Distr.
GENERAL

Working Paper
10 April 2013

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, Bangkok, Thailand, 23-25 April 2013)

Topic (iii): Innovation

An efficient editing and imputation strategy within a corporate-wide data collection system at INE Spain: a pilot experience

Prepared by Rocío López-Ureña, María Mancebo, Silvia Rama, and David Salgado
(National Statistical Institute, Spain)

I. Introduction

1. Data editing and imputation is a crucial phase in survey statistics production. It affects important dimensions of data quality such as accuracy, timeliness, cost effectiveness and response burden. Although a large body of theoretical development is still lacking (at least in comparison with other phases such as sampling strategies or variance estimation), there already exist recognized standards [EDIMBUS (2007); de Waal et al. (2011)] to design editing and imputation (E&I henceforth) strategies combining different editing modalities as interactive editing, automatic editing, (micro)selective editing and macro editing.

2. On the other hand, in national and international statistical offices it is nowadays recognized the necessity to abandon the traditional stovepipe production model in favour of a more industrialized one. This transition involves many different aspects of the production process, including data editing itself.

3. In this context, INE Spain have recently begun to undergo this transition by implementing a corporate-wide metadata production process [Revilla et al. (2011)], which, among other things, adopts the Generic Statistical Business Process Model [GSBPM (2009)] as a generic framework. Given the complexity of this transition and the unavoidable nonstopping production compromise, a step-by-step approach has been undertaken. The first step has been the development and implementation of a global parameterized IT tool called IRIA designed to collect data for all surveys at INE Spain, either business or household surveys. Data collection through the CAWI mode is already in production.

4. IRIA supports the GSBPM in different phases and subprocesses [Revilla et al. (2012)]. In particular, the subprocess 5.3 (review, validate and edit) is partially supported inasmuch as editing during data collection is involved. Post-capture data editing is still undertaken in a stovepipe-like way.

5. In this paper we report the results of a pilot experience regarding the design and development of an efficient E&I strategy for a monthly short-term business statistics. The strategy follows the generic structure of an E&I strategy proposed in EDIMBUS (2007) together with a stage involving editing during data collection. The objective is two-fold. On the one hand, an increase of efficiency is pursued by flagging the minimal number of questionnaires to edit which guarantees the same accuracy achieved with the current strategy. On the other hand, we develop an editing methodology which can be exported to other surveys. The strategy design exploits IRIA's capabilities and versatility.

II. Design of the E&I strategy

A. The Spanish Industrial Turnover Index and Industrial New Orders Received Index survey

6. The short-term Industrial Turnover Index (ITI) and Industrial New Orders Received Index (INORI) survey is part of the action programme for the Development of the Statistical System on Industry compiled by Eurostat, and is governed, pursuant to (EC) Council Regulation No. 1165/98, of 19 May 1998 regarding short-term statistics, modified by (EC) Regulation No. 1158/2005 of the European Parliament and Council, of 6 July 2005. Although the new orders variables have been discarded in the context of European regulations (EC Regulation No. 461/2012), INE Spain keeps its data collection within the Spanish National Statistical Plan.

7. 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.

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

(a) Fixed panel of aprox. 11000 industrial establishments selected by cut-off (originally coincident with the Spanish Industrial Production Index survey sample).
(b) Data collection modes: CAWI, mail, email, fax and telephone. CAWI-mode submitted questionnaires are collected with IRIA; the rest of questionnaires are collected at several provincial delegations.
(c) Laspeyres indices disseminated for 37 publication cells identified as certain divisions and subdivisions[1] of the NACE Rev.2, without geographical breakdown but broken down by national, euro, noneuro and rest of the world markets.

9. For forthcoming comparisons, we detail the current E&I strategy field work flow for each time period (month) t on average:

Day(s)	Action(s)
t+0	Beginning of data collection
t+0 through t+16	Paper questionnaire submissions to respondents and data collection, data recording and interactive editing at provincial delegations
t+16	First edited data transmission from provincial delegations to the central office
t+16 through t+27	Data collection, data recording and interactive editing at provincial delegations
t+27	Final edited data transmission from provincial delegations to the central office
t+27 through t+46	Macro editing and indices computation at central office
t+46	Press release

10. The current E&I strategy comprises traditional format, balance and ratio edits with fixed thresholds. Currently around 55% of questionnaires are monthly flagged for editing. These all undergo interactive editing. Nonflagged questionnaires undergo no microediting at all.

[1] By subdivisions we mean an intermediate level between divisions and groups.

B. The new E&I strategy

11. The new E&I strategy follows the recommendations of the EDIMBUS manual. It is designed according to the generic structure of an E&I strategy in addition to the inclusion of a previous phase for editing during data collection (for the CAWI mode). Thus, according to field work conditions, we distinguish three stages in the strategy:

(a) editing during data collection using IRIA;
(b) interactive editing at the provincial delegations;
(c) macro editing at the central office.

Those questionnaires flagged in the macro stage at the central office are again subjected to interactive editing at the provincial delegations. At most one editing cycle takes place.

12. Each stage comprises a set of check controls for the whole sample. The first two stages are based on the traditional methodology based on edits and score functions [de Waal et al. (2011)]. The third stage is based on the optimization approach to selective editing recently proposed by INE Spain [Arbués et al. (2012)].

13. We describe by and large in the following items the check controls for the two first stages. We distinguish three generic types, namely, survey-specific controls, interval-distance controls and distribution-angle controls. By survey-specific controls we mean those controls which depend very sensibly upon the nature and meaning of the survey variables, such as format edits (e.g. all variables regarding new orders received must be nonnegative) and balance edits (e.g. the total turnover must equal the sum of the turnovers of each market). More specific controls are also included in this type (e.g. a questionnaire is flagged if the total turnover for the current time period equals the total turnover of the preceding time period). This kind of controls can be hardly generalized because each survey carries its own idiosyncrasy.

14. By interval-distance controls we mean the following methodological proposal intended to be used in any kind of business survey. For the reported value of the variable of analysis of each respondent we assign a validation interval. We measure the distance (see below) from the reported value in the questionnaire to this interval: if this distance is greater than certain threshold (see below), the questionnaire is flagged; it is not, otherwise.

15. The construction of these validation intervals depends on the auxiliary information related to the survey. In the ITI and INORI survey, we exploit the fact that the sample is a fixed panel, which entails that we have a time series for each respondent since it became part of the sample. Thus we adjust automatically an ARIMA model using TRAMO-SEATS [Caporello and Maravall (2004)] for each respondent producing a predicted value $\hat{y}_k^{(q)}$ and an estimated standard deviation $\hat{\sigma}_k^{(q)}$ for variable $y_k^{(q)}$, q=1,...,Q. The interval is constructed as $I_k^{(q)}(s) = \left[\hat{y}_k^{(q)} - s \cdot \hat{\sigma}_k^{(q)}, \hat{y}_k^{(q)} + s \cdot \hat{\sigma}_k^{(q)}\right]$. To choose the parameter s, we construct intervals $I_k^{(q)}(s)$ for the preceding time period $t-1$ and for a range of values of s (say, from $s=0$ to $s=15$). Let s_{t-1}^* denote the value of s which maximizes the hit rate (obtained using both the raw and edited versions of the corresponding data). The final value of s for the current period t is chosen as $s_t = \frac{1}{12} s_{t-1}^* + \frac{11}{12} s_{t-s}$. This convex combination is intended to stabilize monthly the length of the intervals, which at the end affects the number of flagged questionnaires and the field work. To account for the potential impossibility to adjust an ARIMA model (due to too short time series or too many missing values in the time series), a ratio edit with respect to the preceding period is used instead driving us to another interval $\bar{I}_k^{(q)}$. The final interval is selected as the most restrictive one: $I_k^{(q)} \cap \bar{I}_k^{(q)}$.

16. The distance measure is indeed a set of distance measures depending on two options. The first option depends on whether the control is to be used as an edit or as a score function. If it is to be used as an edit, the distance $d\left(y_k^{(rep,q)}, I_k^{(q)}\right)$ from a reported value $y_k^{(rep,q)}$ to the interval $I_k^{(q)} = \left[l_k^{(q)}, u_k^{(q)}\right]$ for the questionnaire k is 0 if the value lies inside the interval and ∞ otherwise. If it is to be used as a function score, then we must take into account the second option, namely, whether the variable $y^{(q)}$ is discrete or continuous. If it is discrete, we define (ω_k stands for the sampling weight of unit k):

$$d(y_1k^{\intercal}((rep,q)), I_1k^{\intercal}((q))) = \omega_1k \cdot \{\begin{matrix}0 & \text{if } y_1k^{\intercal}((rep,q)) \in I_1k^{\intercal}((q))\\ @y_1k^{\intercal}((rep,q)) - u_1k^{\intercal}(\end{matrix}$$

If it is continuous, we define instead

17. To determine the threshold used as a benchmark to flag the questionnaires, we compute the distance between each final edited value $y_k^{(edit,q)}$ of the preceding period and their corresponding interval $I_k^{(q)}$. The sample is broken down into domains S_j coincident with the minimal publication cells. In the ITI and INORI survey these are the 37 publication cells referred to above in item 8. For each domain S_j the threshold is established as a subject-matter chosen distance quantile $q_f\left(\{d_k : k \in S_j\}\right)$ over the distribution of distances.

18. These interval-distance control checks are applied as edits in the first stage during the Web data collection and as function scores in the second stage during the editing process at the provincial delegations. They are applied to the total turnover and the total new orders received. We will refer to these two variables as the level variables of the survey.

19. On the contrary, their respective breakdowns into markets (national, euro, noneuro and the rest of the world) will be referred to as distribution variables, since they express how those totals are distributed among the markets. To control these variables we make use of the third type of checks, that is the distribution-angle control. For concreteness' sake let us firstly focus upon the turnover variables. Let us denote by $T_{k,i}$ the turnover of establishment k for market i. Notice that the total turnover T_k satisfies $T_k = T_{k,1} + T_{k,2} + T_{k,3} + T_{k,4}$. Set the vector $d_{T,k} = \frac{1}{T_k}(T_{k,1}, T_{k,2}, T_{k,3}, T_{k,4})$, which represents basically the percentage of turnover of each market. Set analogously $d_{O,k} = \frac{1}{O_k}(O_{k,1}, O_{k,2}, O_{k,3}, O_{k,4})$, with self-explaining notation. Define the (squared cosine of the) angle between two distribution vectors d_1 and d_2 as

$$\widehat{(d_1, d_2)} = \frac{(d_1 \cdot d_2)^2}{(d_1 \cdot d_1)(d_2 \cdot d_2)}.$$

This value is a measure of similarity between two distribution vectors. If the value $\widehat{(d_{k,1}, d_{k,2})}$ is below certain threshold, the questionnaire k is flagged; it is not, otherwise.

20. This control is applied to the pairs $\left(d_{T,k}^{(t)}, d_{O,k}^{(t)}\right)$ and $\left(d_{T,k}^{(t)}, d_{T,k}^{(t-1)}\right)$. That is, similarities between (i) the reported distributions of the turnover and the new orders received at the current time period t, and (ii) the reported distribution of the turnover at the current time period t and the preceding time period $t-1$ are checked.

21. The threshold is determined using a similar procedure to that in item 17. The angles are computed for the preceding time period and subject-matter chosen quantiles are determined for each publication cell.

22. Finally we include a very broad description of the optimization approach to selective editing which we use in the final macro stage of the strategy (see [Arbués et al. (2012)] and references therein for details). As generic principles for selective editing we state

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

We implement these principles as an optimization problem whose solution states which questionnaires are flagged for further interactive editing. This optimization problem seeks to minimize the number of questionnaires to flag restricted to imposing an upper bound to the mean squared measurement error. This measurement error is modelled using what we have called an observation-prediction model, that is, we construct a model for the true values $y_k^{(0,q)}$ and a model for the reported values $y_k^{(rep,q)}$ conditional on the true values $y_k^{(0,q)}$. By an extension of Bayes' theorem, we can compute the second-order moments of the measurement errors $e_k^{(q)} = y_k^{(obs,q)} - y_k^{(0,q)}$ conditional on the reported (observed) values $y_k^{(rep,q)}$. These moments allow us to estimate the mean squared measurement error of each survey variable, which is bounded from above with bounds chosen by subject-matter knowledge. The final result is a selection of units.

23. For convenience of editing field work conditions, it is often preferable to have a prioritization of questionnaires instead of a selection. This can be accomplished by choosing a suitable sequence of bounds (see [Arbués et al. (2012)] for details). Thus, we can achieve a higher degree of control of the field work by fixing the numbers of units to flag. In the ITI and INORI survey we have studied the relation between this number of units selected at the macro editing stage and the precision gain of the indices in each publication cell.

24. The prioritization of questionnaires is applied to each publication cell, so that two possibilities arise: either a fixed number of flagged units is chosen for each publication cell or a global fixed number of flagged units is chosen which have to be allocated among the different cells. The second option favours the possibility to adjust the allocation of units among cells accounting for the accuracy of the indices. This is accomplished by proceeding in three steps:

(a) An initial allocation is chosen according to subject-matter knowledge (possibly assigning no unit to each cell).
(b) The number of allocated units in each cell is fixed by

$$n_j = \left\lfloor \left(\sum_{f=1}^{F} \lambda_f E_{jf} \right) \cdot n \right\rfloor,$$

where n stands for the sampling size; E_{jf}, $f = 1, \ldots, F$, stands for chosen direct or indirect measures of the number of errors in cell j and λ_f are positive numbers such that $\sum_{f=1}^{F} \lambda_f = 1$. For the ITI and INORI survey, we have chosen $F = 4$ measures given by E_{j1}, which is the maximal second-order moment of the measurement errors of the total turnover and total new orders received in cell j; E_{j2}, which is the weight of cell (NACE division) j in the national indices; E_{j3}, which is the fraction of questionnaires with reported total turnover $T_k^{(rep)} = 0$ in cell j; E_{j4}, which is the

proportion of questionnaires in cell j having reported $T_k^{(rep)} = 0$ in the preceding time period but whose final value was imputed ($T_k^{\dagger}((ed)) \neq 0$). Prior to their inclusion in the preceding formula, all these values are to be normalized across the sample (dividing by their sum across cells) so that they satisfy $\sum_j E_{jf} = 1$.

(c) The floor function in the preceding formula does not guarantee that $\sum_j n_j = n$. Thus, the remaining units ($n - 1$ at most) are allocated one by one according to the descending order of one of the factors E_{jf}. In the ITI and INORI survey we have chosen E_{j1}.

C. Simulation and results

25. We have conducted a simulation study using both the raw and edited versions of the ITI and INORI microdata corresponding to 13 months (from December, 2011 to December, 2012). We have used only data collected via the CAWI mode, since production conditions at INE Spain do not allow us to keep raw data (data are edited from the very first moment they are collected and recorded). They represent the 70% of the sampling size on average.

26. We have applied the proposed E&I strategy simulating the editing field work by substituting the flagged raw questionnaires by their edited counterparts, computing both the indices using the traditionally edited values and the indices using the selectively edited values and comparing them for different set of parameters.

27. As a first result, only 15% of the questionnaires are finally flagged in any stage of the strategy. The precision is assessed by comparing both sets of indices through the absolute relative error given by $|\Delta I_j^{sel}| = \left|\frac{I_j^{sel} - I_j^{ed}}{I_j^{ed}}\right| (\%)$ for publication cell j. In figure 1 we show the progressive error reduction of the national ITI (vertical axes) of each NACE division (horizontal axes) for $n_1 = 0, 50, 100, 150, 200$ (from top to bottom) units selected in the macro editing analysis after the first data transmission and for $n_2 = 100, 150, 200$ (from left to right) units selected after the second data transmission.

III. Implementation in actual production conditions

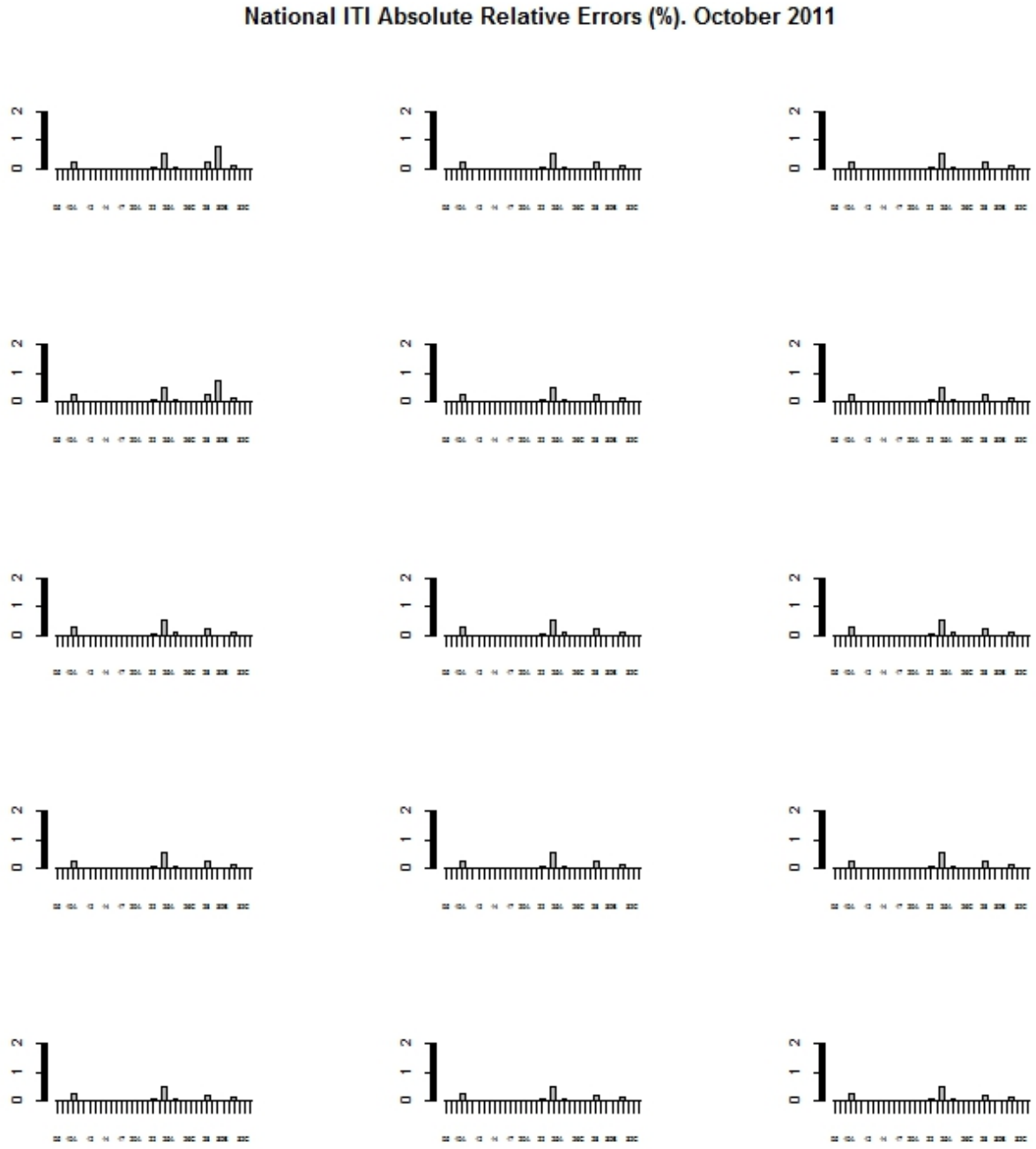
28. The implementation of the above E&I strategy in the production process at INE Spain has taken advantage of the capabilities of IRIA. Firstly, the final strategy must be completed with the editing of those questionnaires not collected via the CAWI mode. To accomplish this, the strategy distinguishes between questionnaires collected via the CAWI mode and the rest. The former ones undergo the process described above; the latter ones undergo the two last stages, but using the interval-distance controls as edits instead of as score functions during the interactive editing at the provincial delegations.

29. Although the strategy depends on parameters which remain stable throughout the successive time periods, it entails new intervals, distance thresholds and angle thresholds in each month. These are timely uploaded in IRIA according to the following modified averaged monthly work flow:

Day(s)	Action(s)
t-7 through t-5	Computation of new intervals, distance thresholds and angle thresholds
t-5	Uploading in IRIA of new intervals, distance thresholds and angle thresholds
t+0	Beginning of data collection
t+0 through t+16	Paper questionnaire submissions to respondents and data collection, data recording and interactive editing at provincial delegations
t+16	First edited data transmission from provincial delegations to the central office
t+16 though t+18	First macro editing analysis. Selection of 100 respondents for further interactive editing

t+16 through t+27	Data collection, data recording and interactive editing at provincial delegations
t+27	Second edited data transmission from provincial delegations to the central office
t+27 through t+29	Second macro editing analysis. Selection of 100 respondents for further interactive editing
t+27 through t+39	Data collection, data recording and interactive editing at provincial delegations
t+39	Final edited data transmission from provincial delegations to the central office
t+39 through t+46	Indices computation at central office
t+46	Press release

Figure 1. Absolute relative errors for the national ITI. October 2011.



30. We notice that the final edited values of data corresponding to the preceding month are not ready for their use in the computation of the new intervals and thresholds. The most recent data set is used instead, which corresponds to two months prior to the reference time period. Accordingly, ARIMA predictions are

made two periods ahead of time and ratios and angles are taken with variables from two periods before the reference month.

31. The computation of the new intervals and thresholds is carried out automatically by SAS macros specifically written for this purpose. These macros are inserted in the information systems of the survey conductor unit so that they can read both the raw and edited versions of the data of each time period. Subject-matter experts choose the different parameters and the macros output the final file to be uploaded to IRIA.

32. Since a preliminary macro editing analysis is carried out after the first data transmission, it is highly convenient to have collected as many questionnaires as possible. INE Spain, through their provincial delegations, has formally asked the respondents to bring forward their response.

33. This strategy has been implemented so as to be applied on data being collected from February, 1 2013.

IV. Conclusions and future prospects

34. Despite the fact that we still lack data to draw conclusions about the performance of this strategy in real production conditions, this pilot experience allows us to envisage modifications to improve this proposal regarding its potential usage as a standard tool in the editing production phase at a large scale, in particular for short-term business statistics.

35. With the three types of controls depicted above, both level and distribution variables can be controlled using the same methodology. Furthermore, the angle-distribution control can be indeed reformulated as an interval-distance control by defining an interval $[q_j, 1]$, where q_j stands for the angle quantile of cell j computed as above.

36. The use of the interval-distance control as a standard eases its software implementation in data collection applications, since for each questionnaire we need only to specify the interval bounds, the distance quantile threshold, the edit/score condition and the discrete/continuous condition for each survey variable of analysis.

37. Seasonality can be accounted for by using time series techniques to determine the interval bounds and the distance quantile thresholds as predicted values. In the above ITI and INORI survey, ARIMA modelling has been used at individual respondent level due to the fact of being a fixed panel.

38. In other surveys (e.g. in rotating panels), we can apply a similar approach in a more aggregated level and descend to respondent level to determine the predicted value. As an illustrating example, let us consider a level variable y in a yearly rotating panel. We compute the ratio $r_k^{(c)} = y_k^{(c)} / y_k^{(c-1)}$ for the last 12 time periods using the historic edited values of the survey variable (leaving out the missing values and/or dropped-off respondents), and construct a time series of the quantiles $q_{j_1}^{(c)}\left(\{r_k^{(c)}\}_{k \in s_j}\right)$ and $q_{j_2}^{(c)}\left(\{r_k^{(c)}\}_{k \in s_j}\right)$ for each domain s_j. Thus we can predict the values $\hat{r}_{j_1}^{(c)}$ and $\hat{r}_{j_2}^{(c)}$, and then the values $\hat{y}_{k,j_1}^{(c)} = \hat{r}_{j_1}^{(c)} \cdot y_k^{(c-1)}$ and $\hat{y}_{k,j_2}^{(c)} = \hat{r}_{j_2}^{(c)} \cdot y_k^{(c-1)}$, which determine the validation interval for unit k.

39. The conjunction of statistical methodology and information technologies entails an increase of data quality, since timeliness is favoured, cost effectiveness is enhanced, response burden is reduced and accuracy is under control. Notice that the parametrisation of the above E&I strategy through the choice of the above quantiles drives us to a good control of the accuracy/cost trade-off.

40. At INE Spain we have recently initiated a programme to apply these proposals to most short-term business statistics.

References

I. Arbués, P. Revilla, and D. Salgado (2012). *Optimization as a theoretical framework to selective editing*. UNECE Work Session on Statistical Data Editing WP2. Available from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/02_Spain.pdf. Accessed February, 15, 2013.

G. Caporello and A. Maravall (2004). *Program TSW. Revised Reference Manual. Banco de España*. Available from http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/04/Fic/do0408e.pdf. Accessed February, 15, 2013.

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

EDIMBUS (2007). *Recommended practices for editing and imputation in cross-sectional business surveys*. ISTAT, CBS, SFSO, EUROSTAT. Available from http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf. Accessed February, 15, 2013.

GSBPM (2009). *Generic Statistical Business Process Model*. UNECE. Available from http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model. Accessed February, 15, 2013.

P. Revilla, J.L. Maldonado, and J.M. Bercebal (2011). *Towards a corporate-wide electronic data collection system at the National Statistical Institute of Spain*. UNECE Work Session on Statistical Data Editing WP12. Available from http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp.12.e.pdf. Accessed February, 15, 2013.

P. Revilla, J.L. Maldonado, F. Hernández and J.M. Bercebal (2012). *Implementing a corporate-wide metadata driven production process at INE Spain*. INE Spain Working Paper 05/12. Available from http://www.ine.es/ss/Satellite?L=es_ES&c=INEDocTrabajo_C&cid=1259939125971&p=1254735839320&pagename=MetodologiaYEstandares%2FINELayout. Accessed February, 15, 2013.