Instituto Nacional de Estadística

Working Papers

04/2014

Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey

R. López-Ureña, M. Mancebo, S. Rama, and D. Salgado

The views expressed in this working paper are those of the authors and do not necessarily reflect the views of the Instituto Nacional de Estadística of Spain

First draft: October 2014

This draft: October 2014

Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey

Abstract

We describe in detail the redesign process of the editing and imputation strategy of the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey. This process incorporates the optimization approach to selective editing in its combinatorial version, which we show to contain the score function approach for output editing as a particular case. We also include considerations about editing during data collection and a standardized expression for edits in short-term business statistics. The process embraces from the design of the new edits to their implementation in production. As a global result, the rate of selected units for interactive editing (the most resource-consuming directly impinging on both costefectiveness and response burden) has been reduced 20 percentage points on average without diminishing data quality.

Keywords

Selective editing, optimization approach, editing and imputation strategy design

Authors and Affiliations

S. Rama and D. Salgado

S.G. Metodología y Desarrollo de la Producción Estadística

Instituto Nacional de Estadística

R. López-Ureña and M. Mancebo

S.G. Estadísticas Coyunturales y Precios

Instituto Nacional de Estadística

INE.WP: 04/2014

Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey

R. López-Ureña, M. Mancebo, S. Rama, and D. Salgado

Abstract

We describe in detail the redesign process of the editing and imputation strategy of the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey. This process incorporates the optimization approach to selective editing in its combinatorial version, which we show to contain the score function approach for output editing as a particular case. We also include considerations about editing during data collection and a standardized expression for edits in short-term business statistics. The process embraces from the design of the new edits to their implementation in production. As a global result, the rate of selected units for interactive editing (the most resource-consuming directly impinging on both cost-efectiveness and response burden) has been reduced 20 percentage points on average without diminishing data quality.

1 Introduction

The editing phase of the statistical production process has been well documented as one of the more resource-consuming, time and money included (FCSM, 1990; Granquist, 1997). More recently, this fact can be thought of as part of the ineludible need of streamlining and optimising the whole production process at Statistical Offices (EFTA/Eurostat/UNECE, 2007; MSIS, 2012, 2013).

In this realm, several efforts to improve the editing phase at Statistics Spain have been undertaken in last years. These efforts have entailed some theoretical proposals (Arbués et al., 2012; Salgado et al., 2012), which have recently crystalised in an optimization approach to selective editing (Arbués et al., 2012; Arbués et al., 2013). The ultimate objective in this approach is to optimise the selection of influential units, both at the input and output editing phases.

During the years 2011 and 2012 a thorough study was conducted using both (nearly) raw and edited data from the Spanish Industrial Turnover Index (ITI) and Industrial New Orders Received Index (INORI) survey (INE, 2011). A redesign of its editing strategy was proposed and tested in a simulation exercise using these sets of data. The results were encouraging, since a reduction of recontact and follow-up rates were obtained whereas no reduction of data quality was detected. In October, 2012 the decision of implementing the new strategy was taken and in February, 2013 the first collection of data using this approach was put into practice. Since then, actual (not simulated) recontact and follow-up rates in production with tantamount data quality have been confirming the improvement in this production phase.

This paper depicts this study and its implementation. In section 2, we briefly review the principles of the optimization approach to selective editing for its later adaptation to the editing strategy of the ITI and INORI survey. We show in subsection 3.1 how the strategy was redesigned taking into account the characteristics of the survey, in subsection 3.2 how specific edits for the input editing phase were constructed, in subsection 3.3 how the optimization approach was adapted to this survey, in subsection 3.4 how the simulations were performed and in subsection 3.5 how the new strategy was implemented in production. In section 4 we collect the main conclusions of this pilot experience, which has become the basis at Statistics Spain for a more ambitious programme of redesign of editing strategies of short-term business statistics. As we point out later on, this experience suggests that there still exists ample room for further theoretical work.

2 The optimization approach, briefly revisited

The optimization approach to selective editing begins by explicitly setting out two generic principles for the data editing production phase (Arbués et al., 2013) already present in the literature since some time ago (Latouche and Berthelot, 1992):

i) editing must minimize the amount of resources deployed to recontacts, follow-ups and interactive tasks, in general;

ii) data quality must be ensured.

These principles will be the basis for the formulation of a generic optimization problem whose solution will be the optimal selection of units to be edited interactively.

By and large, we need two ingredients to pose the problem in full detail. On the one hand, we need an important element throughout all the statistical production process: the available information to carry out a task. In particular, we distinguish three dimensions in this concept of available information, namely, longitudinal, cross-sectional and multivariate. 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 signify the information arising from the multidimensional character of the survey. We introduce the concept of available information into the problem by using random variables $\mathbf{Z}$ upon which we will condition the different variables, i.e. we will use conditional expectations on $\mathbf{Z}$ (see Arbués et al. (2013) for details).

On the other hand, we need to specify the classical elements of an optimization problem (see e.g. Nemhauser and Wolsey (1999)), namely, the variables, the objective function and the restrictions.

The variables will constitute the so-called *selection strategy vector* $\mathbf{R} \in \{0,1\}^{\times n}$, where $R_k = 0$ stands for unit k being selected for interactive editing and $R_k = 1$, otherwise (see Arbués et al. (2013) for details about this counterintuitive assignment). Occassionally, a subset of the indices $1 \leq k \leq n$ will be fixed, which we shall denote compactly by $\mathbf{R} \in \Omega \subset \{0,1\}^{\times n}$. The selection strategy vector is random, that is, its probability distribution will be the object of the optimization problem. From a more general standpoint the variables can be thought of as the conditioned random variables $R_k \big| \mathbf{Z}$, $k \in s$.

The objective function is esentially the number of units to edit interactively, so as a first indication we consider $\sum_{k\in s} R_k$ (note that this amounts to the number of units **not** to be edited interactively). But, since the selection strategy vector is a random variable, this objective function must be elaborated a bit further. The optimal expected number of units not to be selected will be given by $\mathbb{E}_m \left[\sum_{k\in s} R_k \big| \mathbf{Z}\right]$, where m stands for any statistical model not embracing the sampling design. This is the function to maximize.

The restrictions will impose an upper bound on the increment of the mean squared error due to the possible presence of measurement errors in the estimators. Results by Arbués et al. (2013) show that we can keep the total mean squared error of an arbitrary linear estimator $\hat{Y}^{(q)}$ under a chosen bound $\eta_q \geq 0$ if we impose

$$\mathbb{E}_m \left[\sum_{k\in s}\sum_{l\in s} R_k R_l \Delta_{kl}^{(q)} \big| \mathbf{Z}\right] \leq \eta_q,$$

where $q \in \{1,\ldots,Q\}$. The loss matrices $\Delta^{(q)} = [\Delta_{kl}^{(q)}]_{1\leq k,l\leq n}$ are conditional moments of the measurement errors $\epsilon_k^{(q)} = y_k^{(q,obs)} - y_k^{(q,0)}$ given by[1] $\Delta_{kl}^{(q)} = \mathbb{E}_m \left[\left|\epsilon_k^{(q)}\right| \big| \mathbf{Z}_k\right]$ if $k = l$ and $\Delta_{kl}^{(q)} = 0$ if $k \neq l$. These moments are computed using a so-called *observation-prediction* model (Arbués et al., 2013) for the observed

[1] We focus on the absolute value loss function. For the squared loss choice see Arbués et al. (2013).

(measured) variables $y^{(q,obs)}$ and true variables $y^{(q,0)}$, succinctly denoted by the subscript m.

The generic optimization problem then reads

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

Now depending on the available information $\mathbf{Z}$ used we arrive at different realizations of problem P_0. In Arbués et al. (2013) we contemplated two extreme cases: either $\mathbf{Z}$ reduces to the longitudinal and multivariate dimensions of the available information or it also includes its whole cross-sectional dimension. Notice that this is equivalent to consider input editing or output editing, respectively. In the first case, the problem P_0 drives us to a stochastic optimization problem, whereas in the second case, it yields a combinatorial optimization problem.

The stochastic version requires to solve an intermediate optimization problem (Arbués et al., 2012) and a standard Matlab optimization routine was used when developing these ideas (Arbués and Revilla, 2014; Arbués et al., 2011). Currently, a specific algorithm for this problem is under development so that it can be more easily inserted in the production chain. Results will be reported elsewhere (Salvador and Salgado, 2014). In the subsequent, no further use of the stochastic version is made.

The combinatorial version reads

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

where $\mathbf{M}^{(q)} = \mathbb{E}_m\left[\Delta^{(q)}\middle|\mathbf{Z}^{cross}\right]$ and the original random selection strategy vector $\mathbf{R}$ reduces to the deterministic vector $\mathbf{r}$. As before, $r_k = 0$ entails that unit k is selected for interactive editing and $r_k = 1$, otherwise. Now the selection is obtained with probability 1.

The solution to problem P_{co} gives the optimal selection of units to be edited interactively. We have proposed a greedy search algorithm (Salgado et al., 2012) whose main characteristics can be read in the appendix A. It is important to keep in mind that this sort of algorithms is heuristic, yielding in this case a suboptimal solution. This suboptimality is equivalent in practice to a little amount of overediting.

From the field work standpoint, it is more adequate to have a prioritization of units instead of a selection (see Arbués et al. (2013) for a discussion). To achieve this we have proposed a prioritization algorithm based on the decremental construction of a sequence of upper bounds $\boldsymbol{\eta}^{(i)}$ so that in each iteration i a new unit is selected for interactive editing, the previously selected ones being fixed.

These two ingredients, the greedy search algorithm and the prioritization algorithm, allow us to prove that we can recover the traditional approach based on score functions (see de Waal et al. (2011) and multiple references therein). The diagonal moments $M_{kk}^{(q)}$ play the role of the local score functions whereas the infeasibility function h used to perform the local search plays a role analogous to that of the global score function. The proof is included in the appendix B. Thus, the combinatorial proposal offers a more solid basis to the traditional heuristic approach, justifying the choice of a global score function as a method of performing the suboptimal selection dealing with the estimators at stake (one per restriction) in different ways. Furthermore, since the optimization approach contains the construction of statistical models (observation-prediction models) in a natural way, a priori it extends the traditional approach by allowing us to consider qualitative variables. This line of work is in progress with the use of logistic regression models in the Spanish National Health Survey and related results will be also reported elsewhere.

Nonetheless, for the ITI and INORI survey we will concentrate upon quantitative variables. For the output editing phase we will use the combinatorial approach. Thus we will construct the diagonal loss matrices $\mathbf{M}^{(q)}$ and choose a particular infeasibility function h to prioritize the units. This procedure will be applied to each (quasi) minimal publication cell. Taking into account field work constraints, we will assign a maximal number of units to be edited for the whole sample. The number of selected units in each cell will be allocated through an algorithm specifically designed for this purpose.

3 The design of the new strategy

3.1 The Spanish ITI and INORI survey

The current Spanish ITI and INORI (INE, 2011) are computed under a joint survey whose aim is to offer an anticipated measurement of the evolution of the industrial production in the country. It complies with different European regulations (EC Council Reg, 1998, 2005). Although the INORI has been recently suppressed from these regulations, Statistics Spain currently continues conducting both surveys.

The ITI has the objective of measuring the evolution of the demand aimed at the industrial branches. The INORI, in turn, has the objective of measuring the evolution of the future demand aimed at these industrial branches. Both the ITI and the INORI are value indicators, in other words, they measure the joint evolution of quantity, quality and price.

From a methodological standpoint, the main pertinent characteristics of the survey are as follows:

(a) It is a fixed panel of aprox. 11000 industrial establishments selected by cut-off (originally coincident with the Spanish Industrial Production Index sample).

(b) There are several data collection modes: CAWI, mail, email, fax and telephone. CAWI-mode received questionnaires are collected with a global parameterized IT tool called IRIA (Bercebal and Maldonado, 2014). This tool has been designed at Statistics Spain to collect data for all surveys, either business or household surveys. It stands up as the first major step towards the streamlining and standardization of the production process at Statistics Spain. The rest of questionnaires are collected, recorded and coded at several provincial delegations.

(c) Estimates correspond to Laspeyres indices disseminated for 37 publication cells identified as certain divisions and subdivisions[2] of the NACE[3] Rev. 2, without geographical breakdown for the total turnover and total new orders received. This dissemination plan has been recently broken down into geographical regions[4]. This change was not present during the redesign process and it will not be taken into account.

The general approach to design the new E&I strategy is that depicted in the EDIMBUS manual (EDIMBUS, 2007) (see also de Waal et al. (2011) for a wider overview), which offers clear guidelines aiming to a harmonization of this phase of the statistical production process within the European Statistical System. This general approach, which embraces post-capture data editing, is complemented with an extra initial stage to edit questionnaires during collection when possible (mainly with computer-assisted collection modes).

The new strategy aims at selecting units for interactive editing more efficiently. It preserves much of its past structure comprising the following stages:

1. Editing during collection. This stage was already present. However, now completely new edits are proposed for the CA modes to flag suspicious

[2] A functional domain-specific splitting of NACE divisions into smaller units, but larger than groups.

[3] Actually of the CNAE2009, the Spanish almost equivalent classification.

[4] The Spanish CCAA, which are equivalent to NUTS2 geographical regions in the European Statistical System.

questionnaires during the CAWI, thus prompting the respondent for a first revision.

2. Longitudinal phase. This stage was also present in the old strategy and conducted at the provincial delegations. Now, a similar set of edits is posed both for CAWI collected units whose data are already in the system and for data collected with different modes to select those questionnaires entering into interactive editing, which is still conducted at the provincial delegations.

3. Cross-sectional phase. This stage is completely new. All questionnaires already subjected to the preceding phases are sent to the central office where a final selection of units is performed for a further revision at the provincial delegations.

4. Macro editing and validation phase. This stage comprises the final macro editing and validation works carried out by the subject matter experts of the unit responsible for the survey. Almost nothing new is added in this stage.

The core of the new strategy lies on the design of the new edits used to flag units for interactive editing. Clearly the objective is to gain efficiency in the editing phase by minimizing the interactive editing work but maintaining the same data quality. This course of action also prevents us from a complete redesign of the strategy which may impinge on more intricate aspects such as a reassignment of positions or tasks within the organization, training personnel on new processes, etc. The goal is to edit interactively more efficiently to optimize resources while keeping the same organizational structure.

In this sense, more efficient and less edits have been proposed and tested. First we present the edits for the first and second stages. Then we show different adaptations of the combinatorial optimization approach for the third stage. All these stages are used together in the simulation exercise, from which decisions are taken for the implementation of the strategy in production. In our view, all these elements can hold interest in their own for the editing strategy redesign or fine-tuning considerations of other surveys.

3.2 The interval-distance edits

The sets of edits both for the first and second stages are indeed similar, except for their parametrization. Thus we will treat them as a unique set.

It comprises two kind of edits. First, we have the traditional format, range and balance edits, all of them hard edits arising from the nature and relation itself of the variables. Only those strictly necessary are posed for having logical coherence in the values. Second, statistical "correction" is pursued. To this end, let us focus on a variable y of interest for their later dissemination. The idea is to critically construct a validation interval I_k and to determine a threshold value t_k for each unit $k \in s$ in the sample. Then if the distance $d_k = d(y_k, I_k)$ between the value y_k in the questionnaire and the interval I_k measured through the distance function d is greater than the threshold value $d_k > t_k$, the questionnaire is flagged. We call it an interval-distance edit.

There is not much new so forth, thus some comments are in due order. Admittedly this is indeed a local score function (see Latouche and Berthelot (1992) and multiple references in de Waal et al. (2011)), but as we have been progressing in the implementation of this idea we have observed its convenience both from theoretical and practical standpoints. Firstly, we bring it closer to the well-known if-then form of edits by de Waal (2005); de Waal et al. (2011). An edit e will be specified by (i) a condition $C(\mathbf{y}, \mathbf{x})$ for the objective $\mathbf{y}$ and possibly auxiliary $\mathbf{x}$ variables, (ii) a function $f(\mathbf{y}, \mathbf{x})$, (iii) an interval $I_k = [l_k, u_k]$, (iv) a threshold t_k, and (v) a distance function d. If the unit k satisfies the condition $C(\mathbf{y}_k, \mathbf{x}_k)$, the edit is applied so that if $d\left(f(\mathbf{y}_k, \mathbf{x}_k), I_k\right) > t_k$, the unit is flagged.

We have observed several advantages of this form of the interval-distance edit. As we will show below, this form gives room for time series modelling techniques in a straightforward way. The use of these techniques is very appropriate to build edits as efficient as possible. Next, we can give a further simplified expression by synthesizing the condition $d\left(f(\mathbf{y}_k, \mathbf{x}_k), I_k\right) > t_k$ into a larger interval I'_k so that the new condition reads $f(\mathbf{y}_k, \mathbf{x}_k) \notin I'_k$. Therefore, each unit k is assigned an interval I'_k and the system only has to check (i) if the condition $C(\mathbf{y}_k, \mathbf{x}_k)$ is satisfied and (ii) in the positive case, if the value $f(\mathbf{y}_k, \mathbf{x}_k)$ lies inside the interval I'_k or not. Finally, the versatility of the if-then form allows us to express other edits. For example, consider a balance edit $y_1 + y_2 = T$ to be fulfilled by all units. Then the condition $C(\mathbf{y}, \mathbf{x})$ is identically true by construction, the function $f(\mathbf{y}, \mathbf{x})$ is $f(\mathbf{y}, \mathbf{x}) = y_1 + y_2$, and the interval I' is degenerate $I'_k = [T, T]$. The strategy is thus expressed as a set of standardized interval-distance edits.

To build the edits, we have exploited the fact that this survey is a fixed panel, so that each respondent has a multivariate time series since the very time period it entered into the panel. We have focused on turnover and new orders received, whose indices are disseminated monthly[5]. Then we automatically adjust an ARIMA model for each unit k and each variable. This is performed using the program TRAMO-SEATS (Caporello and Maravall, 2004). For the next coming month we predict under these models the corresponding values $\hat{y}_k$ for each unit. The actual working conditions force us to predict two months ahead in time, since preceding month values are not completely edited before starting the data collection process for the following month. These values will be the centre $c_k = \hat{y}_k$ of each interval $I_k = [c_k - r_k, c_k + r_k]$ for the total turnover and total new orders received.

[5] However see section 4.

The radius r_k is computed striving for efficiency. We have chosen $r_k = \eta \times \hat{\sigma}_k$, where η is a parameter related to the efficiency of the intervals (see immediately below) and $\hat{\sigma}_k$ is the estimated standard deviation under the same ARIMA model. To give concrete values to η we have proceeded as follows. We have defined the indicator $\widetilde{\mathrm{HR}}$ as the ratio between the number of units which are correctly flagged regarding this variable (the value was changed) and the total number of flagged units. Notice that it is an approximation to the hit rate (EDIMBUS, 2007) (hence the notation). This pseudo-HR is computed using the last completely edited month whose both raw and edited values are available, in our case, $t-2$. In practice this is performed using the set of (nearly) raw values collected and stored by IRIA and the set of completely edited values finally validated for $t-2$. For this double data set, $\hat{y}$ and $\hat{\sigma}$ are computed disregarding the completely edited version of period $t-2$. Then for different values of η we shall have different lengths of the intervals and thus different values of $\widetilde{\mathrm{HR}}_{t-2}$. That is, we can write $\widetilde{\mathrm{HR}}_{t-2} = \widetilde{\mathrm{HR}}_{t-2}(\eta)$. Let us denote $\eta_t^* = \mathrm{argmax}_\eta \widetilde{\mathrm{HR}}_{t-2}(\eta)$, where the maximum is computed numerically for different values of $\eta \in [0,5]$. Let η_t denote the value η for the radius of the intervals at time period t. We have set $\eta_t = (1-\lambda)\cdot\eta_{t-2} + \lambda\cdot\eta_t^*$, where $\lambda \in [0,1]$. Notice that under this choice the radii become shorter if the pseudo-hit rate decreases and vice versa, i.e. they are adjusted according to the efficiency of these edits.

For those units whose time series do not allow TRAMO-SEATS to adjust an ARIMA model automatically (for being too short or holding too many 0's or missing values, ...), we have proceeded as follows. We compute the ratios $r_k^{(t)} = \frac{y_k^{(t)}}{y_k(t-2)}$ for the last available time period t. In our case, if the reference time period is t, then we compute $r_k^{(t-2)}$. Then we calculate the quantiles $q_{li} = q_l\left(\left\{r_k^{(t-2)}\right\}_{k\in s_i}\right)$ and $q_{ui} = q_u\left(\left\{r_k^{(t-2)}\right\}_{k\in s_i}\right)$ of degrees $p_l < p_u \in [0,1]$ over the cells s_i. The interval for time period t is set as $I_k = \left[y_k^{(t-2)} \times q_{li}, y_k^{(t-2)} \times q_{ui}\right]$ if $k \in s_i$. The degrees p_l and p_u are chosen so as to obtain ample intervals (e.g. $p_l \approx 0.05$ and $p_u \approx 0.95$). In practice, indeed we have conservatively set as final validation intervals $I_k^{\mathrm{ARIMA}} \bigcap I_k^{\mathrm{RATIO}}$.

These intervals so computed together with the degenerate distance function $d_1(y, I) = 0$ if $y \in I$ and ∞ if $y \notin I$, are used as interval-distance edits in the CAWI mode. Notice that it is unnecessary to determine the thresholds with this distance function.

However we must compute the thresholds for the longitudinal stage, where we have chosen the function[6]

[6] For completeness, d_2 stands for the usual geometrical distance.

$$d_3(y, I) = \begin{cases} \frac{y-u}{u-l} & \text{if } y \geq u, \\ \frac{l-y}{u-l} & \text{if } y \leq l, \\ 0 & \text{otherwise,} \end{cases}.$$

The function d_3 takes implicitly into account the historical variability of the values in the time series. To this end, first we partition the sample into domains $s = \bigcup_{i \in \mathcal{I}} s_i$. Then for the last available time period, in our case $t-2$, we compute the distance d_3 between the values $y_k^{(ed,t-2)}$ and their corresponding intervals $I_k^{(t-2)}$: $d_{3k}^{(t-2)} = d_3(y_k^{(ed,t-2)}, I_k^{(t-2)})$. We have set $t_k = q\left(\left\{d_{3k}^{(t-2)}\right\}_{k \in s_i}\right)$ for all $k \in s_i$, where q denotes the quantile of a chosen order $p \in [0,1]$. The use of quantiles allows us to have a more direct control on the interval-distance edits and thus on the potential proportion of units to be flagged. The domains must be chosen small enough to have thresholds as individuated as possible but large enough as to make the quantile computation sensible. In our case, no deeper analysis has been carried out in this regard and NACE (sub)divisions have been chosen as domains, that is, the minimal dissemination cells.

For the ITI and INORI survey, the interval computation is undertaken for each unit, each month, each variable (turnover and new orders received) and maximizing the parameter η within each NACE division. The threshold determination is carried out monthly with $p = 0.95$ and for each NACE (sub)division. The intervals together with distance function d_1 are used as interval-distance edits for all units collected through CAWI. These are the interval-distance edits in the editing during collection stage. For the longitudinal stage, the same intervals with distance function d_3 and thresholds t_k conform the corresponding interval-distance edits. A unit is flagged if any of the edits is activated. This tight choice is made in the hope of having built notably efficient edits.

Finally, for the first and second stages of the strategy these interval-distance edits have been complemented with an elementary control for very specific inliers: those values with $y_k^{(obs,t)} = y_k^{(obs,t-a)}$, $a = 1, 12$, will be flagged. Additionally, those respondents whose values have changed between the raw and edited versions in all the last three time periods are also selected for interactive editing.

3.3 The cross-sectional selection

The cross-sectional phase comprises the selection of units after the completion of the first two stages, i.e. of the editing during collection and longitudinal phases. This new phase incorporates the innovative proposal of the combinatorial optimization approach to selective editing. Since the very beginning of the work reported herein this analysis has been incorporating advances proposed in the optimization approach. These advances have been added in different moments of the analysis and there is still more to come with the current investigation of a specific optimization algorithm for the stochastic optimization approach. Here we include those main findings finally taking part in the implementation in production.

In the first proposal containing the stochastic version (Arbués et al., 2012), a quadratic loss function was used driving us to nondiagonal loss matrices $\mathbf{\Delta}^{(q)}$. In the combinatorial version, as a first option this loss function was also considered, but it entailed some algorithmic complexities regarding the resolution of the optimization problem. We changed to the absolute-value loss function, which produces diagonal loss matrices $\mathbf{\Delta}^{(q)}$, thus simplifying and speeding up the resolution of the problem.

This choice of loss function also allowed us to find an analytical expression for the diagonal entries of the loss matrices. Beforehand, we need assumptions for the underlying so-called observation-prediction model (see Arbués et al. (2013) for details). This is a statistical model m for both the observed (reported) and true values of each variable. Since we are dealing with continuous values (turnover and new orders received), we make use of the continuous variable model proposed by Arbués et al. (2013) by which $y_k^{(q,obs)} = y_k^{(q,0)} + \epsilon_k^{(q,obs)}$ and $y_k^{(q,0)} = \hat{y}_k^{(q)} + \epsilon_k^{(q,pred)}$, for $q = 1, 2$, where $\epsilon^{(q,\cdot)}$ stands for observation or prediction errors and $\hat{y}_k^{(q)}$ denotes prediction values under an auxiliary independent model m^*. This is completed with the specifications about the errors $\epsilon_k^{(q,obs)}$ and $\epsilon_k^{(q,pred)}$:

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

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

These assumptions allow us to express the diagonal entries of the loss matrices as

$$\Delta_{kk}^{(q)} = \sqrt{\frac{2}{\pi}}\cdot\omega_k\cdot\nu_k^{(q)2}\cdot{}_1F_1\left(-\frac{1}{2};\frac{1}{2};-\frac{1}{2}\left(\frac{y_k^{(q,obs)}-\hat{y}_k^{(q)}}{\nu_k^{(q)}}\right)^2\right)\cdot\zeta_k^{(q)}\left(\frac{y_k^{(q,obs)}-\hat{y}_k^{(q)}}{\nu_k^{(q)}}\right), \tag{3}$$

where $\zeta_k^{(q)}(x) = \frac{1}{1+\frac{1-p_k^{(q)}}{p_k^{(q)}}\left(\frac{\nu_k^{(q)2}}{\nu_k^{(q)2}+\sigma_k^{(q)2}}\right)^{-1/2}\exp\left(-\frac{1}{2}\frac{\sigma_k^{(q)2}}{\sigma_k^{(q)2}+\nu_k^{(q)2}}\cdot x^2\right)}$. Notice that using the asymptotic properties of the confluent hypergeometric function ${}_1F_1$ (Abramowitz and Stegun, 1972), when $\left|\frac{y_k^{(q,obs)}-y_k^{(q,pred)}}{\nu_k^{(q)}}\right| \gg 1$, we have $\Delta_{kk}^{(q)} \approx \omega_k\left|y_k^{(q,obs)}-\hat{y}_k^{(q)}\right|$. Thus, under very precise predictions, we recover the usual recipe (de Waal et al., 2011) for the local score functions as a limit case. In other words, the usual recipe is valid when the anticipated value is a very good prediction of the true value.

The parameters $p_k^{(q)}$, $\sigma_k^{(q)2}$, $\nu_k^{(q)2}$ are estimated with maximum-likelihood estimators over the double sets of raw and edited values using some simplifying assumptions (Arbués et al., 2013). These estimated values are denoted by $\hat{p}_k^{(q)}$, $\hat{\sigma}_k^{(q)2}$, $\hat{\nu}_k^{(q)2}$, respectively.

The auxiliary model m^* used to compute the predicted values is the best time series model ξ^* among $\{\xi_1, \xi_2, \xi_3\}$ where (s stands for the seasonal period, $s = 12$ in our monthly series):

$$\begin{aligned}
\xi_1: &\quad (1-B)y_t &= a_t,\\
\xi_2: &\quad (1-B^s)y_t &= a_t,\\
\xi_3: &\quad (1-B^s)(1-B)y_t &= a_t.
\end{aligned}$$

The best model ξ^* is selected as that minimizing the mean squared error of the white noise component a_t under each model ξ_j. Given the actual working conditions, the dissemination calendar only allows us to predict two time periods ahead at best. These predicted values are equally denoted by $\hat{y}_k^{(q)}$.

The moments are thus given by $\Delta_{kk}^{(q)} = \Delta_{kk}^{(q)}\left(y_k^{(q,obs)}, \hat{y}_k^{(q)}; \hat{p}_k^{(q)}, \hat{\sigma}_k^{(q)2}, \hat{\nu}_k^{(q)2}\right)$.

Once the loss matrices are computed, we must prioritize the units within each cell (NACE (sub)division). As explained in section 2, to choose a particular prioritization algorithm we must choose an infactibility function h (see appendix B). Equivalently we can choose a global score function $S_k = S\left(\Delta_{kk}^{(1)}, \Delta_{kk}^{(2)}\right)$. Apart from the traditional Minkowskian functions $S = S^{(\alpha)}$ (Hedlin, 2008), which we have tested for $\alpha = 1$ and $\alpha = \infty$, we have also tested

$$\tilde{S}_k^{(\alpha)} = \left(S^{(\alpha)} \circ \left(F^*_{\mathrm{diag}\left(\Delta^{(1)}\right)}, F^*_{\mathrm{diag}\left(\Delta^{(2)}\right)}\right)\right)\left(\Delta^{(1)}_{kk}, \Delta^{(2)}_{kk}\right),$$

where $F^*_{\mathbf{z}}$ stands for the empirical distribution function of the set of values $\mathbf{z}$. This choice has resulted to be more efficient in selecting influential units (see below).

Finally, the chosen number n_{cross} of units to be selected in the cross-sectional stage must be allocated among the different cells. We have applied a simple algorithm in steps. Let n_i denote the number of units to be allocated in cell i:

1. Set $n_i^{(0)} = n_{i0}$ be the initial allocation chosen for some subject matter questions, where $n_{im} \leq n_{i0} \leq n_{iM}$, n_{im} and n_{iM} being lower and upper bounds (respectively) for the allocation of cell i.

2. Set $n_i^{(1)} = \min\left(n_{iM}, \lfloor \Lambda_i \cdot \left(n_{cross} - \sum_i n_i^{(0)}\right) \rfloor\right)$, where $\Lambda_i = \sum_f \lambda_f E_{if}$ is a proportionality constant embracing $f = 1, \ldots, F$ synthetic error measures and/or relevance factors $E_{if} \geq 0$ for cell i with reliability weights $\lambda_f \geq 0$. Both the error measures and the reliability weights are normalized, i.e. $\sum_{f=1}^{F} \lambda_f = 1$ and $\sum_i E_{if} = 1$, for all f.

   The quantities E_{if} for each cell are chosen to be:

   (a) The maximum moment of the measurement errors of both variables ITI and INORI in each cell i.

   (b) The weight of each cell i in the national index.

   (c) The fraction of questionnaires in each cell i with reported total turnover equal to 0.

   (d) The proportion of questionnaires in cell i having reported a null value for the total turnover in the preceding month but whose final value was imputed to a non-null value.

   The values of the reliability weights λ_f were chosen empirically using the three first months of the series as a training set. We have set $\boldsymbol{\lambda} = (0.80, 0.02, 0.09, 0.09)$.

3. If $n_{cross} - \sum_i \left(n_i^{(0)} + n_i^{(1)}\right) > 0$, then allocate one unit in turn in descending order of the values Λ_i provided the allocation does not exceed the corresponding maximum value n_{iM} until no unit is left. This produces the allocation vector with components $n_i^{(2)}$.

The final allocation is then $n_i = n_i^{(0)} + n_i^{(1)} + n_i^{(2)}$. In each cell i the first n_i units according to the prioritization obtained before are to be edited interactively. The number of units n_{cross} is chosen according to the results of the simulation and to the timing imposed by the dissemination calendar (see below).

3.4 Simulations

Simulations are to be understood in a somewhat different way to the usual sense. Data are not simulated at all by a random generation of their values according to a prescribed probability distribution. We have used both the raw and edited versions of the ITI and INORI microdata corresponding to 13 consecutive months (from December, 2010 to December, 2011). Instead we simulate the implementation of the strategy by applying the E&I strategy to the raw data thus producing a selection of units to edit interactively, whose values are then substituted by their corresponding edited version. As any other simulation exercise, assumptions must be made which we comment in detail.

The simulation tries to emulate the actual production conditions, so that the editing of every month is not undertaken until the editing of the preceding month is finished. Edited microdata are accessed by the unit responsible for the survey at two moments t_1 and t_2. At t_1 only around 60% of the sample has been collected, which then undertakes the last macro editing phase in the old strategy. The rest is finally collected at t_2 so that the corresponding macro editing is undertaken and the final validation of the whole sample is carried out.

Firstly we make use for the simulation study only of CAWI-mode collected data, which represents approximately 70% of the sample. Moreover, during the time periods used in the simulation IRIA was not operative and the former data collection system only incorporated soft edits with very loose constraints. This entails that so collected data were indeed nearly completely raw data.

We reproduce the above proposed E&I strategy by firstly imposing the edits of the first stage (editing during collection) to the raw data. Then the respondent behaviour during the CAWI is simulated. We make our first assumption: the selected respondents will react positively with probability $1/2$ thus changing their original values for their edited counterpart in their whole questionnaire. It is admittedly a simplifying assumption and we will make a critical comment in section 4 judging by the results. The probability value $1/2$ was chosen as a maximum ignorance compromise, since the data collection system did not provide us with data to produce an estimate. This simulated stage is carried out in two steps: first for those questionnaires accessed at t_1 and then for those accessed at t_2.

We have a new data set composed of both raw and edited values. We apply the longitudinal phase edits to this data set, producing a new selection of units to edit interactively. This task must be carried out by data editing personnel in the provincial delegations. We simulate this just by substituting the values in the data set for their corresponding edited version in the selected questionnaires. Note that this hypothesis is rather realistic since the difference between raw and edited values comes up indeed as a result of this same task in real production conditions. Again this stage is simulated in two steps: for questionnaires accessed both at t_1 and t_2.

Now we have a new data set with a higher degree of editing. We apply the cross-sectional phase thus producing a new selection of units. Again the edited values are substituted in the selected units. Notice that already edited questionnaires may be selected, thus producing no change in their values. The editing task would amount to confirming the values in those flagged questionnaires. Two cross-sectional selections are carried out: first among the subsample accessed at t_1 and later on at t_2 among the whole sample.

We have another data set, which will be the final data set. The macro editing and validation phase is reduced in the simulation to confirm the present values. Since this avoided phase comprises a non-negligible amount of subject matter judging, this decision makes it possible to carry out the monthly simulations automatically.

With these final data sets we compute the ITI and INORI for each publication cell. In particular, we focus on the minimal cells. Thus we have the set of indices ITI_i^{sel} and INORI_i^{sel} for the $i = 1, \ldots, I$ NACE (sub)divisions. We also compute their corresponding counterparts ITI_i^{ed} and INORI_i^{ed} using the original edited version of the data sets. We examine the absolute relative difference $\Delta_i^{\mathrm{ITI}} = \left|\frac{\mathrm{ITI}_i^{sel} - \mathrm{ITI}_i^{ed}}{\mathrm{ITI}_i^{ed}}\right|$ and $\Delta_i^{\mathrm{INORI}} = \left|\frac{\mathrm{INORI}_i^{sel} - \mathrm{INORI}_i^{ed}}{\mathrm{INORI}_i^{ed}}\right|$ for each cell i. As an illustrative example in figure 1 we show the progressive difference reduction of the national INORI (vertical axes) of each NACE division (horizontal axes) for $n_1 = 0, 50, 100, 150, 200$ (from top to bottom) units selected in the cross-sectional selected at t_1 and for $n_2 = 100, 150, 200$ (from left to right) units selected at t_2.

Besides, for each pair n_1, n_2 we have also computed the yearly rates for both the ITI and the INORI together with their absolute error ${}_{\mathrm{ITI}}\bar{\Delta}_i^t = {}_{\mathrm{ITI}}R_i^{t,sel} - {}_{\mathrm{ITI}}R_i^{t,ed}$ and ${}_{\mathrm{INORI}}\bar{\Delta}_i^t = {}_{\mathrm{INORI}}R_i^{t,sel} - {}_{\mathrm{INORI}}R_i^{t,ed}$, where the yearly rates R_i^t are defined as ${}_{\mathrm{ITI}}R_i^{t,\cdot} = \mathrm{ITI}_i^{t,\cdot} - \mathrm{ITI}_i^{t-12,\cdot}$ and ${}_{\mathrm{INORI}}R_i^{t,\cdot} = \mathrm{INORI}_i^{t,\cdot} - \mathrm{INORI}_i^{t-12,\cdot}$. As another illustrative example in figure 2 we include these quantities for $n_1 = n_2 = 100$ on May, 2011.

These are the figures of merit used by the subject matter experts responsible for the survey to assess the performance of the new strategy in comparison with the old one. This choice is motivated by the fact that both the indices and their yearly rates are the figures published in the dissemination plan.

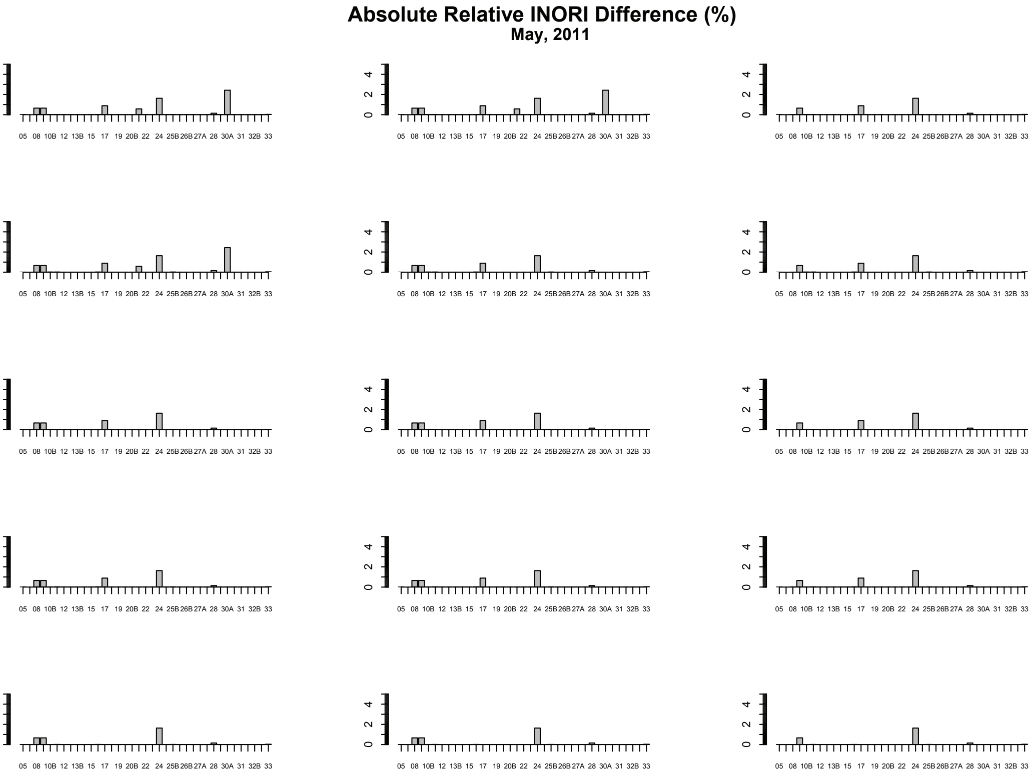

Figure 1: Absolute relative error of the INORI on May, 2011.

Finally the number of selected units both in the longitudinal and cross-sectional phases are computed for each month t. These are the questionnaires to edit interactively, thus incurring in resource consumption for the Institute. It is computed as a fraction of the sample size and compared to the same quantity under the old strategy. See figure 3.

Based on these figures and the time restrictions given by the dissemination calendar on the editing field tasks, the decision was taken to select $n_1 = 100$ and $n_2 = 100$ in the cross-sectional phase. In section 4 we include more detailed conclusions about the results of the simulations.

3.5 Implementation in production

The implementation in production is a critical step, since the simplifying assumptions present in the simulation are not strictly valid any more and actual production conditions enter into play. Currently Statistics Spain is undertak-

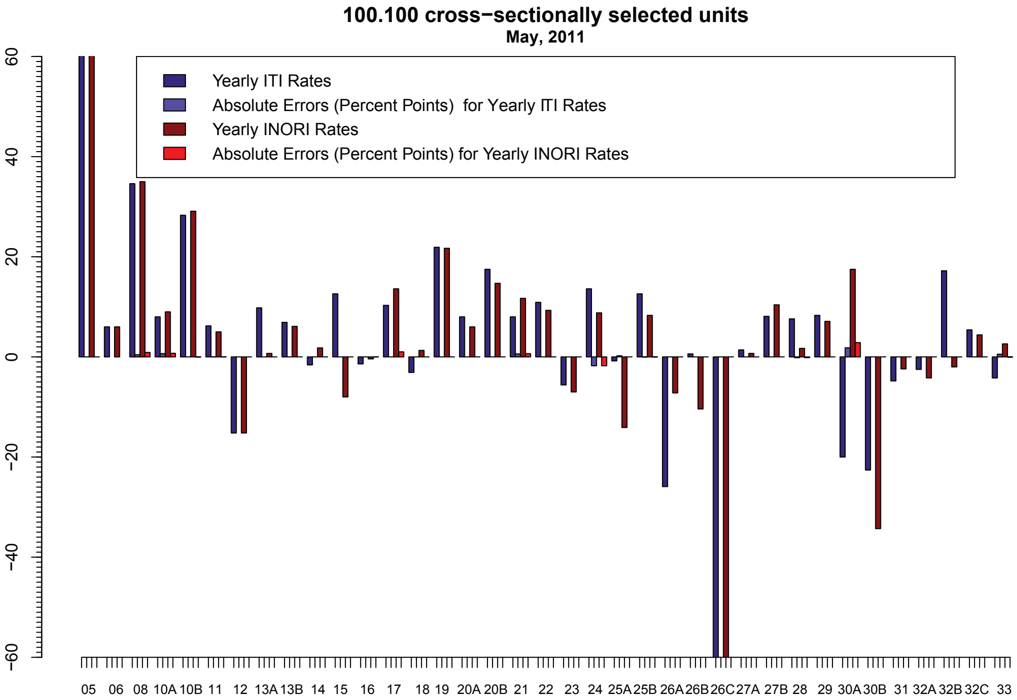

Figure 2: Yearly rates and their absolute error for $n_1 = n_2 = 100$ on May, 2011.

ing a transition from a stovepipe model to an industrialized production process (Bercebal and Maldonado, 2014). As of the date of the decision of implementation of this new strategy, the editing phase of this survey was still conducted under that traditional stovepipe model. Thus, the implementation was undergone in these conditions.

The new control edits were implemented in IRIA for the CAWI collection mode and in the provincial delegations for the longitudinal phase. The parameters for these edits were chosen according to the following idea. First we impose somewhat severe soft edits during the CAWI flagging around 40% of these questionnaires. This severity was substantiated in the choice of the distance function d_1 (see subsection 3.2). The soft character entails that an appropriate confirmation message will appear on screen for those suspicious data. Thus, the response burden rests on the usability of the self-administered electronic questionnaire, which becomes an important aspect not only in data collection but also in data editing (Nichols et al., 2005). In this sense we take

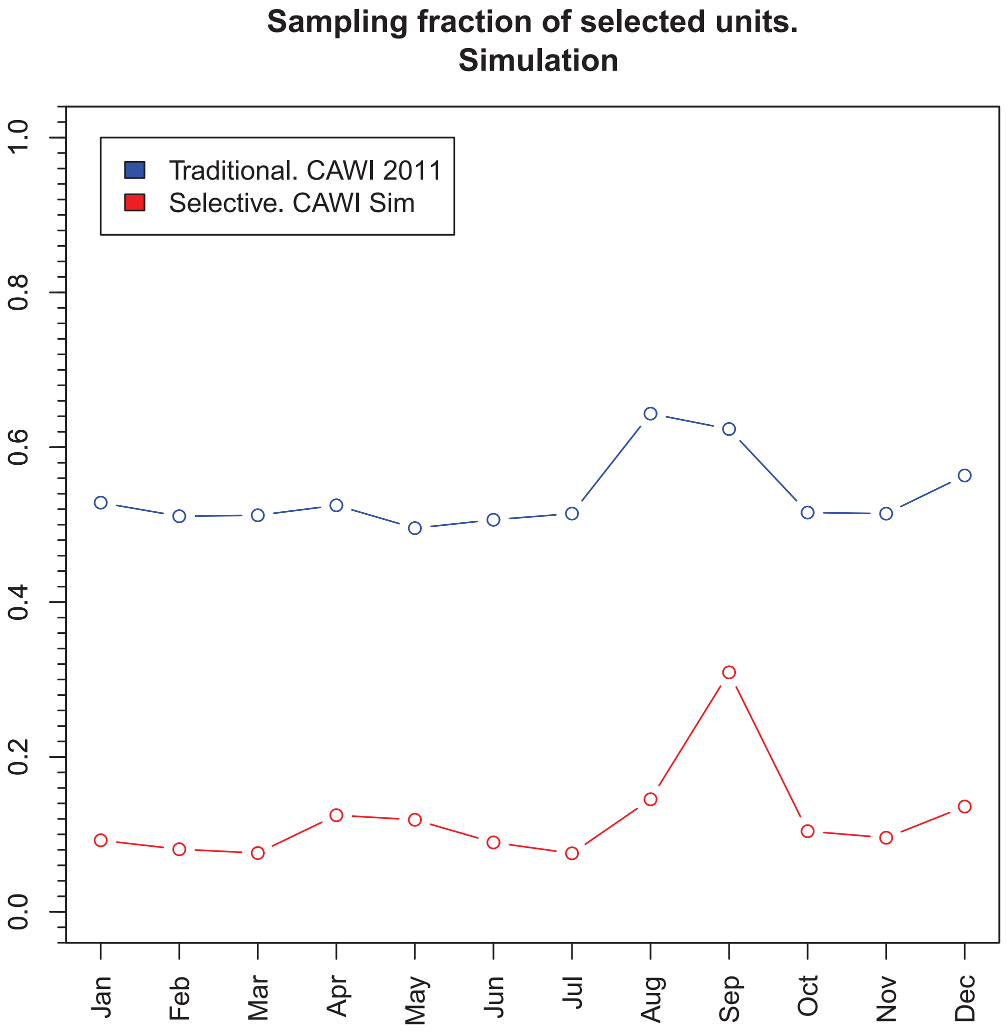

Figure 3: Sampling fraction of selected units under both the old and the new strategies.

advantage of the versatility of the data collection tool IRIA at Statistics Spain. Once data are in the system, the edits in the longitudinal phase are applied with looser parameters, so that typically only highly suspicious questionnaires will be flagged. In this case, the distance function d_3 is used.

Note that a decision must be taken for those questionnaires not received through the CAWI mode since they will not be subjected to the previous double data checking. Furthermore, the simulation exercise can only partially assist in this decision. The distance function d_1 has been chosen, so that foreseeably some overediting is expected.

A further complication arose since the dissemination plan changed during the implementation work and a breakdown of both the turnover and new orders received was to be published. This breakdown entails that the number of variables to be under control increases from 2 (total turnover and total new orders received) to 10 (the same total variables and their breakdown into 4 markets: national, euro zone, non-euro zone, rest of the world).

Priority was given to the implementation in production over a new simulation study. Thus the following two simple edits were included in the strategy. Let $\mathbf{y}_k^t = (y_{1k}^t, y_{2k}^t, y_{3k}^t, y_{4k}^t)$ denote the breakdown of variable $y = \mathrm{ITI}, \mathrm{INORI}$ in the four markets for unit k at time period t. Denote the (squared cosine of the) angle between $\mathbf{a}$ and $\mathbf{b}$ by $\alpha\left(\mathbf{a}, \mathbf{b}\right) = \frac{(\mathbf{a}\cdot\mathbf{b})^2}{\|a\|^2\|b\|^2}$. We define the first edit for the turnover breakdown as: unit k will be flagged if $\alpha_k^t = \alpha\left(\mathbf{ITI}_k^t, \mathbf{ITI}_k^{t-2}\right) < t_k$, where t_k is a threshold value (see below). The second edit is defined as: unit k will be flagged if $\bar{\alpha}_k^t = \alpha\left(\mathbf{ITI}_k^t, \mathbf{INORI}_k^t\right) < \bar{t}_k$, where $\bar{t}_k$ is a different threshold (see below). Notice that the first edit controls the coherence between the current breakdown of the turnover and that of the preceding available time period. In turn, the second edit controls the coherence between the breakdown of the new orders received in relation to that of the turnover for the same time period.

The thresholds are set computing the quantities α and $\bar{\alpha}$ for the last available period and calculating their quantile over the publication cells i so that $t_k = q_i\left(\{\alpha_k^{t-2}\}_{k\in s_i}\right)$ and $\bar{t}_k = q_i\left(\{\bar{\alpha}_k^{t-2}\}_{k\in s_i}\right)$.

The angle function α has been also included in the macro editing phase to detect erroneous exchange of variables in the components of $\mathbf{y}$. Occasionally some respondents mistake the component y_3 for the component y_4 (also y_2 for y_3). If we focus on the subvector $\mathbf{y}_{34} = (y_3, y_4)$, then for those vectors with one of the components equal to 0 the following edit detect this wrong exchange: flag unit k if $\alpha(\mathbf{y}_{34,k}^t, \mathbf{y}_{34,k}^{t-2}) = 0$. This is applied to both sets of variables.

The strategy was finally implemented and entered into production on January, 2013 (month of reference). In figure 4 we represent the actual results regarding the fraction of selected questionnaires for interactive editing. Data quality was not diminished. These results triggered the generalization of this proposal to other short-term business statistics. This generalization takes into account the conclusions compiled in the next section.

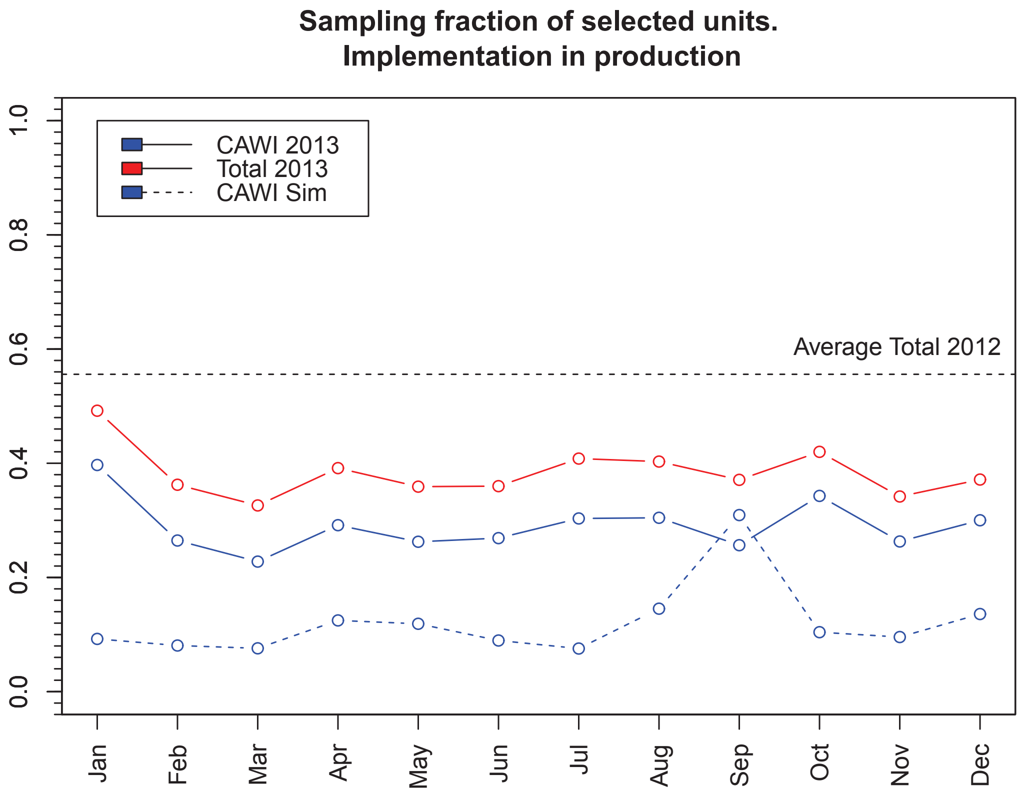

Figure 4: Fraction of selected units in production. Year 2013.

4 Conclusions

The study depicted in the preceding sections has triggered the redesign of the E&I strategy of most of the short-term business statistics at Statistics Spain after extracting some suggestions about what to do and not to do in this production phase. We include here the most important.

As a first conclusion we underline the remarkable importance of having a common business architecture for this production phase among the different surveys in which the new strategy will be used. In particular, the new E&I strategies must be given a generic form compatible with the top-down view of the statistical production process provided by standards like the GSBPM (UNECE, 2013a) and the GSIM (UNECE, 2013b). In our opinion, the concepts of statistical production function and standard production step (Renssen and Camstra, 2011) go indeed beyond this top-down standards and stand as very adequate tools in this respect since they introduce the reusability of resources in a very natural way. In the programme of redesign of this production phase at Statistics Spain all the preceding ideas are being reformulated in terms of production functions and activities.

In our opinion, only after the generic E&I strategy has been defined in terms of reusable production functions, we should substantiate these functions with the different mathematical proposals. In particular, notice that the observation-prediction model m and the auxiliary prediction model m^* are indeed very simple (hence also robust for their exploitation in production), thus leaving room for more sophisticated proposals. For example, instead of the three time series models ξ_1, ξ_2 and ξ_3 we can use ARIMA models or even multivariate time series models to compute the values $\hat{y}_k$ and $\hat{\nu}_k$. All these possible choices are indeed different exchangeable statistical functions for the same tasks.

The new strategy incorporates the editing during collection phase since its very design, thus underlining the integration of statistical functions in the production process already pinpointed by Pierzchala (1990). This stresses the importance in data editing itself of aspects as the usability of self-administered electronic questionnaires, originally thought of as part only of the data collection phase.

In this same line of integrating production functions, in our view the interval-distance form of edits show several advantages. Given its versatility to express different kinds of edits in this form (see section 3.2), the entire strategy can be expressed as a set of interval-distance edits. This eases both the computer implementation of the strategy and its dissemination in a very transparent way. However, this form of expression of edits is currently used for short-term business statistics, which comprise short questionnaires with just a few continuous variables. More complex questionnaires such as those of structural business statistics embracing also semicontinuous and/or discrete variables are still to be tested under this proposal.

From a statistical methodology perspective, we have proved that the traditional approach based on score functions for output editing is embraced by the combinatorial optimization approach. In particular, the local score function for a variable $y^{(q)}$ naturally arising from the latter is given by

$$s_k^{(q)} = M_{kk}^{(q)}(y_k^{obs}, \hat{y}_k; \hat{p}_k, \hat{\nu}_k^2, \hat{\sigma}_k^2),$$

where the different quantities $\hat{\cdot}$ are estimated by usual methods using statistical models. The different local scores are put together in the global score function, which for the output editing we choose to be

$$S_k = S^{(\alpha)}\left(F^*_{\mathbf{s}^{(1)}}(s_k^{(1)}), \dots, F^*_{\mathbf{s}^{(Q)}}(s_k^{(Q)})\right).$$

The efficiency of this choice can be seen in figure 1. Firstly, notice that when more units are cross-sectionally selected at time t_1 (top to bottom: 0, 50, 100, 150, 200) more somewhat influential units are detected. The detection of influential units also takes place when cross-sectionally selecting units at t_2 (left to right: 100, 150, 200). Indeed, the latter is more efficient than the former, thus indicating the power of using cross-sectional information: the more you use the more efficient you are. This strongly suggests that this cross-sectional dimension should be exploited in the application of edits upon each unit as soon as possible. This means that this information should be incorporated into the limits and threshold of the interval-distance edits in real-time. This points at deriving an intermediate optimization problem from the generic problem (1) so that the available information $\mathbf{Z}$ comprises the already collected values. This remains to be done.

Both the simulation and its later implementation rest on assumptions that must be critically assessed. Firstly the different interval-distance edits partially incorporate both the longitudinal and cross-sectional dimensions. To this end, there exist different ways of partitioning the sample to compute the involved quantiles. We have directly chosen the dissemination cells, not testing other alternatives. Some of these cells have a critical size for the quantile computation and further testing could bring some improvement. Secondly, to assign the radius to each interval we have chosen the pseudo-hit rate $\widetilde{\text{HR}}$. Other alternatives can also be considered. No further research has been conducted in this line. Conversely, the response simulation during the CAWI is admittedly too simplistic, thus driving us to too optimistic results (see figure 4), and more elaborate choices are currently under use in other surveys. Finally, the quantities E_{if} are admitted too *ad hoc* for this survey and its generalization to other surveys seems difficult to us. We are currently working on more standardized choices which do not depend so much on the particular survey.

This question of choice of the quantities E_{ij} is deeply connected to the allocation algorithm of the n_{cross} units for the cross-sectional phase. Notice that the need for such an algorithm arises from the multidomain character of the dissemination plan: which cell is given greater priority? In this sense, the algorithm is noticeably heuristic and a more rigourous treatment in terms of an optimization problem would be desirable.

According to EDIMBUS (2007) the entire preceding process lacks a final stage: the monitoring and fine-tuning of the strategy using performance indicators. This would allow us to assess the performance of each edit. However, the generalization to other surveys has been prioritized over the monitoring. Thus, it also remains to be done.

From the final output perspective, given the importance of short-term business statistics to measure the day-to-day economic situation and possible changes in its evolution, we see judging by figure 2 that the proposed E&I strategy yields indeed measurements of change rates in the indices with no noticeable error. Thus it does not impinge negatively on the survey quality.

Finally, the figure 4 shows a notable descending in the number of units to edit interactively, thus a gain in efficiency. The predicted rate has not been reached, indicating the need for the preceding revision of some simulation assumptions, in particular the response simulation during the CAWI. Also the final two multivariate edits to control the breakdown into markets need further research. The implementation in production has been again prioritized over further research. The current surveys under implementation will incorporate all future novelties.

A The greedy search algorithm to solve the combinatorial problem

The greedy search algorithm to solve the combinatorial problem P_{co} is based on a very simple idea. Let $I_0 = \{1, \ldots, n\}$ denote the index set of components of the selection strategy vector $\mathbf{R}$. We begin by considering the vector $\mathbf{r}^{(0)} = (1, \ldots, 1)^t$ and choose a component $i^* \in I_0$ whose value is to be transformed into 0: $r_{i^*} = 1 \rightarrow r_{i^*} = 0$. If the new vector $\mathbf{r}^{(1)}$ so obtained satisfies every restriction, this is the (sub)optimal solution; otherwise fix $r_{i^*} = 0$, set $I_1 = I_0 - \{i^*\}$ and choose another component i^* in I_1 whose value is to be also transformed into 0. Proceed iteratively until the vector $\mathbf{r}^{(l)}$ so obtained satisfies every restriction.

To choose the component i_l^* in each iteration l Salgado et al. (2012) proposed to use an infeasibility function $h(\mathbf{r}) = \sum_{q=1}^{Q} \left(\mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r} - \eta_q\right)^+$ so that $i_{l+1}^* = \text{argmin}_{k \in I_l} h(\mathbf{r}^{(l+1)})$. The sooner every restriction is satisfied, the closer the suboptimal solution will be to the exact optimal solution. Indeed, they will occassionally coincide. Notice that we can write

$$h(\mathbf{r}_{l+1}) = \sum_{q=1}^{Q} \left(\mathbf{r}_l^T\mathbf{M}^{(q)}\mathbf{r}_l - M_{i_{l+1}^* i_{l+1}^*}^{(q)} - \eta_q\right)^+,$$

so that the choice of the i^* is indeed a choice of seeking the simultaneous fulfillment of every restriction by finding the joint maximal values $M_{i^* i^*}^{(q)}$ for all q. In this proposal this is undertaken by minimizing the infeasibility function h. But this can be slightly generalized.

If we focus on the selection of the component i^* instead of the infeasibility function, we can choose $i_{l+1}^* = \text{argmax}_{k \in I_l} S\left(M_{kk}^{(1)}, \ldots, M_{kk}^{(Q)}\right)$. For example, we can choose $i_{l+1}^* = \text{argmax}_{k \in I_l} \max_{1 \leq q \leq Q} M_{kk}^{(q)}$ or $i_{l+1}^* = \text{argmax}_{k \in I_l} \sum_{q=1}^{Q} M_{kk}^{(q)}$.

We can think of each choice of S as the minimum of a certain infeasibility function h. In the next appendix we will show that the diagonal entries $M_{kk}^{(q)}$ play the role of local score functions, while the function S is equivalent to a global score function.

B The prioritization algorithm and the equivalence with the score function approach

Firstly, we discuss the prioritization algorithm. Then we will use it together with the results of the preceding appendix to establish the equivalence with the traditional score function approach.

The main idea behind the prioritization of units using the combinatorial optimization P_{co} is the construction of an adequate sequence of upper bounds $\boldsymbol{\eta}$. We start by setting $\eta_q^{(0)} = \sum_{k\in s} M_{kk}^{(q)}$ and $\Omega^{(0)} = \{1,\ldots,n\}$. Then we solve problem $P_{co}(\boldsymbol{\eta}^{(0)},\Omega^{(0)})$, whose solution is $\mathbf{r}^{(0)} = (1,\ldots,1)^T$. No unit is selected, as expected. Then we reduce the upper bounds to $\boldsymbol{\eta}^{(1)}$ and set $\Omega^{(1)} = \Omega^{(0)}$ so that the solution $\mathbf{r}^{(1)}$ to the new problem $P_{co}(\boldsymbol{\eta}^{(1)},\Omega^{(1)})$ is $r_k = 1$ for all k, except for one $r_{k^*} = 0$. Unit k^* has been selected. Again we reduce the upper bounds to $\boldsymbol{\eta}^{(2)}$ and set $\Omega^{(2)} = \Omega^{(1)}\big|_{r_{k^*}=0}$, i.e. the set of possible solutions $\mathbf{r} \in \Omega_1$ with $r_{k^*} = 0$. This procedure is repeated n times. The prioritization is given by the sequence of indices $\{k_l^*\}_{l=1,\ldots,n}$ which in each iteration indicates the new unit selected.

Now, conjugating both this algorithm with an appropriate sequence of bounds $\{\boldsymbol{\eta}^{(l)}\}_{l=1,\ldots,n}$ and the choice of the infeasibility function h we can recover the traditional score function approach to selective editing. If $\eta_q^{(l+1)} = \sum_{k\in I_l} M_{kk}^{(q)} - \max_{k\in I_l} S\left(M_{kk}^{(1)},\ldots,M_{kk}^{(Q)}\right)$ and the infeasibility function is chosen so as to have $i_{l+1}^* = \operatorname{argmax}_{k\in I_l} S\left(M_{kk}^{(1)},\ldots,M_{kk}^{(Q)}\right)$ in each iteration of the greedy algorithm, then the prioritization of units is given by the descending order of the values of $S\left(M_{kk}^{(1)},\ldots,M_{kk}^{(Q)}\right)$.

To prove it, note that in iteration $l+1$ of the prioritization algorithm the restrictions of the problem read

$$\mathbf{r}_{l+1}^T\mathbf{M}^{(q)}\mathbf{r}_{l+1} - \eta_q^{(l+1)} \leq 0 \quad q = 1,\ldots,Q. \tag{4}$$

Now to solve the problem in this iteration we substitute $\eta_q^{(l+1)} = \sum_{k\in I_l} M_{kk}^{(q)} - \max_{k\in I_l} S\left(M_{kk}^{(1)},\ldots,M_{kk}^{(Q)}\right)$ into equation (4) to arrive at

$$\mathbf{r}_{l+1}^T\mathbf{M}^{(q)}\mathbf{r}_{l+1} - \mathbf{r}_l^T\mathbf{M}^{(q)}\mathbf{r}_l + \max_{k\in I_l} S\left(M_{kk}^{(1)},\ldots,M_{kk}^{(Q)}\right) \leq 0, \quad q = 1,\ldots,Q. \tag{5}$$

Now the suboptimal solution $\mathbf{r}_{l+1}$ found with the recipe

$$i^*_{l+1} = \text{argmax}_{k\in I_l} S\left(M^{(1)}_{kk}, \dots, M^{(Q)}_{kk}\right)$$

in the greedy algorithm is precisely $\mathbf{r_l}$ with $r_{i^*_{l+1}} = 0$, since equation (5) reduces to $M^{(q)}_{i^*i^*} \geq \max_{k\in I_l} S\left(M^{(1)}_{kk}, \dots, M^{(Q)}_{kk}\right)$ for all q. Thus $i^* = i^*_{l+1}$ makes $\mathbf{r}_l$ to satisfy every restriction.

Finally the ordering given by $i^*_{l+1} = \text{argmax}_{i\in I_l} S\left(M^{(1)}_{kk}, \dots, M^{(Q)}_{kk}\right)$, $I_{l+1} = I_l - \{i^*_{l+1}\}$ is exactly the decreasing order of the values $S\left(M^{(1)}_{kk}, \dots, M^{(Q)}_{kk}\right)$. If we think of $M^{(q)}_{kk}$ as local score functions and of S as a global score function, we are reproducing the traditional approach.

References

INE (2011). *Industrial Turnover Indices. Industrial New Orders Received Indices. Base 2005. CNAE-09. Methodological Manual*.

Abramowitz, M. and Stegun, I.A. *Handbook of mathematical functions*. Dover.

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

Arbués, I. and Revilla, P. Score functions under the optimization approach. *UNECE Work Session on Statistical Data Editing*, WP**1**, 1–9.

Arbués, I., Revilla, P., and Saldaña, S. (2011). Selective editing as a stochastic optimization problem. *UNECE Work Session on Statistical Data Editing*, WP**18**, 1–10.

Arbués, I., Revilla, P., and Salgado, D. (2012). Optimization as a theoretical framework to selective editing. *UNECE Work Session on Statistical Data Editing*, WP**1**, 1–10.

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

Bercebal, J.M. and Maldonado, J.L. IRIA: statistics production model of the National Statistical Institute of Spain. *Meeting on the Management of Information Systems*, WP**23**, 1–12.

Caporello, G.L. and Maravall, A. Program TSW. Revised Reference Manual. *Documentos Ocasionales* núm. 0408. Banco de España.

de Waal, T. (2005). Solving the error localization problem by means of vertex generation. *Survey Methodology* **29**, 71–79.

de Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley.

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

EFTA/Eurostat/UNECE High-level seminar on streamlining statistical production (2011).

U.S. Federal Committee on Statistical Methodology. (1990). Data Editing in Federal Statistical Agencies. Technical report, Statistical Policy Office: U.S. Office of Management and Budget, Washington, D.C.

Granquist, L. (1997). The new view on editing. *International Statistical Review* **65**, 381–387.

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

EC Council Regulation No. 1165/98, of May 1998.

EC Council Regulation No. 1158/05, of July 2005.

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

Meeting on the Management of Statistical Information Systems, 2012. *Topic ii. Streamlining statistical production*.

Meeting on the Management of Statistical Information Systems, 2013. *Topic ii. Streamlining statistical production*.

Nemhauser, G. and Wolsey, L. (1999). *Integer and combinatorial optimization*. Addison-Wesley.

Nichols, E., Murphy, E., Anderson, A., Willimack, D., and Sigman, R. (2005). Designing interactive edits for U.S. electronic economic surveys and censuses: issues and guidelines. *Research Report Series(Survey Methodology)* 2005-03, 1–10.

Pierzchala, M. (1990). A review of the state of the art in automated data editing and imputation. *J. Official Stat.* **6**, 355–377.

Renssen, R. and Camstra, A. Standard process steps in Statistics. *Meeting on the Management of Statistical Information System*, WP**2**, 1–9.

Salgado, D., Arbués, I., and Esteban, M. (2012). Two greedy algorithms for a binary quadratically constrained linear program in survey data editing. *INE Spain Working Paper* 02/12.

Salvador, A. and Salgado, D. A generalization of the stochastic approach to selective editing. *In preparation*.

UNECE (2013a). Generic statistical business process model. Version 5.0.

UNECE (2013b). Generic statistical information model. Version 1.1.