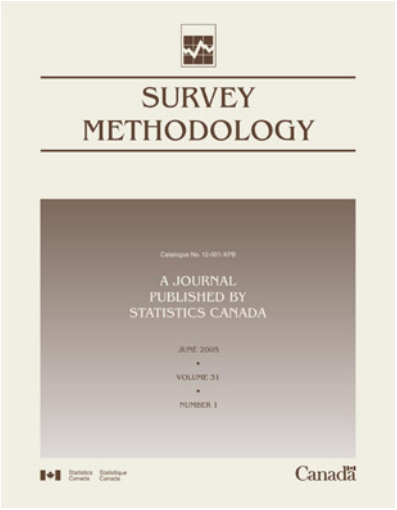
Catalogue no. 12-001-XIE

Survey Methodology

June 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 1, pp. 53-63
Statistics Canada, Catalogue No. 12-001

Editing Systematic Unity Measure Errors Through Mixture Modelling

Marco Di Zio, Ugo Guarnera and Orietta Luzi [1]

Abstract

In Official Statistics, data editing process plays an important role in terms of timeliness, data accuracy, and survey costs. Techniques introduced to identify and eliminate errors from data are essentially required to consider all of these aspects simultaneously. Among others, a frequent and pervasive systematic error appearing in surveys collecting numerical data, is the unity measure error. It highly affects timeliness, data accuracy and costs of the editing and imputation phase. In this paper we propose a probabilistic formalisation of the problem based on finite mixture models. This setting allows us to deal with the problem in a multivariate context, and provides also a number of useful diagnostics for prioritising cases to be more deeply investigated through a clerical review. Prioritising units is important in order to increase data accuracy while avoiding waste of time due to the follow up of non-really critical units.

Key Words: Editing; Random error; Systematic error; Selective editing; Model-based cluster analysis.

1. Introduction

Elements determining the quality of an Editing and Imputation (E&I) process are various and have been widely discussed in literature (Granquist 1995). We deal with a particular non-sampling error that highly affects two main competing quality dimensions: timeliness and data accuracy. As far as accuracy is concerned, we adopt the definition suggested in the Encyclopedia of Statistical Sciences, (1999): "accuracy concerns the agreement between statistics and target characteristics". A number of factors can cause inaccuracy along the overall statistical survey process. Inaccuracy can be reduced during the E&I phase, which can be viewed as an "accuracy improvement tool by which erroneous or highly suspect data are found, and if necessary corrected (imputed)" (Federal Committee on Statistical Methodology 1990).

Due to the complexity of investigated phenomena and the existence of several types of non-sampling errors the E&I phase can be a very complex and time consuming task (Granquist 1996). In the specialised literature a common error classification leads to define two different error typologies: *systematic error* and *random error*. The former relates to errors which go in the same direction and lead to a bias in statistics, while the latter refers to errors which spread randomly around zero and affect the variance of estimates (Encyclopedia of Statistical Sciences 1999). Understanding nature of errors is not only useful in order to identify their source and to assess their effects on estimates, but also to adopt the most appropriate methodology to deal with them (Di Zio and Luzi 2002). While the Fellegi–Holt approach (Fellegi and Holt 1976) is a well-established paradigm to deal with random errors, systematic errors are generally treated by means of ad hoc solutions (see for instance Euredit 2003, Vol. 1, Chapter 5). Systematic errors are generally treated before dealing with random errors, particularly when the latter are tackled through automatic software, like for instance the Generalised Editing and Imputation System (GEIS) (Kovar, Mac Millan and Whitridge 1988) and more recently De Waal (2003).

In the family of systematic errors, one that has a high impact on final estimates and that frequently affects data in statistical surveys measuring quantitative characteristics (*e.g.*, business surveys) is the *unity measure error times a constant factor* (*e.g.*, 100 or 1,000). This error is due to the erroneous choice, by some respondents, of the unity measure in reporting the amount of some questionnaire items.

As real examples of surveys affected by this type of error, we selected two ISTAT investigations: the 1997 Italian Labour Cost Survey (LCS) and the 1999 Italian Water Survey System (WSS).

The LCS is a periodic sample survey that collects information on employment, worked hours, wages and salaries and labour cost on about 12,000 enterprises with more than 10 employees. In Figure 1 the logarithm of Labour Cost (LCOST), Number of Employees (LEMPLOY), Worked Hours (LWORKEDH) are represented in a scatter plot matrix. Note that the employment variable at this editing stage is error free because of a preliminary check with respect to information from business registers (Cirianni, Di Zio, Luzi and Seeber 2000). The analysis of Figure 1 shows that Labour Cost is affected by two types of unity measure error (*i.e.*, 1 million and 1,000 factor), while Worked Hours exhibits only the 1,000 factor error. These errors cause the different clusters in Figure 1. Note that the clusters in the low left corners of each scatter plot represent non-erroneous data.

1. Marco Di Zio, Ugo Guarnera and Orietta Luzi, Italian National Statistical Institute, Via Cesare Balbo 16, 00184 Roma, Italy.

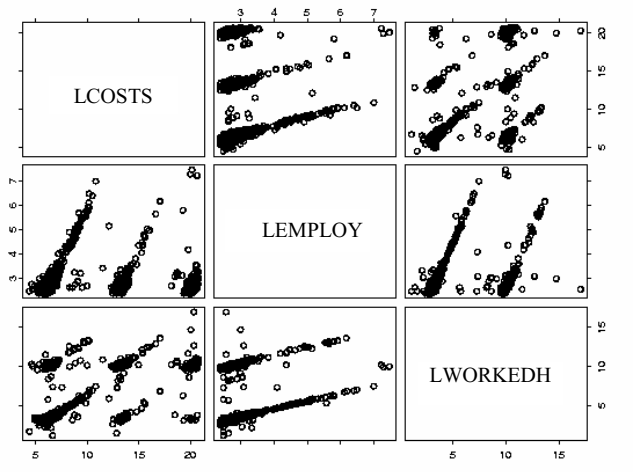


Figure 1. Multiple scatter plot between total labour cost, employees, worked hours (logarithmic scale).

The WSS example will be described in detail in subsection 4.2 where an application of the method proposed in this paper for identifying and treating the unity measure error will be presented.

For the unity measure error, the critical point is the localisation of items in error rather than their treatment. In fact, once an item is classified as erroneous, the optimal treatment is uniquely determined and consists in a deterministic action recovering the original value through an inverse action (*e.g.*, division by 1,000) neutralising the error effect.

The unity measure error is generally tackled through ad hoc procedures using essentially graphical representations of marginal or bivariate distributions, and *ratio edits*. A ratio edit is a rule stating that the value of a ratio between two variables must lie within a predefined interval. The interval bounds are generally determined through a priori knowledge or via exploratory data analysis, possibly using reliable auxiliary information. For this type of error, ratio edits are effective when one of the two variables is error free. Furthermore ratio edits allow taking into account only bivariate relationships between variables and even using interactive graphical inspection (*e.g.*, scatter plot matrix), no more than a pairwise analysis can be performed, disregarding more complex interactions between variables. Finally, we notice that adopting pairwise analyses implies that variables are to be treated in a pre-defined hierarchy, thus increasing the complexity of the error localisation procedure.

With traditional approaches, the error localisation problem is not only complex, but also time and cost consuming. Time and cost are mainly affected by: 1) the complexity of designing and implementing automatic deterministic *ad hoc* procedures, and 2) the resources spent in manually editing observations having low probabilities of being in error and/or low impact on target estimates (*over-editing*).

In this paper we propose a probabilistic formalisation of the problem through finite mixture models (McLachlan and Basford 1988; McLachlan and Peel 2000).

This modelling can provide a principled statistical approach, allowing an estimate of the conditional probability that an observation be affected by unity measure error. The advantage of the proposed approach is that it represents a general method allowing a multivariate data analysis, and providing elements that can be used to optimise the balance between the automatic and interactive components of the editing procedure, *i.e.*, between time and accuracy (Granquist and Kovar 1997).

This work is organised as follows. In section 2 the proposed model is introduced together with the EM algorithm for the estimates of the model parameters. In section 3 diagnostics for selective editing are described. In section 4 the results of the application of the proposed method to both simulated and real data are illustrated. Finally, in section 5 concluding remarks and future research are outlined.

2. The Model

It is hard to give a comprehensive formalisation of random and systematic errors. In this context, we provide a definition that, though not exhaustive, includes many common situations. Let $\boldsymbol{X}^*$ be the vector of the survey target variables, and $(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$ the corresponding mean vector and covariance matrix. Let us suppose that the measurement process is affected by a random error mechanism R having impact on the covariance structure of $\boldsymbol{X}^*$ but leaving the mean vector unchanged, and consequently let $\boldsymbol{X}$ be the corresponding "contaminated" variable, with $E(\boldsymbol{X}) = E(\boldsymbol{X}^*) = \boldsymbol{\mu}$, $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Also, we assume that $\boldsymbol{X}$ can in turn be affected by a systematic error mechanism S acting only on its expected value: $\mu \xrightarrow{S} \varphi(\boldsymbol{\mu})$ for some function φ (*e.g.*, if an additive error mechanism is assumed, $\varphi(\boldsymbol{\mu}) = \boldsymbol{\mu} + constant$). As a consequence of the two error mechanisms, assumed to be independent of one another, observed data can be described by a random vector $\boldsymbol{Y}$ whose distribution, conditional on $\boldsymbol{X}$, depends only on the systematic error mechanism. Our approach to the treatment of systematic errors consists of building up a model for $\boldsymbol{Y}$ focusing only on the detection of systematic errors, thus aiming at recovering the randomly contaminated data represented by the random vector $\boldsymbol{X}$. This is the approach generally adopted in editing procedures, where systematic errors and random errors are dealt with separately and hierarchically.

The previous definition of systematic error includes unity measure error, once data have been transformed in logarithmic scale. In fact, unity measure error generally acts multiplying variables by a constant factor. Hence data in error appear in log-scale as translated by a vector of constants, that depends on which items are in error ("error pattern"), while the covariance structure is the same for each error pattern. Moreover, as matter of fact, in business surveys variables are frequently considered log-normal. Thus in logarithmic scale the Gaussian setting can be adopted.

Following the formalisation so far introduced, our goal becomes to assign each single observation to a specific "error pattern", that corresponds to localise items in error. If we interpret each single error pattern as a "cluster", the error localisation problem is transformed in a cluster analysis problem, and we can exploit experiences from the model-based cluster analysis theory (Fraley and Raftery 2002).

More in detail, let us suppose we have n independent observations $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iq})$, $i = 1, ..., n$, corresponding to the $q-$dimensional vectors $\boldsymbol{X}_i = (X_{i1}, ..., X_{iq})$ with p.d.f. $f(x_1, ..., x_q; \boldsymbol{\theta})$, such that $E(X_1, ..., X_q) = (\mu_1, ..., \mu_q) = \boldsymbol{\mu}$, and $\mathrm{Var}(X_1, ..., X_q) = \boldsymbol{\Sigma}$.

Based on the assumption that systematic errors affect the random vector $\boldsymbol{X}$ only by transforming its expected value $\boldsymbol{\mu}$ into $\varphi_g(\boldsymbol{\mu})$, where $\varphi_g(\cdot): \mathrm{R}^q \to \mathrm{R}^q$, for $g = 1, ..., h$, are a set of known functions, the functions φ_g characterise univocally h distinct clusters (error patterns), differing each other only on the location parameter. For instance, if the systematic error possibly affects all the variables X_s for $s = 1, ..., q$, in the same manner by transforming their expected values μ_s according to $\mu_s \to \mu_s + C$, where C is a known constant, the number of clusters will be $h = 2^q$, *i.e.*, the number of different combinations of error occurrence on the q variables (including the case of no error). In this case, each function φ_g and each corresponding cluster, is associated with one of the 2^q possible sub-sets of variables affected by the error; *e.g.*, the group G characterised by the mean vector $\boldsymbol{\mu}_G = (\mu_1, \mu_2 + C, \mu_3, \mu_4, ..., \mu_q)$, is a cluster of units with error affecting only the variable X_2. We remark that we assume a common covariance matrix because we make the hypothesis that the possible random error acts in the same way on all the data.

For the error localisation purpose we follow a model-based approach based on finite mixture models, where each mixture component G_g, $g = 1, ..., h$, represents a single error pattern. Formally, we assume that $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{iq})$, for $i = 1, ..., n$, are iid w.r.t $\sum_{t=1}^{h} \pi_t f_t(\cdot; \boldsymbol{\theta}_t)$, where $\sum_t \pi_t = 1$ and $\pi_t \geq 0$. The mixing parameters π_t represent the probability that an observation belongs to the t^{th} mixture component.

In order to classify an observation $\boldsymbol{y}_i$ in one of the h groups, we compute the posterior probability $\tau_g(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) = \mathrm{pr}(i^{\text{th}}$ observation $\in G_g \mid \boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi})$, that is

$$\tau_g(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) = \pi_g f_g(\boldsymbol{y}_i; \boldsymbol{\theta}_g) \Big/ \sum_{t=1}^{h} \pi_t f_t(\boldsymbol{y}_i; \boldsymbol{\theta}_t) \qquad g = 1, ..., h. \quad (1)$$

The i^{th} observation is assigned to the cluster G_t, if

$$\tau_t(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi}) > \tau_g(\boldsymbol{y}_i; \boldsymbol{\theta}, \pi) \quad g = 1, ..., h; \; g \neq t.$$

The previous allocation rule is the optimal solution for the classification problem, in the sense that it minimises the overall error rate (Anderson 1984, Chapter 6).

Since, in place of the parameters $(\boldsymbol{\theta}, \pi)$, generally unknown, we use the maximum likelihood estimates $(\hat{\boldsymbol{\theta}}, \hat{\pi})$, the classification rule becomes:

$$\tau_t(\boldsymbol{y}_i; \hat{\boldsymbol{\theta}}, \hat{\pi}) > \tau_g(\boldsymbol{y}_i; \hat{\boldsymbol{\theta}}, \hat{\pi}) \quad g = 1, ..., h; \; g \neq t. \quad (2)$$

We assume that the $f_t(\boldsymbol{y}; \boldsymbol{\theta}_t)$ is a multivariate normal density $MN(\boldsymbol{\mu}_t, \boldsymbol{\Sigma})$ and that each function $\varphi_g(\cdot)$ acts on the mean vector $\boldsymbol{\mu}$ as a translation: $\varphi_g(\boldsymbol{\mu}) = \boldsymbol{\mu} + \boldsymbol{C}_g$, where $\boldsymbol{C}_g$ represents the translation vector for the mean of the g^{th} cluster, and it is supposed to be known. This setting, as already noticed, is suitable for dealing with unity measure error. In order to compute the likelihood estimates, we use the EM algorithm as suggested in McLachlan and Basford (1988). Nevertheless, an additional effort is necessary to adapt the algorithm to our particular situation, where the mean vectors of the mixture components are linked by a known functional relationship. Thus, while in the non-constrained case (McLachlan and Basford 1988) a different mean vector has to be estimated for each mixture component, in our constrained situation only one mean vector needs to be estimated. The resulting modified EM algorithm consists of defining some initial guess for the parameters to be estimated $\hat{\pi}_g^{(0)}$ for $g = 1, ..., h$, $(\hat{\boldsymbol{\mu}}^{(0)}, \hat{\boldsymbol{\Sigma}}^{(0)})$ and applying until convergence the following recursive scheme:

i) compute the posterior probabilities $\tau_{gi}^{(k)} = \tau_g^{(k)}(\boldsymbol{y}_i; \boldsymbol{\theta}, \boldsymbol{\pi})$ under the current estimates $\hat{\pi}^{(k)}$, $\hat{\boldsymbol{\mu}}^{(k)}$, $\hat{\Sigma}^{(k)}$ (k is the index referring to the k^{th} cycle)

$$\hat{\tau}_{gi}^{(k)} = \frac{\hat{\pi}_g^{(k)} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_g^{(k)}\right)' \left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1} \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_g^{(k)}\right)\right\}}{\sum_{t=1}^{h} \hat{\pi}_t^{(k)} \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_t^{(k)}\right)' \left(\hat{\boldsymbol{\Sigma}}^{(k)}\right)^{-1} \left(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_t^{(k)}\right)\right\}}$$

ii) calculate the new estimates by the following recursive equations:

$$\hat{\pi}_g^{(k+1)} = \sum_{i=1}^{n} \hat{\tau}_{gi}^{(k)} / n$$

$$\hat{\boldsymbol{\mu}}^{(k+1)} = \sum_{g=1}^{h} \sum_{i=1}^{n} \hat{\tau}_{gi}^{(k)} \boldsymbol{y}_i / n - \sum_{g=1}^{h} \boldsymbol{C}_g \hat{\pi}_g^{(k+1)}$$

$$\hat{\boldsymbol{\Sigma}}^{(k+1)} = \sum_{g=1}^{h} \sum_{i=1}^{n} \hat{\tau}_{gi}^{(k)} \left(\boldsymbol{y}_i - \boldsymbol{\mu}_g^{(k+1)}\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_g^{(k+1)}\right)' / n\hat{\pi}_g^{(k+1)}.$$

We remark that $\hat{\boldsymbol{\mu}}_{\boldsymbol{g}}^{(k)}$ stands for $\hat{\boldsymbol{\mu}}^{(k)} + \boldsymbol{C}_g$.

In practical applications, it turns out that a crucial role is played by the choice of starting points, as usual in the EM algorithms (see Biernacki, Celeux and Govaert 2003). To overcome this problem, we use an initialisation strategy, following Biernacki *et al.* (2003), consisting of several short runs in terms of number of iterations, of the algorithm from random initialisations followed by a long run of EM from the solution maximising the observed log-likelihood.

It is worth to mention that, due to the location constraints, the parameters to be estimated are sensibly fewer than those in a usual mixture problem. Actually the higher is the number of variables analysed the bigger is this difference; for instance in the case of three variables and 8 clusters we need to estimate 16 parameters instead of 37. This aspect is particularly important when we deal with small samples. Moreover, constraints on cluster locations make easier to identify "rare clusters". In fact, being the relative distances between mean vectors fixed, the estimation problem reduces to estimate the location of the convex polyhedron whose vertices are the cluster centroids. In other words, since the location of one centroid univocally determines the positions of all the others, small cluster parameters are more easily estimated than if they were not constrained.

Since the introduced modelling is based on the assumption that observations are normally distributed, model validation is an issue to take into account. The problem of assessing normality in mixture models is well described in McLachlan and Basford (1988). It is essentially based on the quantities $\hat{a}_{gi}$ described in the following. Let $\boldsymbol{y}_{gi}$ for $i = 1, \ldots, \hat{m}_g$ be the observations assigned to the g^{th} cluster for $g = 1, \ldots, h$, according to the estimated model. Let $\hat{p}_{gi}$ be the value calculated using the estimated parameters, following the formula:

$$\hat{p}_{gi} = \frac{(\nu \hat{m}_g / q) D\left(\boldsymbol{y}_{gi}, \hat{\boldsymbol{\mu}}_g; \hat{\boldsymbol{\Sigma}}\right)}{(\nu + q)(\hat{m}_g - 1) - \hat{m}_g D\left(\boldsymbol{y}_{gi}, \hat{\boldsymbol{\mu}}_g; \hat{\boldsymbol{\Sigma}}\right)}, \qquad (3)$$

where $\mathrm{D}(\cdot, \cdot; M)$ is the Mahalanobis squared distance based on the metric M, and $\nu = n - h - q$. We define $\hat{a}_{gi}$ as the area to the right of the $\hat{p}_{gi}$ value under the $F_{q,\nu}$ distribution (for details see McLachlan and Basford 1988, Chapter 2).

Under the normality assumption, $\hat{a}_{gi}$ for $i = 1, \ldots, \hat{m}_g$ is approximately uniformly distributed on (0,1). Hawkins (1981) suggests using the Anderson–Darling statistic for assessing the uniform distribution of $\hat{a}_{gi}$. The $\hat{a}_{gi}$ are also useful to detect outliers, *i.e.*, atypical observations with respect to the model. In McLachlan and Basford (1988) the lower is $\hat{a}_{gi}$ the higher is the probability of $\boldsymbol{y}_{gi}$ of being atypical, thus all observations with $\hat{a}_{gi} < \alpha$, where α is a specified threshold, can be considered as atypical. Suggested threshold levels range from $\alpha = 0.05$ to $\alpha = 0.005$, depending on which outlying observations (more or less extreme values) are to be selected.

3. Diagnostics for Selective Editing

Once the parameters of the mixture have been estimated, we are able to classify data into the different clusters, *i.e.*, for each observation we can assess whether it is in error or not, and which variables are in error. However, different types of critical observations can be identified after the modelling phase: units classified in a cluster, but having a non-negligible probability of belonging to another cluster, and observations that are outliers with respect to the model.

In order to increase data accuracy it would be useful to make a double check on critical observations (through either a clerical review or, in the most difficult cases, a follow-up). On the other hand, in order to reduce possible over-editing and editing costs, the manual review and/or follow up should be concentrated on the most critical observations. The proposed mixture model directly provides diagnostics that can be used to this aim.

A first type of critical units is represented by possibly misclassified observations. In order to measure the degree of belief in the class assigned to an observation $\boldsymbol{y}_i$ we can consider the corresponding probability resulting from (2). Observations, for which this probability is not very close to one, have a non-negligible probability to belong to another cluster. These observations are those in the region where the mixture components overlap each other.

In addition to the previous type of critical units, there are other observations that are far from all the clusters (all the mixture components), *i.e.*, outliers with respect to the model. Also these observations represent critical situations. In order to identify this kind of outlier we refer to the quantities $\hat{a}_{ij}$ described in the previous section.

Classification probability and atypicality index $\hat{a}_{gi}$ should be used, according to a selective/significance editing approach (Latouche and Berthelot 1992; Lawrence and McKenzie 2000), to build up appropriate score functions to

prioritise critical units. An example of how to use these diagnostics to this aim is given in subsection 4.2.

4. Illustrative Examples

In this section some experiments carried out in order to investigate the peculiarities of the proposed method are presented. Firstly, through a simulation study, we analyse the performance of the proposed model when applied to data that depart from normality. Secondly, through an application on real data, we describe how this approach can be applied in Official Statistics.

All the experiments are performed using the R environment for statistical computing (http://www.r-project.org/).

4.1 Simulated Example: Departure from Normality

In this experiment we describe the results obtained by applying the mixture approach to the three different populations depicted in the first line of Figure 2. The first distribution is a bivariate normal (MN), hence it represents the case when the model is correctly specified. The second one corresponds to a bivariate t distribution (MT), *i.e.*, it mimes the situation when the departure from normality is essentially in having heavier tails. The last one is a bivariate skew–t distribution (ST) (Azzalini and Capitanio 2003, Azzalini, Dal Cappello and Kotz 2003), and it represents a population distributed according to an asymmetric distribution with heavy tails.

From these distributions we build a four components mixture model by adding to each unit one of the four translation vectors $\mathbf{C}_1 = (0, 0)$, $\mathbf{C}_2 = (0, \log(1{,}000))$, $\mathbf{C}_3 = (\log(1{,}000), 0)$, $\mathbf{C}_4 = (\log(1{,}000), \log(1{,}000))$ with probabilities $\pi_1 = 0.5$, $\pi_2 = 0.1$, $\pi_3 = 0.1$, and $\pi_4 = 0.3$ respectively. These parameters represent the mixing proportions of the mixture model and refer respectively to the probabilities of no translation in the variables, translation in only one of the two variables, and translation in both variables. From each mixture, we draw 100 samples of 1,000 observations. In the second line of Figure 2, we report one of these samples (MN–Mixt, MT–Mixt, ST–Mixt), corresponding to the three different populations MN, MT, ST respectively.

For each sample, we compute the number of correct classifications obtained by using the mixture approach described in section 2. The mean number of correct classifications over the 100 samples is reported in Table 1.

As it can be seen in Table 1, the frequency of correct classifications decreases with the departure from normality. However it seems acceptable also in the critical case ST, where the population is characterised by both asymmetry and heavy tails.

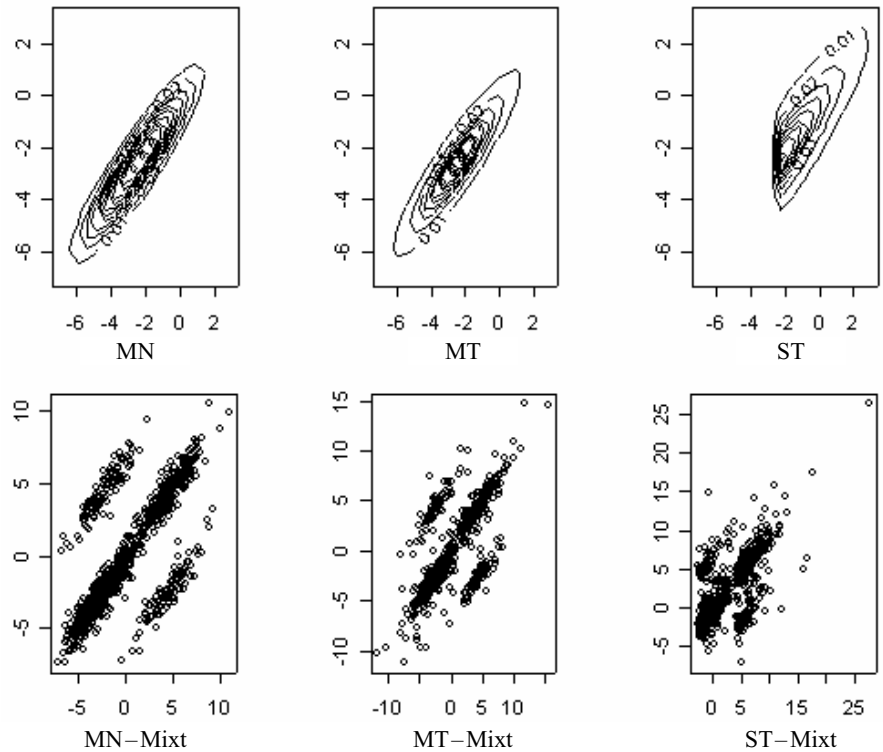


Figure 2. Contour plots of the three bivariate distributions multinormal (MN), t–student (MT), skew–t (ST), and scatter plot of the corresponding mixtures MN–Mixt, MT–Mixt, ST–Mixt.

Table 1
Frequency of Correct Classifications

	MN	MT	ST
% correctly classified	98.5	97.5	95.6

As discussed in section 3, the mixture approach provides elements (such as the degree of atypicality and the classification probability) that can be used in order to prioritise units to be clerically reviewed. Therefore, an overall assessment of the procedure should consider also the results obtained through a selective editing approach based on these model diagnostics.

In order to analyse the characteristics of atypicality index and classification probability, we examine a single sample of 1,000 observations drawn from the three populations so far introduced. In Figure 3, the three samples MN–Mixt(a), MT–Mixt(a), ST–Mixt(a) are represented, furthermore the misclassified units are depicted with a cross in the same graph. The number of misclassified units is 19 for MN–Mixt, 20 for MT–Mixt, and 36 for ST–Mixt.

On this sample, we focus on the impact of different threshold levels both for atypicality (α) and classification probability (β). For each threshold, we report in Table 2 and Table 3 the number of units below that threshold, *i.e.*, the number of critical observations (*N. Atyp, N. Pr. Class*), and among them the number of misclassified units (*Atyp - Misclas, Pr. Class - Misclas*).

As far as atypicality is concerned, we note that when the model is correctly specified, the importance of the atypicality index in recovering misclassified units is negligible, while the classification probabilities are more effective. On the other hand the degree of atypicality is important when the model departs from normality. It is clear that the number of observations selected for a given combination of thresholds α and β is not equal to the sum of the frequencies obtained in Table 2 and Table 3. Thus, in order to evaluate the joint impact of these two indices we choose the two following thresholds $\alpha = 0.005$ and $\beta = 0.975$. We report in Figure 3 (second line) the units selected only for the atypicality value (squares), only for the classification probability (triangles), and for both of them (crosses). From these figures we see how the impact of atypicality is mainly on outliers identification while the classification probability works on the overlapping regions. In Table 4 the number of selected units and, out of them the number of misclassified units are shown.

We note that for population MN–Mixt, apart one observation, all the misclassified units are selected. For MT–Mixt, we are able to select 14 out of the 20 misclassified units, and in the most critical sample ST–Mixt we select 24 out of the 36 misclassified units.

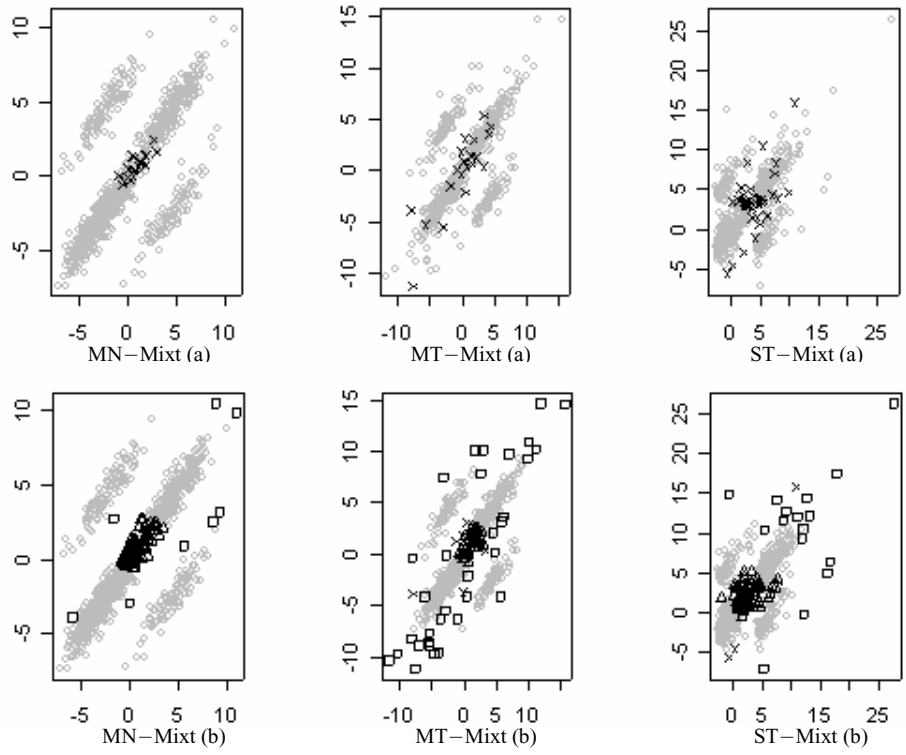


Figure 3. Misclassified units (crosses) in MN–Mixt(a), MT–Mixt(a), ST–Mixt(a). Critical units for atypicality (square), for classification probability (triangle), and for both of them (cross), in MN–Mixt(b), MT–Mixt(b), ST–Mixt(b).

Table 2
Number of Critical Observations and Misclassified Units with Respect to Three Different Thresholds for Atypicality

	MN–Mixt		MT–Mixt		ST–Mixt	
α	*N. Atyp*	*Atyp – Misclas*	*N. Atyp*	*Atyp – Misclas*	*N. Atyp*	*Atyp – Misclas*
0.05	50	1	84	9	68	14
0.01	15	0	50	7	33	8
0.005	8	0	39	7	20	5
0.001	4	0	25	4	14	2

Table 3
Number of Critical Observations and Misclassified Units with Respect to Three Different Thresholds for Classification Probability

	MN–Mixt		MT–Mixt		ST–Mixt	
β	*N. Pr. Class*	*Pr. Class – Misclas*	*N. Pr. Class*	*Pr. Class – Misclas*	*N. Pr. Class*	*Pr. Class – Misclas*
0.99	119	19	63	12	182	26
0.975	76	18	46	11	82	26
0.95	55	14	35	9	66	21

Table 4
Number of Critical Observations and Misclassified Units with Respect to Atypicality and Classification Probability

	MN–Mixt		MT–Mixt		ST–Mixt	
Thresholds	*N.Crit. Units*	*N. Misclas*	*N.Crit. Units*	*N. Misclas*	*N.Crit. Units*	*N. Misclas*
$\alpha = 0.005,\ \beta = 0.975$	84	18	79	14	98	24

4.2 An Application to Real Data: The 1999 Italian Water Survey System

In this section we describe an application of the mixture model approach to real survey data. The data are taken from the 1999 Italian *Water Survey System* (WSS). The WSS is a census that collects information on water abstraction, supply and usage for the 8,100 Italian municipalities. We restrict our analysis to the municipalities belonging to one of the data domains defined by altimetry (2,041 observations) and to the main variables *Total Invoiced Water* (*TI*) and *Total Supplied Water* (*TS*). Both these variables refer to water volumes and the respondents are requested to provide them in thousands of cubic meters. The scatter plot on log-scale of per capita water invoiced (*WI*) versus per capita water supplied (*WS*) (Figure 4) shows the presence of four clusters corresponding to unity measure error in one, both, or none of the target variables. This is probably due to the misunderstanding of some respondents that expressed water volumes in litres or in cubic meters rather than thousands of cubic meters, as requested. As expected, the two most populated clusters are those corresponding to non-erroneous units and to units where both variables are in error. Nevertheless, we can note the presence of two rare clusters corresponding to observations where the unity measure error affects only *TI* or only *TS* respectively.

In Table 5 a label is assigned to each group associated with a specific error pattern. For the sake of simplicity we introduce two flags E_{TS} and E_{TI} assuming value 1 or 0, depending on whether the corresponding variables are affected by the unity measure error or not, respectively.

In order to identify and correct the unity measure error we apply the procedure described in sections 2 and 3. We classify each observation according to a specific error pattern, *i.e.*, we assign each unit to one of the clusters G_t, for $t = 1, \ldots, 4$. The results are reported in Table 6.

For each unit the atypicality index is also calculated and the threshold $\alpha = 0.005$ is chosen in order to flag atypical units. According to this threshold, 71 observations are selected as atypical, marked by "crosses" in Figure 7. Once the values $\hat{a}_{gi}$ are computed according to Formula (3), a test assessing the normality assumption can be performed. Actually, following McLachlan and Basford (1988, Chapter 2), the Anderson–Darling test on the uniformity of $\hat{a}_{gi}$ on each single estimated cluster is performed. The p–values are below 0.001 for the two largest clusters. Since the test is based on asymptotical approximations, we do not take into account the results on the other two rare populations. In Figure 5 we report the empirical sample quantiles versus the normal quantiles of the variables log(WI) and log(WS), focusing only on the subset of data classified as non-erroneous. We notice that departure from normality is mainly due to heavy tails. Based on the results obtained in section 4.1, where the method performed satisfactorily also in non-gaussian setting, we are confident about the good performance of the mixture approach on the survey data. This expected behaviour is confirmed by the application results showed in the following.

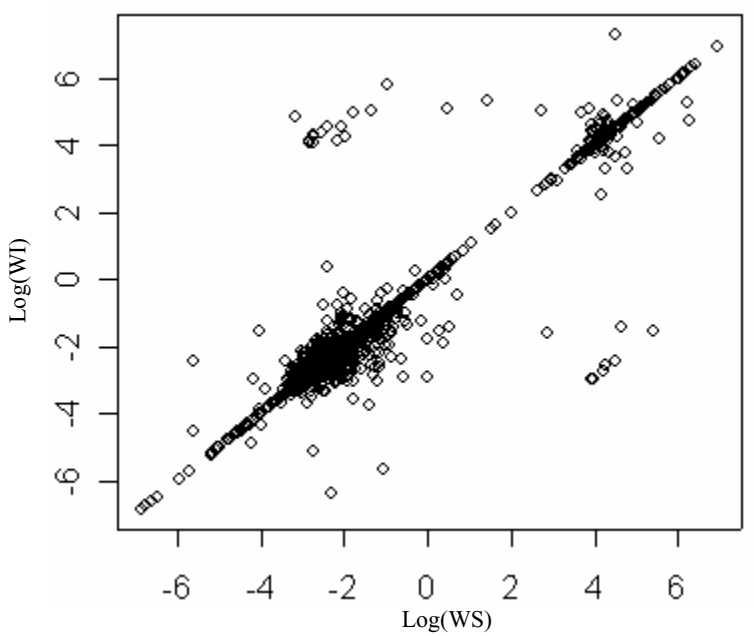


Figure 4. Scatter plot of log(WS) and log(WI).

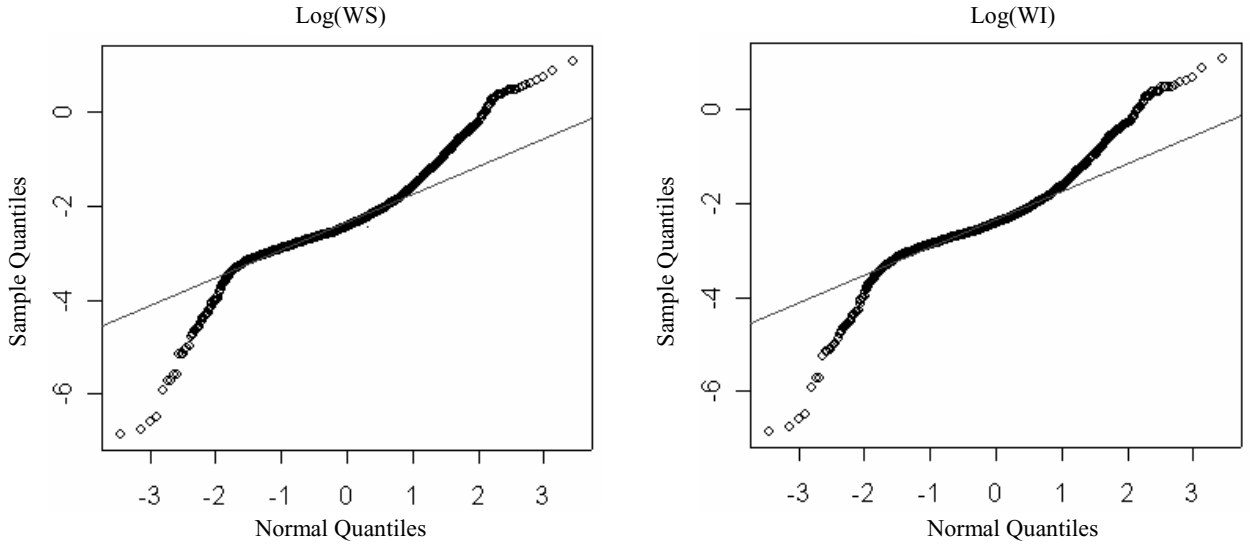


Figure 5. Normal qq−plot of log(WS) and log(WI).

Table 5
Error Patterns and Error Labels

Error pattern	$E_{TS} = 0$ $E_{TI} = 0$	$E_{TS} = 0$ $E_{TI} = 1$	$E_{TS} = 1$ $E_{TI} = 0$	$E_{TS} = 1$ $E_{TI} = 1$
Cluster label	*G1*	*G2*	*G3*	*G4*

Table 6
Number of Units Assigned to Each Cluster

Cluster label	**G1**	**G2**	**G3**	**G4**
N. of units	1,800	16	10	215
%	88.2	0.8	0.5	10.5

In the remaining part of this section, it is shown how the posterior probabilities can be used to prioritise units to be reviewed which are likely to provide the greatest editing benefit, taking into account the potential impact of the clerical editing on the estimates. To this aim, note that a wrong classification of an observation causes that the final values of at least one variable differ from the corresponding true values by a multiplicative factor. These discrepancies can seriously affect the accuracy of the estimates leading to a strong bias. In order to select the potentially erroneous units that most likely have a strong impact on the target estimates, we follow the *selective editing approach*. Let X_1, X_2 denote the variables *TS*, *TI* respectively. For each unit u_i, $i = 1, ..., n$, and for each variable X_j, $j = 1, 2$, let us define:

X_{ij} : data free of systematic error;

Y_{ij} : observed data;

$\widetilde{X}_{ij}$: data after the treatment of systematic error based on the classification through mixture model (*i.e.*, $\widetilde{X}_{ij} = Y_{ij}$ or $\widetilde{X}_{ij} = Y_{ij}/1{,}000$ depending on the cluster the unit u_i is assigned to).

Let us suppose that the target estimates refer to population totals $T(X_j) = \sum_i X_{ij}$. Further, denote by $E_\xi(\cdot)$ the expectation over the distribution of the random variable X_j conditional on the observed data Y_{ij} and the data after correction $\widetilde{X}_{ij}$. Then, from the inequality $|\sum_i E_\xi(X_{ij} - \widetilde{X}_{ij})| \leq \sum_i E_\xi |X_{ij} - \widetilde{X}_{ij}|$ it follows that the quantity on the right hand side can be viewed as an upper bound for the expected bias of the total estimate for the variable X_j based on the corrected values $\widetilde{X}_{ij}$. The last consideration suggests a method for selecting the most "influential" units with respect to the estimate $T(X_j)$: in order to guarantee the requested level of accuracy and to minimise costs due to manual check, we define a local score function $S_{ij} = (E_\xi |X_{ij} - \widetilde{X}_{ij}|)/\hat{T}(X_j)$, where $\hat{T}(X_j)$ is a reference estimate for $T(X_j)$, for instance the estimate from a previous survey, or a robust estimate. In our case, in order to robustify the preliminary estimate we first exclude from the data the atypical observations, then compute the mean value on this subset, and then multiply it by the total number of units.

The local score S_{ij} measures the impact of the potential unity measure error associated to the unit u_i on the target estimate $T(X_j)$. Then, units can be sorted by their score S_{ij} and, starting from the highest values, the first units can be selected until the sum of the remaining S_{ij} values is lower than a predefined threshold.

If both the variables TS and TI are considered simultaneously, a global score S_i, for $i = 1, \ldots, n$, can be obtained by suitably combining the local score functions S_{ij}, $j = 1, 2$. Possible choices are $S_i = (S_{i1} + S_{i2})/2$, or $S_i = \max_{j=1,2} S_{ij}$. The latter function, for instance, ensures that the impact of the potential unity measure error associated with u_i on each estimate is not greater than S_i.

In order to compute the scores S_{ij} the conditional expected value $E_\xi |X_{ij} - \widetilde{X}_{ij}|$ is to be estimated for each unit u_i, $i = 1, \ldots, n$, and for each variable X_j for $j = 1, 2$. This can be easily done through the posterior probabilities. For instance, suppose that the unit u_i has been assigned to the cluster G_2. This means that, for this unit, the observed value of TS (Y_{i1}) has been considered correct, while the observed value of TI (Y_{i2}) has been flagged as affected by unity measure error (*i.e.*, multiplied by 1,000). The correction consists of dividing by 1,000 the observed value of TI, *i.e.* $(\widetilde{X}_{i1} = Y_{i1},\ \widetilde{X}_{i2} = Y_{i2}/1{,}000)$. The conditional expected value $E_\xi |X_{ij} - \widetilde{X}_{ij}|$ can be computed as follows:

$$
\begin{aligned}
E_\xi\left|X_{i1} - \widetilde{X}_{i1}\right| &= \left|Y_{i1} - Y_{i1}\right| \Pr(u_i \in G_1 \cup G_2) \\
&\quad + \left|\frac{Y_{i1}}{1{,}000} - Y_{i1}\right| \Pr(u_i \in G_3 \cup G_4) \\
&= \frac{999}{1{,}000} Y_{i1}(\hat{\tau}_{3i} + \hat{\tau}_{4i})
\end{aligned}
$$

$$
\begin{aligned}
E_\xi\left|X_{i2} - \widetilde{X}_{i2}\right| &= \left|\frac{Y_{i2}}{1{,}000} - \frac{Y_{i2}}{1{,}000}\right| \Pr(u_i \in G_2 \cup G_4) \\
&\quad + \left|Y_{i2} - \frac{Y_{i2}}{1{,}000}\right| \Pr(u_i \in G_1 \cup G_3) \\
&= \frac{999}{1{,}000} Y_{i2}(\hat{\tau}_{1i} + \hat{\tau}_{3i}),
\end{aligned}
$$

where $\hat{\tau}_{gi}$ is the estimated probability that unit u_i belongs to cluster G_g. In a similar manner the score functions can be calculated for all the units.

In practice, in our application we sort the units by their global score S_i, $\max_{j=1,2} S_{ij}$ (ascending order). Then we exclude from clerical review all the first observations such that their cumulative sum of S_i is below δ, where δ is a specified tolerance level for the impact on the estimates due to errors remaining in data. In Figure 6 the behaviour of the cumulative sum of S_i, $S_{(i)} = \sum_{k \leq i} S_k$, is shown for the first most critical 10 observations. We remark that for the sake of clarity we have not reported all the observations because for most of them $S_{(i)}$ is close to zero causing an unreadable picture for their different magnitude. Note that a residual relative error less than $\delta = 0.001$ is expected by selecting only the first two units (drawn with crosses).

In Figure 7 all the units selected because of their atypicality (71) and/or the relative impact on estimates of their potential errors (2) are shown: crosses correspond to observations that are critical for atypicality, squares indicate the other two types of critical units.

A comparison with the results obtained by the official procedure is made. Out of the 1,968 units not selected for clerical review, 1,911 observations are error free or affected by unity measure error only. For all of them the classification of the mixture model is correct. Out of the remaining 57 units characterised by other error typologies, 45 are classified as non-affected by the unity measure error, while 12 as units with the 1,000 error in both the variables. This last misclassification can be explained by the presence of another systematic error (times 100, 10,000 factors) that is not taken into account in the model used for this example.

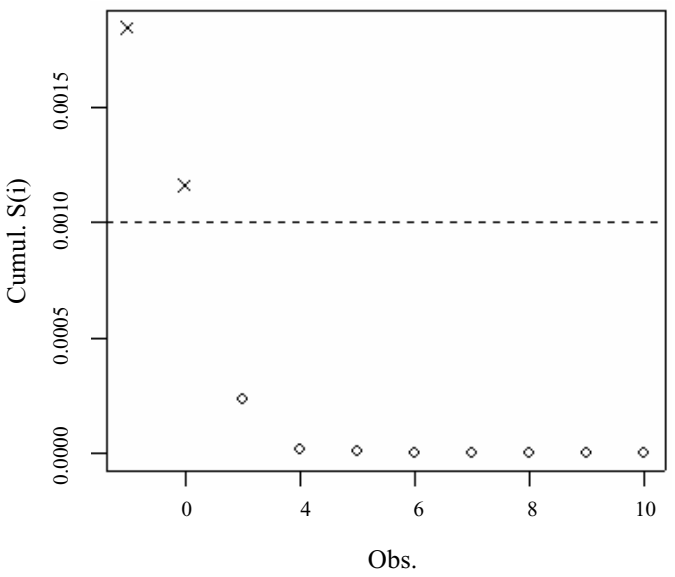


Figure 6. Plot of the cumulative score $S_{(i)}$ for the first most critical 10 observations.

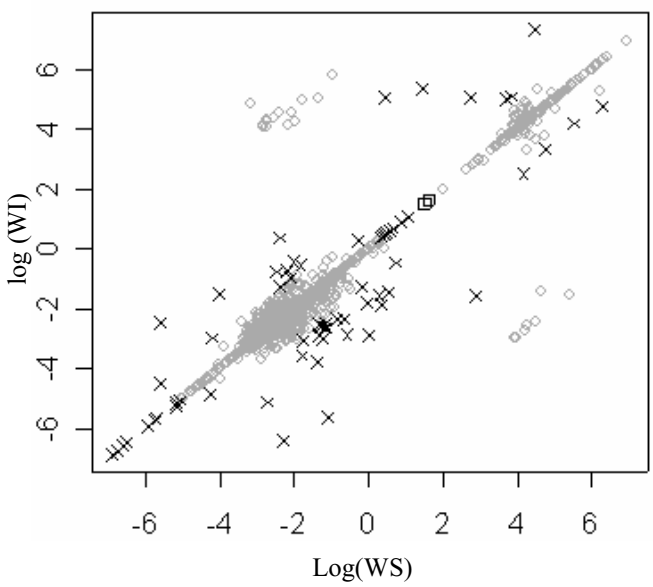


Figure 7. Scatter plot of log(WS) *vs* log(WI). Crosses indicate critical units for atypicality, squares mark critical units for the impact of their potential error.

A further comparison is about the estimate of the totals. Under the hypothesis that the values selected for a clerical review are correctly restored, the relative differences between the "true" total values according to the official procedure $T(X_j)$ and the model estimate $\hat{T}(X_j)$ as $B(X_j) = (|\hat{T}(X_j) - T(X_j)|)/T(X_j)$, for $j = 1, 2$ are $B(X_1) = 0.005$ and $B(X_2) = 0.002$. These values are not directly comparable with the tolerance level $\delta = 0.001$, in fact this threshold relates only to impact of the remaining unity measure errors, while $B(X_j)$ is also affected by other kind of errors. Thus, for a more direct comparison, we replace for these units the wrong values with the "true" ones obtaining $B(X_1) = B(X_2) = 0$. This particularly high performance of the model is justified by the low degree of overlapping of the clusters as clear in Figure 7.

5. Final Remarks and Further Research

In this paper we propose a finite mixture model to deal with a particular type of systematic error that frequently affects numerical continuous survey data: the unity measure error times a constant factor. The proposed approach has the advantages, with respect to the traditional ones, to formally state the problem in a multivariate context, to be easily implemented in generalised software, and to naturally provide useful diagnostics for prioritising doubtful units possibly containing influential errors. The latter characteristic is particularly important when the situation is critical, *i.e.*, when different error patterns overlap each other or in other words when unity measure errors are among plausible observations. In these circumstances a clerical review is needed. Hence, it is important to optimise the selection of critical observations in order to save time and costs. All these advantages are the natural consequence of the introduction of a model-based technique. On the other hand, it is clear that the use of a model-based approach implies problems related to model assumptions. However, based on the experiments illustrated in the paper, it seems that also in cases of departure from the normality assumption, the proposed technique performs satisfactorily. Nevertheless, it is worth to mention that for extreme departure from normality, *e.g.*, when the distribution is not unimodal, the method is expected to fail. This can happen in real situations when true data contain different clusters, for instance differences in men and women income might cause a bimodal distribution for the income itself. In some cases the problem could be overcome by stratifying data with respect to some explicative variables, *e.g.*, sex in the previous example. An alternative approach to this specific problem could be based on modelling each cluster in turn as a Gaussian mixture, thus obtaining a "mixture of mixture models" (McLachlan and Peel 2000; Di Zio, Guarnera and Rocci 2004).

Finally, a last concern is about the number of variables that can be treated simultaneously. Actually, the number of clusters and then the number of mixing parameters π_t can have an exponential growth with respect to the number of variables, making the parameter estimation a critical task. However it is worthwhile noting that the number of parameters related to the mean vector and covariance matrix increases much slower, due to the constraints characterising our model.

Acknowledgements

We are grateful to the referees and the Associate Editor for their helpful comments.

References

Anderson, T.W. (1984). *An introduction to Multivariate Statistical Analysis*. Second Edition. New York: John Wiley & Sons, Inc.

Azzalini, A., and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew–*t* distribution. *Journal of the Royal Statistical Society* (*B*), 65, 367-389.

Azzalini, A., Dal Cappello, T. and Kotz, S. (2003). Log-skew-normal and log-skew–*t* distributions as models for family income data. *Journal of Income Distribution*, 11, 13-21.

Biernacki, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41, 561-575.

Cirianni, A., Di Zio M., Luzi O. and Seeber, A.C. (2000). The new integrated data editing procedure for the Italian Labour Cost survey: Measuring the effects on data of combined techniques. *Proceedings of the International Conference on Establishment Surveys II*, Buffalo, 7-21.

De Waal, T. (2003). Solving the error localization problem by means of vertex generation. *Survey Methodology*, 29, 1, 71-79.

Di Zio, M., Guarnera, U. and Rocci, R. (2004). A mixture of mixture models to detect unity measure error. *Proceedings in Computational Statistics*, (Ed. Antoch Jaromir), 919-927, Physica Verlag, Prague, August 23-28.

Di Zio, M., and Luzi, O. (2002). Combining methodologies in a data editing procedure: an experiment on the survey of Balance Sheets of Agricultural Firms. *Italian Journal of Applied Statistics*, 14, 1, 59-80.

Encyclopedia of Statistical Sciences (1999). New York: John Wiley & Sons, Inc. Update 3, 621-629.

Euredit (2003). *Towards Effective Statistical Editing and Imputation Strategies – Findings of the Euredit project*, 1, 2. Forthcoming. Now available at http://www.cs.york.ac.uk/euredit/

Federal Committee on Statistical Methodology (1990). *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18.

Fellegi, I.P., and Holt, D. (1976). A systematic approach to edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Fraley, C., and Raftery, A. (2002). Model-Based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97, 611-631.

Granquist, L. (1995). Improving the traditional editing process. In *Business Survey Methods*, (Eds. B.G. Cox and D.A. Binder).

Granquist, L. (1996). The new view on editing. *International Statistical Review*, 65, 3, 381-387.

Granquist, L., and Kovar, J. (1997). Editing of survey data: How much is enough? In *Survey Measurement and Process Quality*, (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 415-435.

Hawkins, D.M. (1981). A new test for multivariate normality and homoscedasticity. *Technometrics*, 23, 105-110.

Kovar, J.G., Mac Millian, I.H. and Whitridge, P. (1988). Overview and strategy for the generalized edit and imputation system, (updated February 1991). Statistics Canada, Methodology Branch Working Paper, BSMD-88-007E/F.

Latouche, M., and Berthelot, J.M. (1992). Use of a score function to prioritise and limit recontacts in business surveys. *Journal of Official Statistics*, 8, 389-400.

Lawrence, D., and McKenzie, R. (2000). The general application of significance editing. *Journal of Official Statistics*, 16, 243-253.

McLachlan, G.J., and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker.

McLachlan G.J., and Peel D. (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.