLE 3. Real data

| Source | $\hat{\alpha}$ | $\hat{\pi}$ |
|---|---|---|
| *FS* | 4.28 | 0.02 |
| *SDS* | 6.12 | 0.01 |
| *UNI* | 0.02 | 0.08 |

From Table 3, it results that the source UNI has the highest estimated percentage of errors, but the variance inflation parameter is very small (0.02), so that this source can be considered quite accurate for the set of enterprises under analysis. On the other hand, the measurement process for SDS seems to be characterized by large errors occurring with low probability. The FS data source presents an intermediate situation. An important finding from the application on real data is that the procedure seems to be quite sensitive with respect to the choice of the tolerance parameters (results are not reported here). In fact, if the normality assumption for the error model were adequate, using a tolerance parameter would not be necessary. However, in real economic data gross errors are apparently superimposed to a "ground noise" (due to rounding, unknown source-specific characteristics, etc. ) causing small discrepancies among the sources, that one tries to "filter out" through the tolerance value. However, it is not easy to distinguish noise from gross errors so that there is not a clear threshold value for the tolerance, and different tolerance values corresponds to changing the "definition of error".

15. Another important concern is about the maximum likelihood estimation of the model parameters. An extensive exploratory analysis (not reported here) of the likelihood surface based on running the EM algorithm from different starting points showed the presence of several local maxima. This probably reflects the difficulty of the methodology in identifying the observed values which are closest to the unknown unobserved true values, when strongly predictive covariate are not available (see Figure 3 showing the scatter plot matrix of earnings vs employees).

## C. Conclusions

16. In this paper we presented a model-based approach to deal with multiple source data for different purposes, such as statistical estimation, quality assessment and editing. The advantage of this approach is that it exploits the information from all data sources, instead of relying on a predetermined hierarchy among sources. The studies based on simulated error mechanisms showed that the proposed methodology can be usefully applied even in presence of moderate departure from the normality assumption for the true data distribution. However, as remarked at the end of the previous section, non Gaussianity of the errors causes sensitivity of the procedure with respect to some set-up parameters such as the tolerance value and the starting points for the EM. Further research should be devoted to extend the methodology to non Gaussian errors. A possibility could be modeling errors through a two-component Gaussian mixture where the components are associated with small and gross errors respectively. At the moment, algorithm and software have been fully developed in the univariate case. Thus, a further direction for future research should be the extension to the multivariate setting where different versions are possible depending on whether errors on different variables are to be considered independent one of each other or not. In any case, the estimation procedure becomes much more complex. Finally, an interesting point is the one concerning the identification of the model in case of more complex error mechanisms such as, for instance, models that account for possible biased measurement processes. It is plausible that the intermittent nature of the error, which implies a partially

supervisioned estimation problem, makes the identification problem less severe with respect to error models based on properly continuous distribution.

## References

B.F.M. Bakker. Estimating the validity of administrative variables. *Statistica Neerlandica*, 66(1):8–17, 2012.

M. Di Zio and U. Guarnera. A contamination model for selective editing. *Journal of Official Statistics*, 29(4):539–555, 2013.

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. *Paper presented at The 3rd European Establishment Statistics Workshop 2013. Nuremberg, Germany*, 2013.

D. Pavlopoulos and J.K. Vermunt. Measuring temporary employment. do survey or register data tell the truth? *Survey Methodology*, 2015.

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. *Statistics Netherlands, The Hague/Heerlen*, 02, 2013.

A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling*. Chapman & Hall, 2004.