WP.
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchatel, Switzerland, 5-7 October 2009)

Topic (vi): New and emerging methods

A SELECTIVE EDITING APPROACH BASED ON CONTAMINATION MODELS: AN APPLICATION TO AN ISTAT BUSINESS SURVEY

Invited paper

Submitted by ISTAT[1]

I. INTRODUCTION

One of the most important phases of an editing and imputation (E&I in the following) strategy for business surveys is the identification of outliers, i.e. observations which deviate from a given data model, and which may have a strong impact on the parameters of interest (see, among others, Barnett and Lewis, 1994; Lee, 1995). Outliers may be caused by errors in the survey process, such as data entry errors or misinterpretation of some variables (in this case they are also referred to as *non representative outliers*, see Chambers, 1986): these observations are to be edited and corrected at the E&I stage, before computing the survey estimates. On the other hand, outliers may correspond to observations which deviate from the model assumed for the rest of the data, in this case they are referred to as *representative outliers* and are to be treated at the estimation stage. In the case of numerical variables, the concept of outlier naturally mixes up with that of influential observation, defined as an observation that has a large impact on a particular result of a survey, such as the estimate of some parameter. The generally adopted approach used for identifying influential observations is *selective editing* (Latouche and Berthelot, 1992; Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994). In this class of methods, potentially influential observations are ranked based on the values of a *score function* expressing the potential impact of a potential error on the estimate of the parameter of interest. Limiting the recontact/manual revision to observations having score function values above a given threshold, reduces time and resources spent while taking under control the potential (residual) error on the parameter's estimate. The definition of the *score function* implies the determination of at least two elements: the probability of the observation to be in error, and the magnitude of the error (see for example Jäder and Norberg, 2005; Hedlin, 2003).
One of the most commonly used methods to determine these elements is based on the comparison of the observed value of a certain variable for a given unit with the value of the variable for the same unit predicted by a specific model. The differences between observed and predicted values are then used to calculate appropriate score functions which allow to identify the observations having the largest impact

[1]Prepared by Diego Bellisai (bellisai@istat.it), Marco Di Zio (dizio@istat.it), Ugo Guarnera (guarnera@istat.it), Orietta Luzi (luzi@istat.it).

on the target estimates. Di Zio et al. (2008) have proposed an alternative method for estimating both the probabilities and the magnitude of the errors for continuous variables in a multivariate context. The method is based on the use of contamination normal models (see Little, 1988), where it is assumed that the distribution of the erroneous data can be obtained from the distribution of the error free data by inflating the variance. Under this modeling, scores are estimated as expected values of errors, where errors are the differences between the observed values and the corresponding "true" unobserved ones, estimated by using the model.
Following this idea, in this paper we propose a selective editing approach based on contamination normal models in a regression context. The approach is used to identify influential outliers and influential observations in the Istat *Quarterly Survey on Job Vacancies and Hours Worked*, and in particular on the effectively *worked hours per capita*. We assume a contaminated normal distribution for the hours worked conditioned on the number of occupied posts, where variances of acceptable and erroneous observations are different and to be estimated. Furthermore, we assume that good data and bad data have same mean. The resulting model is a mixture of two linear regressions having the same intercept and slope and different residual variances, where the two components represent good (smaller variance) and bad (larger variance) data, respectively. Again, it is possible to define the selective editing scores as expectations, with respect to the estimated conditional distribution, of the difference between the observed values and the corresponding predicted "true" ones.
The performance of the proposed selective editing approach is assessed in terms of the number of observations selected for the interactive editing revisions by survey experts.
The paper is structured as follows. In section II, the *Quarterly Survey on Job Vacancies and Hours Worked* is shortly described. Sections III and IV describe the adopted contamination model and the selective editing approach respectively. In section V the proposed method is applied on the ISTAT *Quarterly Survey on Job Vacancies and Hours Worked*. Section VI contains final considerations and ideas for future work.

II. THE SURVEY

The Italian National Statistical Office (Istat) has been carrying out the *Quarterly Survey on Job Vacancies and Hours Worked* (in the following, VELA) since the third quarter of 2003. The main variables collected by VELA are: occupied posts on the last day of the previous and the reference quarter; hirings and separations in the reference quarter; job vacancies on the last day of the reference quarter; hours worked as normal time and overtime and hours not worked but paid by the employer in the quarter. All variables are collected separately for manual and non manual workers (managers are excluded). The target population is composed by all enterprises with at least 10 employees classified, until the end of 2007, in NACE Rev.1.1 sections C-K and, from the beginning of 2008, in NACE Rev.2 sections B-N (http://ec.europa.eu/eurostat/ramon/). The sample includes around 13,000 enterprises and is drawn on the basis of a stratified random scheme with economic activity, size and geographical area as stratification variables. All population enterprises with at least 500 employees belong to take-all strata. Moreover, in order to reduce the response burden for smaller enterprises, around one third of the sample enterprises with 10-499 employees are rotated once a year. The E&I methods are based on an integration of the VELA microdata with those of other two Istat surveys: the Monthly Survey on Employment, Hours Worked, Wages and Labour Cost in the largest enterprises (form now on, LE survey) and OROS (a quarterly survey on Employment, Wages and Social Contributions based on administrative data). The LE survey collects monthly from a panel of about 1,100 enterprises (which had at least an average of 500 employees in the base year and which are classified, until 2008, in NACE Rev.1.1 sections C-K, and, from 2009, in NACE Rev.2 sections B-N) the number of occupied posts at the end of the month, hirings and separations, hours worked (with the same definitions as VELA), as well as variables related to wages and labour cost. On the other hand, the OROS survey relies on the

whole population of the DM10 forms which enterprises must monthly fill in and send to the Italian Social Security Institute (INPS), in order to declare the compulsory social contributions. By integrating the information included in these forms with that collected by the LE survey for the panel enterprises, OROS produces quarterly indicators on wages, labour cost and occupied posts for, until 2008, NACE Rev.1.1 sections C-K, and, from 2009, NACE Rev.2 sections B-N classification. The occupied posts variable measured by OROS is the quarterly average of the monthly number of employees. The aggregate data on job vacancy rates are transmitted to Eurostat and, since January 2009, they are also published in Italy. As far as the hours worked variables are concerned, the publication of the aggregate data on the per capita variables is not yet planned since the editing and imputation methods are still under construction. In this paper will concentrate on the analysis of a model for the outlier identification in the hours worked per capita as normal time, having assumed that the denominator of such variable, i.e. average occupied posts, is not affected by any error. This can be seen as a consequence of the fact that editing and imputation on occupied posts has already been performed, in order to produce estimates for the job vacancy rate.

III. THE MODEL

Let us now describe the model which we will apply in our paper. The idea is to consider the observed data as coming from a mixture of two distributions corresponding to "true" (i.e. correct) data and contaminated data respectively. This is generally referred to as "contamination model" (see Little, 1988) and often used for automatic outlier identification. In the present approach, we treat the case of two variables (W, Z), where only Z is assumed to be affected by measurement error. We denote by Z^* and Z the random variables corresponding to true and contaminated values respectively of the response variable and we define the corresponding variables in log-scale by Y^* and Y. No distinction is made for the explanatory variable W which is supposed error free. In particular, we model the conditional distribution of $Y = log(Z)$ given $X = log(W)$ through standard linear regression and assume that the measurement error is generated by a random "intermittent" mechanism which determines the inflation of the residual variance in the regression model. Specifically, we assume that, conditional on $\{X = x\}$, Y^* is normally distributed with mean $\alpha + \beta x$ and variance σ^2, where α, β and σ^2 are parameters to be estimated. If we make the additional assumption that the error is normally distributed, the additive error mechanism can be described by: $Y^* \rightarrow Y^* + \epsilon$ with $\epsilon \sim N(0, \sigma_\epsilon^2)$. Due to the intermittent nature of the error, the probability density function of Y, conditional on Y^* is given by the mixture density:

$$f_{Y|Y^*}(y|y^*) = (1-p)\delta_{y^*} + pN(y; y^*, \sigma_\epsilon^2) \tag{1}$$

where p (*mixing weight*) represents the "a priori" probability of contamination and δ_t is the delta-function with mass at t . From the previous assumption, it follows that, conditional on x, the observed values of Y are distributed according to:

$$f_{Y|X}(y|x) = (1-p)N(y; \alpha + \beta x, \sigma^2) + pN(y; \alpha + \beta x, \sigma_c^2) \tag{2}$$

where $\sigma_c^2 = \sigma^2 + \sigma_\epsilon^2$ is the variance of the contaminated data. It is straightforward to notice that expression (2) represents a mixture of two regression models having same intercept and same slope but different residual variances.

In the following we describe the estimation procedure of the model parameters based on a sample of n bivariate observations (x_i, y_i) $i = 1, \ldots, n$

The parameters to be estimated are the mixing weight p, and the regression parameters $(\alpha, \beta, \sigma^2, \sigma_c^2)$; we compute the maximum likelihood estimates (MLEs) using an Expectation Conditional Maximization (ECM) algorithm (see Meng and Rubin, 1993) consisting in the iterative application of an E-step

and a CM-step described in the following. The E-step simply consists in updating the posterior probabilities $\tau_i = \tau(y_i; x_i),\ (i = 1, \ldots, n)$, where $\tau(y; x)$ is defined as:

$$\tau(y; x) = \frac{pN(y\,; \alpha + \beta x, \sigma_c^2)}{(1-p)N(y\,; \alpha + \beta x, \sigma^2) + pN(y\,; \alpha + \beta x, \sigma_c^2)}. \tag{3}$$

In order to simplify the notation we define:

$$\tau_{i1} = \tau_i,\ \ \tau_{i2} = 1 - \tau_i,\ \ \sigma_1^2 = \sigma^2,\ \sigma_2^2 = \sigma_c^2 \quad ,$$
$$w_{ig} = \tau_{ig}/\sigma_g^2 \quad (g = 1, 2) \quad ,$$

and

$$\tilde{\tilde{y}} = (\sum_{ls} w_{ls})^{-1} \sum_{ig} w_{ig} y_i, \quad \tilde{\tilde{x}} = (\sum_{ls} w_{ls})^{-1} \sum_{ig} w_{ig} x_i \quad .$$

The CM-step consists in the following updating rules:

(M1) *update the mixing weight* (p)

$$p = \frac{1}{n} \sum_{i=1}^{n} \tau_i \quad , \tag{4}$$

(M2) *update intercept and slope* (α, β)

$$\beta = \frac{\sum_{i=1}^{n} \sum_{g=1}^{2} w_{ig} (y_i - \tilde{\tilde{y}})(x_i - \tilde{\tilde{x}})}{\sum_{l=1}^{n} \sum_{s=1}^{2} w_{ls} (x_l - \tilde{\tilde{x}})^2} \quad , \qquad \alpha = \tilde{\tilde{y}} - \beta \tilde{\tilde{x}} \quad ,$$

(M3) *update residual variances* (σ^2, σ_c^2)

$$\sigma_g^2 = (\sum_{l=1}^{n} \tau_{lg})^{-1} \sum_{i=1}^{n} \tau_{ig} (y_i - \alpha - \beta x_i)^2 \quad (g = 1, 2).$$

Note that, in (M1)-(M3), maximization with respect to model parameters is not simultaneous but conditional. This makes the algorithm convergence slower than it would be in a genuine EM algorithm. In order to initialize the algorithm we use as starting points for α, β, σ^2 the estimates of the

corresponding parameters obtained through ordinary linear regression on all data. A random value for p in the interval $[0.6, 1]$ is chosen, and σ_c^2 is initialized as $\lambda\sigma^2$ with λ in $[5, 100]$.

IV. THE SELECTIVE EDITING APPROACH

In this section we illustrate how to use the contamination model in the context of selective editing. Starting from the error model (1) and the assumed (conditional) distribution of the error-free data Y^*, for each observation (x_i, y_i), it can be easily derived, via Bayes formula, the following conditional distribution of y_i^* given y_i:

$$f_{Y^*|Y}(y_i^*|y_i) = [1 - \tau(y_i; x_i)]\delta_{y_i} + \tau(y_i; x_i)N(y_i^*; \tilde{\mu}_i, \tilde{\sigma}^2) \tag{5}$$

where

$$\tilde{\sigma}^2 = (\sigma^{-2} + \sigma_\epsilon^{-2})^{-1} \quad \text{and} \quad \tilde{\mu}_i = \tilde{\sigma}^2 \left(y_i/\sigma_\epsilon^2 + (\alpha + \beta x_i)/\sigma^2\right)$$

From (5) it is possible to derive the corresponding distribution $f_{Z^*|Z}$ for the data in the original scale. In fact, as it can easily verified, for each observation (w_i, z_i), the conditional distribution of z_i^* given z_i is a mixture of a delta function with mass in z_i and a log-normal distribution $LN(z_i; \tilde{\mu}_i, \tilde{\sigma}^2)$ with parameters $\tilde{\mu}_i, \tilde{\sigma}^2$. The mixing weights are the same as in the mixture (5):

$$f_{Z^*|Z}(z_i^*|z_i) = [1 - \tau(log(z_i); x_i)]\, \delta_{log(z_i)} + \tau(log(z_i); x_i) LN(z_i^*; \tilde{\mu}_i, \tilde{\sigma}^2) \tag{6}$$

The explicit expression (5) for the distribution of the correct data conditional on the observed data allows one to estimate the *expected error* $E\left((z_i - z_i^*)|z_i\right)$ by substituting the parameters $p, \alpha, \beta, \sigma^2, \sigma_\epsilon^2 = \sigma^2 - \sigma_c^2$ with the corresponding ECM estimates. From (6) it follows that

$$E\left(z_i - z_i^*|z_i\right) = \tau_i \left[z_i - exp\left(\frac{1}{2}\tilde{\sigma}^2 + \tilde{\mu}_i\right)\right].$$

Based on these estimates, both robust estimation of finite population quantities and selective editing can be done. In the following, we assume that the target estimate is given by the total T_z of the variable Z, *i.e.* $T_z = \sum_i z_i$. Thus, if we have some estimator of T_z based on the sample $z_1, \ldots, z_n$, we can compute a robust version of it, T_z^*, by substituting the observed values z_i with the "predictions" $E(z_i^*|z_i)$. Furthermore, the absolute value of the estimated expected errors can be seen as score functions for selective editing. Interpreting the score function as expected error is particularly useful in that it makes it possible to estimate the residual error remaining in the data after the interactive editing of the units with the highest expected error. Thus, a number of units to be interactively reviewed can be chosen such that the residual error is below a prefixed threshold. It is natural to define the threshold in term of ratio between expected residual error and a (robust) reference estimate. We give details below. Let S_i be the relative error $E(z_i - z_i^*|z_i)/T_z^*$, and R_M the absolute value of the expected residual percentage error remaining in data after removing errors in the units belonging to the set M : $R_M = \left|\sum_{i \in \bar{M}} S_i\right|$ where $\bar{M}$ denotes the complement of M in $(1, \ldots, n)$. Let η be a chosen "accuracy" threshold. We propose a selective editing procedure consisting in:

(1) sorting the observations in descending order according to the value of $|S_i|$;
(2) selecting the first $\bar{k}$ units for reviewing, where $\bar{k} = min\left\{k \in (1, \ldots, n) \mid R_{M_j} < \eta \ \ \forall j > k\right\}$ where M_m denotes the set composed of the first m units.

TABLE 1. number of selected units by SECT and threshold

SECT	Threshold					No.Out.	Size
	0.001	0.002	0.005	0.01	0.02		
C	17	12	4	2	0	21	294
D	72	42	10	2	0	89	1820
E	9	6	3	1	0	19	287
F	67	50	33	15	0	71	610
G	56	41	23	6	0	53	645
H	103	79	53	32	7	71	1442
I	55	49	29	18	2	62	568
J	7	1	0	0	0	17	526
K	18	13	7	4	1	17	612

V. THE APPLICATION

In this section we describe an application of the proposed method to a subset of data from the Istat survey VELA. The aim is to show how the selective editing approach allows to select a minimal set of observations to be interactively revised in order to keep the expected residual error below a given threshold. The data used for the application are taken from the responding enterprises of the second quarter 2007 wave. The target variable is the number of hours worked per capita (number of hours worked divided by the average number of occupied posts). As already mentioned in section II, we assume that the average occupied posts variable is not affected by any error, due to the fact that it has been edited and imputed in a previous phase of the data E&I process based on external auxiliary information.

The analysis of potentially influential outliers on the target variable has been performed separately on all the economic activity sections (SECT) in NACE Rev 1.1 classification in the field of observation of the survey.

More in detail, the study has consisted in the following steps:

- for the subset of units of a given SECT, the values of hours worked (HW) and occupied posts (OCC) have been transformed in log scale to make normality assumption more realistic;
- the contaminated model parameters have been estimated, and for each unit the score function has been computed using the estimated predicted errors, as described in Section IV;
- the observations have been ordered by descending values of their score functions, and the units affected by the most influential expected errors have been selected on the basis of a set of chosen thresholds ranging from 0.001 to 0.02 (see Table 1). The stopping criterion is based on the value of the ratio between the expected residual error and the robust estimate of the total number of hours worked as described in the previous section.

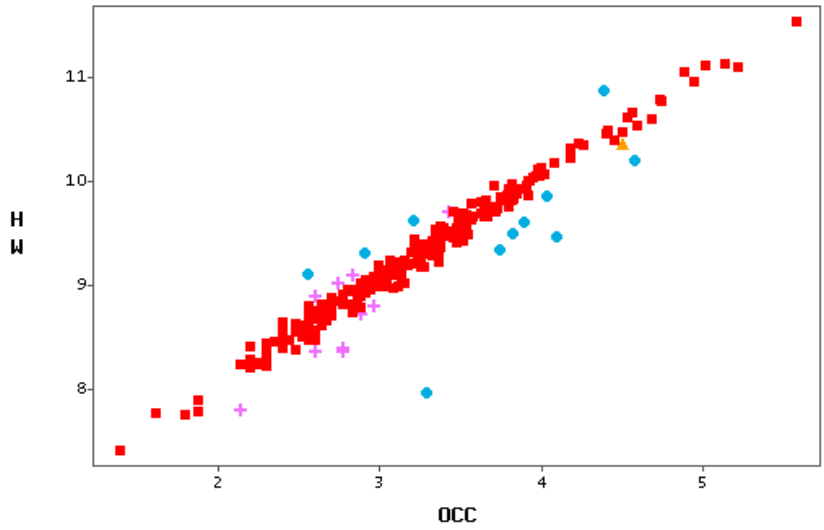

FIGURE 1. Scatter plot of Hours Worked vs. Occupied Posts (log scale) with threshold 0.002 . Dots are both outliers and influential units, crosses are outliers but not units to be revised. The triangle represents an observation with probability of being in error less than 0.5 but potentially influential

In Table 1 the number of units selected for interactive review are reported for different thresholds and different economic sections. For each section, No.Out. represents the number of units whose probability of belonging to the cluster of corrupted data is greater than 0.5, i.e. the units that would be selected for reviewing based only on outlyingness. In the last column the sample size is reported. Note that, for each considered threshold, the number of outliers is generally greater than the number of units with influential errors. As an example, in Figure 1 outliers and influential units are represented for Section C (mining and quarrying) and threshold 0.002. It is to be noted that there are sections, such as F (Constructions), I (Transport, Storage and Communication) and particularly H (Hotels and Restaurants) where there is a relatively large number of units to be revised. This is due to the fact that, because of the specific characteristics of the economic activities of these sections, the dependence relation between HW and OCC is not well represented by a linear regression. Many enterprises in these sections are in fact characterized by a quite large turnover rate (defined as the percentage of hiring and separations with respect to the quarterly average number of occupied posts), and this rate tends to be larger for smaller enterprises. Since the HW variable reflects the total amount of hours worked in the quarter, it includes also the hours worked by employees and workers who started or finished working during the quarter. On the other hand, the average number of occupied posts does not take into account all those employees or workers who started and finished working in the same quarter. Therefore, it is clear that this effect results in a much larger variability of the number of hours worked per capita for smaller enterprises than for larger enterprises (where the turnover rate is naturally smaller).

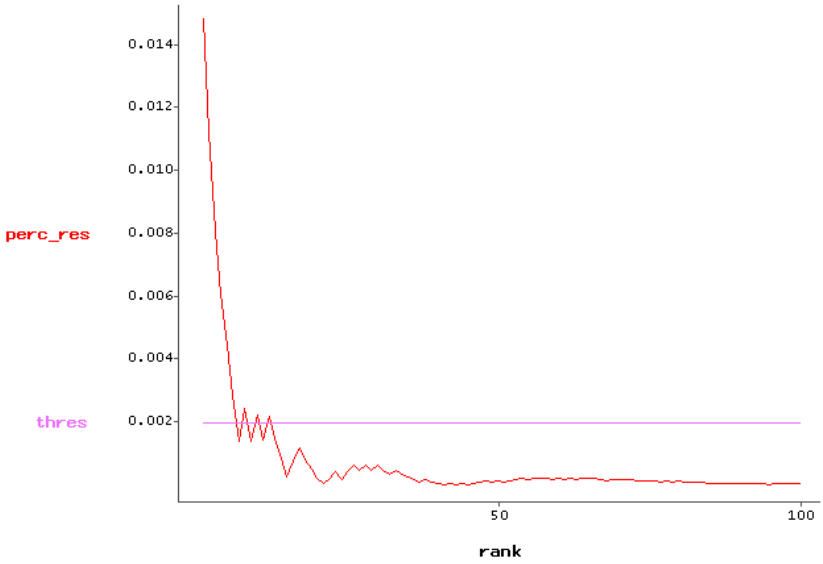

FIGURE 2. Estimated relative residual error (perc_res) for Section C vs number of selected ordered units (rank). The reference line corresponds to a 0.002 threshold (thres). Twelve units suffice to keep the estimated residual error below the threshold.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper a selective editing approach based on a contamination normal model for conditional distributions is proposed. In this approach the error free data are modelled through a linear regression with gaussian residuals and it is assumed that in a portion of the data, the dependent variable Y is contaminated by an additive gaussian error with zero mean. The independent variable X is always supposed to be error free. The explicit modelling of the correct data and the error mecchanism allows to represent the conditional distribution of the true data given the observed data as a mixture of a delta function and a normal distribution. Based on this distribution, it is possible to define the score function of each unit as expected error conditional on the observed value of the corrupted variable on that unit. Furthermore, differently from other traditional approaches, the proposed method allows to define a stopping criterion in terms of expected residual error left in the data. It is worthwhile noting that, the requirement that the percentage residual error R_{M_k} is less than a fixed threshold η, for $k = \bar{k}, \ldots, n$, ensures that each single (expected) error is not greater than 2η. On the other hand, due to the choice of this criterion, it may happen that the residual error be below the given threshold before stopping (see Fig. 2). Given the multivariate nature of economic variables, one of the future developments will consist in the extension of the proposed approach to multivariate conditional models. This seems to be particularly useful for those business surveys which make use of external auxiliary information (e.g. administrative data), which is usually considered (almost) error free. Furthermore, the proposed method can be easily generalized to account for possible missing values in the data. A last concern is the application of the proposed method in the context of sample survey with unequal sampling weights. In fact, in this paper we have ignored the sample design. However, it is natural to incorporate sample weights in the procedure by simply multiplying each expected error by the appropriate weight and using a weighted estimate in the denominator of the expression defining the score function.

References

Barnett V., Lewis T. (1994). *Outliers in Statistical Data*, New York: Wiley.

Chambers R. L. (1986). Outlier robust finite population estimation. *J. Am. Statist. Ass., 81, 1063-1069.*

Di Zio M, Guarnera U., Luzi O. (2008). Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data.*UN/ECE Work Session on Statistical Data Editing, Vienna* (http://www.unece.org/stats/documents/2008.04.sde.htm).

Hedlin D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics, *Journal of Official Statistics, Vol. 19, No. 2, 177-199.*

Jäder A., Norberg A. (2005). A Selective Editing Method considering both suspicion and potential impact, developed and applied to the Swedish Foreign Trade Statistics, *UN/ECE Work Session on Statistical Data Editing, Ottawa* (http://www.unece.org/stats/documents/2005.05.sde.htm).

Latouche M., Berthelot J.M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics, 8, n.3, 389- 400.*

Lawrence D., McKenzie R. (2000). The General Application of Significance Editing. *Journal of Official Statistics, 16, n. 3, 243-253.*

Lawrence D., McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics, Vol. 10, No. 4, pp. 437-447.*

Lee H. (1995). *Outliers in Business Surveys, in: Business Survey Methods*, Cox B.G., Binder D.A., Chinappa B.N., Christanson A., Colledge M.J. and Kott P.S. (Eds), John Wiley and Sons, Inc. 503-526.

Little, J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values, *J. R. Stat. Soc., Ser. C, Vol. 37, No. 1, 23-38.*

Meng X.L. and Rubin D.B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: a General Framework, *Biometrika, Vol. 80, 267-278.*