Working Papers

04/2011

Exploiting auxiliary information: selective editing as a combinatorial optimization problem

David Salgado

First draft: April 2011

This draft: April 2011

Exploiting auxiliary information: selective editing as a combinatorial optimization problem

Abstract

We formulate selective editing as a combinatorial optimization problem whose solution establishes which sampled units contain influential errors and thus must undergo interactive editing within a generic editing and imputation strategy. This optimization problem arises naturally from considerations on editing resources savings and estimates accuracy control. Cross-sectional auxiliary information is taken into account through linear mixed models assisting the construction of the problem's feasibility region. We provide a general algorithm for the univariate version of this problem, i.e. for editing one single variable. By applying this proposal to each questionnaire variable we illustrate its use upon the Spanish industrial turnover index and industrial new orders received index surveys. A reduction of interactive editing with a controllably increase of estimates error is observed.

Keywords

Selective editing, combinatorial optimization, auxiliary information, linear mixed models

Authors and Affiliations

David Salgado

D.G. Metodología, Calidad y Tecnologías de la Información y las Comunicaciones

Instituto Nacional de Estadística

Exploiting Auxiliary Information: Selective Editing as a Combinatorial Optimization Problem

David Salgado

D.G. Metodología, Calidad y Tecnologías de la Información y de las Comunicaciones

Instituto Nacional de Estadística

April 28, 2011

Keywords: Selective editing, combinatorial optimization, auxiliary information, linear mixed models

Abstract

We formulate selective editing as a combinatorial optimization problem whose solution establishes which sampled units contain influential errors and thus must undergo interactive editing within a generic editing and imputation strategy. This optimization problem arises naturally from considerations on editing resources savings and estimates accuracy control. Cross-sectional auxiliary information is taken into account through linear mixed models assisting the construction of the problem's feasibility region. We provide a general algorithm for the univariate version of this problem, i.e. for editing one single variable. By applying this proposal to each questionnaire variable we illustrate its use upon the Spanish industrial turnover index and industrial new orders received index surveys. A reduction of interactive editing with a controllably increase of estimates error is observed.

1 Introduction

"If I am to select one issue [on calibration], let me focus on the concept of auxiliary information. It is the pivotal concept [...]" (Särndal, 2007). This prominent role of auxiliary information in survey sampling is traditional both in the design and the estimation phases. Long-established sampling designs (Deming, 1950; Cochran, 1977; Särndal et al., 1992) constitute indeed different ways to incorporate this information and even more recent proposals (Deville and Tillé, 2005; Tillé, 2006) also pursue this view. The calibration approach (see e.g. Särndal (2007) and multiple references therein) is perhaps the most outstanding example. But these are not the only stages of a survey sampling estimation process where auxiliary information is clearly useful. Small area estimation (Rao, 2003)

is another clear example, to name just one more.

On the other hand, somehow implicit above is the fact that the production of survey sampling estimates, and official statistics in general, is a complex integration of processes, most of them of an eminent statistical nature. An excellent description of this view is offered by the so-called Generic Statistical Business Process Model (GSBPM hereafter) described in UNECE (2010), where 53 processes are identified. Among these one can clearly find data editing and imputation (E&I henceforth) stages usually carried out in practice jointly.

In the last two decades different methods of statistical data editing have been clearly recognized (de Waal, 2008, 2009), which have indeed been proposed to be integrated in a whole so-called E&I strategy (EDIMBUS, 2007). Here we will concentrate upon selective editing (Hidiroglou and Berthelot, 1986; Latouche and Berthelot, 1992; Lawrence and McDavitt, 1994; Lawrence and McKenzie, 2000; Hoogland, 2002; Hedlin, 2003, 2008; Hoogland and Smit, 2008) whose goal is to identify those sampled units which contain influential errors with noticeable effects on the estimated aggregates. By and large selective editing runs through four stages (Lawrence and McKenzie, 2000), namely (i) the choice of an expected amended value for a questionnaire variable (a value more likely than the actual reported value according to a chosen editing model), (ii) the determination of local scores, (iii) the combination of these to produce a respondent global score, and (iv) the choice of cut-off thresholds (to establish a limit under which sampled units will not enter the critical stream of units to be further queried). This clearly involves many decisions by the survey conductor.

Under the general philosophy portrayed above, we propose an approach to the selective editing stage of the E&I phase which seeks to exploit auxiliary information under a reduced set of assumptions by the statistician conducting the survey. Here the selective editing stage is always considered as a subprocess integrated in the whole E&I strategy. No score function is needed and the identification of influential units is performed according to their a priori effects on the accuracy of the estimation of aggregates, always aiming at reducing the amount of editing work (hence of editing costs). As it will be discussed below, selective editing is thus posed as an optimization problem minimizing editing costs subject to estimates accuracy being under control. Previous work in this direction has already been carried out in Arbués et al. (2009, 2010), where auxiliary information from preceding time periods for each sampled unit has been successfully exploited in the editing task.

The work is organized as follows. In section 2 we pose the general assumptions under which this formulation develops. In section 3 we formulate selective editing as a combinatorial optimization problem. In section 4 we explain how cross-sectional auxiliary information shaping the feasibility region of this problem is taken into account through linear mixed models and in section 5 an algorithm to solve it is provided. In section 6 we apply our proposal to the Spanish industrial turnover index and industrial new orders

received index surveys. In section 7 we discuss the work both theoretically and in the light of the preceding application. Finally in section 8 we collect some concluding remarks.

2 General assumptions

Before proceeding to formulate selective editing as an optimization problem, we must make clear the general assumptions under which this formulation is drawn. Firstly, the concept of auxiliary information is extremely wide (even vague) and some concretion is desirable. For our purposes we shall conceive auxiliary information as a three-dimensional concept fully deploying its utility when all these three dimensions are jointly exploited. These are (i) the longitudinal, (ii) the cross-sectional and (iii) the multivariate information. By longitudinal we mean the value of variables for each unit in previous time periods. This implicitly assumes that the survey is periodical. By cross-sectional we refer to the information stemming out from the whole sample at the current period. Finally, by multivariate we signify the information arising from the multidimensional character of the survey (usually several variables are investigated). In previous works (Arbués et al., 2009, 2010) selective editing was formulated as a stochastic optimization problem focusing upon the longitudinal dimension of the auxiliary information. In the present work we concentrate complementarily upon the cross-sectional one. Our final goal is to find an integration of all dimensions in the same E&I strategy.

Secondly, we adopt the generic E&I strategy outlined in EDIMBUS (2007), which is represented in figure 1, so that our selective editing approach is thought to identify both a critical stream of units with data to be edited interactively and a noncritical stream to be edited automatically. We make clear that we focus only on the error detection stage of the E&I process. No imputation or error treatment issues whatsoever are tackled with in this work.

In the third place, in consonance with the quality view of selective editing (Biemer and Lyberg, 2003) (see also Granquist and Kovar (1997); Granquist (1997)), our formulation is based upon the goal of *minimizing data editing costs keeping the survey estimates accuracy under control*. This will allow us to reallocate resources from pure data cleaning to identifying and collecting information about error causes in order to improve survey quality. Thus we adhere to the increasingly extended view of building quality into every stage of survey production.

Finally, we focus on measurement errors, which are dealt with in the usual form (Lessler and Kalsbeek, 1992; Särndal et al., 1992): the observed value of a variable y by respondent k, denoted by y_k^{obs}, will be written as

$$y_k^{\text{obs}} = y_k + \epsilon_k, \tag{1}$$

where y_k stands for the true value and ϵ_k denotes the individual error of unit k. In rigour we should denote instead $\epsilon_k^{(y)}$ (or something similar) to account for the variable which the

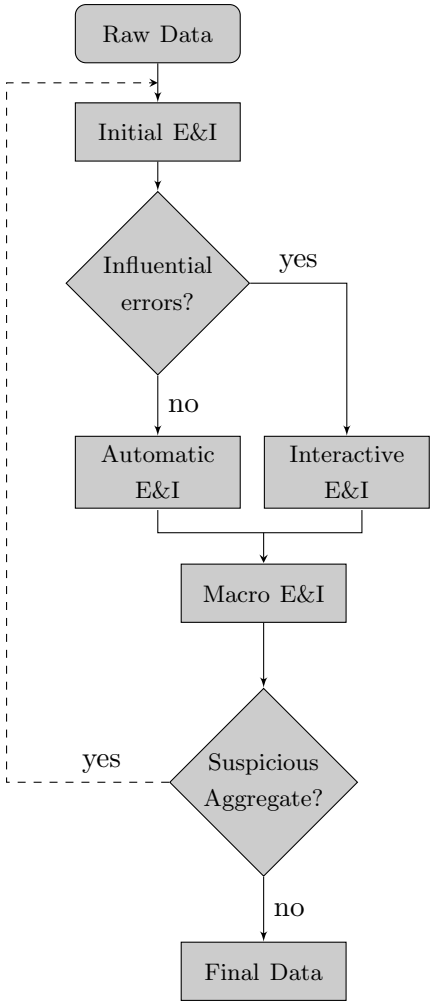


Figure 1: Generic E&I strategy for economic surveys taken from EDIMBUS (2007). The selective editing phase corresponds to the detection of influential errors.

error is affecting to, but since we will not deal with multivariate questions, we keep our notation as simple as possible. To be exhaustive, we state the following remarks:

- As in existing approaches of selective editing, the variable y is quantitative. We leave the analysis and adaptation of this formulation to qualitative variables for future research.
- For the time being the true value y_k is a fixed non-random value. As we shall see below, we will promote y_k to a random variable (in consonance with Arbués et al. (2009, 2010)) in order to pave the way for model building tools.
- The concept of measurement error should be understood in the editing problem in

a generalized way, i.e. as an observational error in the sense of Groves' classification of survey errors (Groves, 1989). Thus errors due to the interviewer, the respondent, the measurement instrument and the data collection mode are considered.

- The error ϵ_k is a random variable. The random nature of the measurement error in sample surveys can be subject to debate (Lessler and Kalsbeek, 1992), but here it is stated in a pragmatic fashion, encompassing possible nonrandom situations as degenerate random variables.

- In general, as we will discuss, scarcely necessary is the distribution of ϵ_k and always unattainable. The first and second moments will generally suffice, in particular, for the forthcoming formulation. This follows the spirit of the simple measurement model (Särndal et al., 1992).

- The moments are allowed to depend on the particular sample selected for the survey. That is, we will need the expected value $\mathbb{E}_m[\epsilon_k|s] = \mu_{ks}$, the variance $\mathbb{V}_m[\epsilon_k|s] = \sigma^2_{ks}$, and covariance $\mathbb{C}_m[\epsilon_k, \epsilon_l|s] = \sigma_{kls}$, where the subscript m refers to the chosen error model and s stands for the selected sample. This implies that with a different sample these moments may change.

- Auxiliary information is a key concept in our formulation, as in Arbués et al. (2009, 2010). With this purpose we denote by $\mathcal{G}$ the $\sigma-$algebra generated by all auxiliary variables potentially pertinent to the determination of the model (1). In other words, $\mathcal{G}$ represents all information available at the moment of carrying out the survey which can be used to construct the error model[1]. Thus, the quantities of interest will actually be $\mathbb{E}_m[\epsilon_k|s, \mathcal{G}] = \mu_{ks}$, $\mathbb{V}_m[\epsilon_k|s, \mathcal{G}] = \sigma^2_{ks}$, and $\mathbb{C}_m[\epsilon_k, \epsilon_l|s, \mathcal{G}] = \sigma_{kls}$.

3 Selective editing as a combinatorial optimization problem

To begin with, we define for each sampled unit $k \in s$ a binary variable $r_{ks} \in \{0, 1\}$ such that

$$r_{ks} = \begin{cases} 0 & \text{if unit } k \text{ enters interactive editing,} \\ 1 & \text{if unit } k \text{ does not enter interactive editing.} \end{cases}$$

This counterintuitive assignment will be clear immediately below. Also, we allow r_{ks} to be sample-dependent, so that we write more completely r_{ks} instead of simply r_k. This gives room for sampling variations and it is in consonance with the concept of influential error, since an error in a unit may be influential in a given sample (and thus edited interactively), but noninfluential in another.

Using r_{ks} we can readily define the (interactively) edited value of a variable y for each unit $k \in s$ as

$$y^{\text{ed}}_{ks} = y_k + r_{ks}\epsilon_k.$$

[1] With the sole exception of the selected sample s.

Note that $y_{ks}^{\text{ed}} = y_k$ when unit k is edited and $y_k^{\text{ed}} = y_k^{\text{obs}}$ otherwise. This reflects the working assumption that editing drives us to the true value of the variable. It is a strong hypothesis, but a fairly acceptable one. Not in vain interactive editing is the most resource-demanding.

Once introduced y_{ks}^{ed}, one can readily distinguish three related estimators for, say, a population total Y_U:

1. The true estimator $\widehat{Y} = \sum_{k\in s} \omega_{ks} y_k$;
2. The unedited estimador $\widehat{Y}^{\text{un}} = \sum_{k\in s} \omega_{ks} y_k^{\text{obs}} = \widehat{Y} + \sum_{k\in s} \omega_{ks}\epsilon_k$;
3. The edited estimator $\widehat{Y}^{\text{ed}}(\boldsymbol{r}) = \sum_{k\in s} \omega_{ks} y_k^{\text{ed}} = \widehat{Y} + \sum_{k\in s} \omega_{ks} r_{ks}\epsilon_k$,

where $\{\omega_{ks}\}_{k\in U}$ are sampling weights, which can be sample-dependent[2]. We assume henceforth that $\widehat{Y}$ is unbiased or asymptotically unbiased. Now we resort to the mean squared error to express the loss of accuracy due to unedited measurement errors.

Proposition 3.1. Let $\widehat{Y}$ and $\widehat{Y}^{\text{ed}}$ be the true and edited estimators of the population total Y_U with sampling weights $\{\omega_{ks}\}_{k\in U}$, respectively, under both the sample design $p(\cdot)$ and the error model m. Then

$$\text{MSE}_{pm}\left[\widehat{Y}^{\text{ed}}|\mathcal{G}\right] = \text{MSE}_p\left[\widehat{Y}\right] + \mathbb{E}_p\left[\widehat{\mathbb{V}}_s^{\text{ed}} + \left(\widehat{\mathbb{B}}_s^{\text{ed}}\right)^2 + 2\left(\widehat{Y} - Y\right)\widehat{\mathbb{B}}_s^{\text{ed}}\right],$$

where $\widehat{\mathbb{B}}_s^{\text{ed}} = \sum_{k\in s} \omega_{ks} r_{ks}\mu_{ks}$ and $\widehat{\mathbb{V}}_s^{\text{ed}} = \sum_{k\in s}\sum_{l\in s} \omega_{ks}\omega_{ls} r_{ks} r_{ls}\sigma_{kls}$.

Proof. Writing out

$$\text{MSE}_{pm}\left[\widehat{Y}^{\text{ed}}|\mathcal{G}\right] = \mathbb{E}_{pm}\left[\left(\widehat{Y}^{\text{ed}} - \widehat{Y}\right)^2|\mathcal{G}\right] + \mathbb{E}_{pm}\left[\left(\widehat{Y} - Y\right)^2|\mathcal{G}\right] + 2\cdot\mathbb{E}_{pm}\left[\left(\widehat{Y}^{\text{ed}} - \widehat{Y}\right)\left(\widehat{Y} - Y\right)|\mathcal{G}\right],$$

we identify:

- $\mathbb{E}_{pm}\left[\left(\widehat{Y}^{\text{ed}} - \widehat{Y}\right)^2|\mathcal{G}\right] = \mathbb{E}_p\left[\mathbb{E}_m\left[\left(\sum_{k\in s}\omega_{ks} r_{ks}\epsilon_{ks}\right)^2|s,\mathcal{G}\right]\right] = \mathbb{E}_p\left[\widehat{\mathbb{V}}_s^{\text{ed}} + \left(\widehat{\mathbb{B}}_s^{\text{ed}}\right)^2\right]$.
- $\mathbb{E}_{pm}\left[\left(\widehat{Y} - Y\right)^2|\mathcal{G}\right] = \mathbb{E}_p\left[\mathbb{E}_m\left[\left(\widehat{Y} - Y\right)^2|s,\mathcal{G}\right]\right] = \mathbb{E}_p\left[\left(\widehat{Y} - Y\right)^2\right] = \text{MSE}_p\left[\widehat{Y}\right]$.
- $\mathbb{E}_{pm}\left[\left(\widehat{Y}^{\text{ed}} - \widehat{Y}\right)\left(\widehat{Y} - Y\right)|\mathcal{G}\right] = \mathbb{E}_p\left[\left(\widehat{Y} - Y\right)\mathbb{E}_m\left[\left(\widehat{Y}^{\text{ed}} - \widehat{Y}\right)|s,\mathcal{G}\right]\right] = \mathbb{E}_p\left[\left(\widehat{Y} - Y\right)\widehat{\mathbb{B}}_s^{\text{ed}}\right]$.

□

[2]This entails, for example, that $\hat{Y}$ is not necessarily the Horvitz-Thompson estimator, but can possibly be a ratio or regression estimator.

From this expression one can readily read the contribution to the mean squared error coming from the sampling design and that coming from the partial editing work. Note that should every unit be edited, we would have $r_{ks} = 0$ for all $k \in s$ and $\mathrm{MSE}_{pm}\left[\widehat{Y}^{\mathrm{ed}}|\mathcal{G}\right] = \mathrm{MSE}_{p}\left[\widehat{Y}\right]$, as expected: no further editing whatsoever can possibly reduce the mean squared error due to sampling variability.

Now, given the (asymptotic) unbiasedness of $\widehat{Y}$, it is immediate to find the following bound in the limit[3] $n_s \to \infty$:

$$\mathrm{MSE}_{pm}\left[\widehat{Y}^{\mathrm{ed}}|\mathcal{G}\right] \leq \mathrm{MSE}_{p}\left[\widehat{Y}\right] + \mathbb{E}_{p}\left[\widehat{\mathbb{V}}_{s}^{\mathrm{ed}} + \left(\widehat{\mathbb{B}}_{s}^{\mathrm{ed}}\right)^{2}\right].$$

We will use the second term as the figure of control for the loss of accuracy due to allowing some errors in the data. In particular, if we write $\widehat{\mathbb{V}}_{s}^{\mathrm{ed}} + \left(\widehat{\mathbb{B}}_{s}^{\mathrm{ed}}\right)^{2} \leq v^2$, we immediately arrive at

$$\mathrm{MSE}_{pm}\left[\widehat{Y}^{\mathrm{ed}}|\mathcal{G}\right] \leq \mathrm{MSE}_{p}\left[\widehat{Y}\right] + v^2.$$

To be concrete, we will refer to v^2 as a accuracy control parameter. Were we considering several aggregates $Y_U^{(p)}$, $p = 1, \ldots, P$, we would have to introduce P accuracy control parameters v_p^2 in the corresponding bounds $\widehat{\mathbb{V}}_{s}^{(p)\mathrm{ed}} + \left(\widehat{\mathbb{B}}_{s}^{(p)\mathrm{ed}}\right)^{2} \leq v_p^2$.

The use of the moments μ_{ks} and σ_{kls} allows us to view $\widehat{\mathbb{B}}_{s}^{(p)\mathrm{ed}}$ and $\widehat{\mathbb{V}}_{s}^{(p)\mathrm{ed}}$ as estimators of the bias and variance contributions, respectively, due to unedited measurement errors. In this sense, $\widehat{\mathbb{V}}_{s}^{(p)\mathrm{ed}} + \left(\widehat{\mathbb{B}}_{s}^{(p)\mathrm{ed}}\right)^{2}$ can be considered as the corresponding mean squared error contribution arising from unedited measurement errors. This makes clear the role of the bounds v_p^2 from a theoretical standpoint.

On the other hand, as discussed in section 2, our goal is to minimize the amount of interactive editing, which we translate as maximizing the function $\sum_{k \in s} r_{ks}$. Thus, given the set of variables $\{r_{ks}\}_{k \in s}$ we have the two ingredients to formulate an optimization problem, namely (i) an objective function and (ii) a feasibility region. We formulate the selective phase of the E&I strategy as the following combinatorial optimization problem:

$$\begin{aligned} & \max \sum_{k \in s} r_{ks} \qquad p - a.s. \\ \text{such that} \quad & \widehat{\mathbb{V}}_{s}^{(p)\mathrm{ed}} + \left(\widehat{\mathbb{B}}_{s}^{(p)\mathrm{ed}}\right)^{2} \leq v_p^2 \qquad p - a.s. \quad p = 1, \ldots, P \\ & r_{ks} \in \{0, 1\} \qquad k = 1, \ldots, n_s. \end{aligned} \tag{2}$$

[3] This limit is always understood as in Isaki and Fuller (1982).

This first formulation embraces all kind of auxiliary information and does not focus upon any of its particular dimensions. However, note that the feasibility region of problem (2) would be defined only if we knew the moments μ_{ks} and σ_{kls}, which is never the case. Thus we will estimate them, denoting $\widehat{\mu}_{ks}$ and $\widehat{\sigma}_{kls}$, and consequently $\widehat{\widehat{\mathbb{V}}}_s^{(p)\text{ed}} = \sum_{k\in s}\sum_{l\in s}\omega_{ks}\omega_{ls}r_{ks}r_{ls}\widehat{\sigma}_{kls}$ and $\widehat{\widehat{\mathbb{B}}}_s^{(p)\text{ed}} = \sum_{k\in s}\omega_{ks}r_{ks}\widehat{\mu}_{ks}$ for the editing variance and editing bias estimators. The combinatorial optimization problem will then be formulated as

$$\begin{aligned} & \max \sum_{k\in s} r_{ks} \qquad p-a.s. \\ \text{such that} \quad & \widehat{\widehat{\mathbb{V}}}_s^{(p)\text{ed}} + \left(\widehat{\widehat{\mathbb{B}}}_s^{(p)\text{ed}}\right)^2 \leq v_p^2 \qquad p-a.s. \quad p=1,\ldots,P \\ & r_{ks}\in\{0,1\} \qquad k=1,\ldots,n_s. \end{aligned} \tag{3}$$

This is the completely general formulation. The auxiliary information is embraced in the estimation procedures chosen to set up the feasibility region. In the present work we focus only on the case $P = 1$, ruling out the multivariate dimension (see however the example in section 6), and we will only exploit the cross-sectional information when estimating $\widehat{\mathbb{V}}_s^{(p)\text{ed}} + \left(\widehat{\mathbb{B}}_s^{(p)\text{ed}}\right)^2$ through the error model. In Arbués et al. (2009, 2010), the longitudinal dimension is used to estimate the moments μ_{ks} and σ_{kls}, apart from a generalization of the mathematical nature of the variables r_{ks}, which are treated as continuous random variables thus turning the problem into a stochastic optimization one.

Thus, in the following we will concentrate upon the following one-aggregate combinatorial problem, written in a simplified and more usual notation in optimization theory (Beasley, 1996):

$$\begin{aligned} & \max \mathbb{1}^T\boldsymbol{r} \\ \text{s.t.} \ & \boldsymbol{r}^T B\boldsymbol{r} \leq v^2 \\ & \boldsymbol{r}\in\{0,1\}^{\times n}, \end{aligned} \tag{4}$$

where $\mathbb{1}$ is the vector $(1,\ldots,1)^T$, the quadratic form $\boldsymbol{r}^T B\boldsymbol{r}$ stands for $\widehat{\widehat{\mathbb{V}}}_s^{\text{ed}} + \left(\widehat{\widehat{\mathbb{B}}}_s^{\text{ed}}\right)^2$ for the chosen aggregate to be estimated and we have dropped out the sample s subscript and the references to the sampling design $p(\cdot)$.

4 Exploiting cross-sectional auxiliary information: linear mixed models

Firstly we include a brief description in very general terms of our proposal to use linear mixed models to incorporate auxiliary information in the preceding optimization problem.

Later on, we adapt this general recipe to the exploitation of cross-sectional information.

In principle we must estimate the moments μ_{ks} and σ_{kls}. Since the variables are quantitative, we will make use of linear models including both fixed and random effects in order to gain flexibility. As a possible a priori linear mixed model in vector notation, we could think of $\boldsymbol{\epsilon} = X\boldsymbol{\beta} + \sum_{q=1}^{Q} Z_q \boldsymbol{u}_q + \boldsymbol{e}$, where X is the matrix of regressor values, $\boldsymbol{\beta}$ is a vector of parameters (fixed effects), Z_q are incidence matrices, $\boldsymbol{u}_q$ are random effects and $\boldsymbol{e}$ denotes the residual errors of the model fitting. However it is clear that we lack the values of $\boldsymbol{\epsilon}$, which are never known. Instead we promote the original true values y_k from fixed but unknown numbers to random variables and include them in the modelling so that our linear mixed model will be

$$\boldsymbol{y}^{\text{obs}} = X^{(\boldsymbol{y})}\boldsymbol{\beta}_{(\boldsymbol{y})} + \sum_{\bar{q}=1}^{\bar{Q}} Z_{\bar{q}}^{(\boldsymbol{y})} \boldsymbol{u}_{\bar{q}}^{(\boldsymbol{y})} + X^{(\boldsymbol{\epsilon})}\boldsymbol{\beta}_{(\boldsymbol{\epsilon})} + \sum_{q=1}^{Q} Z_q^{(\boldsymbol{\epsilon})} \boldsymbol{u}_q^{(\boldsymbol{\epsilon})} + \boldsymbol{e}.$$

As in any linear mixed model, apart from the choice of effects, we must specify the distribution of $\boldsymbol{u} = (\boldsymbol{u}_1^{(\boldsymbol{y})}, \ldots, \boldsymbol{u}_{\bar{Q}}^{(\boldsymbol{y})}, \boldsymbol{u}_1^{(\boldsymbol{\epsilon})}, \ldots, \boldsymbol{u}_Q^{(\boldsymbol{\epsilon})}, \boldsymbol{e})$ in terms of its variance components[4] $\boldsymbol{\theta}$, which must then be estimated by some point estimation method (see Searle et al. (1992)), representing this as usually by $\hat{\boldsymbol{\theta}}$. We propose to use the so-obtained Empirical Best Linear Unbiased Predictor of any combination of effects and, in particular, of that representing the measurement error $\widehat{\boldsymbol{\mu}}_s = \text{EBLUP}(X^{(\boldsymbol{\epsilon})}\boldsymbol{\beta}_{(\boldsymbol{\epsilon})} + \sum_{q=1}^{Q} Z_q^{(\boldsymbol{\epsilon})} \boldsymbol{u}_q^{(\boldsymbol{\epsilon})})(\hat{\boldsymbol{\theta}})$. For the estimation of $\sigma_s = [\sigma_{kls}]_{1 \le k,l \le n_s}$ we propose to use the estimation of the mean squared error of the EBLUP, denoted as $\widehat{\text{MSE}}(\hat{\boldsymbol{\mu}}_s)$, as in small area estimation techniques (see e.g. Kackar and Harville (1984); Prasad and Rao (1990)).

Following in general terms this recipe we concentrate upon exploiting cross-sectional auxiliary information to set up the feasibility region of the optimization problem and, in particular, the matrix B in (4). Having economic surveys in mind, we will consider a population U of establishments, companies or economic units in general and a variable y reflecting some aspects of the result of their economic activity (production, turnover, income, expenses, ...). The population U accepts a nested classification according to their elements' economic activity, as e.g. the European Statistical Classification of Economic Activities (NACE) (Eurostat, 2008). To simplify our arguments let us consider that the population U is divided into $j = 1, \ldots, J$ disjoint subsets of this classification.

We have also in mind Groves' classification of survey errors, which are due to the interviewer, the respondent, the measurement instrument and the data collection mode (Groves, 1989). For concreteness' sake we consider $i = 1, \ldots, I$ interviewers and $m = 1, \ldots, M$ data collection modes[5]. Thus we model the observed value of variable y of unit

[4] E.g. $\boldsymbol{u} \simeq N(0, V(\boldsymbol{\theta}))$, being $V = V(\boldsymbol{\theta})$ the unconditional model variance of $\boldsymbol{y}^{\text{obs}}$.

[5] The measurement instrument is typically irrelevant in economic surveys and can be considered part of the collection mode (CATI, CAPI, ... (see e.g. Biemer and Lyberg (2003)), not alike in, say, some agricultural surveys where crop surface measurements are an important part of the field work.

k in branch j collected by interviewer i using mode m as

$$y_{imjk} = \beta_{0j} + \beta_{1j} x_{imjk} + u_{2m} + u_{3i} + u_{4imjk} + e_{imjk}, \tag{5}$$

where

- β_{0j} is a random intercept term with normal distribution $N(\beta_0, \sigma^2_{0j})$, often written as $\beta_{0j} = \beta_0 + u_{0j}$, with $u_{0j} \simeq N(0, \sigma^2_{0j})$;
- β_{1j} is a random slope with normal distribution $N(\beta_1, \sigma^2_{1j})$, often written as $\beta_{1j} = \beta_1 + u_{1j}$, with $u_{1j} \simeq N(0, \sigma^2_{1j})$;
- x_{imjk} is the value of an auxiliary variable x of unit k in branch j collected by interviewer i using mode m;
- u_{2m} is a random effect due to the collection mode with distribution $N(0, \sigma^2_{2m})$;
- u_{3i} is a random effect due to the interviewer with distribution $N(0, \sigma^2_{3i})$;
- u_{4imjk} is a random effect due to the respondent with distribution $N(0, \sigma^2_4)$;
- e_{imjk} is the residual term with distribution $N(0, \sigma^2_e)$;
- all these random terms are independent among themselves.

We comment some assumptions. The random nature of the intercept and slope terms are assumed on the basis of taking into account possible differences in the linear relation between the y and x variable values for the different branches j. Implicit in the structure is the hypothesis that both objective and auxiliary variables for each unit are collected by the same agent i using the same mode m. Were this not the case, a much more complex expression with respect to the subscripts i and m would be necessary. Each source of error is assumed to have its own variability, except for the respondent random effect, where the same variability has been assigned throughout the population. We are aware that less simple hypotheses do exist, e.g., by recognizing different variability in the respondents' random effect. This could be taken into account by dividing U into $g = 1, \ldots, G$ groups according to their response behaviour[6]. In the latter case we would need a new subscript $g = 1, \ldots, G$ denoting instead u_{4gimjk} with $u_{4gimjk} \simeq N(0, \sigma^2_{4g})$, as well as y_{gimjk}, x_{gimjk} and e_{gimjk}. However, within a reasonable generality, we keep our arguments as simple as possible. This has an important consequence: u_{4imjk} and e_{imjk} are statistically indistinguishable. Thus we rewrite (5) as

$$y_{imjk} = \beta_{0j} + \beta_{1j} x_{imjk} + u_{2m} + u_{3i} + e_{imjk}, \tag{6}$$

with the caveat that the new residual term e_{imjk} also contains the respondent contribution to the random error ϵ_k. In this sense, in the model (6) a high residual term is not necessarily interpreted as a deficient fitting, but as an a priori high random nonsampling

[6]Equivalently we are assuming $G = 1$.

error in the reported value y_k^{obs}.

Now instead of estimating μ_{ks} and σ_{kls} we write

$$\widehat{\mathbb{V}}_s^{\text{ed}} + \left(\widehat{\mathbb{B}}_s^{\text{ed}}\right)^2 = \sum_{k\in s}\sum_{l\in s} \omega_{ks}\omega_{ls} r_{ks} r_{ls} \mathbb{E}\left[\epsilon_k \epsilon_l | s, \mathcal{G}\right],$$

and we focus on estimating ϵ_k by $\hat{\epsilon}_k = y_k^{\text{obs}} - \hat{y}_k$, where $\hat{y}_k = \hat{y}_{imjk} = \hat{\beta}_{0j} + \hat{\beta}_{1j} x_{imjk}$. Using general techniques in linear mixed models (see Searle et al. (1992), chap. 7) we write in vector notation $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}$, where $\hat{\boldsymbol{y}}$ is the vector of predicted values of the modelled true values of the variable y; $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0\ \hat{\beta}_1)^T$ and $\hat{\boldsymbol{u}} = (\boldsymbol{u}_0^T\ \boldsymbol{u}_1^T)^T$ are estimators/predictors of the fixed $\boldsymbol{\beta}$ and random $\boldsymbol{u}$ effects under the model (6); and where the structure of the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ depends very sensitively on the details of the survey. In section 6 we will explicitly express them after applying these arguments to the industrial turnover index and industrial new orders received index surveys conducted at INE Spain. Then we have

$$\begin{aligned} \mathbb{E}\left[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T | s, \mathcal{G}\right] &= \mathbb{E}\left[(\boldsymbol{y}^{\text{obs}} - \hat{\boldsymbol{y}})(\boldsymbol{y}^{\text{obs}} - \hat{\boldsymbol{y}})^T | s, \mathcal{G}\right] \\ &= (\boldsymbol{y}^{\text{obs}} - X\hat{\boldsymbol{\beta}})(\boldsymbol{y}^{\text{obs}} - X\hat{\boldsymbol{\beta}})^T + X\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}]X^T + \boldsymbol{Z}\widehat{\mathbb{V}}[\hat{\boldsymbol{u}}]\boldsymbol{Z}^T, \end{aligned} \tag{7}$$

where $\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}]$ and $\widehat{\mathbb{V}}[\hat{\boldsymbol{u}}]$ are empirical estimators of the respective variances[7]. Denoting the matrix (7) by $E = \mathbb{E}\left[\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T | s, \mathcal{G}\right]$, we have indeed the positive semidefinite matrix B of our optimization problem (4):

$$[B]_{kl} = \omega_{ks}\omega_{ls}E_{kl}.$$

5 Solution to the univariate combinatorial problem

The combinatorial optimization problem to be solved is

$$\begin{aligned} & \max\ \mathbb{1}^T \boldsymbol{r} \\ \text{s.t.} \quad & \boldsymbol{r}^T B \boldsymbol{r} \leq v^2, \\ & \boldsymbol{r} \in \{0,1\}^{\times n}. \end{aligned} \tag{8}$$

Beforehand, we need some notation. We will denote by I the index set of components of $\boldsymbol{r} \in \{0,1\}^{\times n}$. Once and again we will consider a disjoint partition of I in two subsets $I = I_0 \cup I_1$ such that $I_0 = \{i \in I : r_i = 0\}$ and $I_1 = \{i \in I : r_i = 1\}$. The cardinality

[7] Those obtained after substituting the variance components $\boldsymbol{\theta}$ by their point estimation $\hat{\boldsymbol{\theta}}$ (e.g. if $\mathbb{V}[\hat{\boldsymbol{\beta}}] = \mathbb{V}[\hat{\boldsymbol{\beta}}](\boldsymbol{\theta})$, then $\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}] = \mathbb{V}[\hat{\boldsymbol{\beta}}](\hat{\boldsymbol{\theta}})$). These are often computed in statistical modelling software packages. See Searle et al. (1992), page 272 for general expressions of variances and covariances of BLUPs of both fixed and random effects.

of both subsets will be denoted by n_0 and n_1, respectively, so that $n = n_0 + n_1$. Note that having a particular partition (I_0, I_1) is equivalent to having a particular binary vector $\boldsymbol{r} \in \{0,1\}^{\times n}$. We denote by Π_k the set of binary vectors with k components equal to 1 and $n-k$ components equal to 0, that is, $\Pi_k = \{\boldsymbol{r} \in \{0,1\}^{\times n} : \sum_{i \in I} r_i = k\}$. Note that the set Π of all vertices of the hypercube $\{0,1\}^{\times n}$ can be decomposed as $\Pi = \bigcup_{k=0}^{n} \Pi_k$. The feasibility region will be denoted by $\mathcal{R} = \{\boldsymbol{r} \in \{0,1\}^{\times n} : \boldsymbol{r}^T B \boldsymbol{r} \leq v^2\}$. For later convenience we also introduce the convex hull of $\mathcal{R}$, which we denote by $\mathcal{R}_c = \{\boldsymbol{r} \in [0,1]^n : \boldsymbol{r}^T B \boldsymbol{r} \leq v^2\}$. Note that $\mathcal{R} = \mathcal{R}_c \cap \Pi$. Finally, given any real vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ we denote by $\bar{\boldsymbol{x}} = (x_{(1)}, \ldots, x_{(n)})$ the ordered vector such that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$, where this is a non-decreasing sequence of the values x_i of the components of $\boldsymbol{x}$.

In very general terms, we propose to reach an optimal solution of problem (8) in two steps:

1. Choose an initial binary vector $\boldsymbol{r}^{(0)}$ in Π_M, where $M \leq n$ is such that $\Pi_j \cap \mathcal{R} = \emptyset$ for all $j > M$.

2. Search Π_M for a binary solution such that it is also contained in $\mathcal{R}$. If it exists, then this is an optimal solution $\boldsymbol{r}^*$. If if does not exist, then jump down to Π_{M-1} and repeat this search recursively until finding an optimal solution.

We must make these steps concrete.

5.1 Initial binary vector

To choose the initial binary vector consider the convexified problem

$$\begin{aligned} & \max \mathbb{1}^T \boldsymbol{r}_c \\ \text{s.t.} \quad & \boldsymbol{r}_c^T B \boldsymbol{r}_c \leq v^2, \\ & \boldsymbol{r}_c \in [0,1]^{\times n}. \end{aligned} \tag{9}$$

If $\boldsymbol{r}^*$ denotes an optimal solution to the original problem (8) and $\boldsymbol{r}_c^*$ is an optimal solution to the convexified version (9), it is clear that $\mathbb{1}^T \boldsymbol{r}^* \leq \lfloor \mathbb{1}^T \boldsymbol{r}_c^* \rfloor$, since $\mathcal{R} \subset \mathcal{R}_c$, where $\lfloor \cdot \rfloor$ stands for the floor function.

Now use the spectral decomposition of B to write $B = ODO^T$ so that the problem (9) can be rewritten as

$$\begin{aligned} & \max \mathbb{1}^T \boldsymbol{s} \\ \text{s.t.} \quad & \boldsymbol{s}^T D \boldsymbol{s} \leq v^2, \\ & \boldsymbol{0} \leq \boldsymbol{s} \leq \boldsymbol{z}, \end{aligned} \tag{10}$$

where $\boldsymbol{s} = O^T \boldsymbol{r}_c$ and $\boldsymbol{z} = O^T \mathbb{1}$. Define the index sets J_0 and J_1 by $J_0 = \{i \in I : D_{ii} = 0\}$ and $J_1 = \{i \in I : D_{ii} \neq 0\}$. These two sets allow us to decompose any vector $\boldsymbol{a}$ as $\boldsymbol{a} = [\boldsymbol{a}_0 \ \boldsymbol{a}_1]$, with evident definitions of $\boldsymbol{a}_k$, $k = 0, 1$ and the diagonal matrix D as $D = D_0 \oplus D_1 = \operatorname{diag}\{D_{ii}\}_{i \in J_0} \oplus \operatorname{diag}\{D_{ii}\}_{i \in J_0}$. Problem (10) can be again rewritten as

$$\begin{aligned} & \max \ \boldsymbol{z}_1^T \boldsymbol{s}_1 + \boldsymbol{z}_0^T \boldsymbol{s}_0 \\ \text{s.t.} \quad & \boldsymbol{s}_1^T D_1 \boldsymbol{s}_1 \leq v^2, \\ & \mathbf{0} \leq \boldsymbol{s}_1 \leq \boldsymbol{z}_1, \\ & \mathbf{0} \leq \boldsymbol{s}_0 \leq \boldsymbol{z}_0, \end{aligned}$$

which can be divided into two separate problems:

$$\begin{aligned} & \max \ \boldsymbol{z}_0^T \boldsymbol{s}_0 \\ \text{s.t.} \quad & \mathbf{0} \leq \boldsymbol{s}_0 \leq \boldsymbol{z}_0, \end{aligned}$$

and

$$\begin{aligned} & \max \ \boldsymbol{z}_1^T \boldsymbol{s}_1 \\ \text{s.t.} \quad & \boldsymbol{s}_1^T D_1 \boldsymbol{s}_1 \leq v^2, \\ & \mathbf{0} \leq \boldsymbol{s}_1 \leq \boldsymbol{z}_1. \end{aligned}$$

The first one has an immediate optimal solution:

$$s_{0i}^* = \begin{cases} 0 & \text{if } z_{0i} \leq 0, \\ z_{0i} & \text{if } z_{0i} > 0. \end{cases}$$

We find an upper bound for the maximum value of the objective function of the second problem. Consider the relaxed problem

$$\begin{aligned} & \max \ \boldsymbol{z}_1^T \boldsymbol{s}_1 \\ \text{s.t.} \quad & \boldsymbol{s}_1^T D_1 \boldsymbol{s}_1 \leq v^2, \\ & \boldsymbol{s}_1 \in \mathbb{R}^{m_1}, \end{aligned}$$

where $m_1 = \operatorname{card} J_1$. Denote by $\mathcal{R}_1$ its feasibility region and by $\boldsymbol{s}_1^*$ its optimal solution. Note that D_1 is a positive definite matrix. Using Lagrange multipliers we readily have

$$\boldsymbol{s}_1^* = \sqrt{\frac{v^2}{\boldsymbol{z}_1^T D_1^{-1} \boldsymbol{z}_1^T}} D_1^{-1} \boldsymbol{z}_1,$$

which yields a maximal value of the objective function given by $\boldsymbol{z}_1^T \boldsymbol{s}_1^*$. Now, since $\{\boldsymbol{s}_1 \in [0,1]^{\times m_1} : \boldsymbol{s}_1^T B_1 \boldsymbol{s}_1 \leq v^2\} \subset \mathcal{R}_1$, we conclude that

$$\mathbb{1}^T \boldsymbol{r}^* \leq \lfloor \boldsymbol{z}_0^T \boldsymbol{s}_0^* + \boldsymbol{z}_1^T \boldsymbol{s}_1^* \rfloor.$$

This is the initial upper bound, which we will denote by $n_1^{(0)} = \lfloor \boldsymbol{z}_0^T \boldsymbol{s}_0^* + \boldsymbol{z}_1^T \boldsymbol{s}_1^* \rfloor$, determining the first set Π_M where we must seek an initial binary vector $\boldsymbol{r}$. Note that by construction we have $\Pi_k \cap \mathcal{R} = \emptyset$ for all $k > M$. Given $n_1^{(0)}$ define $I_1 = \{i \in I : B_{ii} \leq \overline{\text{diag}(B)}_{n_1^{(0)}}\}$ and $I_0 = I - I_1$. This is our initial binary vector. Note that if it is feasible (i.e. if $\sum_{i \in I_1} \sum_{j \in I_1} B_{ij} \leq v^2$), then it is an optimal solution. Otherwise we proceed to search consecutively $\Pi_{M-1}, \Pi_{M-2}, \Pi_{M-3}, \ldots, \Pi_0$ for a feasible binary vector.

5.2 Search in Π_k

In order not to run into computational troubles we must fix a criterion to carry out an exhaustive, although efficient, search for an optimal solution. To this end, firstly we look for the binary vector in Π_M yielding the minimum value of $\sum_{i \in I_1} \sum_{j \in I_1} B_{ij}$. If it is a feasible vector, i.e. if $\sum_{i \in I_1} \sum_{j \in I_1} B_{ij} \leq v^2$, then it is an optimal solution. Otherwise we resume the search in the set Π_{M-1}. The two key steps are (i) the choice of the initial vector when the search begins in each Π_k and (ii) the searching process in each Π_k.

We propose to use the following lemmas, which we state without proof, to build up a search algorithm:

Lema 5.1. If $I_1 = I_1^* \cup \{i^*\}$, where $i^* \in I_1$ and $I_1^* = I_1 - \{i^*\}$, then

$$\begin{aligned} \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} &= \sum_{i \in I_1^*} \sum_{j \in I_1^*} B_{ij} + 2 \sum_{i \in I_1^*} B_{ii^*} + B_{i^*i^*} \\ &= \sum_{i \in I_1^*} \sum_{j \in I_1^*} B_{ij} + 2 \sum_{i \in I_1} B_{ii^*} - B_{i^*i^*} \end{aligned}$$

Lema 5.2. If $I_1^* = I_1 \cup \{i^*\}$, where $i^* \in I_0$, then

$$\sum_{i \in I_1^*} \sum_{j \in I_1^*} B_{ij} = \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} + 2 \sum_{i \in I_1} B_{ii^*} + B_{i^*i^*}$$

From these lemmas we propose to search the binary vector with minimum $\sum_{i \in I_1} \sum_{j \in I_1} B_{ij}$ by iteratively redefining I_1 by $I_1 := (I_1 - \{i^*\}) \cup \{j^*\}$, where

$$\begin{aligned} i^* &= \text{argmax}_{i \in I_1} \{2 \sum_{k \in I_1} B_{ki} - B_{ii}\}, \\ j^* &= \text{argmin}_{j \in I_0} \{2 \sum_{k \in I_1 - \{i^*\}} B_{kj} + B_{jj}\}, \end{aligned}$$

until the minimal vector is found. To jump down from Π_k to Π_{k-1} when there is no feasible vector in Π_k we need the following lemma:

Lema 5.3. Let $I_1^{(0)} = I_1^{(k)} \cup J_k$, where $J_k = \{i_1, \ldots, i_k\} \subset I_1^{(0)}$ and $I_1^{(k)} = I_1^{(0)} - J_k$ for $k = 1, \ldots, K \leq n$. Define $\delta(I_1^{(k-1)}, i_k) = 2 \sum_{i \in I_1^{(k-1)}} B_{ii_k} - B_{i_k i_k}$ and $\Delta(I_1^{(k)}) = \sum_{i \in I_1^{(k)}} \sum_{j \in I_1^{(k)}} B_{ij}$ for $k = 1, \ldots, K$. Then

$$\Delta(I_1^{(0)}) = \Delta(I_1^{(k)}) + \sum_{j=1}^{k} \delta(I_1^{(j-1)}, j) \qquad \forall k = 1, \ldots, K.$$

Proof. Use lemma (5.1) recursively. □

In the light of this result, to jump down from Π_k to Π_{k-1}, we choose

$$j_k = \text{argmax}_{j \in I_1^{(k-1)}} \delta(I_1^{(k-1)}, j).$$

Thus we have the ingredients to propose a complete algorithm.

5.3 The algorithm

Putting all preceding steps together we have an algorithm to reach the optimal solution of the univariate combinatorial problem (8). The complete algorithm is detailed in mathematical style pseudocode in appendix A. The analysis of this algorithm and its computational efficiency lies beyond the scope of this paper. Our goal is to show that an algorithm exists and that it works with actual data from an official survey.

Note that the algorithm considered here does not attempt to find all optimal solutions, only one yielding the minimum value of $\sum_{i \in I_1} \sum_{j \in I_1} B_{ij}$ is sought. However, note that once an optimal solution is found with our algorithm, it is immediate to proceed the search for the rest of optimal solutions as follows. Given our optimal solution $I^* = I_1^* \cup I_0^*$, define $I_1^{**} = (I_1^* - \{i^*\}) \cup \{j^*\}$, where

$$\begin{aligned} i^* &= \text{argmax}_{i \in I_1^*} \left\{ 2 \sum_{k \in I_1^*} B_{ik} - B_{ii} \right\}, \\ i^* &= \text{argmin}_{j \in I_0^*} \left\{ 2 \sum_{k \in I_1^*} B_{jk} + B_{jj} \right\}. \end{aligned}$$

If I^{**} is feasible, then it is also optimal and you can apply the same argument iteratively; otherwise there is no more optimal solutions.

6 Application to the Spanish industrial turnover index and industrial new orders received index surveys

We applied the preceding proposal to the Spanish industrial turnover index (ITI) and industrial new orders received index (INORI) surveys conducted at INE Spain as follows. Beforehand we need a brief description of these statistical operations. The Spanish ITI and INORI are short-term statistics whose data are collected monthly and jointly through

Internet and mail questionnaires. The sampling design is an extreme cut-off sampling design where $\omega_{ks} = 1$ for all those industrial establishments included in the sample. Both indexes are Laspeyres-type indexes. Different levels of disaggregation according to Eurostat's so-called main industrial groupings [REF] are published at the national realm. Internet questionnaires are received directly from the respondents in our central premises whereas mail questionnaires are collected locally at provincial delegations where the E&I process begins immediately and then are sent to the central premises, where a final E&I stage is undergone for both the electronic and paper modes.

To apply our preceding proposal to the Spanish ITI and INORI we focused only on those data collected during 2008 through Internet since we had both raw and edited data for these respondents, which amounted up to one third of the total sample. We discarded January data since for some unknown reason the sample size was too reduced in this period. So, starting in February ($I_{Feb,2008} = 100$), which was fully interactively edited, we applied our proposal month to month choosing a somewhat arbitrary value of the accuracy control parameter v^2 when passing from one period to the next.

Firstly we adapted the generic E&I strategy depicted in figure 1 in section 2 to these surveys. The initial E&I phase was designed in four stages, namely (i) we dealt with unit measure errors by taking into account digit numbers difference for each respondent between two consecutive periods and multiplying by a factor 1000 when such a diference was 3; (ii) we turned every missing value of each variable entering a balance equation satisfied by the reported non-missing values into a zero value; if the balance equation was not satisfied by the latter values, we carried out no change; (iii) those respondents with variable values above an index threshold[8] were assigned to interactive E&I; and finally (iv) all respondents with a missing value either on the current or the preceding period were also assigned to interactive E&I. Given that the Internet-mode collected sample is considerably lesser than the whole sample, we focused only on the general index, discarding lower levels of disagreggation in main industrial groupings.

Up to this point we had two data streams, that entering interactive E&I and that entering selective E&I. We applied the preceding selective editing proposal to the latter by choosing y_k as value at the current period of the variable of interest and x_k as the value of the same variable at the preceding month. Note that the initial E&I design ensured that the (x, y) set used to build the linear model was complete, i.e. it had no missing value. The linear model gained simplicity from the facts that there was only one collection mode (Internet questionnaire) and no interviewer. The branch $j = 1, \ldots, J$ was determined by a two-digit NACE variable. Thus the model read in matrix notation

$$\mathbf{y} = \beta_0 \mathbb{1} + \beta_1 X + Z_0 \mathbf{u}_0 + Z_1 \mathbf{u}_1 + \mathbf{e},$$

where[9]

[8] Set arbitrarily to 1% of the index computed up to the current status of editing.

[9] $\{\mathbf{e}_j\}_{j=1,\ldots,J}$ denote the canonical vectors, $\{E_{jj}\}_{j=1,\ldots,J}$ are the diagonal Weyl matrices and $\otimes$ stands

- $\mathbf{y} = \sum_{j=1}^{J} \mathbf{e}_j \otimes (y_{j1}^t, \ldots, y_{jn_j}^t)^T$;
- $X = \sum_{j=1}^{J} \mathbf{e}_j \otimes (y_{j1}^{t-1}, \ldots, y_{jn_j}^{t-1})'$;
- $[\mathbf{u}_p]_j = u_{qj}$, for $q = 0, 1$ and $j = 1, \ldots, J$;
- $Z_0 = \sum_{j=1}^{J} E_{jj} \otimes \mathbb{1}_{n_j}$;
- $Z_1 = \sum_{j=1}^{J} E_{jj} \otimes \mathbf{x}_j$;
- $\mathbb{V}\begin{pmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} G_{2J\times 2J} & 0_{2J\times n} \\ 0_{2J\times n}^T & \sigma_e^2 \mathbb{I}_n \end{pmatrix}$, with $G = \begin{pmatrix} \sigma_{u0}^2 & 0 \\ 0 & \sigma_{u1}^2 \end{pmatrix} \otimes \mathbb{I}_J$.

Then $\hat{y}_{jk}^t$ was computed for each unit using the lme4 package in language R (Bates and Maechler, 2010), so that $\hat{\epsilon}_{jk} = y_{jk}^{obs} - \hat{y}_{jk}$ and the recipe in section 4 driving us to the matrix E applied.

Once we carried out the selective editing phase new respondents were assigned to interactive E&I and the rest entered into automatic E&I. In our study the interactive E&I amounted to recovering the final edited value accepted in the original survey which underwent the traditional E&I mode in 2008 at INE Spain's central premises. On the other hand, we reduced the automatic E&I to making no change whatsoever on the reported raw data. Moreover, no further macro E&I phase was carried out. Both final data stream were joined back together to compute the final index. The selective phase was applied separately both to turnover and new orders received variables but assigning a respondent to interactive E&I whenever it was so assigned in *any* of both variables.

Complete results are depicted in the graphs in appendix B. By and large, a reduction in the amount of interactive E&I is observed associated to a loss of accuracy in the indexes, where this loss of accuracy is understood as a departure from the index computed under the traditional E&I strategy. Furthermore, this association is established in a controlled way, since as a result of the method the set of units entering interactive E&I is flagged while keeping the accuracy under control after choosing the accuracy control parameter v^2.

After the somewhat arbitrarily choice of the accuracy control parameter on each period[10] we computed the general ITI and INORI for both the data edited under the traditional strategy and the present strategy and compared them to the published original indexes obtained with the whole sample, but changing the basis to make it coincide with our previous choice $I_{Feb,2008} = 100$. The comparison is illustrated in figure 2.

for the Kronecker product.

[10]Subject matter expertise must guide this choice; ours in this study was $v^2 = (2 \cdot 10^8, 2 \cdot 10^5, 10^6, 5 \cdot 10^5, 10^5, 10^5, 5 \cdot 10^5, 1.4 \cdot 10^5, 10^5, 2 \cdot 10^5)$ for each period.

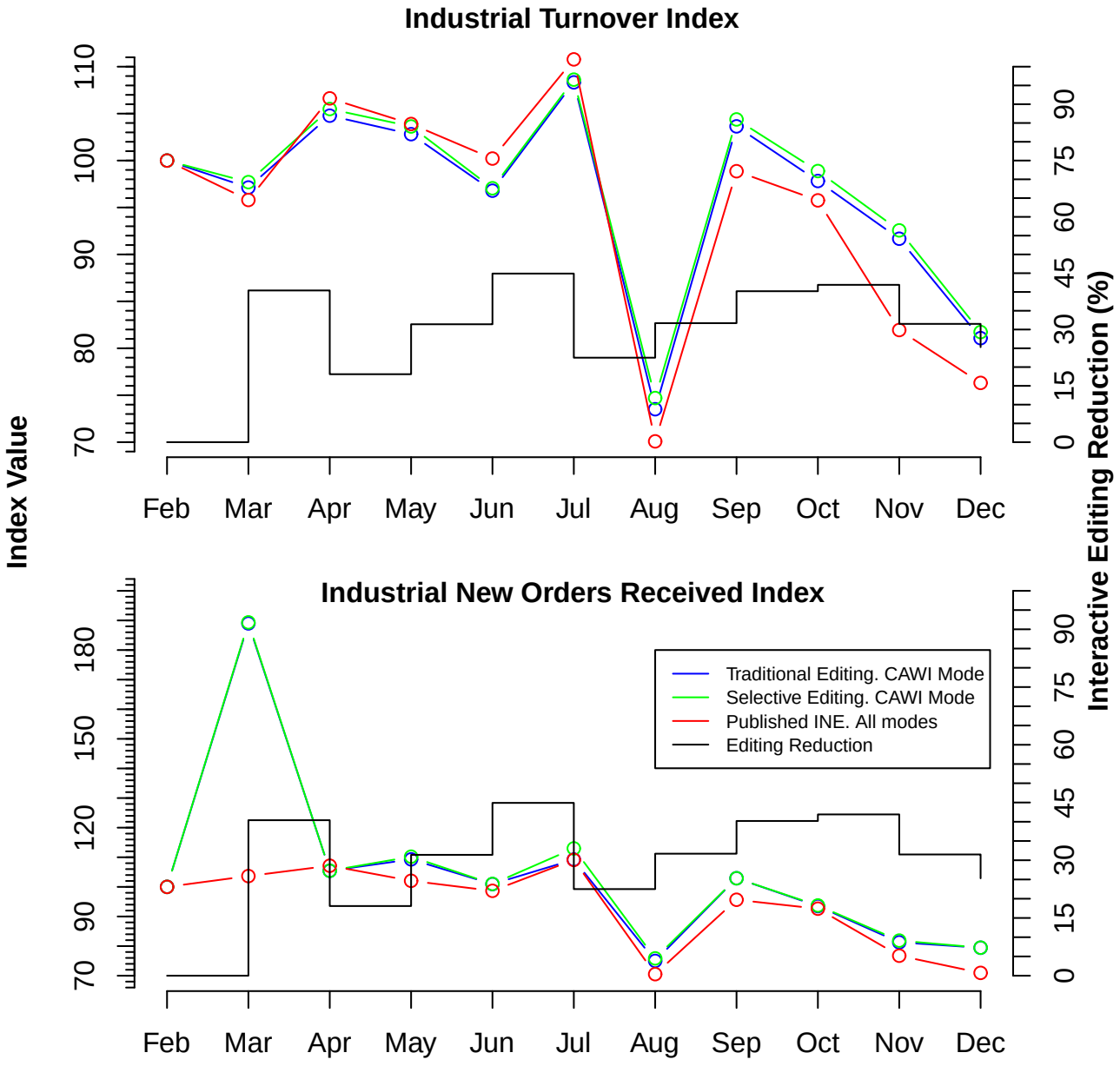


Figure 2: Comparison of ITI and INORI under traditional and selective editing.

7 Discussion

We discuss some relevant issues concerning this proposal both from the theoretical viewpoint and in the light of the preceding application.

7.1 Theoretical considerations

Comparison with the general strategy in the literature.- The preceding approach to selective editing shows both similarities and differences with the four-element general strategy depicted in the introduction (Lawrence and McKenzie, 2000), which we comment in detail.

In first place, the role of the expected amended values is now played by the conditional moments μ_{ks} and σ_{kls}. As in the original approach, the conditional moments should be chosen in line with the general editing model being used, they will usually rely on subject matter knowledge and they do not need to be accurate enough as to be used as imputed values. They provide the basis to establish a comparison of the relative importance, across units, of the errors to be edited and are obtained within an assumed linear mixed model.

Secondly, score functions are not used in this approach. This is an important difference. Score functions are used to provide a final ranking of respondents. Here this ranking is substituted by a direct determination of those units to be recontacted. Despite the fact that in both approaches a final subset of the sample s is obtained (mathematically indicated by $\boldsymbol{r}_s$), in the original scheme we also have information about which units show a greater incidence on the estimates, whereas in the optimization scheme we lack this. In contrast, in the former both local and the global score functions must be chosen, while in the latter this is not necessary. In our view this can be understood as an advantage since the intervention of the statistician is reduced in the survey estimates production. This is especially relevant in important large-scale surveys with far-reaching consequences, "where 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" (Hansen et al. (1983), p. 785).

Finally, the role of cut-off scores is now played by the accuracy control parameters v_p^2, $p = 1, \ldots, P$. The choice of these bounds, as with the cut-off scores, should be based on subject matter knowledge and should follow external considerations related to the accuracy of the survey estimates. In particular a thorough analysis like the one conducted in the preceding section using data from previous periods is highly recommended.

The role of the accuracy control parameters v_p^2, data imputation and variance estimation.- By the choice of the name "accuracy control parameters" for v_p^2 we do not intend to convey the idea that the final variance estimation is completely controlled by them. Let us remark clearly that we are only tackling with the error detection step in the selective editing phase of the E&I strategy. Once influential units entering interactive editing are identified, data imputation must be undergone and consequently variance estimation must take this whole process into account. In this sense the accuracy control parameters v_p^2 are just a small part of the entire E&I strategy whose role in the variance estimation is left as future research.

Disaggregated estimates.- In nearly all official statistics, estimates are required to be published in disaggregated domains (size categories, classification branches, regional areas, etc.). In the preceding applications this disaggregation has not been carried out; only fully aggregated estimates have been considered. If we were to give disagreggated estimates, we would have to identify influential units at such a disagreggated level by splitting up the sample into the corresponding disjoint domains and perform the same preceding analysis in each one.

Multivariate generalization.- Nowadays every survey collects values of several variables. In this respect, two comments must be made. Firstly, the present proposal only considers univariate editing tasks, i.e. trivial questionnaires with only one item. This comes from the optimization problem (8), where the feasible region is determined accounting only for the accuracy control on one aggregate $\hat{Y}$. However in the application to the ITI and INORI surveys we have overcome this strict limitation by applying the selective editing proposal as many times as the number of items to be edited (two in our case). Then every questionnaire with a flagged item is set to enter interactive E&I. Being aware of the practical limitations of this approach we recognize the multivariate generalization of the algorithm as an open problem for future research, considering this proposed shortcut as an upper bound to the solution of the multivariate problem.

Secondly only linear estimators of the form $\hat{Y} = \sum_{k \in s} \omega_{ks} y_k$ have been considered in the formulation of the optimization problem. This is another limitation, since, for example, ratio estimators like $\hat{R} = \hat{Y}^{(1)}/\hat{Y}^{(2)}$, where $\hat{Y}^{(i)}$ are both estimators arising from data collected in the survey, although much used in practice, are ruled out. We propose another roundabout solution by considering separately the selective editing of $\hat{Y}^{(1)}$ and $\hat{Y}^{(2)}$, making each unit k enter interactive E&I whenever it is flagged in any of both cases.

Algorithm efficiency.- Although a complexity analysis of the presented algorithm is beyond the scope of this paper, we point out that an efficiency gain would be obtained if we could enhance the initial binary vector choice step. Note that the step presented herein is not sensitive to the real parameter v^2 when going from the upper bound obtained from the convexified problem to the binary initial vector. We claim that a more clever choice taking into account v^2 will increase the efficiency of the algorithm rendering it more practical.

The model.- Our choice of the model, and in particular that of the exogenous variable $x_k = y_k^{t-1}$, seems to suggest that we are not actually exploiting cross-sectional auxiliary information, but longitudinal information. This is not the case. The landmark in the model is that the estimation/prediction of its coefficients arises from the relation of the exogenous variables to the endogenous ones all throughout the sample[11]. The time dependency of the endogenous variables is accesory: had we an alternative independent variable,

[11] In rigour, the subset of the sample not entering interactive E&I at the stage of construction of the model.

we would have obtained a similar result. In our view, this poses two interesting features of the model we chose. On the one hand, the model is versatile because it does not depend on the particular meaning of the variables involved in these particular surveys and could be applied to other surveys with continuous variables. On the other hand, the model is very simple and could be sophisticated with more variables if necessary. Indeed, keeping the same simple structure, combination of changes referring to the branches and to the independent variables immediately makes the number of possibilities to explore proliferate.

Finally the role of the term e_{imjk} in model (6) as a mixture of the error due to the respondent and the residual of the model looks arguably suspicious. Indeed this is a delicate point since a poor model fitting could be wrongly understood as a measurement error possibly not being the case. The residual e_{imjk} carries both features: lack of model fitting and respondent error. Then it is compulsory for the model to be trustworhty enough as to avoid the misinterpretation of e_{imjk}. This is where the relaxation on the demand about the estimation of the moments μ_{ks} and σ_{kls} within acceptable bounds fits in.

7.2 Considerations in the light of the preceding application

In view of the results reported in the preceding section we can comment the following. Firstly we can observe a gradual stepwise variation in the index error as the accuracy control parameter v^2 increases. This variation, in our view, shows two components: an increasing drift and a random element. The increasing drift reveals an intuitive fact: the lesser we edit, the greater the index error. On the other hand, the random element reflects mimetically the random nature of the effect of measurement errors on aggregates. Secondly, the number of units entering interactive E&I shows a clear monotically decreasing behaviour with respect to the accuracy control parameter v^2. This is also intuitive: the less accuracy demanded, the less number of units needed to enter interactive E&I.

Nonetheless we remark some relevant points. Firstly, despite choosing $v^2 = 0$, which presumably should drive us to no error in the index, this is not actually the case. The reason is that the index computation depends heavily on its value on the preceding period, which already contained the error resulting from any choice such that $v^2 > 0$ on that period. Thus this choice should take this cumulative effect into account. Secondly a great departure from the published full-sample-based INORI is observed in March both for the traditional and selective editing strategies. This is an immediate consequence of the lack of a macroediting phase. This departure arises from the existence of a single respondent with a true outlier, meaning having the same variable value before and after the traditional editing strategy. This shows that this proposal is intended to be part of a whole strategy. Finally, some exceptions apparently appear in some months where the gradual stepwise drift is indeed decreasing. We believe that this is a consequence of a dominance of the random component over the gradual drift together with the cumulative error pointed out above, which entails a nonnegligible effect upon the initial error for $v^2 = 0$.

Finally, the reduction of interactive E&I is not especially significant with those values

of v^2 chosen in our analysis (see figure 2). However judging from the closeness between the indexes under the traditional and the selective editing strategies, we believe that other choices of v^2 on each month would entail a higher interactive editing reduction keeping the indexes within an acceptable margin of error. This is obviously under the survey conductors' judgment.

8 Conclusions

We have proposed to formulate the selective editing stage of an E&I strategy as a combinatorial optimization problem by defining a binary variable $r_{ks} \in \{0,1\}$ for each unit k in the sample s denoting whether it must undergo interactive editing ($r_k = 0$) or automatic editing ($r_k = 1$). After modelling for each variable of interest $y^{(p)}$ the so-called *editing bias* and *editing variance* estimators $\widehat{\mathbb{B}}_s^{\text{ed}} = \sum_{k\in s} \omega_{ks} r_{ks} \mu_{ks}$ and $\widehat{\mathbb{V}}_s^{\text{ed}} = \sum_{k\in s} \sum_{l\in s} \omega_{ks}\omega_{ls} r_{ks} r_{ls} \sigma_{kls}$, respectively, where ω_{ks} is the sampling weight for unit k, μ_{ks} is the expected value of the random measurement error ϵ_k of unit k and σ_{kls} the covariance of this error of both units k and l, the problem reads

$$\begin{aligned} & \max \sum_{k\in s} r_{ks} \\ \text{such that} \quad & \widehat{\widehat{\mathbb{V}}}_s^{(p)\text{ed}} + \left(\widehat{\widehat{\mathbb{B}}}_s^{(p)\text{ed}}\right)^2 \leq v_p^2 \qquad p = 1, \ldots, P \\ & r_{ks} \in \{0,1\} \qquad k = 1, \ldots, n_s. \end{aligned}$$

The notation $\widehat{\widehat{\mathbb{V}}}_s^{(p)\text{ed}}$ and $\widehat{\widehat{\mathbb{B}}}_s^{(p)\text{ed}}$ reflects that μ_{ks} and σ_{kls} must be substituted by their modelled counterparts. This is undergone using a linear mixed model, where cross-sectional auxiliary information is incorporated. Here only the univariate version ($P = 1$) has been tackled, giving a detailed algorithm for its solution. We have proposed a roundabout solution to this limitation by suggesting to apply the univariate algorithm to each of the involved variables.

We have applied the proposal to the Spanish industrial turnover index and industrial new orders received index surveys. As a general result we have obtained a reduction of the amount of interactive editing (hence of editing resources) controllably impinging upon the accuracy of the resulting indexes.

As potential advantages of this approach to selective editing, we can state that the selection of units for interactive editing is simply based upon editing resources reduction and estimates accuracy control and that the survey conductors' intervention in the production process is reduced with respect to more traditional approaches. On the contrary, we lose the ordering of all sampled units according to a global score as in traditional methods. This obliges us to solve the optimization problem for each choice of the accuracy control

parameter v^2.

Several questions keep naturally open. Firstly, a generalization of this formulation is desirable providing a general algorithm to find a solution to the multivariate problem (3) and a treatment of nonlinear estimators $\hat{\theta} = g(\hat{Y}^{(1)}, \ldots, \hat{Y}^{(P)})$. Secondly, once the E&I strategy is completed with a carefully chosen imputation scheme, a further analysis must be undertaken relating the accuracy control parameter v^2 and the estimation of the variance of the final estimate under the whole E&I process. Finally, the use of linear mixed models makes us cherish the hope that arguably a more general formulation of the problem giving room for generalized linear mixed models can pave the way for carrying out selective editing on qualitative variables. All these are left for future research.

A The complete algorithm

This is the pseudocode in mathematical style of the complete algorithm to solve the univariate combinatorial problem (8).

Complete Algorithm

$\boldsymbol{z} := O^T \mathbb{1}$;
$J_0 := \{i \in I : D_{ii} = 0\}$;
$J_1 := I - J_1$;
$\boldsymbol{z}_0 := [z_i]_{i \in J_0}$;
$\boldsymbol{z}_1 := [z_i]_{i \in J_1}$;
$D_1 := \mathrm{diag}\{d_i\}_{i \in J_1}$;
$\boldsymbol{s}_0^* := \mathrm{argmax}_{\mathbf{0} \leq \boldsymbol{s}_0 \leq \boldsymbol{z}_0} \boldsymbol{z}_0^T \boldsymbol{s}_0$;
$\boldsymbol{s}_1^* := \mathrm{argmax}_{\boldsymbol{s}_1^T D_1 \boldsymbol{s}_1 \leq v^2} \boldsymbol{z}_1^T \boldsymbol{s}_1$;
$n_1 := \lfloor \boldsymbol{z}_0^T \boldsymbol{s}_0^* + \boldsymbol{z}_1^T \boldsymbol{s}_1^* \rfloor$;
$I_1 := \left\{ i \in I : d_i \leq \overline{\mathrm{diag}(D)}_{n_1} \right\}$;
$I_0 := I - I_1$;
$\Delta_{act} := \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} - v^2$;
$i^* := \mathrm{argmax}_{i \in I_1} \left\{ 2 \sum_{k \in I_1} B_{ki} - B_{ii} \right\}$;
$j^* := \mathrm{argmin}_{j \in I_0} \left\{ 2 \sum_{k \in I_1 - \{i^*\}} B_{kj} + B_{jj} \right\}$;
$\Delta_{pre} := \Delta_{act}$;
$I_1^{(pre)} := I_1$;
$I_0^{(pre)} := I - I_1^{(pre)}$;
$I_1 := (I_1 - \{i^*\}) \cup \{j^*\}$;
$I_0 := I - I_1$;

$\Delta_{act} := \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} - v^2$;

WHILE $\Delta_{pre} > \Delta_{act}$

DO $i^* := \mathrm{argmax}_{i \in I_1} \left\{ 2 \sum_{k \in I_1} B_{ki} - B_{ii} \right\}$;

$j^* := \mathrm{argmin}_{j \in I_0} \left\{ 2 \sum_{k \in I_1 - \{i^*\}} B_{kj} + B_{jj} \right\}$;

$I_1^{(pre)} := I_1$;

$I_1 := (I_1 - \{i^*\}) \cup \{j^*\}$;

$I_0^{(pre)} := I - I_1^{(pre)}$;

$I_0 := I - I_1$;

IF $\Delta_{act} \leq 0$ THEN RETURN $r_i^* = \begin{cases} 1 & i \in I_1 \\ 0 & i \in I_0 \end{cases}$;

$\Delta_{pre} := \Delta_{act}$;

$\Delta_{act} := \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} - v^2$;

ENDWHILE ;

$I_1 := I_1^{(pre)}$;

$I_0 := I_0^{(pre)}$;

WHILE $\Delta_{act} := \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} - v^2 > 0$

DO $i^* := \mathrm{argmax}_{i \in I_1} \left\{ 2 \sum_{k \in I_1} B_{ki} - B_{ii} \right\}$;

$I_1 := I_1 - \{i^*\}$;

$I_0 := I - I_1$;

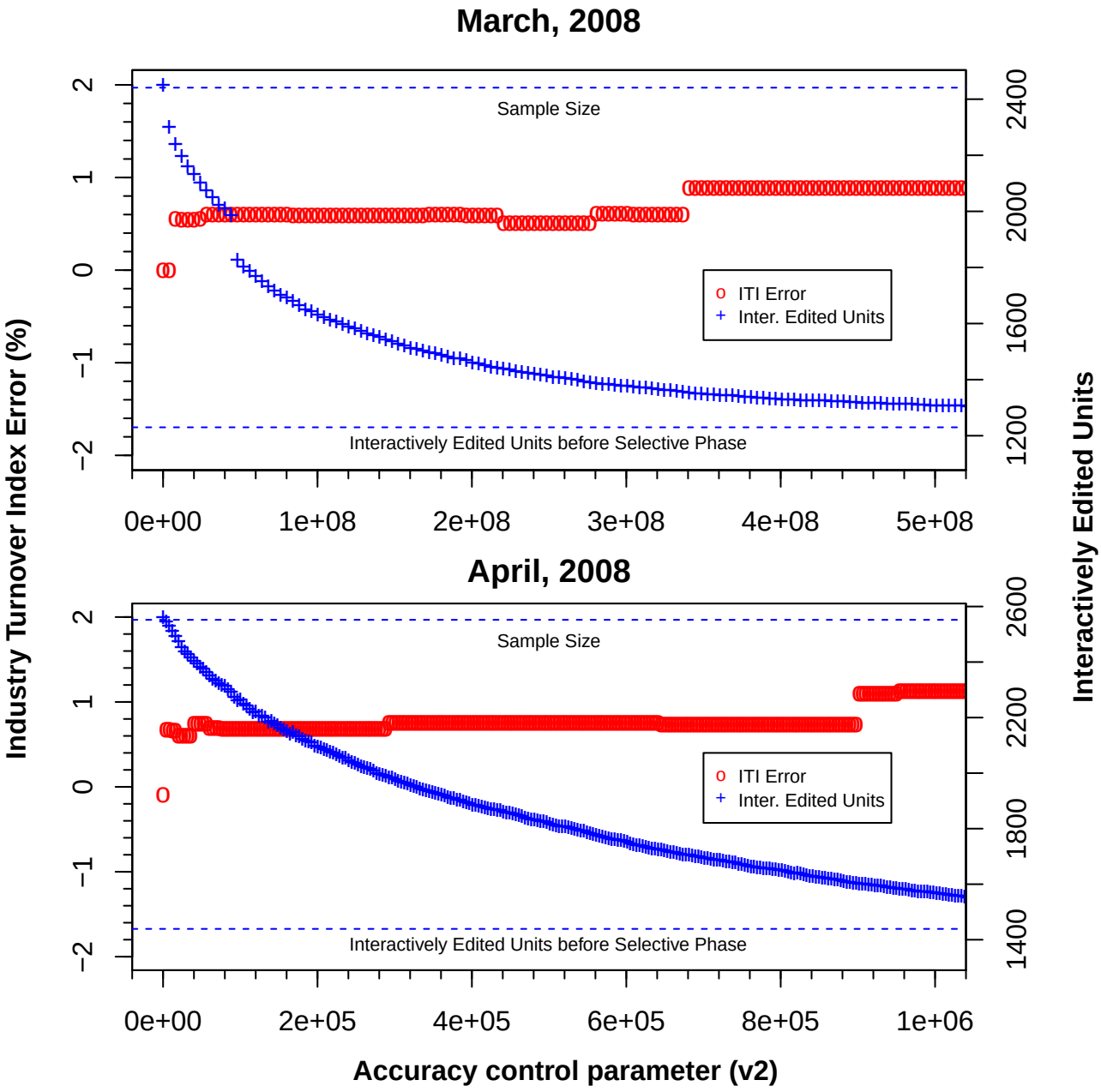
$\Delta_{act} := \sum_{i \in I_1} \sum_{j \in I_1} B_{ij} - v^2$;

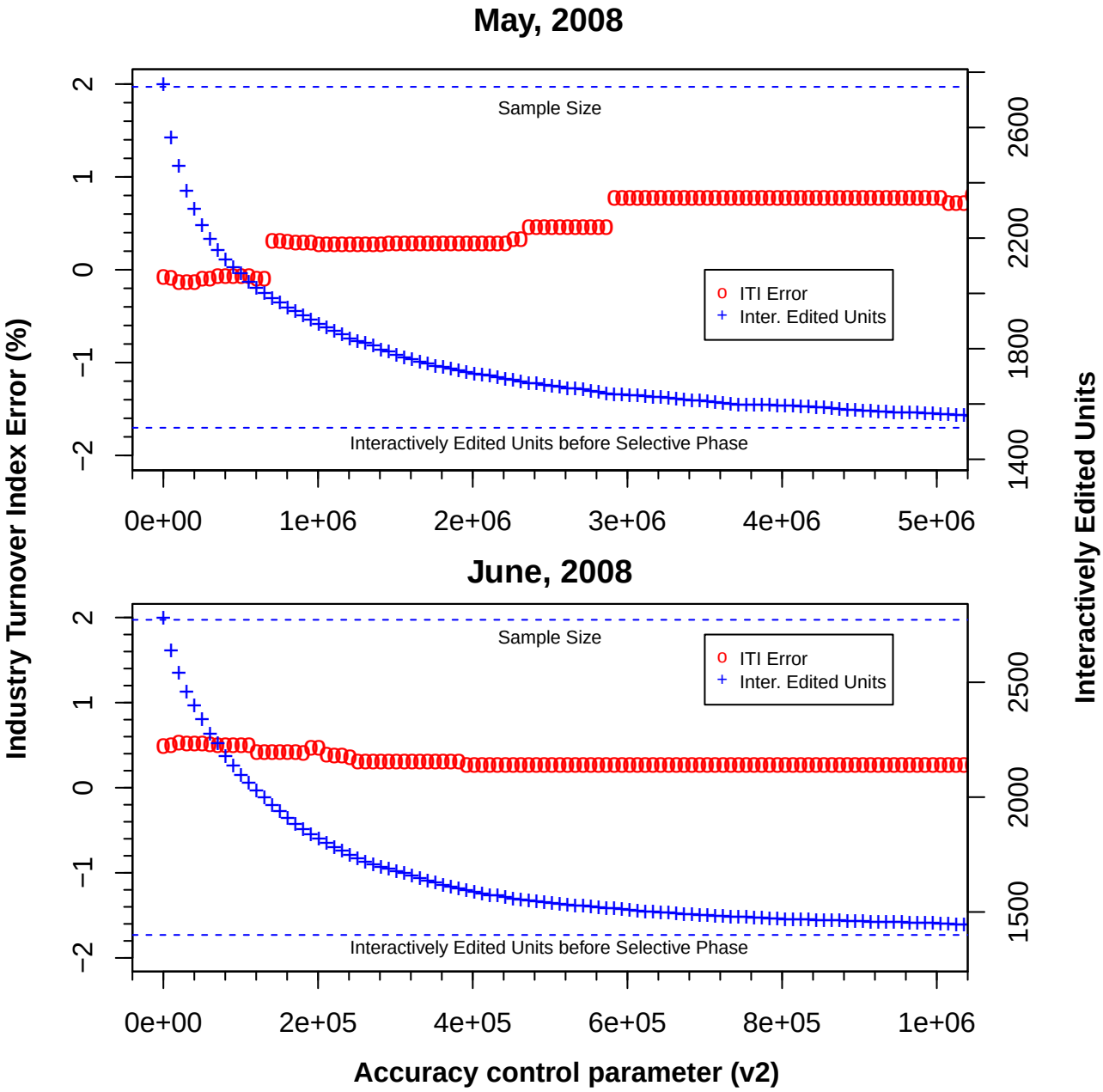
ENDWHILE ;

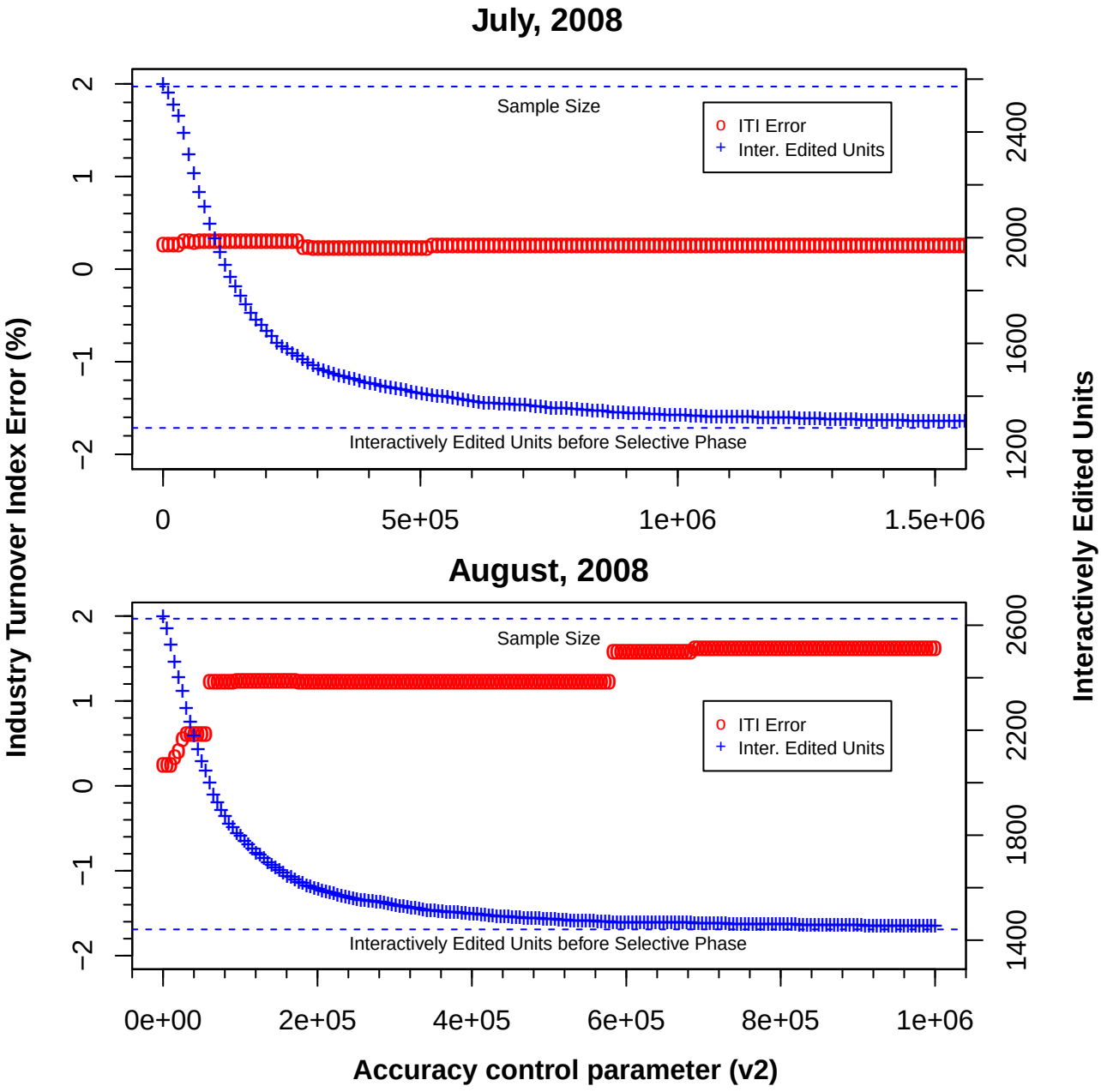
RETURN $r_i^* = \begin{cases} 1 & i \in I_1 \\ 0 & i \in I_0 \end{cases}$.

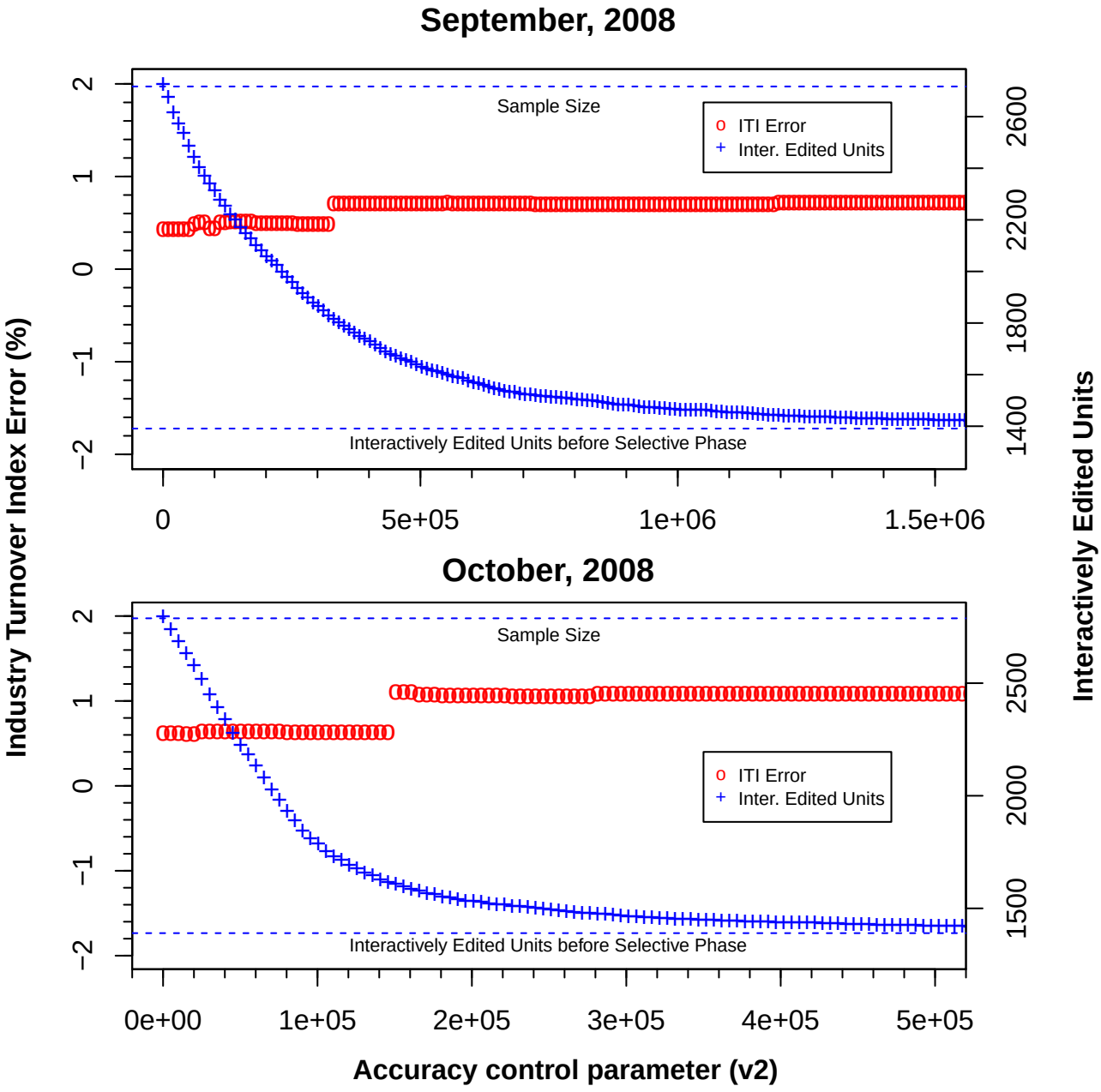
B Graphs

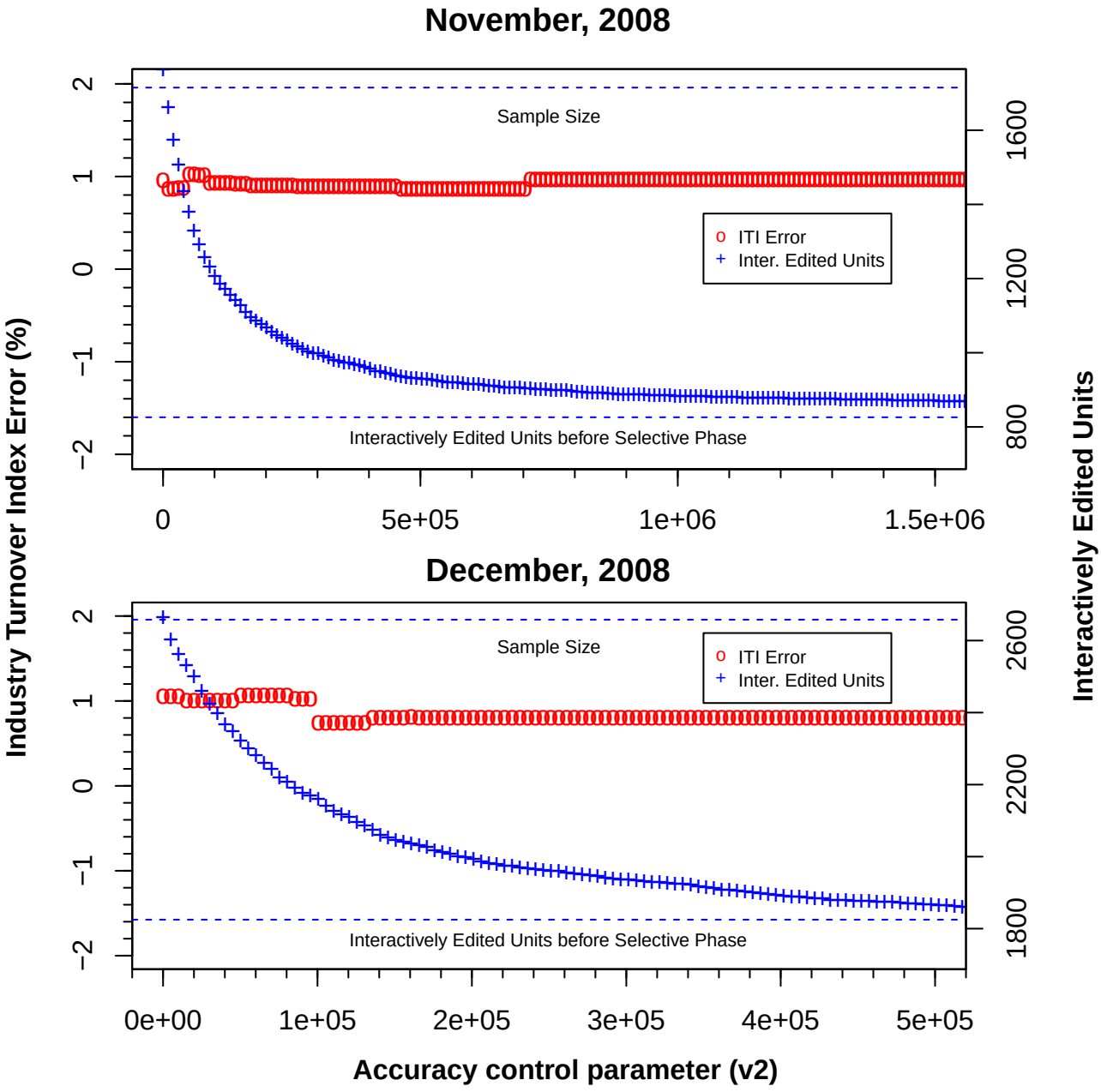
This appendix gathers all graphs resulting from the application of the selective editing proposal to both the industrial turnover index and industrial new orders received index on each month from March, 2008 to December, 2008 (see text for details).

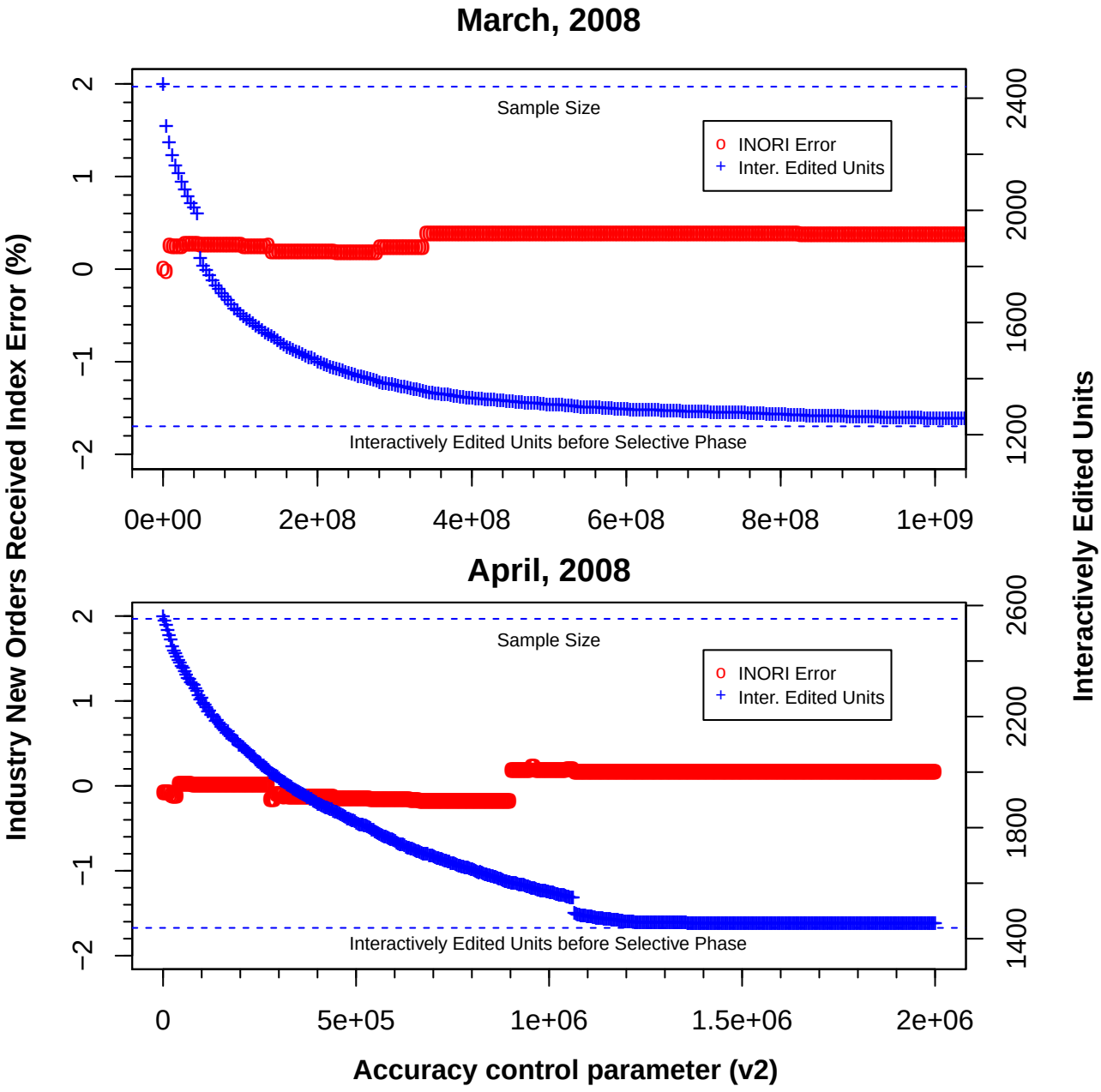


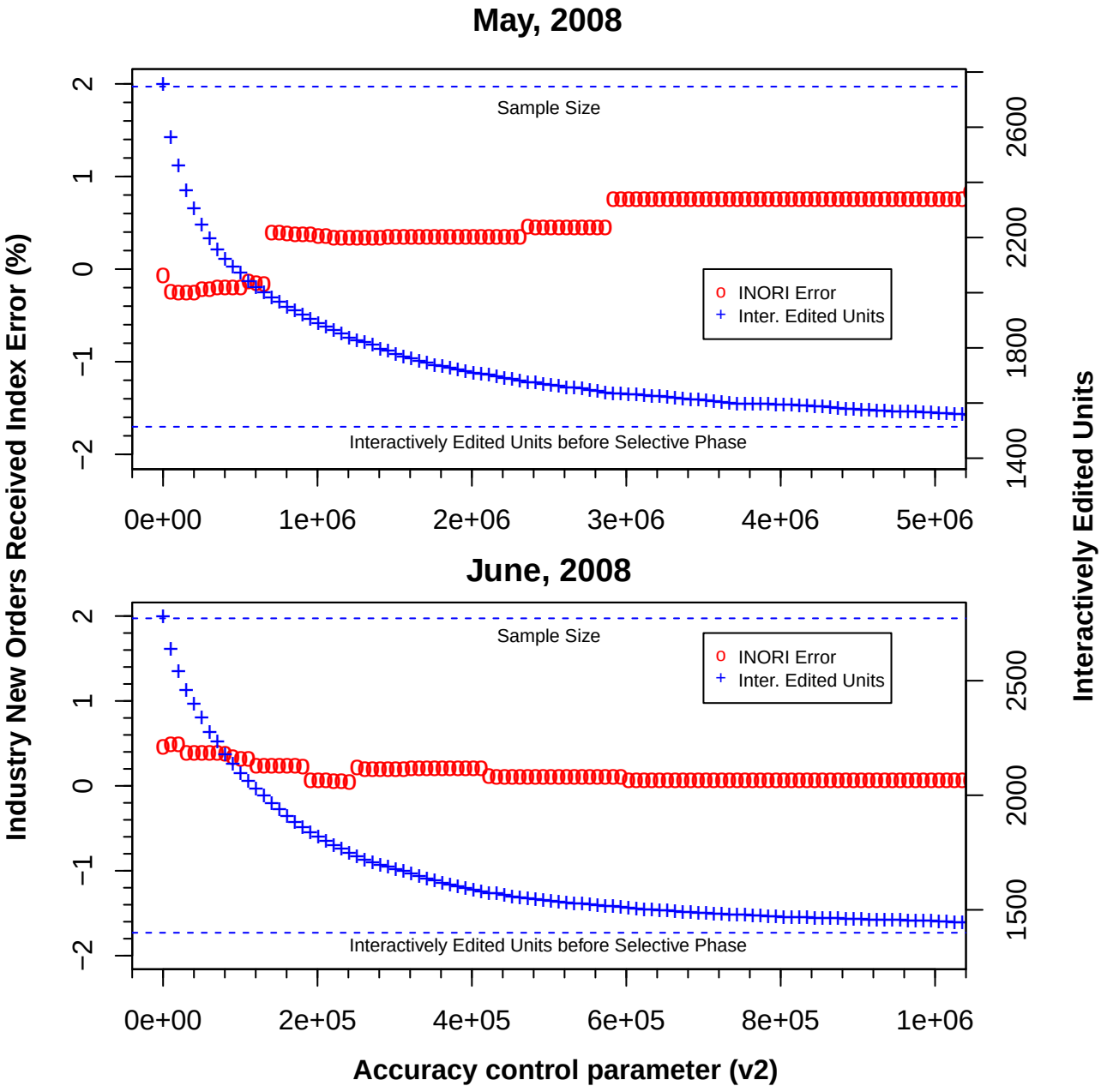

May, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
June, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry Turnover Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)

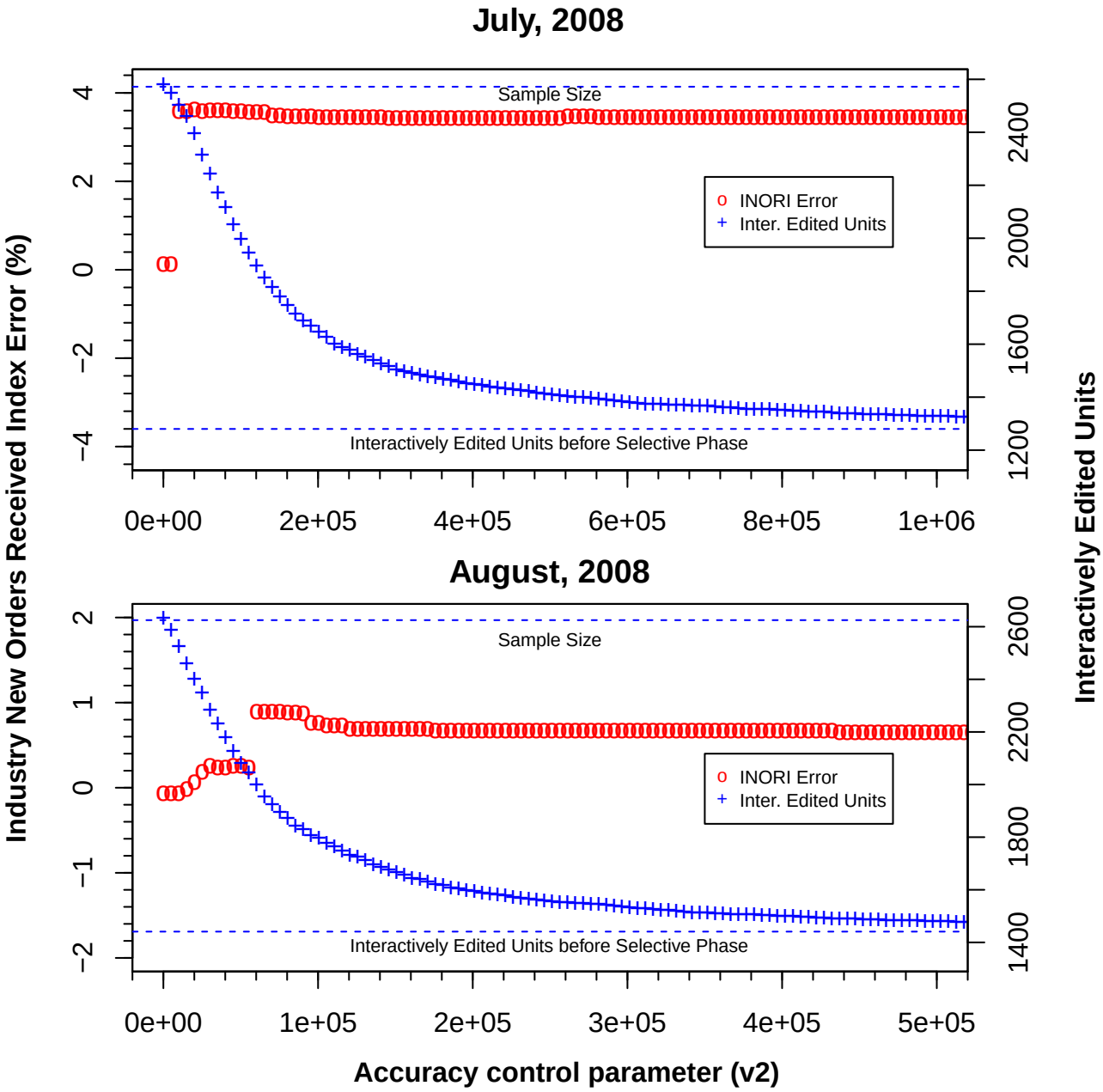

July, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
August, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry Turnover Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)

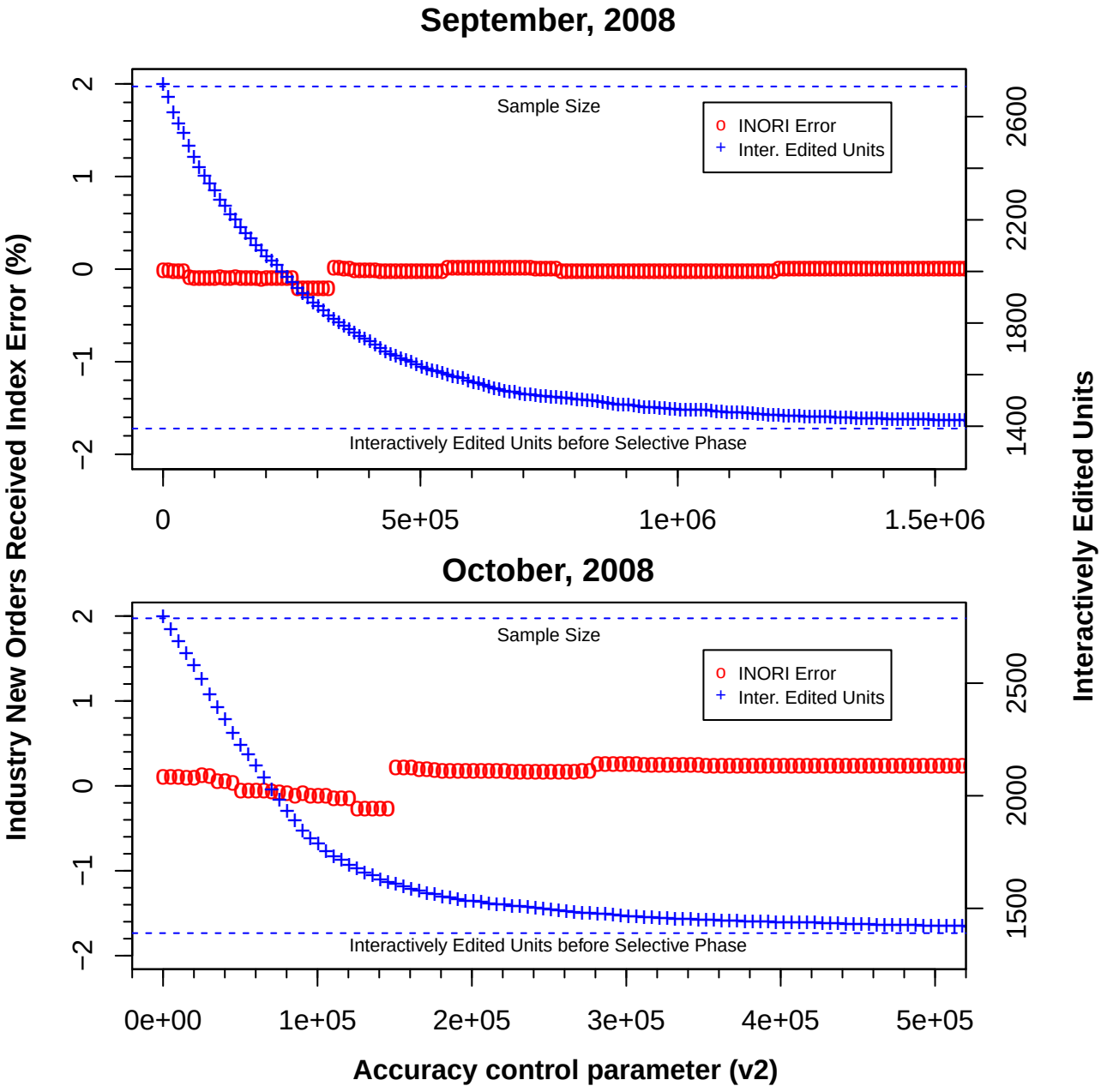

September, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
October, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry Turnover Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)

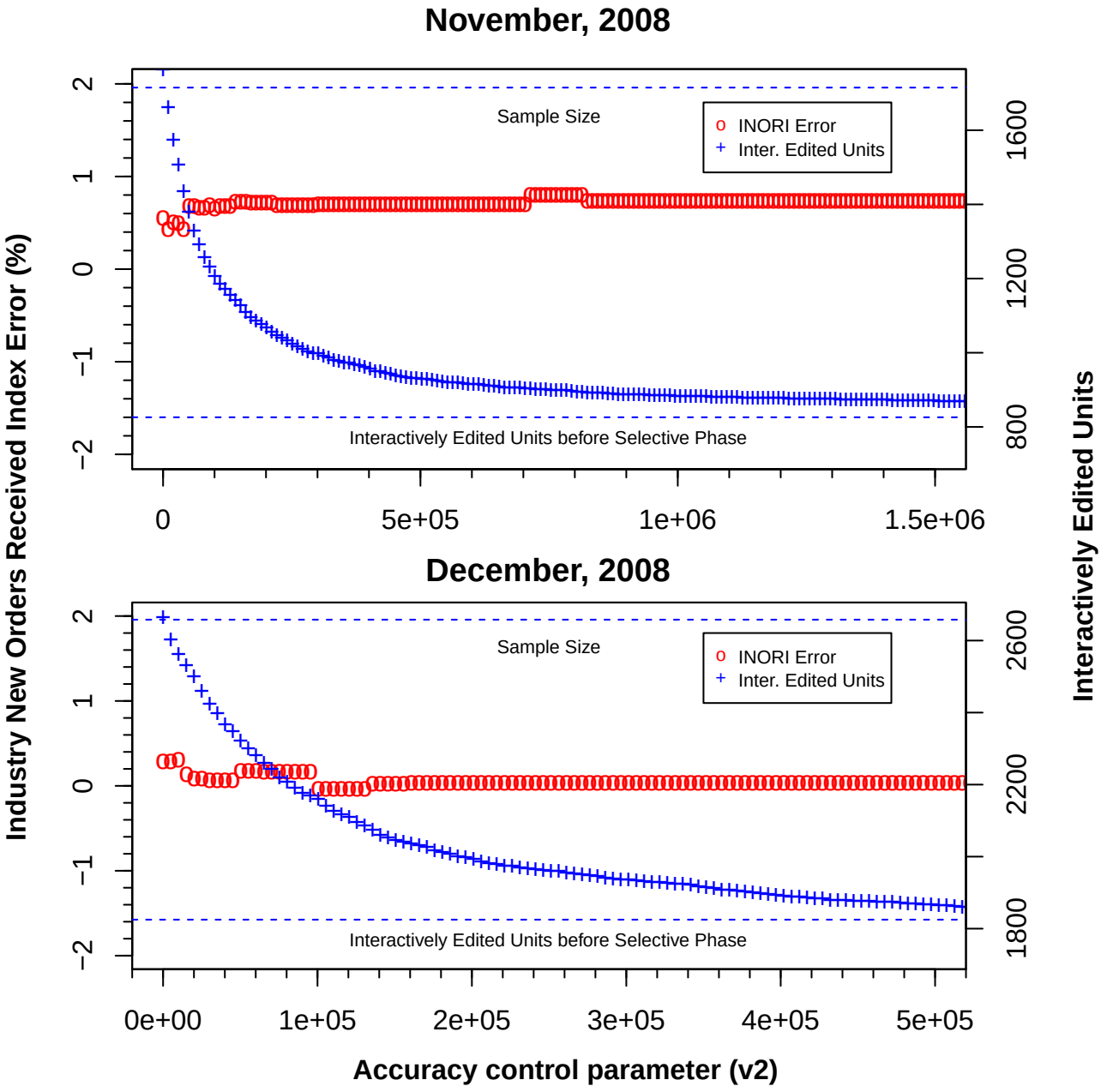

November, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
December, 2008
Sample Size
ITI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry Turnover Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)


March, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
April, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry New Orders Received Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)


May, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
June, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry New Orders Received Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)


July, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
August, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry New Orders Received Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)


September, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
October, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry New Orders Received Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)


November, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
December, 2008
Sample Size
INORI Error
Inter. Edited Units
Interactively Edited Units before Selective Phase
Industry New Orders Received Index Error (%)
Interactively Edited Units
Accuracy control parameter (v2)

Acknowledgments

The author is deeply indebted to I. Arbués for all his comments throughout the execution of this work. I also thank Carlos Pérez and Monserrat Herrador for their invaluable suggestions.

References

I. Arbués, M. González, and P. Revilla (2009). La depuración selectiva como un problema de optimización estocástica. *Boletín de Estadística e Investigación Operativa*, **25**, 32–41.

I. Arbués, M. González, and P. Revilla (2010). A class of stochastic optimization problems with application to selective data editing. *Optimization*, published online on Feb 4, 2010.

D. Bates and M. Maechler (2010). *lme4: Linear mixed-effects models using S4 classes*. `http://CRAN.R-project.org/package=lme4`. R package version 0.999375-36.

J.E. Beasley, ed. (1996). *Advances in linear and integer programming*. Oxford Science Publications, Oxford.

P.P. Biemer and L.E. Lyberg (2003). *Introduction to survey quality*. Wiley, New York.

W.G. Cochran (1977). *Sampling Techniques*, 3rd edition. Wiley, New York.

T. de Waal (2008). An overview of statistical data editing. Discussion paper (08018), Statisics Netherlands.

T. de Waal (2009). *Statistical data editing*, in D. Pfefferman and C.F. Rao, eds. (2009), *Sample Surveys: Design, Methods and Applications*, North Holland, Amsterdam.

W.E. Deming (1950). *Some theory of sampling*. Wiley, New York.

J.-D. Deville and Y. Tillé (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, **128**, 569–591.

EDIMBUS (2007). *Recommended practices for editing and imputation in cross-sectional business surveys*. ISTAT, CBS, SFSO, EUROSTAT, 2007. `http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf`.

Eurostat (2008). Statistical classification of economic activities in the European Community, rev. 2. `http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC`.

L. Granquist (1997). On the current best methods document: edit efficiciently. *UN/ECE Work Session on Statistical Data Editing*, W.P. No. 30.

L. Granquist and J.G. Kovar (1997). *Editing of survey data: how much is enough?*, in L.E. Lyberg *et al.*, eds. (1997), *Survey Measurement and Process Quality*. Wiley, New York.

R.M. Groves (1989). *Survey errors and survey costs*. Wiley, New York.

M.H. Hansen, W.G. Madow, and B.J. Tepping (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, **78**, 776–793.

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.

D. Hedlin (2008). Local and global score functions in selective editing. *UN/ECE Work Session on Statistical Data Editing*, W.P. 31, 1–8.

M.A. Hidiroglou and J.M. Berthelot (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology*, **12**, 73–84.

J. Hoogland (2002). Selective editing by means of plausibility indicators. *UN/ECE Work Session on Statistical Data Editing*, W.P. 33.

J. Hoogland and R. Smit (2008). Selective automatic editing of mixed mode questionnaires for structural business statistics. *UN/ECE Work Session on Statistical Data Editing*, W.P. 2.

C.T. Isaki and W.A. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89–96.

R.N. Kackar and D.A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853–862.

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

D. Lawrence and C. McDavitt (1994). Significance editing in the Australian survey of average weekly earnings. *Journal of Official Statistics*, **10**, 437–447.

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

J.T. Lessler and W.D. Kalsbeek (1992). *Nonsampling error in surveys*. Wiley, New York.

N.G.N. Prasad and J.N.K. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171.

J.N.K. Rao (2003). *Small area estimation*. Wiley, New York.

C.-E. Särndal (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**, 99–119.

C.-E. Särndal, B. Swensson, and J.H. Wretman (1992). *Model assisted survey sampling*. Springer, New York.

C. Scarrott (2007). Feasibility study: a review of selective editing. Technical report, University of Canterbury.

S.R. Searle, G. Casella, and C.E. McCullogh (1992). *Variance components*. Wiley, New York.

Y. Tillé (2006). *Sampling algorithms*. Springer, Berlin.

United Nations Economic Comission for Europe (2010). Generic statistical business process model. version 4.0-april 2009. *Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata*. `http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model`.