Working Paper.
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (i): Selective editing / macro editing

Score Functions under the Optimization Approach

Prepared by Ignacio Arbués and Pedro Revilla, INE, Spain

I. Introduction

1. Nowadays, more and more statistical agencies are researching into new methods and strategies in order to improve the efficiency of statistical production processes. Selective editing constitute an editing approach with increasing use and relevance. It is relevant not only from a methodological point of view to detect influential errors but also from a more global perspective as a key ingredient in the design of efficient editing strategies in the context of streamlining production processes.

2. Many approaches of selective editing have been proposed, sometimes with different names such as macroediting, plausibility indicators, significance editing, etc. (a systematic review can be seen in de Waal, 2013). A frequently used approach to selective editing is that of score functions. Score functions try to reflect the importance of editing a particular record, identifying data records that need to be followed up. A reason why this is convenient is that some records are more likely to improve the quality if edited than some others, either because they are more suspect to have an error or because the error (if it exists) has probably more impact in the aggregated data.

3. Despite this relevance, there is not accepted theory of selective editing, and it lacks of clear methodological principles to be translated into a concrete mathematical formulation. In fact, selective editing is an umbrella term for several methods that identify errors having a substantial impact on aggregate data. Only in recent years, it has been attempted to formalize selective editing, and theoretical frameworks are being developed (e.g. Di Zio and Guarnera 2013, Arbues et al. 2012 and 2013).

4. Since the early work by Granquist and other authors in the 1980s and 1990s, the National Statistical Institute of Spain (INE), like many others statistical institutes, has a significant interest on selective editing. From the meeting in Rome in 1999, INE has presented some papers on this topic in the Work Session, showing its experiences in this field.

5. In this paper, we introduce a theoretical framework to guide in defining a selective editing strategy. Our target is to split the data into a critical and a non-critical stream, as in other selective editing approaches. The critical stream consists of records that are the most likely to contain influential errors and therefore should be edited in an interactive manner. We face a decision problem i.e. which ones of the sample units have to be selected to manual editing. It is formulated as an optimization problem, in which the objective is to minimize the number of sample units to edit, with the constraint that the measurement error due to editing only the selected units and not all of them is below some bound. These ideas are formally translated into a generic optimization problem with two versions. On one hand, if no cross-sectional information is used in the selection of units, we derive a stochastic optimization problem. On the other hand, if that information is used, we arrive at a combinatorial optimization problem. At first

glance, score functions and optimization could be considered quite different approaches. Nevertheless, score functions can be obtained in such a way that they may be considered as embedded in the optimization approach.

6. The remaining of the paper is organized as follows. Section II describes selective editing as an optimization problem and summarizes the theoretical framework. Section III presents the score functions obtained under the optimization approach. Two practical examples are presented in Section IV. The paper ends with some final remarks.

II. Selective editing as an optimization problem

7. In our proposal (for a more detailed description see Arbués et al. 2013) the search for a good selective editing strategy is stated as an optimization problem, in which the objective is to minimize the expected workload with the constraints that the expected error of the aggregates computed with the edited data is below a certain constant. The solution of this problem is the selection of units to be intensively edited. In particular, we want to minimize the number of questionnaires to edit provided that the chosen loss functions of the survey estimators are bounded. To formally set up the optimization problem we need (i) the variables, (ii) the function to optimize, and (iii) the restrictions. We need to introduce the following notation.

8. The true, observed and edited values of a variable for unit k will be denoted, respectively, by y_k^0, y_k^{obs}, and y_k^{edit}. The ultimate variables are the selection strategy vector $\mathbf{R}^T = (R_1, R_2,\ldots, Rn)$ for the sample units $s = \{1,...,n\}$, where $R_k = 0$ if the unit k is selected for interactive editing and $R_k = 1$ otherwise. Since the measurement error $e_k = y_k^{obs} - y_k^0$ is conceived to be random in nature conditional on the realized sample s, and on the available information **Z** chosen to make the selection of units, this selection can vary depending on the realized y_k^{obs}, y_k^0 and **Z**. Thus let **R** denote the stochastic selection strategy so that (i) **R(w) = r** is a realized selection and (ii) $E_m[\mathbf{R}/\mathbf{Z}]$ is the vector of probabilities of nonselection under the specific model m given the chosen information **Z**. The assignment allows us to relate the true, observed and edited values by the equation $y_k^{edit}(r) = (1 - r_k)\, y_k^0 + r_k\, y_k^{obs}$, where we have made explicit the dependence of the edited values upon the selection strategy. Notice that we are implicitly assuming that the editing work drives us from the observed to the true values. We can also write $y_k^{edit}(r) = y_k^0 + r_k\, e_k$

9. The objective function to optimize, given the information Z is then written as $E_m[\Sigma R_i/\mathbf{Z}]$, or in matrix notation, $E_m[\mathbf{1}^T\mathbf{R}/\mathbf{Z}]$, where **1** stands for a vector of ones, whose maximization amounts to minimizing the number of selected units.

10. The constraints derive from the application of a loss function to the survey estimators. Let us concentrate on the two loss functions most used in practice, namely the absolute loss $L=L^{(1)}(a,b) = |a-b|$ or the squared loss $L=L^{(2)}(a,b) = (a-b)^2$. Each constraint controls the loss of accuracy in terms of the chosen loss function L due to nonselected units. For these loss functions, each constraint can always be written as a bound on a quadratic form, denoted by $E_m[\mathbf{R}^T\,\boldsymbol{\Delta}\mathbf{R}|\mathbf{Z}]$ (see Arbués et al., 2013). The n x n matrix Δ of conditional moments of the estimated measurement errors specifies the potential losses at the unit level. Measures of bias and/or MSE seem natural in practice and they stem from the choice of the absolute or the squared loss function respectively. These measures can be heuristic in nature, such as the pseudo-bias for traditional score functions, or explicitly derived under some appropriate measurement-error model. In particular, non-zero off-diagonal terms of **Δ** allow for cross-unit terms to be included in the "overall" loss. The choice of the matrix Δ is naturally linked to the choice of the loss function L, hence the term loss matrix (see Arbués et al. 2013 for details).

11. The entries of the corresponding matrices of moments of the estimated measurement errors Δ are denoted by

$$\Delta_{kl} = \omega_k \omega_l \left(Y_k^{obs} - Y_k^0\right)\left(Y_l^{obs} - Y_l^0\right)$$

12. Taking into account the possibility of multiple constraints, we arrive at the following generic optimization problem:

$$
\begin{aligned}
[P_0] \qquad & \max \ \mathsf{E}_m\left[1^T R \mid \mathbf{Z}\right] \\
& \text{s.t.} \quad \mathsf{E}_m\left[R^T \boldsymbol{\Delta}^{(q)} R \mid \mathbf{Z}\right] \leq \ \eta_q, \quad q = 1,2,...,Q \\
& R \in \Omega_0
\end{aligned}
$$

where $\boldsymbol{\Omega}_0$ denotes the admissible outcome space of **R**, and q refers to the different constraints. Manipulation of $\boldsymbol{\Omega}_0$ creates extra flexibility for adoption. For instance, the problem can be recast for selection conditional on the units that have already been selected, by restricting $\boldsymbol{\Omega}_0$ such that certain R_k s are fixed at 0. The different constraints q may arise from the fact that there are multiple variables of interest inside the questionnaire. In other cases, the constraints may be directed at the different population domains even when there is only a single variable.

13. Choosing the auxiliary information Z and the subset S_0 of sought selection strategies in the general problem P_0, we end up with different concrete optimization problems. If no auxiliary information is used and the sought selection strategies are of the form:

$$
\mathrm{R} \in \mathsf{S} : \mathrm{R}_k = \begin{cases} 1 & \text{if} \quad \xi_k < \mathrm{Q}_k \\ 0 & \text{if} \quad \xi_k > \mathrm{Q}_k \end{cases}
$$

where ξ_k is uniform $(0,1)$ random variable and $Q_k = Q_k(X, Y^{obs}, S)$ is a continuous $(0,1)$ random variable, we have the stochastic version of the optimization problem dealt with in Arbués et al. (2012). In particular, it is a linear optimization problem with quadratic constraints. where the solutions belong to an infinite-dimensional space of random variables. We prove existence of the solutions and show that under convexity conditions, a duality method can be used. The search for a good selective editing strategy is stated as a problem in this class, setting the expected workload as the objective function to minimize and imposing quality constraints

14. If we use both all available auxiliary information and the observed values found in the sample and do not restrict the form of the sought selection strategies $\mathsf{S}_0 = \mathsf{S}$, we have the combinatorial version of the optimization problem. .The concrete form of the combinatorial version can be expressed as

$$
\begin{aligned}
[P_{CO}] \qquad & \max 1^t \mathbf{r} \\
& \text{s.t} \quad \mathrm{r}^t \mathrm{M}^{(q)} \mathrm{r} \leq \mathrm{m}_q^2, \quad \mathrm{q} = 1,\ldots,\mathrm{Q} \\
& \mathrm{r} \in \mathrm{B}^n
\end{aligned}
$$

where **r** stands for the realizeds selection, $\mathbf{M}^{(q)}$ condenses the modelization of the measurement error, m_q are the corresponding bounds chosen by the statistician and B = {0,1}.

15. This problem has been analyzed in Salgado (2011) and two geedy algorithms have been proposed and tested to solve it (Salgado et al., 2012). Since they are heuristics-based, they provide suboptimal solutions, but with polynomial running times, namely in $O(n^4 \cdot P)$ and $O(n^3 \cdot P)$: the slower, the more precise. The degree of approximation is good enough to be accepted in this context. Furthermore, the suboptimality amounts to incurring some overediting to the final selection, but always fulfilling the restrictions upon the mean squared errors

16. The practical application of the optimation approach requires a method to compute the conditional moments of the estimated measurement errors. This method is up to the survey statistician. We propose to use a very general framework which we have named the observation-prediction model (Arbués et al., 2012).

III. Score functions obtained from the optimization approach

17. If in the generic optimization problem P_0 we incorporate some additional assumptions such as neglecting the cross-unit terms, we arrive to a linear version of the problem. In this case, the selection strategy consists of a score function that is built as a linear combination of the conditional expected quadratic errors of each variable of the questionnaire

18. As stated above, the main assumption in this version of problem P_0 is neglecting the cross-unit terms in each constraint. Then these constraints can be rewritten as $E_m[\mathbf{R}^T \boldsymbol{\Delta}\mathbf{R}/\mathbf{Z}] = E_m[\mathbf{R}^T \mathrm{diag}(\boldsymbol{\Delta})/\mathbf{Z}]$. The deduced stochastic optimization problem is solved in Arbués et al. (2012a) by using the duality principle, the sample average approximation and the interchangeability principle. The solution resulting from this linear problem is given in terms of matrices $\mathbf{M}^{(q)} = E_m[\boldsymbol{\Delta}^{(q)}/\mathbf{Z}]$. Since this selection scheme is to be applied unit by unit upon receipt of each questionnaire, and no cross-sectional information except that regarding each unit k separately will be actually used, the formal conditioning reduces effectively to conditioning upon the information of each unit. Thus we write $\mathbf{M}^{(q)} = E_m[\boldsymbol{\Delta}/ Z] = \mathrm{diag}(E_m[\Delta_{kk}^{(q)}/Z_k]) = \mathrm{diag}(M_{kk}^{(q)})$. On the other hand, in order to obtain the optimal Lagrange multipliers λ^* involved in the dual problem, a historic double-data set with raw and edited values is necessary. Putting it all together we arrive at the final solution, which only requires the diagonal entries of the matrices $M^{(q)}$:

$$R_k = \begin{cases} 1 & \text{if} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} \le 1, \\ 0 & \text{if} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} > 1. \end{cases}$$

19. This provides a score function for unit-by-unit selection. In the special case of Q = 1, unit k is selected provided $M_{kk} > 1/\lambda^*$, so that M_{kk} can be regarded as a single score and $1/\lambda^*$ as the threshold value. Equivalently, one may consider $\lambda^* M_{kk}$ as a "standardized" score, in the sense that the threshold value is generically set to 1. The latter extends in a straightforward manner to the setting with multiple constraints, where each $\lambda_q^* M_{kk}^{(q)}$ is a standardized local score, and $\sum_q \lambda_q^* M_{kk}^{(q)}$ is the standardized global score, with the generic global threshold value 1.

20. The global scoring derives from the linear structure of the dual problem and few variations are allowed without a substantial modification of problem P_0. As an exception, if a global score is initially envisaged as the weighted sum of local scores, then one may incorporate each weight into the constraint that generates the corresponding standardized local score to begin with.

21. The stochastic problem thus clarifies the fact that the performance of unit-by-unit selection can only be established over hypothetical repetitions of the selection process. At the end of each selection process, we have the realized selection strategy **r**, and the realized loss $\sum_k r_k M_{kk}^{(q)}$, which can either be higher or lower than the specified bound η_q, for q = 1,…,Q. Upon any hypothetical repetition of the selection process, however, y_k^0 and y_k^{obs} will vary, and so will the corresponding $M_{kk}^{(q)}$ and r_k . It is over such hypothetical repetitions that the constraint $E_m[\mathbf{R}\boldsymbol{\Delta}^{(q)}\mathbf{R}/Z] \le \eta_q$ can possibly be satisfied, but not for each particular realization of the selection process.

IV. Case studies

A. Case study 1: Periodic survey with a simple questionnaire

22. In this heading, we present the results of the application of the methods described to the data of the Spanish Turnover/New Orders Survey. Monthly data from about N = 13,500 units are collected with t ranging from 1 to 57. Only two of the variables requested in the questionnaires are considered in our study, namely, Total Turnover and Total New Orders (q = 2). The total Turnover of unit *j* at period *t is*

x_t^{i1} and Total New Orders is x_t^{i2}. These two variables are aggregated separately to obtain the two indicators.

23. We need a model for the data in order to obtain the conditional moments of the estimated measurement errors. Since the variables are distributed in a strongly asymmetric way, we use their logarithm transform, $y_t^{ij} = log\ (x_t^{ij} + m)$, where m is a positive constant adjusted by maximum likelihood ($m \approx 10^5$ €). The model applied to the transformed variables is very simple. We assume that the variables x_t^{ij} are independent across *(i, j)* and for any pair *(i, j)*, and we choose among the following simple univariate ARIMA models.

$$(1 - B)\, y_t^{ij} = a_t$$
$$(1 - B^{12})\, y_t^{ij} = a_t$$
$$(1 - B^{12})(1 - B)\, y_t^{ij} = a_t$$

where *B* is the backshift operator and a_t are white noise processes. We obtain the residuals $\hat{a}_t$ and then select the model which produces lesser mean of squared residuals, $\sum \hat{a}_t^2/(T-h)$, where h is the maximum lag in the model. With this model, we compute the prediction $\hat{y}_t^{ij}$ and the prediction standard deviation v_{ij}. The *a priori* standard deviation of the observation errors and the error probability are considered constant across units (that is possible because of the logarithm transformation). We denote them by σ_j and p_j with $j = 1,2$ and they are estimated using historical data of the survey.

24. A database is maintained with the original collected data and subsequent versions after possible corrections due to the editing work. Thus, we consider the first version of the data as *observed* and the last one as *true*.

25. We intend to compare the performance of our method to that of the score-function described, for example, in Hedlin (2003), $\delta_i^0 = \omega_i \mid x_i^{obs} - x_i^{pre} \mid$, where x_i^{pre} is a prediction of x according to some criterion. The author proposes to use the last value of the same variable in previous periods. We have also considered the score function δ^1 defined as δ^0 but using the forecasts obtained through the ARIMA models . Finally, δ^2 is the score function computed under the optimization approach.

26. We will measure the effectiveness of the score functions by:

$$E_1^j(n) = \Sigma_{i \geq n} (\omega_i^j)^2 (x_{ij}^{obs} - x_{ij}^0)^2 \qquad E_2^j(n) = [\Sigma_{i \geq n} \omega_i^j (x_{ij}^{obs} - x_{ij}^0)]^2 ,$$

where we consider units arranged in descending order according to the corresponding score function. These measures can be interpreted as estimates of the remaining error after editing the n first units. The difference is that $E_1^j(n)$ is the aggregate squared error and $E_2^j(n)$ is the squared aggregate error. Thus, $E_2^j(n)$ is the one that has practical relevance, but we also include the values of $E_2^j(n)$ because in the linear problem, it is the aggregate squared error which appears in the left side of the expectation constraints. In principle, it could happen that the optimization approach score function was optimal for the $E_1^j(n)$ but not for $E_2^j(n)$, because it is a liner approximation. Nevertheless, the results in table 1 show that δ^2 is better measured both ways.

	Turnover		Orders	
	E_1	E_2	E_1	E_2
δ^0	0.43	0.44	1.16	1.33
δ^1	0.30	0.38	0.36	0.45
δ^2	0.21	0.26	0.28	0.37

Table 1. Comparison of score functions.

B. Case study 2: Cross-sectional survey with a complex questionnaire

27. In this heading we briefly summarize the main results obtained so far from the application of the methodology exposed above to the case in which we deal with cross-sectional data with a large number of variables and no information from previous periods.

28. The results are related to a sample of 7215 questionnaires and 186 quantitative variables extracted from Spanish Agricultural Census of 1999. The variables are strongly skewed. Again, we need a model for the data in order to obtain the conditional moments of the estimated measurement errors. Since we cannot rely in past information to make the predictions, we will use regression models instead of time series models. Specifically, we try as our first model a classical linear regression. For each of the variables in the questionnaire, we build a model in which the log-transformed variable under study $y_j = \log(1+x_j)$ is regressed against a subset of the remaining ones (y_{rem}) as:

$$y_j = \beta_0 + \Sigma\, \beta_{rem}\, y_{rem} + \xi$$

or $y_j = \mathbf{X}'\boldsymbol{\beta} + \xi$. The most difficult and time-consuming task is the selection of the regressors, which is done by an automatic method (a kind of stepwise algorithm).

29. The predictive ability of the linear models varies greatly among the variables. The worst cases are mostly found among those variables whose distribution is less informative because they have very few nonzero values. Let us measure the quality of the regression by its R^2. In the upper plot of figure 1 we present a histogram of the R^2 of all the variables. If we weight the variables according to the number of nonzero values, we see (lower plot) that the quality of the regression is generally good for the variables with more nonzero values.

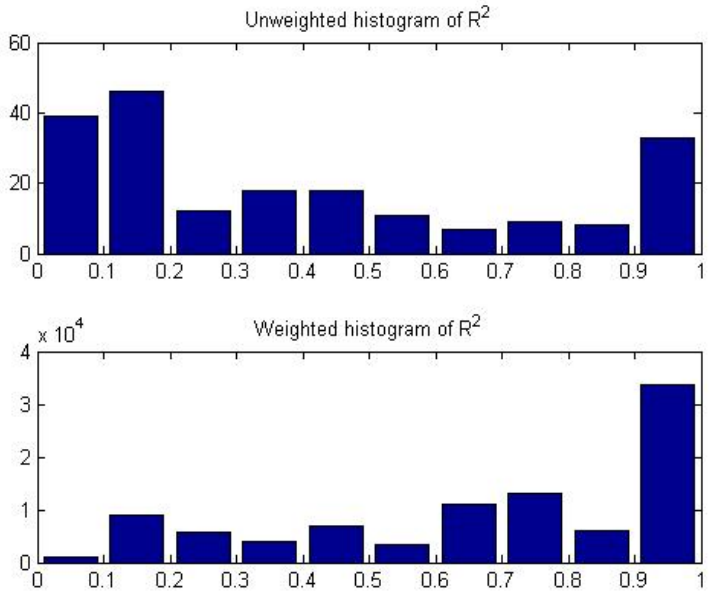

Figure 1: Histograms of the R^2 of the linear regressions, unweighted and weighted.

30. Most of those variables (land area, livestock, etc) take positive values with a continuous distribution but they have also positive probability of a zero outcome, that is, they are semicontinuous variables (see in figure 2 an example). In other words, their are distributed as a mixture of a degenerate distribution in zero and some other continuous distribution in the positive semi-axis. Some models have been proposed in the literature to deal with this kind of data (see Schafer, 1999). In particular, our second model is a two-part model in which a logit is used to predict whether the variable will be equal to zero or not, and a linear regression model conditional to the event $\{y_j > 0\}$. The model has the form

$$y_j = (1-Z)\,(X`\beta + \xi), \qquad \text{with} \qquad \xi \approx N(0, v^2)$$

and where $Z \in \{0, 1\}$. We establish a logistic regression model for the dichotomous event of having zero or positive values.

$$P(Z = 0) = e^{x'\beta} / 1+e^{x'\beta'}$$

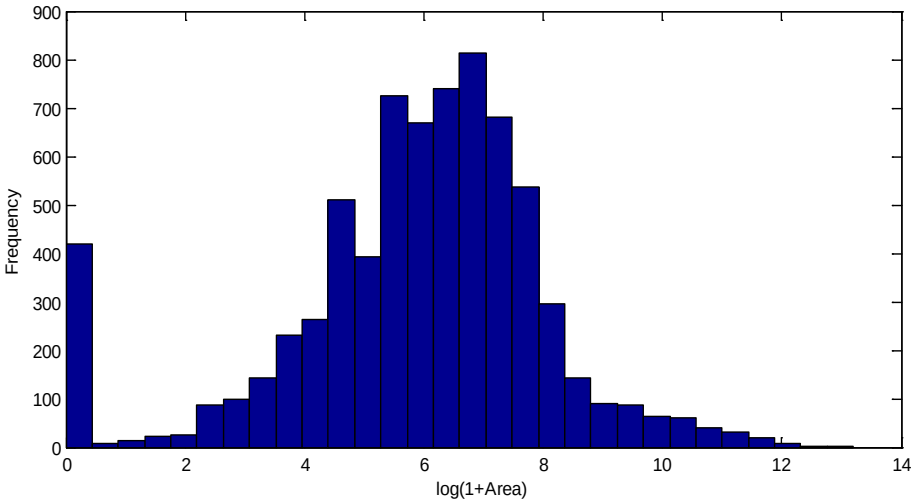

Figure 2. Histogram of the variable "land in property", transformed.

31. In this case, unlike in the one reported in the previous section, the original collected data are not available. Consequently, we had to simulate them by introducing random errors in the final published microdata. Two different settings have been considered. The first one supposes an adverse situation where the errors are quite large and frequent, whereas the other simulates smaller and scarcer errors.

32. In order to assess the quality (or usefulness in detecting errors) of the local score functions (Δ_i^k), the same steps have been followed in both settings: the questionnaires have been sorted from the biggest to the smallest priority of editing relating to each variable and the average across simulation of the remaining relative error has been calculated too. Finally, we have represented the remaining relative errors as functions of the number of edited questionnaires. These errors decrease faster for some variables than for others, but on the whole results are quite good as the summary provided by graphics of some quantiles (Figure 3) across variables shows. Approximately the removed error is about the 80% for the 90% of variables when 40 questionnaires are edited. Consequently, the optimization approach seem to provide an efficient way to detect the errors. On the other hand, not surprisingly, in order to get rid of most of the error in all the variables, it is necessary to edit a very large fraction of the total number of questionnaires.

33. The comparison between the results with the linear regression model and the two-part model shows very little difference. Since the two-part model is much more complex and time-consuming, it seems more convenient to use the linear model, at least until a convenient model selection criteria is found. In that case, a composite method could be used, such that for each variable used one or the other model depending on whether the gains of using the more complex models outweighs the disadvantages or not. That study still remains for future work.

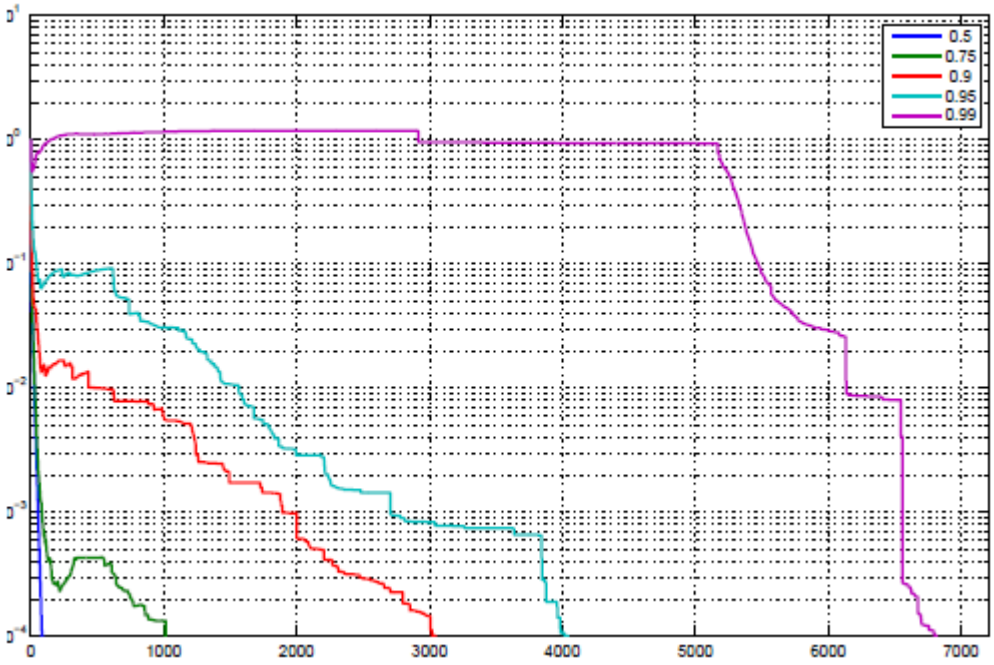

Figure 3. Remaining quadratic error (relative to total error) as a function of the number of questionnaires edited in logarithmic scale. Each curve represents the quantile indicated by is colour: blue=50%, green=75%, red=90%, cyan=95%, purple=99%. Strong error

V. Final remarks

34. We have introduced a theoretical framework to deal with the problem of selective editing. We consider the search for an adequate selection strategy as a generic optimization problem with an stochastic and a combinatorial version. The startig point of the problem is to minimize the number of units selected to manual editing restricted to a bounded increase of the errors of the estimates due to not editing the unselected units. We have shown that a certain score function provides the solution to the problem with linear constraints.

35. Both versions of the problem fit naturally as the head and tail of a time-continuous editing process. The stochastic version corresponds to exploiting longitudinal past information, whereas the combinatorial version arises as an output editing technique focusing upon the cross-sectional information. The stochastic problem, supplemented by the assumption that ignores the cross-unit terms, allows the construction of score functions to be applied independently to each unit. The supplementary assumption amounts to considering the cross-terms more or less constant over time, hence playing no significant role in the selection. Conversely, the combinatorial problem needs a sufficient number of observations available to carry out the selection of all units jointly.

36. Our experiments with real data suggest that the method provides good selection strategies. The results, although still preliminary, are encouraging. The selection obtained outperforms that of traditional score functions in general. Pilot experiences foster our hope to reduce current follow-ups rates and consequently both editing costs and response burden. R packages and SAS macros implementing this approach are under intense development and being currently tested.

37. Much work is still under progress. More methodological research is needed to find generic multivariate models to adjust the observation-prediction model and generalize to qualitative variables. From the practical point of view, all the applications have been carried out using data from traditional surveys. It would be convenient to test the procedures using other sources of information such as administrative and big data.

REFERENCES

Arbués I., Revilla, P. and Salgado D. (2013). *"An optimization approach to selective editing"*. Journal of Official Statistics (JOS). Volume 29, Issue 4. Pages 489–510.

Arbués I., González M. and Revilla, P. (2012). *"A class of stochastic optimization problems with application to selective data editing"*. Optimization. Volume 61, Issue 3, Pages 265-286.

Arbués I., Salgado D. and Revilla, P. (2012). ***"****Selective Editing as a Combinatorial Optimization Problem: a General Overview"*. UN/ECE Work Session on Statistical Data Editing. Oslo, 24-26 September.

Arbués I., Revilla, P., and Saldaña S. (2011). ***"****Selective Editing as a Stochastic Optimization Problem"*. UN/ECE Work Session on Statistical Data Editing. Ljubljana, 16-18 May.

De Waal, T., (2013). *"Selective Editing: A Quest for Efficiency and Data Quality"*. Journal of Official Statistics. (JOS). Volume 29, Issue 4, Pages 473–488.

Di Zio, M. and Guarnera, U. (2013). *"A Contamination Model for Selective Editing"*. Journal of Official Statistics (JOS) Volume 29, Issue 4, Pages 539–555.

Hedlin D. (2003), *"Score Functions to Reduce Business Survey Editing at the U.K"*. Office for National Statistics, *Journal of Official Statistics*, 19, 177–199.

Revilla, P, Rey, P (1999). "*Selective Editing Methods Based on time series modelling"* UN/ECE Work Session on Statistical Data Editing. Rome June.

Salgado, D., Arbués, I. and Esteban, M.E. (2012). "*Two greedy algorithms for a binary quadratically constrained linear program in survey data editing*·. Working Papers. National Statistics Institute.

Salgado, D. (2011). ·"*Exploiting auxiliary information: selective editing as a combinatorial optimization problem"*·. Working Papers. National Statistics Institute.

Schafer J.L., and Olsen M.K. (1999). *"Modeling and imputation of semicontinuous survey variables"*. In Proceedings of Federal Committee on Statistical Methodology (FCSM) Reseach Conference.