WP. 18
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iii): Macro editing methods

SELECTIVE EDITING AS A STOCHASTIC OPTIMIZATION PROBLEM

Key invited paper

Submitted by National Statistical Institute (Spain)[1]

I. INTRODUCTION

1. Efficient editing methods are critical for the statistical offices. In the past, it was customary to edit manually every questionnaire collected in a survey before computing aggregates. Nowadays, exhaustive manual editing is considered inefficient, since most of the editing work has no consequences at the aggregate level and can in fact even damage the quality of the data (see [7] and [4]).

2. Selective editing methods are strategies to select a subset of the questionnaires collected in a survey to be subject to extensive editing. A reason why this is convenient is that it is more likely to improve quality by editing some units than by editing some others, either because the first ones are more suspect to have an error or because the error if it exists has probably more impact in the aggregated data. Thus, it is reasonable to expect that a good selective editing strategy can be found that balances two aims: (i) good quality at the aggregate level and (ii) less manual editing work.

3. This task is often done by defining a score function (SF), which is used to prioritise some units. When several variables are collected for the same unit, different *local* score functions may be computed and then, combined into a *global* score function. Finally, those units with score over a certain threshold are manually edited. Thus, when designing a selective editing process it is necessary to decide: (a) whether to use SF or not; (b) the local score functions; (c) how to combine them into a global score function (sum, maximum, ...); (d) the threshold.

4. Formerly, the points above were dealt with in an empirical way because of the lack of any theoretical support. In [6], [8] and [4] some guidelines are proposed to build score functions, but they rely in the criterion of the practitioner. In this paper, we describe a theoretical framework which, under some assumptions, answers the questions above. For this purpose, we will formally define the concept of *selection strategy*. This allows to state the problem of selective editing as an optimization problem

[1]Prepared by Ignacio Arbués (iarbues@ine.es), Pedro Revilla (previlla@ine.es) and Soledad Saldaña (soledad.saldana.diaz@ine.es).

in which the objective is to minimize the expected workload with the constraint that the expected remaining error after editing the selected units is below some bound.

5. We describe the selective editing as an optimization problem in section II. The problem is presented in two forms and we solve them is sections III and IV respectively. Both solutions depend on the computation of some conditional moments of the error distribution. In section V, we show how to compute them. Then, two practical examples are described in sections VI and VII. We conclude with some remarks in section VIII.

II. THE SELECTION PROBLEM

6. Let us introduce some notation,

- x_t^{ij} is the *true* value of variable j in questionnaire i at period t, with $i = 1, \ldots, N$ and $j = 1, \ldots, q$.
- $\tilde{x}_t^{ij} = x_t^{ij} + \varepsilon_t^{ij}$ is the *observed* value of variable j in questionnaire i at period t, ε_t^{ij} being the observation error.
- $X_t^k = \sum \omega_{ij}^k x_t^{ij}$ is the k-th statistic computed with the true values ($\tilde{X}_t^k$ is computed with the observed ones), according to the weightings ω_{ij}^k with k ranging from 1 to p.

The linearity assumption implies a loss of generality, which is nevertheless not very important in the usual practice of statistical offices. Many statistics are in fact linear aggregates of the data. In some other cases such as indices, they are ratios whose denominator depends on past values that can be considered as constant when editing current values. When the statistic is nonlinear, the applicability of the method depends on the accuracy of a first-order Taylor expansion in $\{x_t^{ij}\}$.

7. Let $(\Omega, \mathcal{F}, P)$ be a probability space. We assume that x_t^{ij} and ε_t^{ij} are random variables with respect to that space. There can be other random variables relevant to the selection process. Among them, some are known at the moment of the selection, such as $\tilde{x}_t^{ij}$, x_s^{ij} with $s < t$ or even variables from other surveys. The assumption that x_s^{ij} is known is equivalent to assume that when editing period t, the data from previous periods have been edited enough and does not contain errors. We will denote by $\mathcal{G}_t$ the $\sigma-$field generated by all the information available up to time t. In order to avoid heavy notation, we omit the subscript t when no ambiguity arises.

8. Our aim is to find a good selection strategy. A selective editing strategy should indicate for any i whether questionnaire i will be edited or not and this has to be decided using the information available. In fact, we will allow the strategy not to determine precisely whether the unit is edited but only with a certain probability.

Definition 1. *A selection strategy (SS) with respect to $\mathcal{G}_t$ is a $\mathcal{G}_t$-measurable random vector $R = (R_1, \ldots, R_N)^T$ such that $R_i \in [0, 1]$.*

We denote by $S(\mathcal{G}_t)$ the set of all the SS with respect to $\mathcal{G}_t$. The interpretation of r is that questionnaire i is edited with probability $1 - R_i$. To allow $0 \leq R_i \leq 1$ instead of the more restrictive $R_i \in \{0, 1\}$ is theoretically and practically convenient because then, the set of strategies is convex and techniques of convex optimization can be used. Moreover, it could happen that the optimal value over this generalized space were better than the restricted case (just as in hypothesis testing a randomized test can have a greater power than any nonrandomized one). If for unit i, $R_i \in (0, 1)$, then the unit is effectively edited depending on whether $\chi_t^i < R_i$, where χ_t^i is a random variable distributed uniformly in the interval $[0, 1]$, and independent from every other variable in our framework. We denote by $\tilde{R}_i$ the indicator variable of the event $\chi_t^i < R_i$ and $\tilde{R} = (\tilde{R}_1, \ldots, \tilde{R}_N)$. If a SS satisfies $R_i \in \{0, 1\}$ a.s.,

then $\tilde{R} = R$ a.s. and we say that R is integer. The set of integer SS is denoted by $S_I(\mathcal{G}_t)$. In our case study, the solutions obtained are integer or approximately integer.

9. It is also convenient to have a formal definition of a Score Function.

Definition 2. *Let R be a SS, $\delta = (\delta_1, \ldots, \delta_N)^T$ a random vector and $\Theta \in \mathbb{R}$, such that $R_i = 1$ if and only if $\delta_i \leq \Theta$. Then, we say that δ is a Score Function generating R with threshold Θ.*

10. In order to formally pose the problem, we will assume that after manual editing, the true values of a questionnaire are obtained. Thus, we have to consider only the observed and true values. We define the *edited* statistic $X^k(R)$ as the one calculated with the values obtained after editing according to a certain choice. We can write $X^k(R) = \sum \omega_{ij}^k (x_t^{ij} + \tilde{R}_i \varepsilon_t^{ij})$.

11. The quality of $X^k(R)$ has to be measured according to a loss function. In this paper, we consider only the Squared Error, $(X^k(R) - X^k)^2$. This choice makes easier the theoretical analysis. It remains for future research to adapt the method for other loss functions. The value of the loss function can be written as

$$(X^k(R) - X^k)^2 = \sum_{i,i'} \epsilon_i^k \epsilon_{i'}^k \tilde{R}_i \tilde{R}_{i'}, \tag{1}$$

where $\epsilon_i^k = \sum_j \omega_{ij}^k \varepsilon_t^{ij}$ or, in matrix form, as $(X^k(R) - X^k)^2 = \tilde{R}' E^k \tilde{R}$, with $E^k = \{E_{i,i'}^k\}_{i,i'}$ and $E_{i,i'}^k = \epsilon_i^k \epsilon_{i'}^k$. For some positive constants $e_k^2 > 0$, we can now state the problem of selection as an optimization problem.

$$\begin{array}{lll} [P_Q] & \max_R & \mathbf{E}[1^T \tilde{R}] \\ & \text{s.t.} & R \in S(\mathcal{G}_t), \mathbf{E}[\tilde{R}^T E^k \tilde{R}] \leq e_k^2, k = 1, \ldots, p. \end{array}$$

12. In section IV we will see the solution to this problem. The vector in the cost function can be replaced for another one in case the editing work were considered different among units (e.g., if we want to reduce the burden for some respondents; this possibility is not dealt with in this paper).

13. Let us now analyse the expression (1). We can decompose it as

$$(X^k(R) - X^k)^2 = \sum_i (\epsilon_i^k)^2 \tilde{R}_i + \sum_{i \neq i'} \epsilon_i^k \epsilon_{i'}^k \tilde{R}_i \tilde{R}_{i'}. \tag{2}$$

The first term in the RHS of (2) accounts for the individual impact of each error independently of its sign. In the second term the products are negative when the factors have different signs. Therefore, in order to reduce the total error, a strategy will be better if it tends to leave unedited those couples of units with different signs. The nonlinearity of the second term makes the calculations more involved. For that reason, we will also study the problem neglecting the second term.

$$\begin{array}{lll} [P_L] & \max_R & \mathbf{E}[1^T \tilde{R}] \\ & \text{s.t.} & R \in S(\mathcal{G}_t), \mathbf{E}[D^k \tilde{R}] \leq e_k^2, k = 1, \ldots, p, \end{array}$$

where $D^k = (D_1^k, \ldots, D_N^k)^T$, $D_i^k = (\epsilon_i^k)^2$

14. This problem is easier than P_Q because the constraints are linear. In section III we will see that the solution is given by a certain score function. Since there is no theoretical justification for neglecting the quadratic terms, the SS solution of the linear problem has to be empirically justified by the results obtained with real data.

III. SOLUTION TO THE LINEAR CASE

15. In [1], it is shown that $[P_L]$ is equivalent to

$$\begin{array}{lll} [P_L^*] & \max_R & \mathbf{E}[1^T R] \\ & \text{s.t.} & R \in S(\mathcal{G}_t), \mathbf{E}[\Delta^k R] \leq e_k^2, k = 1, \ldots, p, \end{array}$$

where $\Delta^k = \mathbf{E}[D^k|\mathcal{G}_t]$ and that under not too restrictive assumptions, if $\bar{\lambda}$ is a solution to the dual problem,

$$[D] \quad \min_{\lambda \geq 0} \quad \varphi(\lambda)$$

with $\varphi(\lambda) = \max_R \mathcal{L}(R, \lambda)$, then $R = \arg\max \mathcal{L}(\cdot, \bar{\lambda})$ is a solution to $[P_L]$. Since $\Delta^k(\omega)$ is known, the maximization of $\mathcal{L}$ with respect to R boils down to solving the deterministic problem

$$\begin{array}{lll} [P_D(\lambda, \omega)] & \max_r & 1^T r - \sum_k \lambda_k (\Delta^k(\omega) r - e_k^2) \\ & \text{s.t.} & r_i \in [0, 1], \end{array}$$

and setting $R(\omega) = r$ the solution to $[P_D(\lambda, \omega)]$.

16. By applying the Karush–Kuhn–Tucker conditions (see [2]), we get a solution to $[P_D(\lambda, \omega)]$ given by

$$r_i = \begin{cases} 1 & \text{if} \quad \lambda^T \Delta_i < 1 \\ 0 & \text{if} \quad \lambda^T \Delta_i > 1, \end{cases} \tag{3}$$

where $\Delta_i = (\Delta_i^1, \ldots, \Delta_i^p)^T$. The case $\lambda^T \Delta_i = 1$ is a zero-probability event when dealing with quantitative data, given that the distribution of Δ_i should be continuous. Then, we arrive to,

Proposition 1. *The solution to $[P_L]$ is the SS generated by the Score Function $\delta_i = \lambda^T \Delta_i$ with threshold equal to 1.*

17. We describe in section V how to use a model for the practical computation of Δ^k. In order to estimate the dual function $\varphi(\lambda) = \mathbf{E}[L(R_\lambda, \lambda,)]$ we replace expectation for the mean value over a sample. Thus, we can seek the optimum of $\hat{\varphi}(\lambda) = \frac{1}{h} \sum_{t=t_0}^{t_0+h-1} L_t(r_\lambda^t, \lambda)$. This method is known as the sample–path optimization or sample average approximation method (SAA, see [3]). The maximization of $\hat{\varphi}$ may be done by numerical methods.

IV. SOLUTION TO THE QUADRATIC CASE

18. The way to solve the quadratic problem is similar, but we face now some further difficulties, in particular, that the constraints are not convex. Therefore, we will replace them by some convex ones in such a way that under some assumptions the solutions remain the same. The reader is henceforth referred to [1] for the proofs and technical details.

19. We first need to express the constraints in a suitable form.

Lemma 1. *It holds, $\mathbf{E}[\tilde{R}^T E^k \tilde{R}|\mathcal{G}_t] = R^T \Gamma^k R + (\Delta^k)^T R$ where $\Gamma^k = \{\Gamma_{ij}^k\}_{ij}$ and,*

$$\Gamma_{ij}^k = \begin{cases} \mathbf{E}[\epsilon_i^k \epsilon_j^k|\mathcal{G}_t] & \text{if} \quad i \neq j \\ 0 & \text{if} \quad i = j. \end{cases}$$

Unfortunately, the matrices Γ^k are indefinite and thus the constraints are not convex. We will overcome this difficulty by using the following lemma.

Lemma 2. *Let $\bar{g}_2$ be a function such that $\forall R \in S_I(\mathcal{G}), \mathbf{E}[\bar{g}_2(R(\omega),\omega)] = \mathbf{E}[g_2(R(\omega),\omega)]$ and $\forall R \in S(\mathcal{G}), \mathbf{E}[\bar{g}_2(R(\omega),\omega)] \leq \mathbf{E}[g_2(R(\omega),\omega)]$, and let $[P'_Q]$ be the problem obtained from $[P_Q]$ replacing g_2 for $\bar{g}_2$. Then, if R is a solution to $[P'_Q]$ and $R \in S_I(\mathcal{G})$, then R is a solution to $[P_Q]$.*

20. We may consider at least the two following possibilities: (i) $\bar{g}_2(r,\omega) = r^T\Sigma^k(\omega)r$, where $\Sigma^k_{ij} = \mathbf{E}[\epsilon^k_i\epsilon^k_j|\mathcal{G}_t]$ and (ii) $\bar{g}_2(r,\omega) = r^T M^k(\omega)r + (v^k(\omega))^T r$, where $M^k_{ij} = m^k_i m^k_j$, $m^k_i = \mathbf{E}[\epsilon^k_i|\mathcal{G}_t]$, $v^k_i = \mathbf{V}[\epsilon^k_i|\mathcal{G}_t]$. The choice (ii) can be used only under the assumption that $\mathbf{E}[\epsilon^k_i\epsilon^k_j|\mathcal{G}_t] = m^k_i m^k_j$ for $i \neq j$ and this will be the one used in our application (section VI). Lemma 2 has practical relevance if we check that the solutions of $[P'_Q]$ are integer. We show in [1] that this holds approximately in our experiment.

21. Now, we have to solve problem $[P'_Q]$, that is, the one with the constraints $\mathbf{E}[r^T A^k r + (b^k)^T r] \leq e^2_k$, where $A^k = \Sigma^k, b^k = 0$ or $A^k = M^k, b^k = v^k$. Since A^k are positive semidefinite, the constraints are convex and we can also replace the original problem by the dual one as we did in the linear case.

22. Once again, the maximization of the Lagrangian function can be reduced to a deterministic optimization problem, in this case a quadratic programming problem.

$$[P_D(\lambda,\omega)] \quad \max_r \quad 1^T r - \sum_k \lambda_k (r^T A^k r + (b^k)^T r - e^2_k) \tag{4}$$

$$\text{s.t.} \quad r_i \in [0,1]. \tag{5}$$

23. An important difference with respect to the linear case is that the problem above does not explicitly provide a Score Function generating the SS as when applying the Karush–Kuhn–Tucker conditions in section III.

24. We describe in section V a practical method to obtain M^k, Σ^k and v^k. $[P_D(\lambda,\omega)]$ was easy to solve in the linear case, but for large sizes (in our case $N > 10,000$), the quadratic programming problem becomes computationally heavy if solved by traditional methods. For $\bar{g}_2$ defined as in (ii), we can take advantage of the low rank of the matrix in the objective function to propose (see [1]) an approximate method to solve it efficiently.

V. MODEL-BASED CONDITIONAL MOMENTS

25. The practical application of the results in previous sections requires a method to compute the conditional moments of the error with respect to $\mathcal{G}_t$. In this section, we drop the index j to reduce the complexity of the notation, but the results can be adapted to the case of several variables per questionnaire.

26. Let $\mathcal{H}_t$ be a $\sigma-$field generated by all the information available at time t with the exception of $\tilde{x}^i_t$. Then, $\mathcal{G}_t = \sigma(\tilde{x}^i_t, \mathcal{H}_t)$. Let $\hat{x}^i_t = \tilde{\pi}(x^i_t)$ be a predictor computed using the information in $\mathcal{H}_t$, that is a $\mathcal{H}_t$-measurable random variable optimal in some way decided by the analyst. The prediction error is denoted by $\xi^i_t = \hat{x}^i_t - x^i_t$. We have to make some assumptions.

Assumption 1. *ξ^i_t and η^i_t are distributed as a bivariate Gaussian with zero mean, variances ν^2_i and σ^2_i and correlation γ_i.*

Assumption 2. *$\varepsilon^i_t = \eta^i_t e^i_t$, where e^i_t is a Bernoulli variable that equals 1 or 0 with probabilities p and $1-p$ and it is independent of ξ^i_t and η^i_t.*

Assumption 3. *ξ_t^i, η_t^i and e_t^i are jointly independent of $\mathcal{H}_t$.*

27. With these assumptions, the conditional moments of the error with respect to $\mathcal{G}_t$ are functions of the sole variable $u_t^i = \hat{x}_t^i - \tilde{x}_t^i$, that is, the difference between the predicted and the observed values. In the next proposition we will also drop i and t in order to simplify notation.

Proposition 2. *Under the assumptions 1–3, it holds*

$$\mathbf{E}[\varepsilon|\mathcal{G}] = \frac{\sigma^2+\gamma\sigma\nu}{\sigma^2+\nu^2+2\gamma\sigma\nu}u\zeta \tag{6}$$

$$\mathbf{E}[\varepsilon^2|\mathcal{G}] = \left[\frac{\sigma^2\nu^2(1-\gamma^2)}{\sigma^2+\nu^2+2\gamma\sigma\nu} + \left(\frac{\sigma^2+\gamma\sigma\nu}{\sigma^2+\nu^2+2\gamma\sigma\nu}\right)^2 u^2\right]\zeta, \tag{7}$$

where,

$$\zeta = \frac{1}{1+\frac{1-p}{p}\left(\frac{\nu^2}{\sigma^2+\nu^2+2\gamma\sigma\nu}\right)^{-1/2}\exp\{-\frac{u^2(\sigma^2+2\gamma\sigma\nu)}{2\nu^2(\sigma^2+\nu^2+2\gamma\sigma\nu)}\}}. \tag{8}$$

VI. CASE STUDY 1: PERIODIC SURVEY WITH A SIMPLE QUESTIONNAIRE

28. In this section, we present the results of the application of the methods described in this paper to the data of the 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, so $p = 2$ and $\omega_{i2}^1 = \omega_{i1}^2 = 0$.

29. We need a model for the data in order to apply proposition 2 and obtain the conditional moments. 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 conditional moments of the original variable can be recovered exactly by using the properties of the log-normal distribution or approximately by using a first-order Taylor expansion, yielding $\mathbf{E}[(\tilde{x}_t^{ij} - x_t^{ij})^2|\mathcal{G}_t] \approx (\tilde{x}_t^{ij} - m)^2\mathbf{E}[(\tilde{y}_t^{ij} - y_t^{ij})^2|\mathcal{G}_t]$. In our study, we used the approximate version. We found that if $\tilde{x}_t^{ij} - m$ is replaced by an average of the last 12 values of $\tilde{x}_t^{ij}$, the estimate becomes more robust against very small values of $\tilde{x}_t^{ij} - m$.

30. 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), we choose among the following simple models.

$$(1-B)y_t^{ij} = a_t, \tag{9}$$

$$(1-B^{12})y_t^{ij} = a_t, \tag{10}$$

$$(1-B^{12})(1-B)y_t^{ij} = a_t. \tag{11}$$

where B is the backshift operator $Bu_t = u_{t-1}$ and a_t are white noise processes. We obtain the residuals $\hat{a}_t$ and then select the model with the smallest squared residual mean, $\sum \hat{a}_t^2/(T-r)$, where r is the maximum lag in the model. With this model, we compute the prediction $\hat{y}_t^{ij}$ and the prediction standard deviation ν_{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.

31. 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*. The coefficient γ_i is assumed zero. Once we have computed σ_j, p_j, ν_{ij} and u_t^{ij}, proposition 2 can be used to obtain the conditional moments and then, Δ^k, Σ^k and v^k.

A. Expectation Constraints

32. We will now check that the expectation constraints in $[P_L]$ and $[P_Q]$ are effectively satisfied. In order to do this, for $l = 1, \ldots, b$ with $b = 20$ we solve the optimization problem with the variance bounds $e_{1l}^2 = e_{2l}^2 = e_l^2 = [s_0^{((l-1)/(b-1))} s_1^{((b-l)/(b-1))}]^2$. The range of standard deviations goes from $s_0 = 0.025$ to $s_1 = 1$.

33. The expectation of the dual function is estimated using a $h-$length batch of real data. For the last 12 periods and for any $l = 1, \ldots, b$, a selection $r(t, l)$ is obtained according the bound e_k^2. The average across t of the remaining squared errors is thus computed as $\hat{e}_{kl}^2 = \frac{1}{12} \sum_{t=t_0}^{t_0+11} r(t,l)^T E^k r(t,l)$.

34. We repeated these calculations for $h = 1, 3, 6$ and 12 both using the linear and the quadratic versions. We report the results for $h = 6$ in table 1 (detailed data can be found in [1]). For each l we present the average number of units edited, the desired bound and for $k = 1, 2$, the quotient $\hat{e}_{kl}/e_{kl}$. In every case, there is a tendency to underestimate the error when the constraints are smaller and to overestimate it when the bounds are larger. The quadratic method produces better results with respect to the bounds but at the price of editing more units.

	Linear			Quadratic		
e_l	$\frac{\hat{e}_{1l}}{e_l}$	$\frac{\hat{e}_{2l}}{e_l}$	n	$\frac{\hat{e}_{1l}}{e_l}$	$\frac{\hat{e}_{2l}}{e_l}$	n
.0250	1.89	3.20	414.8	1.86	2.50	578.4
.0369	2.04	2.43	257.6	1.42	1.82	401.9
.0544	1.58	1.87	157.7	0.96	1.17	390.5
.0801	1.10	1.25	95.9	1.31	1.30	283.4
.1182	0.96	1.18	56.2	1.32	1.22	140.3
.1742	0.97	1.31	30.5	1.24	0.92	77.8
.2569	0.71	1.09	15.0	0.92	0.67	35.4
.3788	0.74	0.77	6.8	0.64	0.66	28.0
.5585	0.57	0.57	2.7	0.56	0.57	6.8
.8235	0.42	0.38	1.2	0.39	0.38	2.9

TABLE 1. Bounds of the linear and quadratic versions.

B. Comparison of Score Functions

35. We intend to compare the performance of our method to that of the score-function described in [5], $\delta_i^0 = \omega_i |\tilde{x}^i - \hat{x}^i|$, where $\hat{x}_i$ 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 models in (9)–(11). Finally, δ^2 is the score function computed using (9)–(11) and proposition 2. The global SF is just the sum of the two

	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 2. Comparison of score functions.

local ones. We will measure the effectiveness of the score functions by $E_l^j = \sum_n E_l^j(n)$, with

$$E_1^j(n) = \sum_{i \geq n}^{N} (\omega_i^j)^2 (\tilde{x}^{ij} - x^{ij})^2 \qquad E_2^j(n) = \Big[\sum_{i \geq n}^{N} \omega_i^j (\tilde{x}^{ij} - x^{ij}) \Big]^2,$$

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 $[P_L]$, it is the aggregate squared error which appears in the left side of the expectation constraints. In principle, it could happen that our score function was optimal for the $E_1^j(n)$ but not for $E_2^j(n)$. Nevertheless, the results in table 2 show that δ^2 is better measured both ways.

VII. CASE STUDY 2: CROSS-SECTIONAL SURVEY WITH A COMPLEX QUESTIONNAIRE

36. In this section 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.

37. The results are related to a sample of 7215 questionnaires and 186 quantitative variables extracted from Spanish Agricultural Census of 1999. 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 as

$$y_j = \beta_0 + \sum_{\ell} \beta_\ell y_\ell + \xi, \tag{12}$$

or $y_j = X'\boldsymbol{\beta} + \xi$. 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).

38. Most of those variables take positive values with a continuous distribution but they have also positive probability of a zero outcome, that is, they are semicontinuous variables. 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 [9]). 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

$$y_j = (1 - Z)\big(X'\boldsymbol{\beta} + \xi\big) \quad \text{with} \quad \xi \approx N(0, \nu^2) \tag{13}$$

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

$$P(Z = 0) = \frac{e^{X'\beta}}{1 + e^{X'\beta}}. \tag{14}$$

39. The theoretical framework and results developed in sections II through IV are applicable to this kind of models, but it is necessary to adapt the expressions of section V. The details of this adaptation will be reported in a forthcoming article.

40. 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.

41. In order to assess the quality (or usefulness in detecting errors) of the conditional moments (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 1) across variables shows. Approximately the removed error is about the 80% for the 90% of variables when 40 questionnaires are edited. Consequently, the conditional moments 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.

42. 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 remains for future work.

VIII. FINAL REMARKS

43. We have described a theoretical framework to deal with the problem of selective editing, defining the concept of selection strategy. We describe the search for a good selection strategy as an optimization problem. This problem is a linear optimization problem with quadratic constraints. We show that there is a score function that provides the solution to the problem with linear constraints. We also show how to solve the quadratic problem.

44. In this framework, in order to adapt the method to a certain survey, it is only necessary to build a tailored model to make the predictions and then, the selection is the result of applying mechanically the theory of sections II through V (with the exception of the two-part model for semicontinuous data, that required some more work). This flexibility is one of the assets of our method.

45. The experiments with real data suggest that the method provides good selection strategies. The quadratic method seems to be more conservative and then, the bounds are better fulfilled, but

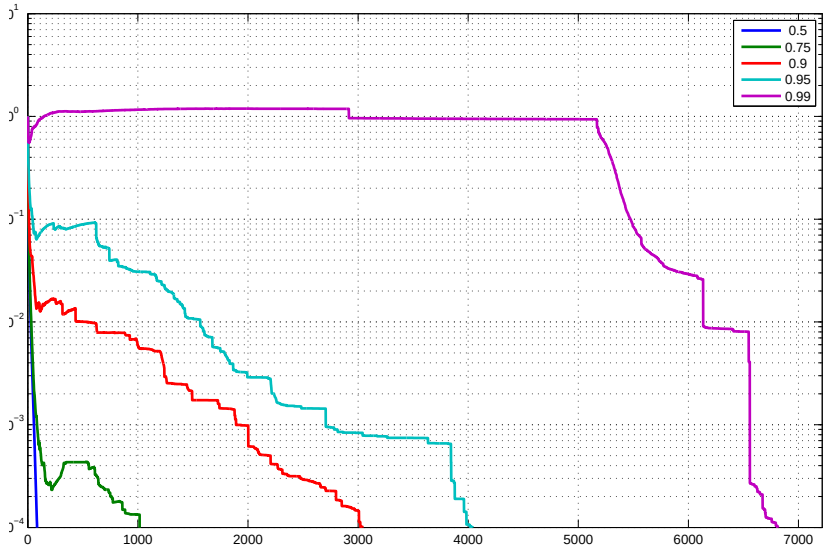

FIGURE 1. Remaining quadratic error (relative to total error) as a function of the number of questionnaires edited in logarithmic scale. Each curve represents the quantile indicated by its color: blue=50%, green=75%, red=90%, cyan=95%, purple=99%. Strong error scenario.

more units are edited. On the other hand, the implementation of the linear method is easier and computationally less demanding.

References

[1] I. Arbués, M. González, P. Revilla (2010), A Class of Stochastic Optimization Problems with Application to Selective Data Editing, *Optimization*, DOI: 10.1080/02331934.2010.511670.

[2] M. S. Bazaraa, H. D. Sherali y C. M. Shetty (1993), *Nonlinear Programming: Theory and Algorithms*, Nueva York: Wiley.

[3] J. Dupačová and R. J. -B. Wets (1988), Asymptotic Behavior of statistical estimators and of optimal solutions of stochastic optimization problems, *The annals of statistics* 16, 1517–1549.

[4] L. Granquist (1997), The new view on editing, *International Statistics Review*, **65**, 381–387.

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

[6] M. Latouche y J. Berthelot (1992), Use of a score function to prioritize and limit recontacts in editing business surveys, *Journal of Official Statistics*, **8**, 389–400.

[7] J. Berthelot y M. Latouche (1993), Improving the efficiency of Data Collection: A Generic Respondent Follow-Up Strategy for Economic Surveys, *Journal of Business and Economic Statistics*, **11**, 417–424.

[8] D. Lawrence y R. McKenzie (2000), The general application of Significance Editing, *Journal of Official Statistics*, **16**, 243–253.

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