UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working paper
15 February 2015

Workshop on the Modernisation of Statistical Production
Meeting, 15-17 April 2015

Topic (iii): Innovation in technology and methods driving opportunities for modernisation

Data collection and selective data editing in a systematized and integrated way: an experience in progress at Statistics Spain

Prepared by J.M. Bercebal, J.L. Maldonado, M.A. Martínez-Vidal, and D. Salgado

josemanuel.bercebal.gomes@ine.es, joseluis.maldonado.cecilia@ine.es, miguelangel.martinez.vidal@ine.es, david.salgado.fernandez@ine.es

Statistics Spain, Spain

I. Introduction

1. Statistics Spain is currently undergoing a strong effort to evolve from a highly scattered stove-pipe production model to an industrialised and standardised model where resources are more efficiently used.

2. Given the complexity of the overall statistical production (up to 250 statistical operations currently at Statistics Spain) and the increasing pressure from budgetary restrictions, we are involved in a stepwise procedure where production processes to be refurbished have been prioritised.

3. Among the first processes to be renovated both data collection and data editing appear as more resource-consuming. Each of these processes is the object of two large projects within the purpose to modernize its production.

4. On the one hand, data collection is being completely revamped through the design, development and usage of a modular and extensible parameterized IT tool called IRIA. In section II we describe the general details of this tool.

5. On the other hand, data editing is under an intense overhaul to exploit selective editing techniques as much as possible. In particular, an optimization approach to selective editing has been proposed which allows us to automatize a great deal of the selection of questionnaires to be edited. Its general principles are included in section III.

6. Finally, in section IV we show how we have combined both initiatives into the actual production process to optimize the available resources.

II. IRIA: a parameterised IT tool

A. Main features

7. Statistics Spain has been implementing and using a new computer system for data collection in different statistical operations in the last three years: IRIA (Bercebal and Maldonado, 2014).

8. IRIA is a modular, configurable and extensible system which allows survey conductors to design, build, edit, exploit and manage both household and business surveys under different collection modes (CAPI, CATI, CAWI, etc.) reusing already created modules. Ease of use is pursued and survey conductors do not require deep computer skills. The system reduces the participation of IT personnel to a minimum.

9. **Modular and configurable:** IRIA comprises a set of elements or modules so that the way in which selected modules couples among them constitutes a specific data collection application. Moreover, these modules are configurable, thus showing different behaviour accounting for a wider variability in their functionalities, depending on the type of survey, survey conductors' preferences, data collection modes …

10. **Extensible:** IRIA can be straightforwardly extended without needing complex programming developments or source code recompiling tasks:

a) It allows the automatic creation of clients to use web services during an interview. This client is managed thorough a JavaScript function.
b) New visual components can be created to be used in the questionnaire design and thus in the interviews.
c) The application used to manage and follow up data collection is built through portlets or components. A basic catalogue of components is available which is growing with proposals from different survey conductors accounting for their own needs.
d) The different stages of each respondent are determined by a workflow, which can be configured and modified along the data collection procedure.

11. **Survey conductor:** IRIA has been designed to be used by survey conductors without deep computer skills. As a main feature it incorporates simple interfaces with help menus allowing the user to check errors.

12. **Types of surveys:** IRIA is prepared to collect data for both household and business surveys, either structural or short-term and using different collection modes (CAPI, CATI, CAWI …).

13. **Reuse:** In a system like IRIA it is essential for a critical reduction of deployment times in new surveys to reuse elements, modules, and components already present in ongoing surveys.

a) **Element reuse:** The IRIA component MANAGER (see below) allows us among other things to assign to each survey all its properties, to determine which elements it will use and to decide and configure its workflow. It is important to remark that it also allows us to copy elements from other ongoing surveys so that the time reduction for the creation of new similar statistical operations is noticeable.
b) **Question reuse:** After the completion of an interview design, every question is stored in the system so that they can be used again either in their current form or as a supporting guide in the design of new interviews.

c) **Information reuse:** IRIA is a common system for different surveys, thus the reuse and sharing of information among them stand up as a natural possibility. Furthermore, IRIA also stores metadata about storage information such as access features and structure of this information.

14. **Minimal computer skills:** At Statistics Spain, personnel with computer skills only takes part in the general management of IRIA and the construction of complex elements.

B. IRIA components

15. IRIA is a computer system designed in a modular fashion comprising applications which are developed under the same programming language and are integrated so that they interact among each other to assist survey conductors in the survey life cycle from design to data collection.

16. IRIA comprises four general applications:

a) An application to manage the properties of a survey called MANAGER.
b) An application to design interviews called DESIGNER.
c) An application to conduct interviews called ENGINE.
d) An application to manage the daily continuous data collection called DATA COLLECTION.

17. These general applications are complemented with specific applications to assist in the data collection procedures under different collection modes.

18. **IRIA MANAGER: the system management application**

This component allows the survey conductor to assign any required property to the survey at stake. The information managed by IRIA MANAGER appears at two levels:

1. "Above survey", i.e. general information for any survey.
2. "Survey specific", i.e. specific information for the survey at stake and its elements.

Within the "above survey" level we outline the management of security. IRIA users are conferred different user roles with IRIA MANAGER, which registers new users defining their characteristics and granting their access permissions. A hierarchy among user roles is determined so that each user role will have supervising permissions over users with an offspring role. Although permissions are generally attached to user roles, it is possible to modify them specifically for a given survey according to concrete needs. Thus tailor-made access hierarchies are achieved even at this general level.

Within the "survey-specific" level, IRIA MANAGER confers surveys different properties thus giving them the desired functionalities either from simple elements such as the survey name itself, languages to use, data collection modes, schedules and human teams to more complex elements. Among these we outline a menu for data collection management and possibly survey-customized screens for self-administered computer-assisted collection modes containing together with the questionnaire information about the sampling unit (respondent) such as contact information, his/her access to the interview and its completion status, etc. Also, it is possible to define different status of the survey data collection phase and workflows thereof.

IRIA MANAGER also allows survey conductors to assign to each collection mode different properties such as customized screens for self-administered questionnaires, diverse aspects related to telephone-administered interviews and tailor-made configurations of portable devices (tablets, laptops …).

19. **IRIA DESIGNER: the interview designer**

This component embraces both the design and the construction of interviews and assists in the composition of their specifications, since it allows the survey conductor to adjust the specifications in detail by partially conducting the survey in test mode.

IRIA DESIGNER is an interface giving survey conductors the flexibility necessary to program the interviews. They, as subject matter experts, can use the interface to create different screens (font specifications, screen background …), questions, workflows or validation rules.

Users can enjoy support in different ways: they can launch the interview, can modify questions during after design phases and during execution stages, view the appearance of certain screens, view the structure of the programmed workflow in graphical mode with the additional aid of a small console to perform interactive editing and imputation, if necessary.

Additionally, IRIA DESIGNER allows the survey designer to make interviews mode-dependent. Thus questions wording, their workflows and their associated edits can vary according to the data collection mode.

Finally, IRIA DESIGNER allows the survey designer to reuse (group of) questions from other surveys or from a library. Therefore, it is not necessary to create them from scratch thus impinging upon homogeneity across surveys. This enables in a natural way both standardization and rules to standardize the production process. As a prominent example we outline that at Statistics Spain there exists a core of sociodemographic variables for household surveys.

20. **IRIA ENGINE: the interview engine**

This is the component allowing us to administer the interview designed previously. It is a key piece in the system since it interprets the logics of all interviews, the language of the designer and applies the presentation, either the default one or that included in the design.

During the interview administration, each element has multiple properties enabling a simple integration with complementary processes, as e.g. the selective editing techniques depicted in section III. IRIA ENGINE arranges different access modes according to the task to be carried out (interviewing, editing, coding, viewing, etc.). The interview will develop following the design established with IRIA DESIGNER according to its status, the access mode, the current user, the collection mode…

In the particular case of data editing, IRIA ENGINE enjoys several recommended features:

a) It enables three types of edits: (i) hard, upon failure of which it is impossible to proceed the interview; (ii) soft, upon failure of which a confirmation from the user is needed by prompting the user for it with a message, and (iii) informative.
b) Edits parameters can vary along the interview or according to the collection mode, the interviewer, or the status of the interview.
c) Edits can be associated with single questions, group of questions or at the end of the interview.
d) Edits can use information stored in the system itself or in auxiliary systems (e.g. past historic data of the same survey).

21. **IRIA DATA COLLECTION: the application for data collection**

This is the component to be finally used by the personnel in direct charge of data collection, from its very beginning to its completion, either for occasional or periodical surveys. The execution of IRIA DATA COLLECTION follows the configuration established with IRIA MANAGER. Therefore, as indicated above, this execution will vary from survey to survey according to the specific needs taken into account in the design phase.

This application manages the coordination and integration of data from every collection mode, allowing the survey conductor to access data from each channel and to open or close channels depending on the particular conditions of the interview at stake. In this way, data can be received from different modes either simultaneously or exclusively through one channel at will.

Complementarily, IRIA DATA COLLECTION manages the contact with the respondents, either through postal letters (e.g. official survey fill-in requests, interviewer visit appointments …) or through emails, SMS, labels, etc.

Also, IRIA DATA COLLECTION manages the execution of the designed workflow. Each sampling unit (respondent) will have an interview administration status, among the set of which he/she will evolve upon executing different tasks.

Since IRIA ENGINE is necessary for different process as to entry a collected paper questionnaire into the system, to carry a telephone survey, etc., both IRIA ENGINE and IRIA DATA COLLECTION are integrated. This integration can follow any logics required by the user. IRIA DATA COLLECTION can deliver information to IRIA ENGINE to carry out an interview, which, in turn, delivers back to IRIA ENGINE information necessary to change the status in the workflow (e.g. the interview is valid, or it requires further revision, or it is incomplete, etc.).

III. Selective data editing under the optimization approach

22 The selection of questionnaires to edit is a crucial step in the statistical production process, since it is documented that the data editing phase consumes up to 40% of the total resources of the survey (Granquist, 1997).

23 In Statistics Spain the selection of questionnaires has been posed as an optimization problem which tries to minimise the amount of resources used in terms of the number of selected units while assuring data quality understood as a control over the measurement non-sampling errors.

24. Besides, this optimization approach allows us to reduce the amount of manual procedures and to implement the methodological proposal as an automatic process. In conjunction with the versatility of IRIA this has been exploited to establish a refurbished data editing workflow in the production chain which is currently reducing the respondents' recontact rates. Without entering into many mathematical details, in this section we depict those particular features of this methodological proposal which have allowed us to exploit it in a standardised manner.

25. To begin with, it is important to recognise the available information to make the selection of units as a key ingredient in the problem. As a matter of fact, this available information will drive us either to a stochastic optimization problem or to a combinatorial optimization.

26. Beforehand, we need some notation. We denote by $\boldsymbol{r}$ the selection strategy vector whose components r_k determine whether each unit k is selected ($r_k = 0$) or not ($r_k = 1$) to be edited (see Arbués et al., 2013 for details on this counterintuitive choice). Let $\Delta^{(q)}(\boldsymbol{r})$ be the loss functional associated to the estimator $\hat{Y}^{(q)}(\boldsymbol{r})$ due to the possible presence of measurement errors in the values of the variable $y^{(q)}$. Let $\boldsymbol{Z}$ denote the set of variables comprising the available information when performing the selection. Some of the units may be selected in advance (because they are influential units, etc.). This is denoted by introducing the set of selection strategy vectors $\Omega(\boldsymbol{Z}) \subset \{0,1\}^{\times n}$. The exploitation of the available information is expressed by means of the conditional expectation. Thus we can compactly express the generic optimization problem as

$$\begin{gathered} \max \mathbb{E}\left[\sum_{k\in s} r_k \mid \boldsymbol{Z}\right] \\ \mathbb{E}\left[\Delta^{(q)}(\boldsymbol{r})|\boldsymbol{Z}\right] \leq\ b_q, q = 1, \dots, Q, \\ \boldsymbol{r} \in\ \Omega(\boldsymbol{Z}). \end{gathered}$$

27. Basically, the available information comprises the sets of historical raw and edited data of the survey at stake. The sets of raw data are those data sets containing the values of the object variables before any editing task is undertaken upon them. The sets of edited data are those data sets containing the final values of the variables entering into the estimators to be used to produce the released aggregates (estimates, yearly rates, etc.). Complementarily, as we shall indicate below, we also exploit the values of the variables right after the input editing phase is completed but before the output editing phase begins.

28. The exploitation of the available information has direct consequences in the design of the production chain. On the one hand, the information must be available as soon as possible when it is generated. On the other hand, this information must be punctually accessed to make the selection of units on time for the subsequent field work.

29. So far, the methodological proposal has focused on two extreme cases: either no cross-sectional information on the current time period data set is available or the values of the (almost) complete sample of units has been collected and can be used. In the first case only the historical data values of each unit can be exploited independently. This is denoted by $\boldsymbol{Z} = \boldsymbol{Z}_{long}$. On the contrary, when all cross-sectional information is available, both the

historical data values of each unit and all values of the current time period can be taken advantage of. This is denoted by $\boldsymbol{Z} = \boldsymbol{Z}_{cross}$.

30. When $\boldsymbol{Z} = \boldsymbol{Z}_{long}$, the generic optimization problem reduces to a stochastic optimization problem (see Arbués et al, 2013, Arbués and Revilla, 2014), which is solved in several steps (see Arbués et al, 2012 for details). This approach drives us to the construction of traditional score functions (de Waal et al, 2011) in a natural way. Specific algorithms adapted for this particular problem to be used in actual production conditions instead of general optimization routines are currently under development. In the remaining we will focus on the second choice.

31. Nonetheless, for the input editing phase we make use of a highly intuitive, though computationally demanding, approach by which we assign to each variable $y_k^{(q)}$, $q = 1, \dots, Q$, of each unit k a validation interval $I_k^{(q)}$, a value-interval distance function $d_k^{(q)}$ and a threshold value t_k, so that should $d_k^{(q)}(y_k^{(q,obs)}, I_k^{(q)}) > t_k$, then questionnaire k will be flagged for editing. We have called it *interval-distance edit* (López-Ureña et al, 2013, 2014).

32. Needless to say the key issue in the interval-distance edits is the computation of both the interval and the threshold value for each variable and each unit. In the first case, this is accomplished by exploiting the historical data sets and automatically adjusting time series models which are then used to make predictions either for both the interval centres and radii or directly for both extremes of the each interval (see López-Ureña et al, 2013, 2014). In the second case, a time series of validation intervals is constructed for each past time period, a distance value is computed for each time period, each variable and each unit and a predicted value thereof is computed. Quantiles of these predicted values upon a carefully chosen set of population cells are calculated to be used as threshold values for each variable and each unit of the corresponding cell (see López-Ureña et al, 2013, 2014).

33. Once interval-distance edits are computed for each variable and each unit, these are loaded into the data collection system, which just checks in turn upon each questionnaire arrival whether each collected value $y_k^{(q,obs)}$ satisfies the edit or not and consequently flags the questionnaire or not.

34. When $\boldsymbol{Z} = \boldsymbol{Z}_{cross}$ the generic optimization problem reduces to a combinatorial optimization problem (see Arbués et al, 2012, 2013). For practical reasons the loss functional is chosen to be $\Delta^{(q)}(r) = \sum_{k \in s} r_k \omega_k |y_k^{(q,obs)} - Y_k^{(q)}|$, where ω_k stands for the sampling design weights, $y_k^{(q,obs)}$ denotes the reported (observed) value of variable $y^{(q)}$ for unit k and $Y_k^{(q)}$ is the modelled random value of the same variable $y^{(q)}$. Thus, the problem reads

$$\max \sum_{k \in s} r_k$$
$$\sum_{k \in s} \sum_{l \in s} r_k M_{kl}^{(q)} r_l \leq \eta_q, q = 1, \dots, Q,$$
$$r_k \in \{0,1\},$$

where η_q are upper bounds to control the bias increase for each variable $y^{(q)}$ and $[M_{kl}^{(q)}]_{1\leq k,l\leq n}$ are so-called diagonal loss matrices whose entries can be expressed in terms of the observed values $y_k^{(q,obs)}$, of the predicted values $\hat{y}_k^{(q)}$ according to a so-called observation-prediction model and of its (estimated) parameters (Arbués et al, 2013).

35. For each set of upper bounds η_q the resolution of this problem yields a particular selection strategy $\boldsymbol{r}^*$ stating which units k are selected ($r_k = 0$) or not ($r_k = 1$). From a field work standpoint (see Arbués et al., 2013 for a discussion) it is more convenient to have a prioritization of units instead of a selection. This prioritization is achieved by using decreasingly running upper bounds η_q from $\eta_q = \sum_{k\in s} M_{kk}^{(q)}$ to $\eta_q = 0$.

36. Furthermore, it can be proven (López-Ureña et al, 2014) that each sequence of running bounds is equivalent to define a global score function $S_k = S(M_{kk}^{(1)}, \dots, M_{kk}^{(Q)})$ of the loss components $M_{kk}^{(q)}$, which can then be considered as local score functions. This global score function plays the role of an infeasibility function in the heuristic resolution of the optimization problem (López-Ureña et al, 2014). This heuristic approach is chosen striving for speed while in practice it only introduces a small amount of overediting.

37. Thus not only does the optimization approach recover the traditional score function approach but it also extends the latter and puts it in a firmer basis by introducing prediction models in a natural way, since $M_{kk}^{(q)} = M_{kk}^{(q)}(y_k^{(q,obs)}, \hat{y}_k^{(q)})$. The introduction of statistical models to make predictions makes us cherish the hope to apply this formalism not only to quantitative variables but also to qualitative and semicontinuous variables. In other words we pursue the ambition to standardise this proposal both for business and household surveys.

38. In actual production conditions, the survey conductor only needs to choose both the global score function S and the parameters ruling the construction of each predicted value $\hat{y}_k^{(q)}$. Then the system automatically computes the loss components $M_{kk}^{(q)}$, the global scores S_k and the prioritisation of units. Depending on time restrictions the survey conductor decides how many of them are to be edited in this output editing phase.

IV. Executing selective data editing with IRIA

39. The implementation of the optimization approach with IRIA is depicted in figure 1, which uses the business process modelling language BPMN 2.0 to express the production workflow.

40. The first step comprises the computation of the predicted values for each variable and each unit. Access to the historical microdata database is crucial. Firstly, only the historical data of each unit is available: the longitudinal phase of the editing strategy is executed. Following the proposal above we construct the interval-distance edits for each variable and each unit with the computed predictions. These edits configure IRIA to control variable values of each received questionnaire. Data are collected following the corresponding input editing procedure if edit failures are detected.

41. The microdata database is updated and new predictions are computed to initiate the cross-sectional phase. Information from (almost) all sampled units is available at this point. This information is exploited as indicated above to prioritize units. Survey conductors usually decide in advance how many of them are to be edited, since time restrictions and available resources very rarely change. IRIA is configured with this selection and units are subjected to a new editing procedure, possibly driving us to recontact them. Finally the microdata database is updated and data processing continues usually with the next (macro) editing stage.

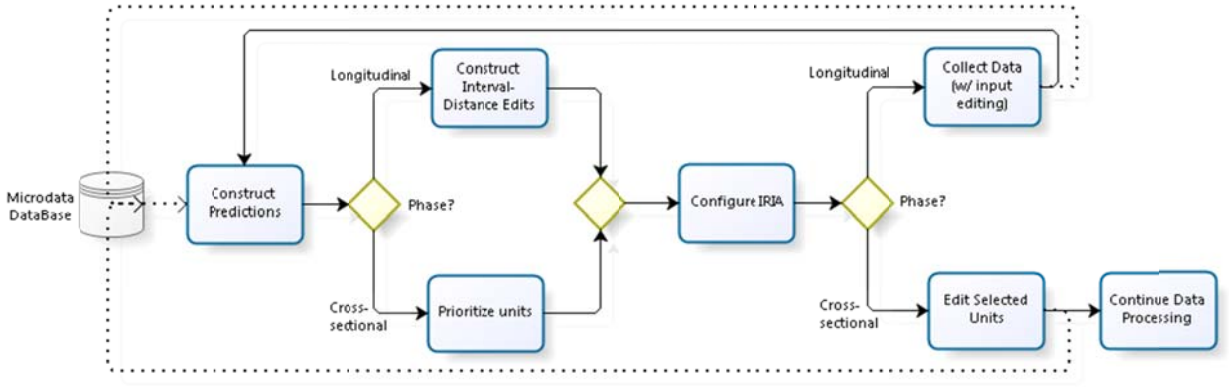

Figure 1. Production workflow for the implementation of the optimization approach with IRIA.

42. Currently not all steps in figure 1 are implemented in IRIA. Let us differentiate among (i) the microdata database, (ii) the prediction construction, (iii) the interval-distance edits construction, (iv) the prioritization of units, (v) the data collection procedures and (vi) the data editing procedures.

43. The microdata database is currently under deployment being completed for the CAWI collection mode for all business surveys and for CAPI, CATI, and CAWI collection modes for some household surveys. The rest of modes and/or surveys, as well as the editing functions will be hopefully integrated within the next 12-18 months.

44. In the meantime, to integrate the new editing strategies in IRIA we have created a microdata repository with the information of each survey following the new proposal. This repository follows a NoSQL database design, in particular, it constitutes a key-value store containing both the raw and edited values of the variable of each sampled unit for the past 24 months. This store is fed monthly with the new information.

45. The computation of predictions, the construction of interval-distance edits and the prioritization of units are carried out in an independent system. The computation of predictions is dealt with using TRAMO-SEATS (Caporello and Maravall, 2008) to automatically model time series. This is complemented with *ad hoc* routines to tackle with non-standard instances. Both the construction of interval-distance edits and the prioritization of units is accomplished with specific routines which implement the preceding methodological proposals.

46. The key feature of these procedures is that the survey conductor is in total control of the parameters ruling both the construction of the interval-distance edits and the

prioritization of units. The system works as a black box producing both the edits and the prioritization.

47. The expression of the interval-distance edits has been standardized by specifying for each one the following elements: (i) a name for the edit, (ii) the name of the variable upon which to apply the edit, (iii) a standard interval $\hat{I}_k^{(q)}$ computed with the original interval, the distance type and the threshold value, and (iv) the data collection mode in which the edit must be used.

48. The name of the edit allows us to make a posteriori analyses of its behaviour in the execution of the editing strategy. The name of the variable to control and the associated interval are the minimal ingredients to implement these edits. The standardization of the intervals by including both the distance types and the threshold values simplifies later computations. The system only needs to check upon data collection whether the reported value is within the standard interval or not. No more computations are needed. Finally, mixed-mode collection procedures are taken into account by attaching the corresponding edit parameters to each mode.

49. The set of selected units resulting from their prioritization is complemented with information about the main three variables with higher local scores $M_{kk}^{(q)}$. In particular, instead of these values, both the predicted error value (basically $\hat{y}_k^{(q)} - y_k^{(q,obs)}$) and its estimated standard deviation $\hat{v}_k^{(q)}$ are included together with the ID variables of the units. This is intended to facilitate the editing field work.

50. All this information is managed by IRIA under a previously agreed schedule. IRIA DESIGNER is used to implement the design of the editing strategy involving both interval-distance edits and the selected set of prioritized units. This design needs to take into account the periodical load of all edits parameters for all variables under control of all units.

51. Data collection along with data editing during collection is accomplished with IRIA. The collection is executed under the design established previously. Both IRIA ENGINE and IRIA DATA COLLECTION execute the editing tasks, as well as others, as designed.

52. As main results of this revamped process the respondents' response rates are being reduced up to 20 percentage points upon the sample sizes. This procedure has already been implemented in the Industrial Turnover Indices (ITI) and Industrial New Orders Received Indices (INORI) survey, the Services Sector Activity Indicators (SSAI) survey, and the Retail Trade Indices (RTI) survey. All of them are short-term business statistics monthly released.

53. The ITI and INORI survey was the first to follow this new editing strategy. The simulation studies embraced the entire year 2012. These studies could only use data collected through the CAWI mode. In figure 2 we represent comparatively the averaged recontact rate during 2012 for all collection modes, the predicted rates for the CAWI mode obtained with the simulations and the actual recontact rates during 2013 for both the CAWI mode alone and all collection modes. In 2014 similar results were obtained.

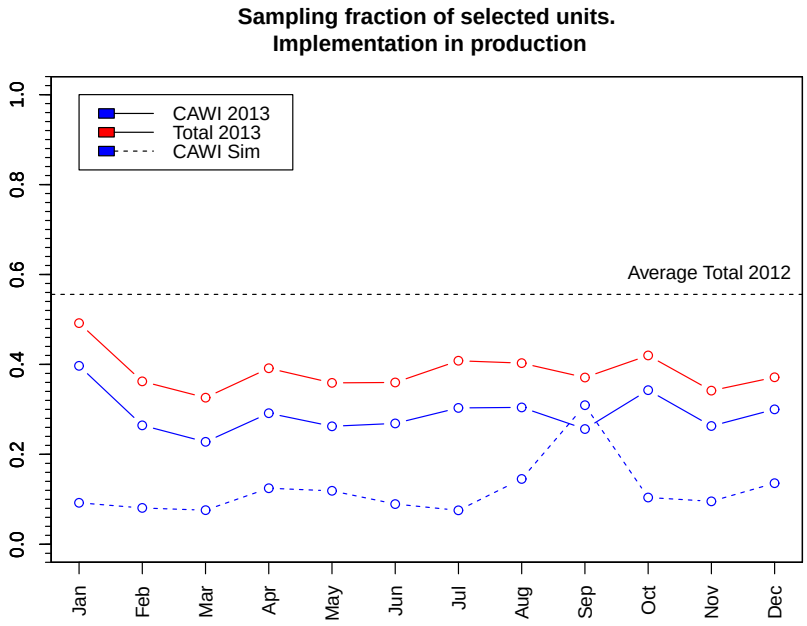

Figure 2. Respondent recontact rates for 2012 and 2013.

54. For both the SSIA and RTI surveys, we are currently analysing the results from the first 6 months of implementation. Strong differences appear to be between the CAWI mode and the rest of modes. In the first case a reduction up to 20% is again noticeable. However, in the latter case, this reduction has not taken place and, rather on the contrary, too tight bounds for the edits have resulted from the choice of parameters. These are being progressively loosed to converge to the CAWI results.

55. From the other point of view, subject matter experts conducting the surveys have not detected a reduction on data quality apart from the necessary and expected fine-tuning of the edit parameters according to actual production conditions.

56. As of this writing, the preliminary tasks of design and numerical simulations to choose optimal parameters have also been finished for Industrial Price Indices survey and the Export and Import Price Indices for Industrial Products survey. These simulations suggest a reduction of up to 10 percentage points in the recontact rates. The implementation in production will be undertaken along the present year.

57. To sum up, different components of the information system of the statistical production process at Statistics Spain are being refurbished. We are undertaking a stepwise modular approach focusing on different phases of the production for its later integration under a common system. The new processes allow survey conductors to have a finer control on the editing tasks, keeping data quality standards while optimizing resources. First results on short-term business statistics are notably positive and further work is under way to extend this proposal to structural business statistics and hopefully household surveys.

References

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. (2014). Score functions under the optimization approach. *UNECE Work Session on Statistical Data Editing, WP1*, 1 – 9.

Arbués, I., Revilla, P., and Salgado, D. (2012). Optimization as a theoretical framework to selective editing. *UNECE Work Session on Statistical Data Editing, WP1*, 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. (2014). IRIA: Statistics Production Model of the National Statistical Institute of Spain. *Meeting on the Management of Statistical Information Systems, WP*, 1-12.

Caporello, G.L. and Maravall, A. (2008). Program TSW. Revised Reference Manual. *Documentos Ocasionales* núm. 0408. Banco de España.

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

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

López-Ureña, R., Mancebo, M., Rama, S., and Salgado, D. (2013). An efficient editing and imputation strategy within a corporate-wide data collection system at INE Spain: a pilot experience. *Meeting on the Management of Statistical Information Systems, WP 10*, 1-10.

López-Ureña, R., Mancebo, M., Rama, S., and Salgado, D. (2014). Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index Survey. *Statistics Spain Working Paper* 04/2014.