VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 489–510, http://dx.doi.org/10.2478/jos-2013-0037

An Optimization Approach to Selective Editing

Ignacio Arbués[1], *Pedro Revilla*[1], *and David Salgado*[1]

We set out two generic principles for selective editing, namely the minimization of interactive editing resources and data quality assurance. These principles are translated into a generic optimization problem with two versions. On the 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. These problems are substantiated by constructing a so-called observation-prediction model, that is, a multivariate statistical model for the nonsampling measurement errors assisted by an auxiliary model to make predictions. The restrictions of these problems basically set upper bounds upon the modelled measurement errors entering the survey estimators. The bounds are chosen by subject-matter knowledge. Furthermore, we propose a selection efficiency measure to assess any selective editing technique and make a comparison between this approach and some score functions. Special attention is paid to the relationship of this approach with the editing fieldwork conditions, arising issues such as the selection versus the prioritization of units and the connection between the selective and macro editing techniques. This approach neatly links the selection and prioritization of sampling units for editing (micro approach) with considerations upon the survey estimators themselves (macro approach).

Key words: Selective editing; optimization; observation-prediction model; selection efficiency measure.

1. Introduction

Data editing is a crucial step in the survey statistics production process. It impinges on several dimensions of survey quality such as accuracy, timeliness, response burden or cost effectiveness. This production phase comprises both the detection and treatment of nonsampling errors, mainly of nonresponse and measurement errors. Over time, a typology of errors has been developed, identifying systematic errors, random errors, influential errors, outliers, inliers or missing values, not to mention particular errors within these classes as measurement unit errors or rounding errors. This diversity has given rise to different techniques and algorithms to detect and treat them, such as interactive editing, automatic editing, selective editing, macro editing, and so on (see De Waal et al. 2011 for a comprehensive overview). Nowadays it is widely accepted that no single technique can

[1] D.G. Metodología, Calidad y TIC, Instituto Nacional de Estadística, Paseo de la Castellana, 28071 Madrid, Spain. Emails: ignacio.arbues.lombardia@ine.es, pedro.revilla.novella@ine.es, and david.salgado.fernandez@ine.es
Acknowledgments: We acknowledge graphics computing support from S. Saldaña. We are indebted to C. Pérez-Arriero and M. Herrador for their invaluable suggestions regarding the selection efficiency measure. The authors are grateful to T. de Waal, M. Di Zio, U. Guarnera, J. Pannekoek, M. van der Loo and S. Scholtus for comments and suggestions. We express special thanks to L.-C. Zhang for invaluable suggestions to improve the readability of the article.

© Statistics Sweden

deal with all kinds of errors. Thus they must be conveniently combined in so-called editing and imputation (E&I henceforth) strategies, specifically designed and fine-tuned for a given survey.

Selective editing focuses upon influential errors so that a selection of influential units is performed to thoroughly treat their errors (mostly with interactive editing), underlining the importance of recognizing and analyzing their source in order to prevent them when the survey is conducted on future occasions (Granquist 1997). In the last two decades, this editing modality has been recognized as a key element in E&I strategies. However, its principles are heuristics. By and large, selective editing comprises four stages (Lawrence and McKenzie 2000), namely (i) the construction of anticipated values $\hat{y}_k$ for each sample unit k according to an editing model; (ii) the construction of local score functions; (iii) the construction of a global score function; and (iv) the choice of cut-off values below which no further unit is selected. In general terms, the rationale is that those questionnaires k with a large discrepancy between the anticipated values $\hat{y}_k$ and the reported values y_k will be selected.

As a first general remark, our proposal can be succinctly described using the recent taxonomy of data editing functions by Pannekoek et al. (in this issue). They identify six types of editing tasks, called editing functions, according to the accomplishment of either error detection only (as data quality verification or field/record selection) or also including error treatment. These six editing functions are (i) rule checking, (ii) compute scores, (iii) field selection, (iv) record selection, (v) amend observations, and (vi) amend unit properties (see Pannekoek et al. in this issue for details). In this context, our proposal is to be understood as a record selection editing function.

We set out two general principles to approach selective editing (Arbués et al. 2012b). In keeping with Latouche and Berthelot (1992), who stated that "*in the development of an effective recontact and follow-up strategy, we have to minimize the amount of resources used without affecting the overall data quality and timeliness of the survey*", we claim that

i) editing must minimize the amount of resources deployed to recontacts, follow-ups and interactive tasks, in general;
ii) data quality must be ensured.

This framework is ample enough to give room to the preceding score function approach, but its rigorous derivation seems difficult to us. In this article we propose a mathematical translation of these principles into a general optimization problem, whose solution is the selection of units. In our formulation, interactive editing resources are tantamount to the number of selected questionnaires, whereas data quality is reduced to the accuracy of estimators. Thus a general optimization approach is to minimize the number of selected units, subjected to bounds on loss functions defined for a chosen number of variables of interest. These loss functions may be targeted at the bias, mean squared error (MSE), variance or other measures of the estimation uncertainty. They may be heuristic in nature, such as the so-called pseudo-bias related measures traditionally used for score functions, or they may be explicitly derived under some measurement-error models that are suitable for the data. One example is the contamination model (Di Zio and Guarnera in this issue), which is specified in terms of the full distribution of the true data and the conditional distribution of the observations given the true data.

Two versions of the optimization problem are provided, corresponding to the two typical scenarios for the implementation of selective editing. In the first case, selection is carried out unit by unit, in such a way that whether a given unit is selected or not does not depend on the selection of the other units. This mode of execution is suitable for input editing, where in principle the selection can be made in real time on arrival of each questionnaire. We refer to this as the stochastic optimization problem, because the real-time performance of the solution can only be established with respect to hypothetical repetitions of the selection process. In the second case, selection is carried out jointly for all (or a group of) units. This mode of execution is suitable for output (or macro) editing, which takes place at a later stage of the data collection after a sufficient number of observations have become available. We refer to this as the combinatorial optimization problem, where the performance of the solution can be established conditional on the actual sample observations under some specified measurement-error model.

Selection of units does not produce an order of priority by which the units are sorted according to their respective "urgency" to be edited. But prioritization of units is helpful for coping with the contingency of editing fieldwork. It is intrinsically related to selection since it should be possible in some sense to regard the highest prioritized unit as the optimal selection of a single unit, the second highest prioritized unit as the optimal selection of a single unit given that the highest prioritized unit has been selected, and so on. The combinatorial optimization problem can be adapted to yield prioritization. Not only is this a useful variation for practice, but sometimes it is theoretically necessary for obtaining a unique optimization solution, as we shall explain.

To perform a comparison with any other selective editing technique, we propose a selection efficiency measure. The rationale of this measure is to choose as an input the number of units to select and to compare our selection with an averaged random selection of this number of units. The comparison is based on the reduction of the absolute relative pseudo-bias of the survey estimators. In our view, the sooner the influential units are selected (hence the faster the reduction of the absolute relative pseudo-bias), the more efficient the technique will be. We perform a comparison with some score functions in the literature (Latouche and Berthelot 1992) using real data from the Spanish Industrial Turnover Index (ITI) and Industrial New Orders Received Index (INORI) survey.

The article is organized as follows. In section 2 we formulate the generic optimization problem as a mathematical translation of the above two principles. After fixing the notation and setting out the problem in general terms in Subsection 2.1, we show how the choice of the actual information used in this problem drives us either to a stochastic optimization version (Subsection 2.2) or to a combinatorial optimization version (Subsection 2.3). In Section 3 we show the general principles of the construct of any observation-prediction model, as well as a general proposal for continuous variables. In Section 4 we deal with the editing fieldwork and show how to choose the bounds and how to go from the selection to the prioritization of units under the combinatorial optimization approach. In Section 5 a selection efficiency measure is proposed and a comparison with several score functions is carried out using real data from the Spanish ITI and INORI survey. Finally we include an ample discussion in Section 6 in an attempt to assess this proposal in the current framework of selective editing with score functions.

2. The Optimization Problem

Before identifying the variables, the objective function and the restrictions of our optimization problem, we need to introduce the following notation. The sampling design according to which a probability sample s is selected will be denoted by $p(\cdot)$. The sample size will be denoted by n and the corresponding sampling weights by w_{ks}. The sample dependence of the sampling weights implicitly assumes that they do not need to be the design weights. For example, in a ratio estimator of the form $\hat{Y}^{rat} = X \cdot \frac{\hat{Y}^{HT}}{\hat{X}^{HT}}$, where x is a known auxiliary variable from the sampling frame, $X = \sum_{k \in U} x_k$ is a known population total, and $\hat{Y}^{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$ (analogously for $\hat{X}^{HT}$) stands for the Horvitz-Thompson estimator of the population total $Y = \sum_{k \in U} y_k$, the sampling weights are given by $w_{ks} = \frac{X}{\hat{X}^{HT}} \frac{1}{\pi_k}$, where π_k is the first-order inclusion probability for unit k. More complex situations are embedded under this notation. The true, observed and edited values of a variable $y^{(q)}$, $q = 1, \ldots, Q$ (for ease of notation we drop the superscript (q) hereafter except when strictly necessary), for unit k will be denoted, respectively, by y_k^0, y_k and y_k^*. We assign a binary variable $r_k \in \{0, 1\}$ to each unit k to indicate whether it is selected ($r_k = 0$) or not ($r_k = 1$). The vector $\mathbf{r} = (r_1, \ldots, r_n)^t$ for the whole sample will be referred to as the *selection strategy*. The counterintuitive assignment allows us to relate the preceding three values by the equation $y_k^*(\mathbf{r}) = (1 - r_k) \cdot y_k^0 + r_k \cdot y_k$, where we have made explicit the dependence of the edited values upon the selection strategy. Note that we are implicitly assuming that the editing work drives us from the observed to the true values. If we denote the corresponding measurement error by $\boldsymbol{\epsilon}_k = y_k - y_k^0$, then we can write $y_k^*(\mathbf{r}) = y_k^0 + r_k \boldsymbol{\epsilon}_k$. Note that these edited values are in fact those to be plugged into the survey estimators at this point of the E&I strategy. That is, if we are to estimate the population domain total $Y_{U_d} = \sum_{k \in U_d} y_k^0$ (for ease of notation we will drop the subscript U_d hereafter), then we denote the corresponding chosen estimator by $\hat{Y}^*(\mathbf{r}) = \sum_{k \in s_d} w_{ks} y_k^*(\mathbf{r})$. However, note that this estimator will not be the final estimator after the whole E&I strategy has been executed. Some later procedures such as weight adjustment or outlier treatment may follow. The selection of units proposed herein divides the sample into a critical and a noncritical stream, the treatments of which are decided by the statistician. We will restrict ourselves to population totals and linear estimators. All auxiliary covariates not included in the questionnaire for unit k will be denoted by $\mathbf{x}_k$.

So far the preceding variables are numeric. To use statistical modelling techniques, we promote these numeric variables to random variables according to a model m in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. As usual, this promotion will not be specifically indicated in the notation, except for the selection strategy, so that $\mathbf{R}$ will denote the random selection strategy, and $\mathbf{R}(w) = \mathbf{r}$, with $w \in \Omega$, will be a particular numeric realization called the *selection*. A predicted value of variable y_k according to the chosen model m will be denoted by $\hat{y}_k$. Note that the statistical model m embraces all promoted random variables different from the probability sample s itself. When random variables are used in survey estimators, we write indistinctly $\hat{Y}^0 = \sum_{k \in s_d} w_{ks} y_k^0$, $\hat{Y} = \sum_{k \in s_d} w_{ks} y_k$ and $\hat{Y}^*(\mathbf{R}) = \sum_{k \in s_d} w_{ks} y_k^*(\mathbf{R})$ for the survey estimators targeted at Y. We will denote by $\mathbf{Z}$ the set of random variables actually used by the statistician to select the units in the E&I strategy.

In particular, we will consider two options, namely, either $\mathbf{Z} = \mathbf{Z}^{long} \equiv s$ or $\mathbf{Z}^{long} \equiv \{s, \mathbf{X}\}$ for the stochastic problem (see below for the difference) or $\mathbf{Z} = \mathbf{Z}^{cross} \equiv \{s, \mathbf{X}, \mathbf{Y}\}$

for the combinatorial version. When this cross-sectional information is restricted to unit k, we shall write accordingly $\mathbf{Z}_k^{cross} = \{s, \mathrm{x}, y_k\}$. The use of information is represented as conditioning upon the corresponding random variables. The auxiliary covariates $\mathbf{X}$ are chosen by the statistician according to the chosen statistical model to be used in the problem (see below). They play a similar role to the auxiliary variables in the sampling design or the known auxiliary variables in the weight calibrating process. Indeed, they may coincide partially or totally with these auxiliary variables used in other parts of the estimation process.

2.1. The General Optimization Problem

As stated in the introduction, we want to minimize the number of questionnaires to edit provided that the chosen loss functions of the survey estimators $\hat{Y}^*$ targeted at the population total Y are bounded. To formally set up the optimization problem we need (i) the variables, (ii) the function to optimize, and (iii) the restrictions. Apart from identifying these elements, it is important to show how the available information enters into the formulation of the problem.

The ultimate variables are the selection strategy $\mathbf{r}^T = (r_i, \ldots, r_n)$ for the sample units $s = \{1, \ldots, n\}$, where $r_k = 0$ if the unit k is selected and $r_k = 1$ otherwise. However, since the measurement error $\boldsymbol{\epsilon}_k = y_k - y_k^0$ is conceived to be random in nature conditional on the realized sample s, and given the available information $\mathbf{Z}$ chosen to make the selection of units, this selection can vary depending on the realized $\mathbf{y}$, $\mathbf{y}^0$ and $\mathbf{Z}$. Thus let $\mathbf{R}$ denote the stochastic selection strategy so that (i) $\mathbf{R}(w) = \mathbf{r}$ is a realized selection and (ii) $\mathbb{E}_m[\mathbf{R}|\mathbf{Z}]$ is the vector of probabilities of nonselection under the specific model m given the chosen information $\mathbf{Z}$. The objective function to optimize, given the information $\mathbf{Z}$, is then written as $\mathbb{E}_m[\mathbb{1}^T\mathbf{R}|\mathbf{Z}]$, whose maximization amounts to minimizing the number of selected units.

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 $L = L^{(2)}(a, b) = (a - b)^2$. Then it is straightforward to prove (see appendix A) that $\mathbb{E}_m[L^{(r)}(\hat{Y}^*(\mathbf{R}), Y)|\mathbf{Z}] \leq \eta$ warrants $\mathbb{E}_{pm}[L^{(r)}(\hat{Y}^*(\mathbf{R}), Y)] \leq \left(\eta^{1/r} + \mathbb{E}_{pm}^{1/r}\left[L(\hat{Y}^0, Y)\right]\right)^r$, where $O(\cdot)$ stands for the well-known big O. In other words, each constraint controls the loss of accuracy in terms of the chosen loss function L due to nonselected units, up to sampling design variability.

For these loss functions, each constraint can always be written as a bound on a quadratic form, denoted by $\mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$ (see Appendix A). Particular forms suitable for the stochastic and combinatorial problems will be explained in Subsection 2.2 and 2.3. The $n \times n$ matrix Δ 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 Appendix A for details). Thus, if Δ is diagonal with entries

$|w_{ks}\epsilon_k|$, then we are choosing the absolute loss so that $\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0)|\mathbf{Z}]$ is also bounded by η (up to sampling design factors). This is targeted at the bias. Similarly, if $\Delta_{kl} = w_{ks}w_{ls}\epsilon_k\epsilon_l$, then we are choosing the squared loss so that $\mathbb{E}_m[(L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0)|\mathbf{Z}]$ is also equally bounded. In turn, this is targeted at the mean squared error. In both cases, model-based techniques using data from the current time period can be applied in the combinatorial version, whereas in the stochastic version we are obliged to resort to auxiliary information from other periods.

For instance, the (local) score for a given y-variable is usually conceived as the product of a "risk" component and an "influence" component. A generic measure can be given using a model-based approach. Let $p_k = P\left(y_k^0 \neq y_k|y_k\right)$, that is, the posterior probability that the true value is different from the observed one. Let $\tilde{\mu}_k = \mathbb{E}_m\left(y_k^0|y_k, y_k^0 \neq y_k\right)$, that is, the conditional expectation of the true value given that it is different from the observed one. Then, we have

$$\mathbb{E}_m\left(y_k^0|y_k\right) = (1 - p_k)y_k + p_k\tilde{\mu}_k \quad \text{and} \quad \delta_k = y_k - \mathbb{E}_m\left(y_k^0|y_k\right) = p_k(y_k - \tilde{\mu}_k)$$

It follows that $w_k\delta_k$ can be used to construct the local score of unit k with respect to y, which is the product of "risk" measured by p_k and "influence" measured by $w_k(y_k - \tilde{\mu}_k)$, where w_k can be the sample weight, for example. Di Zio and Guarnera (in this issue) derive such a measure under the contamination model, which is suitable for the combinatorial problem. For the stochastic problem, where scoring does not use observations other than the unit at hand, $\tilde{\mu}_k$ cannot be evaluated for the current sample data and instead information from preceding realizations of this survey or similar surveys must be used. It is customary to replace it with some reference value, such as y_k from a previous time point, giving rise to a pseudo-bias. Nor can the "risk" component be assessed properly, and some heuristics measure might be used, such as in the SELEKT approach of Statistics Sweden (see for example Lindgren 2011). The auxiliary information, which we exploit in the observation-prediction model (see Section 3), is fundamental.

The main difference between both versions arises when considering their actual application. 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 these cross-terms more or less constant over time, hence playing no significative role in the selection. Conversely, the combinatorial problem needs a sufficient number of observations available to carry out the selection jointly for all units.

Taking into account the possibility of multiple constraints, we now arrive at the following general optimization problem:

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

where Ω_0 denotes the admissible outcome space of $\mathbf{R}$, and q refers to the different constraints. Manipulation of Ω_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 Ω_0 such that certain R_k s are fixed at 0. The different constraints

may arise from the fact that there are multiple y-variables of interest, or the constraints may be directed at the different population domains even when there is only a single y-variable. In particular, the loss matrices $\Delta^{(1)}, \ldots, \Delta^{(Q)}$ may all be derived under a single multivariate model for the joint data, even when the bounds are marginally specified for each target quantity on its own.

Variations of the optimization problem stated above are possible, by either adopting a different function for optimization and/or different forms of constraints. For instance, maximization may be changed to minimization as long as suitable alterations of the selection variables and the loss functions are provided. Alternatively, one may for example use $w_k\delta_k$ in Δ but state the constraint as $\mathbb{E}_m[|\mathbf{R}^T\Delta\mathbf{R}|\,|\mathbf{Z}] \leq \eta$. We do not explicitly consider such variations of the problem in this article, but note that (i) it is possible to adapt the solutions presented below, should such variations be desirable in practice, and (ii) the expounded optimization approach can be carried out in the same spirit.

2.2. *The Stochastic Optimization Problem*

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 $\mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\mathrm{diag}(\Delta)|\mathbf{Z}]$. Furthermore, the distinction between $\mathbf{Z}^{long} = s$ and $\mathbf{Z}^{long} \equiv \{s, \mathbf{X}\}$ is a matter of choice. In the former case, the restrictions are required to be fulfilled only on average for all realizations of the survey, whereas in the latter case they are imposed on the current realization, given the realizations of preceding time periods. 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 $\boldsymbol{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$. This dependence on $\mathbf{Z}^{cross}$ may seem misleading, but only momentarily. 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 upon $\mathbf{Z}^{cross}$ reduces effectively to conditioning upon the information $\mathbf{Z}_k^{cross} = \{s, \mathbf{x}, \mathbf{y}_k\}$ of each unit. Thus we write $\boldsymbol{M}^{(q)} = \mathbb{E}_m[\Delta|\mathbf{Z}^{cross}] = \mathrm{diag}\left(\mathbb{E}_m\left[\Delta_{kk}^{(q)}|\mathbf{Z}_k^{cross}\right]\right) = \mathrm{diag}\left(M_{kk}^{(q)}\right)$. On the other hand, in order to obtain the optimal Lagrange multipliers λ_q^* 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 $\boldsymbol{M}^{(q)}$:

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

Note that since the scheme is "trained" on the historic data, the evaluation of $M_{kk}^{(q)}$ given the observations in the current sample necessarily yields a pseudo-measure, regardless of the definition of the loss matrices.

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 $\sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)}$ is the standardized global score, with the generic global threshold value 1.

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.

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 $\mathbf{r}$, and the realized loss $\sum_{k=1}^{n} r_k M_{kk}^{(q)}$, which can either be higher or lower than the specified bound η_q, for $q = 1, \ldots, Q$. Upon any hypothetical repetition of the selection process, however, y_k and y_k^0 will vary, and so will the corresponding $M_{kk}^{(q)}$ and r_k. It is over such hypothetical repetitions that the constraint $\mathbb{E}_m[\mathbf{R}\Delta^{(q)}\mathbf{R}|\mathbf{Z}] \leq \eta_q$ can possibly be satisfied, but not for each particular realization of the selection process.

2.3. The Combinatorial Optimization Problem

The combinatorial problem deals with the selection among all (or a group of) units. Cross-unit terms are now allowed and the information actually used is that given by the sample s, the auxiliary covariates $\mathbf{X}$ and the variables of interest $\mathbf{Y}$, that is by $\mathbf{Z} = \mathbf{Z}^{cross}$. Notice that all this information is available only after all questionnaires have been collected, thus it is only applicable as a form of output editing. It is easily proved that each constraint reduces to $\mathbb{E}_m[\mathbf{R}^T\Delta^{(q)}\mathbf{R}|\mathbf{Z}^{cross}] = \mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r}$, where $\mathbf{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$, which can now be possibly evaluated under some measurement-error model. Consequently, it becomes possible to establish the performance of the realized selection strategy directly. The optimization problem can be rephrased as

$$
\begin{aligned}
[P_{co}(\mathbf{M}, \eta, \Omega_0)] \quad & \max \mathbf{1}^T\mathbf{r} \\
& \text{s.t.} \quad \mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r} \leq \eta_q, \quad q = 1, 2, \ldots, Q, \\
& \mathbf{r} \in \Omega_0
\end{aligned}
$$

Note that a more direct derivation can be obtained by not promoting the selection strategy vector $\mathbf{r}$ to a random vector $\mathbf{R}$ when modelling the measurement errors.

This combinatorial problem is solved in two different forms using two greedy algorithms, which run in $n^4 \cdot Q$ and $n^3 \cdot Q$ times, respectively. The solution of both algorithms is not exact a priori but suboptimal with a good degree of approximation. The faster algorithm is noticeably less precise than the slower one. This lack of precision entails a small amount of overediting in practice, that is, more units than those optimally obtained will be selected. The fourth and third power dependence on n may appear discouraging for practical applications. However, firstly, the input size P in problem P_{CO} is actually $P = O(n^2)$, thus the algorithms run in $O(P^2)$ and $O(P^{3/2})$, which are acceptable speeds for combinatorial problems. Secondly, in practice the problem is intended to be

applied not to entire samples but to their breakdowns into publication cells, which are the figures upon which precision is called for (see Section 6). These heuristic algorithms locally search the optimum in each iteration until the current solution satisfies all the restrictions. To do this we introduce infeasibility functions $h_i(\mathbf{r})$ for each algorithm $i = 1$, 2 (see Salgado et al. 2012 for details) indicating whether a solution satisfies all the restrictions ($h(\mathbf{r}) = 0$) or not ($h(\mathbf{r}) > 0$). Both algorithms start from the initial solution $\mathbf{r} = 1$ and in each iteration select the next unit in a locally optimal way until all restrictions are satisfied. The infeasibility functions will also be used later when constructing the prioritization of units.

Finally, we can regard both versions as related to two different approaches to the problem of optimization under uncertainty (see e.g., Wets 2002). The combinatorial version is consistent with the wait-and-see approach, since it puts off all decisions until all the information is available. The stochastic version is, at least partially, a here-and-now approach, since the decision about the procedure or rule of selection (although not the selection itself) is made before the data collection.

3. The Observation-Prediction Model

To substantiate the constraints in both versions of the optimization problem, we need to compute the loss matrices $\mathbf{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$ and to choose the bounds η_q. We now show how to undertake the former whereas the latter is dealt with in the next section.

To compute the loss matrices we make use of the standard model-based techniques, but not in a conventional way. Let us digress very briefly. When facing the editing tasks and, in particular, the selection of units, one resorts to the very best auxiliary information available at that precise moment. With full generality, this will comprise (i) the reported values of the variables of analysis $\mathbf{y}_k^{(t)}$ for the present ($t = T$) and preceding ($t < T$) time periods, (ii) the true values of these variables $\mathbf{y}^{(0,\,t)}$ for those edited units in the past $t < T$, (iii) and the values of auxiliary covariates $\mathbf{x}_k^{(t)}$ for all time periods. In the notation of preceding sections, we have $\mathbf{y}_k = \mathbf{y}_k^{(T)}$, $\mathbf{y}_k^0 = \mathbf{y}_k^{(0,T)}$ and $\mathbf{x}_k = \mathbf{y}_k^{(t_1)}, \mathbf{y}_k^{(0,t_1)}, \mathbf{x}_k^{(t_2)}$, with $t_1 < T$ and $t_2 \leq T$. Note that some of these values can be coincidentally equal (e.g., when the measurement error is null) and that $\mathbf{y}_k^0$ is only known after accomplishing the editing work. But this is not everything. We also know (at least we can know) a point prediction $\hat{y}_k$ for each y-variable based on these auxiliary variables. For instance, we can make use of a time series model $\left\{y_k^{(0,t)}\right\}_{t<T}$ to make a point prediction $\hat{y}_k^{(T)}$. Different choices arise depending on the amount and type of auxiliary information. These predictions will enter into the selection problem as auxiliary covariates, so that $\mathbf{x}_k = \mathbf{y}_k^{(t_1)}, \mathbf{y}_k^{(0,t_1)}, \hat{\mathbf{y}}_k^{(T)}, \mathbf{x}_k^{(t_2)}$, with $t_1 < T$ and $t_2 \leq T$.

Let us denote by m^* the auxiliary model used to make the predictions $\hat{\mathbf{y}}_k$, not to be confused with the measurement error model m considered throughout this paper. This measurement error model m is given as usual in terms of (i) the conditional distribution of the predicted values $\mathbf{y}$ upon the true values $\mathbf{y}^0$, and (ii) the distribution of $\mathbf{y}^0$ conditional on the available auxiliary information $\mathbf{X}$. To be specific, for a y-variable we will assume $y_k = y_k^0 + \epsilon_k^{obs}$ and $y_k^0 = \hat{y}_k + \epsilon_k^{pred}$. In other words, we are using the predicted value computed according to the auxiliary model m^* as an exogenous variable for the model regarding y^0. In this sense we refer to this proposal as an observation-prediction model.

Generalizing these ideas, let us consider

i) an observation model $\mathbb{P}_{obs|0}(\mathbf{y}|\mathbf{y}^0)$, that is, a conditional probability distribution for the observed values $\mathbf{y}$ given the true values $\mathbf{y}^0$;
ii) a prediction model $\mathbb{P}_{0|pred}(\mathbf{y}^0|\hat{\mathbf{y}})$, that is, a conditional probability distribution for the true values $\mathbf{y}$ given the predicted values $\mathbf{y}^0$ according to an auxiliary model m^*.

Now let us denote by $\mathbb{P}_{obs|pred}$ the probability distribution of $\mathbf{y}$ conditional on the predicted values $\hat{\mathbf{y}}$ and by $\mathbb{P}_{0|obs,pred}$ the probability distribution of the true values $\mathbf{y}^0$ conditional on the observed values $\mathbf{y}^{obs}$ and the predicted values $\hat{\mathbf{y}}$. Then by Bayes' theorem or a generalization thereof, we can write

$$\mathbb{P}_{0|obs,pred} = \frac{\mathbb{P}_{obs|0} \times \mathbb{P}_{0|pred}}{\mathbb{P}_{obs|pred}} \tag{2}$$

The product must be understood in a suitable generalized form when the distributions are completely general. As usual, if the probability distributions are absolutely continuous with density functions $f.(\cdot)$, Equation (2) can be easily recognized as

$$f_{0|obs,pred}(\mathbf{y}^0) = \frac{f_{obs|0}(\mathbf{y}|\mathbf{y}^0, \hat{\mathbf{y}}) f_0(\mathbf{y}^0|\hat{\mathbf{y}})}{\int_{\mathbb{R}^Q} f_{obs|0}(\mathbf{y}|\mathbf{y}^0, \hat{\mathbf{y}}) f_0(\mathbf{y}^0|\hat{\mathbf{y}}) \mathrm{d}\mathbf{y}^0}.$$

The discrete case also boils down to applying Bayes' theorem. Once we have the distribution $\mathbb{P}_{0|obs,pred}$, the loss matrices can be computed as

$$\mathbf{M}^{(q)} = \mathbb{E}_{0|obs,pred}[\Delta^{(q)}|S, \mathbf{Y}, \hat{\mathbf{Y}}]. \tag{3}$$

To illustrate this proposal, let us consider the following generic example with a continuous variable y. Let us define the observation model $y_k^{obs} = y_k^0 + \epsilon_k^{obs}$ and the prediction model $y_k^{obs} = \hat{y}_k + \epsilon_k^{pred}$, with the following specifications:

1. $\epsilon_k^{obs} = \delta_k^{obs} e_k$.
2. $e_k \simeq Be(p_k)$, where $p_k \in (0, 1)$.
3. $\left(\epsilon_k^{pred}, \delta_k^{obs}\right) \simeq N\left(\mathbf{0}, \begin{pmatrix} \nu_k^2 & \rho_k \sigma_k \nu_k \\ \rho_k \sigma_k \nu_k & \sigma_k^2 \end{pmatrix}\right)$.
4. ϵ_k^{pred}, δ_k^{obs} and e_k are jointly independent of $\mathbf{Z}_k^{cross}$.
5. e_k is independent of ϵ_k^{pred} and δ_k^{obs}.

These are equivalent to stating that unit k has a probability $1 - p_k$ of reporting a value without measurement error $\left(y_k = y_k^0\right)$ and, when reporting an erroneous value, the measurement error distributes as a normal random variable with zero mean and variance σ_k^2. On the other hand, the prediction error distributes as a normal random variable with zero mean and variance ν_k^2. Both errors distribute jointly as a bivariate normal random variable with correlation ρ_k. Reporting an erroneous value is independent of both types of errors.

For the time being let us assume that the parameters $\boldsymbol{\theta} = \left(p_k, \sigma_k^2, \nu_k^2, \rho_k\right)^T$ are known. Let us focus on the squared loss function. Then it is easy to prove (Arbués et al. 2012a) that

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)|s_k, y_k, \hat{y}_k\right] = \nu_k \cdot \frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k} \cdot \left(\frac{y_k - \hat{y}_k}{\nu_k}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{4}$$

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)^2|s_k, y_k, \hat{y}_k\right] = \nu_k^2 \cdot \left(\frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^2 \cdot$$

$$\left[\frac{\sigma_k^2\left(1 - \rho_k^2\right)\left(\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k\right)}{\left(\sigma_k^2 + \rho_k\sigma_k\nu_k\right)^2} + \left(\frac{y_k - \hat{y}_k}{\nu_k}\right)^2\right] \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right),$$

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)\left(y_l - y_l^0\right)|s_k, y_k, \hat{y}_k\right] = \mathbb{E}_m\left[\left(y_k - y_k^0\right)|s_k, y_k, \hat{y}_k\right]\mathbb{E}_m\left[\left(y_l - y_l^0\right)|s_k, y_k, \hat{y}_k\right],$$

$$k \neq l,$$

where

$$\zeta_k(x) = \frac{1}{1 + \dfrac{1 - p_k}{p_k}\left(\dfrac{\nu_k^2}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^{-1/2} \exp\left(-\dfrac{1}{2}\dfrac{\sigma_k^2 + 2\rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}x^2\right)}.$$

Should we choose the absolute loss function, then, under the same hypotheses, we would have (see Appendix A):

$$\mathbb{E}_m\left[|y_k - y_k^0||s_k, y_k, \hat{y}_k\right] = \sqrt{\frac{2}{\pi}} \cdot \nu_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; -\frac{(y_k - \hat{y}_k)^2}{2\nu_k^2}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{5}$$

where ${}_1F_1(a; b; x)$ denotes the confluent hypergeometric function of the first kind.

The estimation of the parameters $\boldsymbol{\theta}$ depends on the scenario. For the stochastic problem, as before, we are obliged to use some reference values or heuristic measures. Once more we resort to the auxiliary information. Our choice depends very much on the amount and type of auxiliary information. From the historic double-data sets comprising τ past time periods (e.g., a fixed panel) we can compute

$$\hat{p}_k = \frac{1}{\tau}\sum_{t=1}^{\tau} I_{y_k^{(t)} \neq y_k^{(0,t)}},$$

$$\hat{\sigma}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)^2,$$

$$\hat{\nu}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)^2, \text{ where } \epsilon_k^{(t)} = \hat{y}_k^{(t)} - y_k^{(0,t)},$$

$$\hat{\rho}_k = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right).$$

In case of rotating panels or sampling designs with too short a continuity in the sample for a number of units, we are forced to make simplifying assumptions such as partitioning

the sample $s = \cup_{i=1}^{I} s_i$ and positing $\boldsymbol{\theta}_k = \boldsymbol{\theta}_i$ if $k \in s_i$. We can also adopt these assumptions for some of the parameters. The extreme case would $\boldsymbol{\theta}_k = \boldsymbol{\theta} = (p, \sigma^2, \nu^2, \rho)^T$ for all $k \in s$, which can be further supplemented with extra hypotheses such as $\rho = 0$.

On the other hand, for the combinatorial problem we do have (almost) the complete current sample so that we can make use of these data, although with important limitations. It is clear that it is impossible to estimate each $\boldsymbol{\theta}_k$ using only the current sample. We are obliged to make some simplifying assumptions, as above. In practice, however, it is advisable to use not only data from the current time period ($t = T$), but also from preceding periods ($t < T$). The stationarity across time periods of the response mechanism supports this course of action.

Alternatively, the contamination model by Di Zio and Guarnera (in this issue) is a relevant example of a model-based technique which uses exclusively data from the current period (except for the covariates for the model) to estimate the model parameters. The usage of statistical models to make the selection of units allows us to cherish the hope of extending this approach to qualitative and semicontinuous variables, thus paving the way for the use of selective editing in household surveys.

4. Fieldwork: Selection and Prioritization of Units

The problem is not completely specified until we choose the bounds η_q to formulate the optimization problem completely. The bound η on a given constraint $\mathbb{E}_m[\mathbf{R}^t \Delta \mathbf{R} | \mathbf{Z}] \leq \eta$ can be set either absolutely or relatively in terms of a chosen figure of merit or reference value. This can be, for example, the a priori variance used in the sampling design phase so that the constraint establishes a bound for the loss of accuracy as a fraction of the desired precision. The decision will necessarily involve some subject-matter knowledge.

So far, the formulation of the selective editing problem as an optimization problem is complete, providing a *selection* of units expressed by the solution $\mathbf{r}$. However, in practice having a selection of units must be confronted with the actual conditions of fieldwork. In particular, both controllability and availability of resources, such as person hours for example, are important issues in this respect. Given a particular selection, either we may run out of resources and cannot edit all selected units or we may finish the editing field work ahead of time and thus miss the opportunity to achieve better accuracy. In this sense it seems natural to have at our disposal a set of selections to optimize the actual use of resources. We achieve this by having a *prioritization* of units. Next we show how to prioritize units in the optimization approach. In Section 6 we discuss in more detail this issue of the selection/prioritization of units in relation with the fieldwork.

From the preceding sections it is clear that it does not make sense to prioritize units in the stochastic formulation. On the other hand, to prioritize units in the combinatorial version we propose combining different selections by choosing a sequence of appropriate values as bounds. The basic idea is to choose large initial bounds which drive us to select no unit, then to decrease the bounds until one unit is selected and to flag this unit for future selections. Then we again decrease the bounds until a new unit is selected and flagged for future selections. The procedure is repeated until all units have been flagged.

Let $f^{[k]} \subset s = \{1, \ldots, n\}$ denote the set of flagged units at iteration k and $\Omega_0^{[k]}$ the outcome space of the combinatorial problem at iteration k. For any given strategy vector $\mathbf{r}$

we denote by $\mathcal{I}^{-1}(\mathbf{r})$ the set of strategy vectors $\bar{\mathbf{r}}$ obtained from $\mathbf{r}$ transforming exactly a component 1 into 0. For example, $\mathcal{I}^{-1}((1,1,0)^T) = \{(0,1,0)^T, (1,0,0)^T\}$. Let h denote the infeasibility function used in the greedy algorithms (see Subsection 2.3).

The algorithm of prioritization reads as follows:

1. Set $f^{[0]} = \emptyset$, $\Omega_0^{[0]} = \{0,1\}^{\times n}$, $\mathbf{s}^{[0]} =$ and $\boldsymbol{\eta}^{[0]} = (\mathbf{s}^{[0]T}M^{(1)}\mathbf{s}^{[0]}, \ldots, \mathbf{s}^{[0]T}M^{(Q)}\mathbf{s}^{[0]})^T$.
2. FOR $k = 0$ TO $k = n$
   i. Set $\mathbf{s}^{[k+1]} = \arg\min_{\mathbf{s}\in\mathcal{I}^{-1}(\mathbf{s}^{[k]})}(h(\mathbf{s}))$. In case of multiple $\mathbf{s}^{[k+1]}$ choose one at random.
   ii. Set $l^* \in s$ such that $s_{l^*}^{[k+1]} \neq s_{l^*}^{[k]}$.
   iii. Set $f^{[k+1]} = f^{[k]} \cup \{l^*\}, \Omega_0^{[k+1]} = \Omega_0^{[k]} - \{\mathbf{s}^{[k]}\}$ and $\boldsymbol{\eta}^{[k+1]} = (\mathbf{s}^{[k+1]T}M^{(1)}\mathbf{s}^{[k+1]}, \ldots, \mathbf{s}^{[k+1]T}M^{(Q)}\mathbf{s}^{[k+1]})^T$.
3. FOR $k = 0$ TO $k = n$
   i. Set $\mathbf{r}^{[k]} = \arg\max[P_{co}(\mathbf{M}, \boldsymbol{\eta}, \Omega_0^{[k]})]$.
4. Set $\boldsymbol{s} = \sum_{k=0}^{n}\mathbf{r}_k$.

The vector $\mathbf{s}$ provides the prioritization: unit k must be edited in the s_kth place. Notice that steps 1 and 2 provide a sequence of bounds $\boldsymbol{\eta}^{[k]}$ and a sequence of outcome sets $\Omega_0^{[k]}$ which are used in step 3 to solve $n+1$ concatenated combinatorial problems. Two comments are in place here. On the one hand, in practice, Step 3 indeed reduces to the first point in Step 2 since $\mathbf{r}^{[k]} = \mathbf{s}^{[k]}$ because h is the infeasibility function of the optimization algorithm.

On the other hand, this invites us to reconsider the role of the infeasibility function in the prioritization of units: this depends on the choice of h. Should we choose, instead of the original infeasibility function $h_1(\mathbf{r}) = \sum_{q=1}^{Q}\left(\mathbf{r}^t M_{kl}^{(q)}\mathbf{r} - m_q^2\right)^+$ of algorithm 1, the function $h(\mathbf{r}) = \sum_{q=1}^{Q} w_q\left(\mathbf{r}^t M_{kl}^{(q)}\mathbf{r} - m_q^2\right)^+$, where $w_q \geq 0$ are positive weights expressing the different priority given to the accuracy of each variable $y^{[q)}$, we would arrive at a different prioritization. This can also be viewed more geometrically. To produce a sequence of bounds we begin by having no selected units, that is, by $\boldsymbol{\eta}_0 = (\mathbb{1}^t M^{(1)}\mathbb{1}, \ldots, \mathbb{1}^t M^{(Q)}\mathbb{1})^t$, and we need to produce a sequence of points in $\mathbb{R}^Q$ such that its final point is $\mathbf{0}$. There exist infinitely many possibilities (see Figure 1). In this context, the prioritization amounts to choosing a path from $\boldsymbol{\eta}_0$ to $\mathbf{0}$. This path expresses the priority which the statistician gives to the accuracy of the different estimators along the process of prioritization of units. The original infeasibility functions of the algorithms confer the same relevance on every estimator $\hat{Y}^{(q)}$.

5. A Selection Efficiency Measure: Comparison with the Score Function Approach

To make a comparison of the selection undertaken under any approach, we propose the following selection efficiency measure for an estimator $\hat{Y}$. Beforehand, we need a double data set with raw and edited values according to a gold standard so that when a unit is selected, its raw values are substituted by their corresponding edited counterparts, considered true. We will denote by $\hat{Y}^{sel}(n_{ed})$ the estimator obtained when n_{ed} questionnaires have been selected according to a selective editing technique *sel* and edited correspondingly. Note that $\hat{Y}^{sel}(n_{ed} = n) = \hat{Y}^0$. As a figure of merit for the

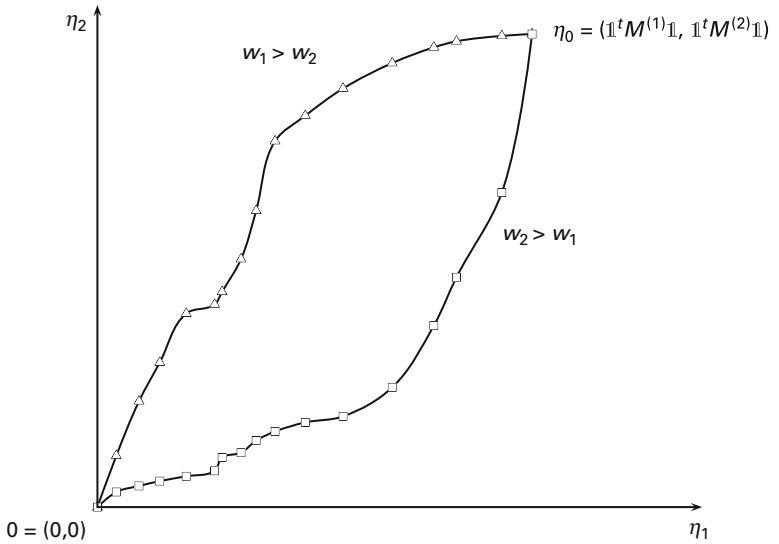

Fig. 1. Example of two different sequences of bounds with $Q = 2$ arising from different choices of the weights w_q.

selection of units we will focus upon the absolute relative pseudo-bias of an estimator $\hat{Y}$, given by $\widetilde{\mathrm{ARB}}\ (\hat{Y}^{sel}(n_{ed})) = \left|\frac{\hat{Y}^{sel}(n_{ed}) - \hat{Y}^0}{\hat{Y}^0}\right|$.

The rationale of the proposed measure is the comparison with a random selection of units. The idea is to compare $\widetilde{\mathrm{ARB}}(\hat{Y}^{sel}(n_{ed}))$ for a selective editing technique *sel* with $\widetilde{\mathrm{ARB}}(n_{ed}) \equiv \widetilde{\mathrm{ARB}}(\mathbb{E}[\hat{Y}^{ran}(n_{ed})])$, where *ran* stands for an equal-probability selection and $\mathbb{E}$ is the expectation with respect to this random selection. It is immediate to show that $\widetilde{\mathrm{ARB}}_0\,(n_{ed}) = \left(1 - \frac{n_{ed}}{n}\right)\widetilde{\mathrm{ARB}}(\hat{Y}^{ran}(0))$. Let us denote by $\gamma_0(n_{ed})$ and $\gamma^{sel}(n_{ed})$ the straight and polygonal lines with vertices $\gamma_0(n_{ed}) \simeq \{(0, \widetilde{\mathrm{ARB}}_0\,(0)), (n_{ed}, \widetilde{\mathrm{ARB}}_0)(n_{ed}))\}$ and $\gamma^{sel}(n_{ed}) \simeq \{(0, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(0))), (1, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(1))), \ldots, (n_{ed}, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(n_{ed})))\}$, respectively. Let us also denote by $A_\gamma(n_{ed})$ the signed area of the surface between the curve γ and the horizontal axis to the left of the vertical line at n_{ed} (see Figure 2). The area is agreed to be positive if the polygonal line lies below the straight line and is otherwise negative. We propose the following definition for the efficiency of the technique *sel*:

$$\epsilon^{sel}(n_{ed}) \equiv (A_{\gamma_0}(n_{ed}) - A_{\gamma^{sel}}(n_{ed}))/A_{\gamma_0}(n_{ed}) = 1 - \frac{A_{\gamma^{sel}}(n_{ed})}{A_{\gamma^0}(n_{ed})}.$$

Note that this measure depends on the number of units to select. This allows us to recognize those techniques which prioritize the most influential units first. A typical situation is depicted in Figure 2.

We have carried out a comparison of the preceding proposal of prioritization of units with that obtained from some score functions in the literature. In order to avoid possible interferences with missing data and units recently added to the sample, we have used a rectangular subset of the sample data of the Spanish ITI and INORI surveys (INE Spain 2010). For clarity's sake we shall concentrate on one particular score function, illustrate the corresponding results and make some comments regarding the similar behavior of all of them. We have used a slightly enhanced version of the RATIO function of Latouche and

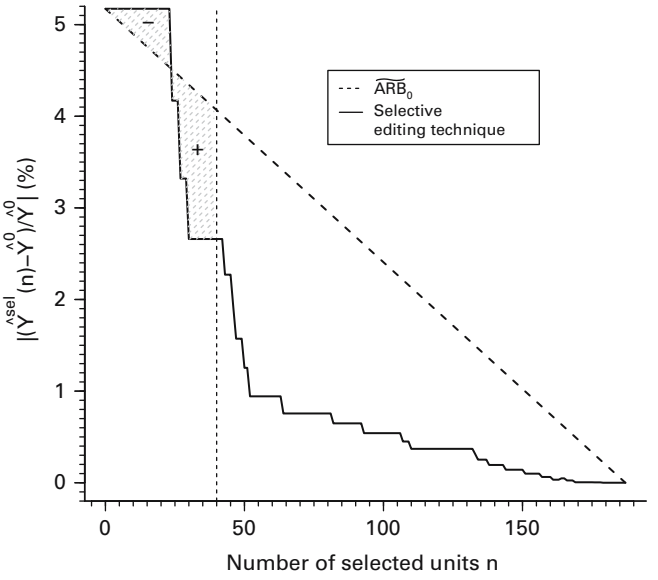

Fig. 2. Absolute relative pseudo-bias vs. number of selected units

Berthelot (1992). Let $r_k^{(t)} = \frac{y_k^{(t)}}{y_k^{(t-1)}}$ and define

$$\bar{r}_k^{(t)} = \begin{cases} \left| \dfrac{r_k^{(t)}}{\text{median}_k\left(r_k^{(t)}\right)} - 1 \right| & \text{if } r_k^{(t)} > \text{median}_k\left(r_k^{(t)}\right), \\ \left| 1 - \dfrac{r_k^{(t)}}{\text{median}_k\left(r_k^{(t)}\right)} \right| & \text{otherwise.} \end{cases}$$

Also define $g_k^{(t)} = w_{ks} \times \bar{r}_k^{(t)} \times \sqrt{\max\left(y_k^{(t)}, y_k^{(t-1)}\right)}$ and then the local score $s_k^{(t)} = \frac{|g_k^{(t)} - \text{median}_k\left(g_k^{(t)}\right)|}{\text{IQR}_k\left(g_k^{(t)}\right)}$, where IQR stands for the interquartile range. For $q = 1, \ldots, Q$ variables, these combine in the global score function defined as RATIO2$(k,t) = S_k^{(t)} = \sum_{q=1}^{Q} s_k^{(q,t)}$. The enhancement arises due to the fact that only data from the time period $t-1$ is used and not from $t-2$, as in the original proposal. Thus this function RATIO2 can only be used as a form of output editing after all data have been collected (as the combinatorial approach, which we are making the comparison with).

Regarding the prioritization of units computed under the combinatorial approach, firstly we must specify the auxiliary model m^* to find the predicted values $\hat{y}_k$. For each unit we have fitted three alternative time series models $\xi_1 : (1-B)_{z_t} = a_t$, $\xi_2 : (1-B^{12})_{z_t} = a_t$ and $\xi_3 : (1-B)(1-B^{12})_{z_t} = a_t$, where B stands for the backshift operator, $zt = \log\left(m + y_t^0\right)$ (m being a nuisance parameter estimated by maximum likelihood) and a_t denotes white noise. Each predicted value $\hat{y}_k$ is computed according to the corresponding best model ξ^* (in terms of the minimal estimated mean squared error). Since the sample is a fixed panel selected by cut-off, the sampling weights w_{ks} are all equal to 1.

Next, we have applied the generic univariate observation-prediction model illustrated in Section 3 to the logarithmic transforms of the turnover and the new orders received

independently. The common error probability $p_k = p$ and observation variance $\sigma_k^2 = \sigma^2$ have been estimated from the past three months using a double-data set. The prediction variance ν_k^2 has been computed according to the corresponding chosen best model ξ^* for each unit. As loss matrices, we have chosen both the squared and the absolute loss function with entries given by Equations (4) and (5), respectively.

Finally, to make the comparison with a random selection of units, we have computed the absolute relative pseudo-bias for 50 equal-probability random selections. We have calculated the mean and first and third quartiles of the corresponding distribution. This provides a confidence-like interval for each number of selected units (see Figure 3). The motivation is to provide an insight not only into the average random selection but also of its distribution.

We have carried out this comparison for 23 NACE Rev. 2 divisions and subdivisions (aggregations of groups according to subject-matter knowledge). Firstly, RATIO2 showed a better performance than the rest of score functions (RATIO, DIFF, FLAG ITI, FLAG INORI;

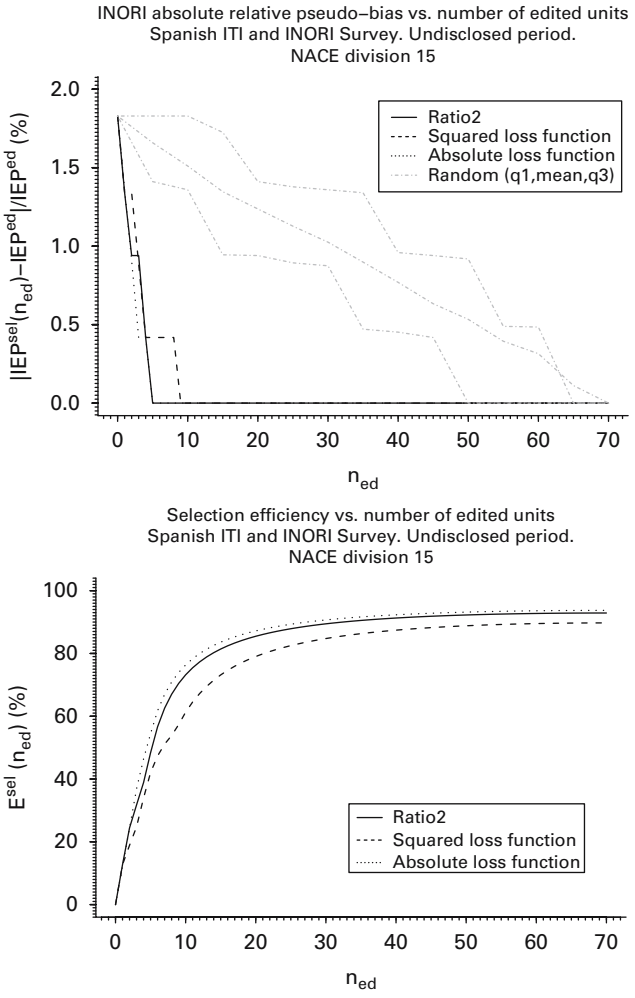

Fig. 3. Absolute relative pseudo-bias and editing efficiency vs. number of selected units

see Latouche and Berthelot 1992). In 15 cases the absolute loss yielded the most efficient prioritization, with nine of these cases having the RATIO2 score function as more efficient than the squared loss choice (Figure 3 illustrates this behavior). However in five cases it is the squared loss function that outperforms the other two choices, in which the absolute loss also did better than the score function. In the remaining three cases, RATIO2 slightly overcame the absolute loss, which in turn performed better than the squared loss.

Thus, in general, the absolute loss is more efficient than the squared loss in terms of the pseudo-bias, as expected. This also happens with the score function RATIO2, since it is also targeted at the bias. In general, the absolute loss is also more efficient than the score functions. However, in actual production conditions, both missing data and respondents newly added to the sample must be taken into account. In these cases, in the optimization approach the prediction values $\hat{y}_k$ must be imputed or fixed under some supplementary scheme, since the considered time series models fail to produce these values. As an elementary test, we assigned $\hat{y}_k = y_k$ in these cases in order for them not to be selected at first positions. The general result was a slight deterioration of the performance of the score functions for all values of n_{ed}, while in the optimization approach, the behavior was as good as before for the most influential units ($n_{ed} = 1, 2, \ldots$), but noticeably poorer for the last units ($n_{ed} \geq n/2$). We have not considered these issues in the preceding comparison, since they belong to sophistications of the observation-prediction model.

In our opinion it is important to note that the above results have been obtained with crude time series models and extremely simplifying assumptions, and they do not incorporate any subject-matter knowledge. Thus there is more room to elaborate further on them (using better parameters, building multivariate models, etc.). In this line of thought the most attractive point will arise if working models can be built for discrete or semicontinuous variables, paving the way for the use of selective editing techniques also in household surveys. The possibility of using well-established tools such as time series models or statistical models in general, reinforces the statistical defensibility of the data editing work.

6. Discussion and Concluding Remarks

Once we have detailed the methodological proposal, we now proceed to discuss several issues regarding this optimization approach from different perspectives. As two immediate objections, a cautious reader can point out the limitation to linear estimators and the polynomial running time of the algorithms. Firstly, the limitation to linear estimators, which contrasts with the common use in practice of some nonlinear estimators such as ratio estimators or regression estimators, can be easily overcome as follows. In practice most nonlinear estimators $\hat{Y}^{nl}_{U_d}$ are functions of linear estimators $\hat{Y}^{nl}_{U_d} = f\left(\hat{Y}^{(1)}_{U_d}, \ldots, \hat{Y}^{(M)}_{U_d}\right)$. Then instead of considering the corresponding restriction for the MSE of $\hat{Y}^{nl}_{U_d}$, we consider a restriction for each linear estimator $\hat{Y}^{(m)}_{U_d}$, $m = 1, \ldots, M$. The rationale amounts to expecting an accurate nonlinear estimator if each linear estimator is accurate. Moreover, a bounded growth in the number of restrictions is expected, since nonlinear estimators are usually built from different combinations of survey variables, whose number is fixed by the questionnaire. Secondly, the polynomial running time of the selection algorithms is not a practical concern, at least in Spanish sampling sizes standards, as we will now explain. On the one hand, the estimation problem in a finite population U is

essentially a multivariate problem seeking accurate and numerically consistent estimations in given partitions of the population U. These partitions are fixed according to the breakdown established by the statistical dissemination plan of each survey. Thus the selection or prioritization should be applied to each of these publication cells, since no lack of accuracy is rightfully allowed in any published figure. On the other hand, we have applied this approach to the Spanish ITI and INORI survey as a pilot experience at INE Spain (details will be published elsewhere). In these monthly short-term business statistics, the sampling size amounts to around 12,000 industrial establishments broken down into 37 publication cells with sizes ranging up to 1,500 units at most. The prioritization of units in all cells took a total of three hours on a desktop PC, which is a reasonable working time.

As a deeper concern, one can inquire why the roles of the two basic principles of our formulation are not interchanged, that is, why data quality is not optimized (minimizing the loss function) restricting the amount of resources used (number of questionnaires to recontact). We give two reasons to support our proposal. From a broad perspective, in a statistical office it appears desirable to minimize the cost of each survey in order to optimize resources to face and embrace as many other surveys in the statistical production as possible. In our view, this is a natural decision given the increasing demand for information from stakeholders. From a more methodological standpoint, the multivariate feature of the problem again arises. If we interchanged the roles of both principles, we would need to minimize the loss function of the different variable estimators corresponding to each publication cell restricted to the number of questionnaires to be recontacted. As a matter of fact this is a multiobjective optimization problem, which ineludibly needs some decisions to compute a solution (see e.g., Marler and Arora 2004). In this respect, our position in official statistics production is to minimize the number of decisions taken by the survey conductor, which is clearly expressed in the following citation by Hansen et al. (1983): "[. . .] it seems desirable, to the extent feasible, to avoid estimates or inferences that need to be defended as judgments of the analysts conducting the survey".

As a matter of fact, the question of the number of decisions is a first relevant point to establish a comparison with the score function approach. Nowadays the score function approach is undisputedly the favored technique for selecting influential units in the editing production phase. Thus it provides the framework to assess advantages and disadvantages of any other technique. Furthermore, in our opinion, a comparison will help us reveal fundamental aspects of the editing production phase irrespective of the particular techniques. Regarding the number of decisions, let us recall that the score function approach comprises four main decisions to determine a selection of units (Lawrence and McKenzie 2000), namely (i) an editing model to construct the anticipated values, (ii) each local score function, (iii) a global score function, and (iv) a cut-off value. On the one hand, in the optimization approach the first three decisions are jointly substituted and integrated into a single step: the construction of the observation-prediction model or an alternative statistical error-modelling technique, and the subsequent formulation of the optimization constraints. Furthermore, in our view, this integration renders this selection procedure more natural within the statistical language, in contrast to a score function, which can seem extraneous. In this sense, let us point out that the construction of an observation-prediction model is a multivariate exercise, so the integration of the choices of both local and global score functions comes naturally together

with the construction of the statistical model. On the other hand, the choice of the cut-off value is now substituted for the choice of the bounds in the optimization problem. In the score function approach, this value must be chosen normally using data from previous realizations of the survey and using a heuristic or empirical connection between this value and the chosen loss function of the survey estimators. In the optimization approach, the choice of the bounds makes use of a priori values of variances (or some other similar measure) as in the survey design stage and shows a neater connection with the loss function, thus fitting again more naturally into the whole survey statistics production process. Indeed, we have shown how the prioritization of units under the score function approach can be reproduced and slightly overcome with a very simple model. Furthermore, although admittedly still too far, this proposal points toward enlarging the traditional sampling strategy $(\mathcal{D}, T)$ comprising the sampling design $\mathcal{D}$ and the construction of the estimator T (see e.g., Hedayat and Sinha 1991) with a selection strategy R, so that we would have a triplet $(\mathcal{D}, R, T)$. This follows the spirit of the total survey design.

The selection/prioritization issue goes hand in hand with the double version of the optimization approach. This issue arises mainly from resource availability and controllability, mainly of timeliness and person-hours in the editing fieldwork. When having a selection of units in practice we face two situations: Either we run out of resources to accomplish the interactive editing of all selected units, or we end up ahead of time and then we miss the opportunity to gain more accuracy. Now, since editing near the source is a must for this production phase, it is advisable to have a real-time selection mechanism on each questionnaire, as pointed out in the introduction, independently of the rest of the sample. Conversely, on later stages it is preferable to prioritize units to edit (interactively) the most influential first. In this line of thought, the stochastic approach suits the selection whereas the combinatorial approach suits the prioritization. Furthermore, since both approaches derive from a common general framework focused on the exploitation of auxiliary information, we envisage a more complex, although unified, editing process. Let us parameterize the auxiliary information used in the editing work in terms of its longitudinal, cross-sectional and multivariate dimensions. By longitudinal we mean the value of variables for each unit in previous time periods. By cross-sectional we refer to the information stemming from the sample at the current period. Finally, by multivariate we mean the information arising from the multidimensional character of the survey (always several variables are investigated). If we focus on the longitudinal and cross-sectional dimensions of the auxiliary information, Figure 4 represents the transition from micro-selective to macro editing as the data collection is being completed. In our view, these two editing techniques appear as the head and tail of a time-continuous process driven by the evolution of the data collection. We envisage that intermediate techniques combining both the longitudinal and available cross-sectional information as a time-continuous process during the data collection will be of practical usefulness.

Regarding the optimization approach, we want to point out that both versions fit naturally as the head and tail of this time-continuous editing process, so that the stochastic version corresponds to exploiting longitudinal information as in traditional selective editing techniques, whereas the combinatorial version arises as a macro editing technique focusing upon the cross-sectional information. In contrast, the score function approach and traditional macro editing techniques can hardly be seen under the same methodological

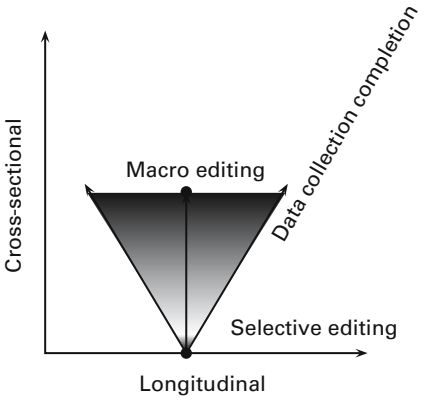

Fig. 4. Schematic representation of the transition from micro-selective to macro editing as data collection is completed. As data collection is completed, more cross-sectional information is available

principles. It remains open for future work to find a more general formulation for this proposed time-continuous process embedding both optimization versions.

A complementary comparison can be made with the automatic data editing techniques based on the Fellegi-Holt methodology, in particular with the different approaches to the error localization problem, which also make an extensive use of optimization techniques (see De Waal et al. 2011). The common points reduce to the fact that mathematical optimization appears as a natural translation of the proposed data editing principles. Conversely, the Fellegi-Holt methodology focuses upon each questionnaire, seeking to minimize the number of items to change satisfying all edits. In this approach we focus upon the whole sample, seeking to minimize the number of units to be recontacted satisfying restrictions upon the loss functions using a statistical model instead of edits.

To conclude, as immediate future prospects, we have recently begun to analyze the inclusion of these techniques in the current E&I strategies in most business surveys in INE Spain. A pilot experience with the ITI and INORI survey fosters our hope to reduce current recontact rates and consequently both editing costs and the response burden at our office. R packages and SAS macros implementing this optimization approach are under intense development and being tested in these pilot experiences. Apart from this, more methodological research is needed to find generic multivariate models fitting the observation-prediction model and to generalize them to both qualitative and semicontinuous variables. In this context, multivariate models already present in the literature for data editing (Di Zio and Guarnera, in this issue) appear as a fruitful alternative. In addition, we already have a first adaptation of the preceding greedy algorithms to be applied to surveys with self-weighting samples and qualitative variables. We are collaborating with experts from the Spanish National Health Survey to produce an observation-prediction model adapted to these variables.

A. Mathematical Appendix

We include some mathematical proofs. Firstly we prove how the constraints imply a control on the loss of accuracy. In particular, if $L = L^{(r)}$ denotes the absolute ($r = 1$) or

squared loss ($r = 2$) function, we prove that $\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R})\hat{Y}^0)|\mathbf{Z}] \leq \eta$ (where $\mathbf{Z} = \mathbf{Z}^{st}$ or $\mathbf{Z}^{cross}$) implies $\mathbb{E}_{pm}[L(\hat{Y}^*(\mathbf{R}), Y)] \leq \left(\eta^{1/r} + \mathbb{E}_{pm}^{1/r}[L(\hat{Y}^0, Y)]\right)^r$. It is straightforward to prove that $d(A, B) = \mathbb{E}_{pm}^{1/r}[L^{(r)}(A, B))]$ is a metric. Then, by the triangle inequality, we have

$$d(\hat{Y}^*(\mathrm{R}), Y) \leq d(\hat{Y}^*(\mathrm{R}), \hat{Y}^0) + d(\hat{Y}^0, Y).$$

Now, using properties of the conditional expectation, we can write

$$d^r(\hat{Y}^*(\mathrm{R}), \hat{Y}^0) = \mathbb{E}_{pm}[\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0|\mathbf{Z}]] \leq \eta,$$

where $\mathbf{Z} = \mathbf{Z}^{st}$ or $\mathbf{Z}^{cross}$. The result follows immediately.

Secondly we show the connection between the loss matrices and the loss function. In the absolute loss case, we have $\mathbb{E}_m[|\hat{Y}^*(\mathbf{R}) - \hat{Y}^0|\,|\mathbf{Z}] = \mathbb{E}_m\left[\left|\sum_{k\in s} R_k w_{ks}\epsilon_k\right|\right]|z] \leq \left[\sum_{k\in s} R_k^2 |w_{ks}\epsilon_k|\,|\mathbf{Z}\right]$, since $R_k^2 = R_k$. Thus we can write $\mathbb{E}_m[|\hat{Y}^*(\mathbf{R}) - \hat{Y}^0|\,|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$, where Δ is diagonal with entries $\Delta_{kk} = |w_{ks}\epsilon_k|$. In the squared loss case, in turn we have $\mathbb{E}_m[(\hat{Y}^*(\mathbf{R}) - \hat{Y}^0)^2|\mathbf{Z}] = \mathbb{E}_m\left[\sum_{k\in s}\sum_{k\in s} R_k R_l w_{ks}\epsilon_k w_{ls}\epsilon_l|\mathbf{Z}\right]$. Thus, we can also write $\mathbb{E}_m[(\hat{Y}^*(\mathbf{R}) - \hat{Y}^0)^2|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$, where $\Delta_{kl} = w_{ks}\epsilon_k \cdot w_{ls}\epsilon_l$.

The conditional moments (4) and (5) are found along similar lines. Under the hypotheses assumed in Section 3 regarding the observation-prediction model, it follows that $y_k - \hat{y}_k = \epsilon_k^{obs} + \epsilon_k^{pred}$ and $\mathbb{E}_m\left[\left(y_k - y_k^0\right)^r|s_k, y_k, \hat{y}_k\right] = \mathbb{E}_m\left[\delta_k^{(obs)r}|s_k, y_k, \hat{y}_k\right]$. $\mathbb{E}_m[e_k|s_k, y_k, \hat{y}_k]$, with $r = 1, 2$. Conditioning on $s_k, y_k, \hat{y}_k$ amounts to conditioning on $s_k, \epsilon_k^{obs}, \hat{y}_k$, thus we can rewrite these conditional expectations as $\mathbb{E}_m[\cdot|s_k, y_k - \hat{y}_k, \hat{y}_k]$. Now the second term is computed using Bayes' theorem, so that $\mathbb{E}_m[e_k|s_k, y_k - \hat{y}_k, \hat{y}_k] = \zeta k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right)$. For the first term, we notice that the random vector $\left(\delta_k^{obs}, \delta_k^{obs} + \epsilon_k^{pred}\right)^T$ is normally distributed with expectation $\mu = 0$ and variance $\Sigma = \begin{pmatrix} \sigma_k^2 & \sigma_k^2 + \rho_k\sigma_k\nu_k \\ \sigma_k^2 + \rho_k\sigma_k\nu_k & \sigma_k^2 \end{pmatrix}$. The conditional moments follow then from standard properties of the multivariate normal distribution.

7. References

Arbués, I., González, M., and Revilla, P. (2012a). A Class of Stochastic Optimization Problems with Application to Selective Data Editing. Optimization, 61, 265–286. DOI: http://www.dx.doi.org/10.1080/02331934.2010.511670

Arbués, I., Revilla, P., and Salgado, D. (2012b). Optimization as a Theoretical Framework to Selective Editing. UNECE Work Session on Statistical Data Editing, 24–26 September. WP2, 1–10.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: Wiley.

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-dependent and Probability Sampling Inferences in Sample Surveys. Journal of the American

Statistical Association, 78, 776–793. DOI: http://www.dx.doi.org/10.1080/01621459.1983.10477018

Hedayat, A.S. and Sinha, B.K. (1991). Design and Inference in Finite Population Sampling. New York: Wiley.

INE Spain (2010). Industrial Turnover Indices. Industrial New Orders Received Indices. Base 2005. CNAE-09. Methodological Manual. Available at http://www.ine.es/en/metodologia/t05/t0530053_en.pdf. (accessed January 10, 2013).

Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing 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.

Lindgren, K. (2011). Selective Editing in the International Trade in Services. Working Paper 19 of 2011 UNECE Metting on Statistical Data Editing. May 9–11, Ljubljana. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp.19.e.pdf (accessed October 2013).

Marler, R.T. and Arora, J.S. (2004). Survey of Multi-Objective Optimization Methods for Engineering. Structural and Multidisciplinary Optimization, 26, 369–395. DOI: http://www.dx.doi.org/10.1007/s00158-003-0368-6

Salgado, D., Arbués, I., and Esteban, M.E. (2012). Two Greedy Algorithms for a Binary Quadratically Constrained Linear Program in Survey Data Editing. INE Spain Working Paper 02/12. Available at http://www.ine.es. (accessed January 10, 2013).

Wets, R.-B. (2002). Stochastic Programming Models: Wait-and-see Versus Here-and-now. Institute for Mathematics and Its Applications, 128.

Received February 2013
Accepted September 2013