
**Working Papers**

05/2014

# Standardising the editing phase at Statistics Spain: a little step beyond EDIMBUS

Silvia Rama and David Salgado



First draft: October 2014

This draft: October 2014

# Standardising the editing phase at Statistics Spain: a little step beyond EDIMBUS


## Abstract

We propose a slight generalization of the generic EDIMBUS editing and imputation strategy based on the notion of statistical production function and the inclusion of editing during data collection therein. Some first consequences are introduced such as the parametrization of the strategy in terms of the amount of cross-sectional information available for the execution of these functions and a minimal set of specification rules for them (already present in the literature). Also, we pose specific examples of the editing function whose goal is the selection of units for interactive editing so as to optimise resources. The whole proposal fits within the efforts for the modernisation of the statistical production process conducted at Statistics Spain.



## Keywords

Editing strategy, EDIMBUS strategy, production function



## Authors and Affiliations

S. Rama and D. Salgado

S.G. Metodología y Desarrollo de la Producción Estadística

Instituto Nacional de Estadística

# Standardising the editing phase at Statistics Spain: a little step beyond EDIMBUS

S. Rama and D. Salgado
S.G. Metodología y Desarrollo de la Producción Estadística
Instituto Nacional de Estadística
Paseo de la Castellana, 183
28046 Madrid (Spain)

**Abstract**

We propose a slight generalization of the generic EDIMBUS editing and imputation strategy based on the notion of statistical production function and the inclusion of editing during data collection therein. Some first consequences are introduced such as the parametrization of the strategy in terms of the amount of cross-sectional information available for the execution of these functions and a minimal set of specification rules for them (already present in the literature). Also, we pose specific examples of the editing function whose goal is the selection of units for interactive editing so as to optimise resources. The whole proposal fits within the efforts for the modernisation of the statistical production process conducted at Statistics Spain.

## 1 Introduction

Recently, Statistics Spain has decided to redesign the editing and imputation (E&I) strategies of its surveys to pursue gains in cost effectiveness and an amelioration of response burden, thus also possibly impinging on timeliness while keeping accuracy in the estimates. This decision was partially adopted as a consequence of the results of a pilot experience (López-Ureña et al., 2013, 2014) on the application of the optimization approach to selective editing (Arbués et al., 2013) developed at Statistics Spain.

This pilot experience did not take into account necessary aspects for the standardisation of the new proposal. The present paper presents the general lines along which the standardisation of the editing phase is being undertaken to implement the redesigned E&I strategies across different surveys. The key goal is to bring together the top-down view proffered by general standards such as both the GSBPM (UNECE, 2013a) and the GSIM (UNECE, 2013b) and the unavoidable set of day-to-day tasks comprising the production process. In this regard we share the view that the notion of statistical production function (Pannekoek et al., 2013) is an excellent tool to connect both extremes. Hereafter a statistical production function is understood precisely as the set of tasks to be executed to carry out (a phase of) the statistical production process.

Furthermore, in the editing phase realm, we recognize the generic E&I strategy proposed in the EDIMBUS manual (EDIMBUS, 2007) as a first major step in this direction, especially

within the harmonization of the European Statistical System. This generic strategy comprises several E&I modalities (de Waal, 2009) which as a matter of fact can be somewhat understood as large standard process steps (Camstra and Renssen, 2011).

We propose an extension of the generic EDIMBUS E&I strategy to fully-fledged embrace both the notion of production function and the new modality of editing during collection. Regarding the first goal we need to introduce a parametrisation of the E&I strategy which will allow us to decompose it into clearly identifiable editing functions. Regarding the second goal we need to take into account the new features brought in by this new data editing modality and the great degree of integration of distinct production functions as already pointed out by Pierzchala (1990). Lastly, we describe in detail the unit selection functions within this scheme. These are indeed the editing functions being used at Statistics Spain to streamline the data editing phase. In particular this extended strategy has been used to implement the Business Turnover Index survey and the Retail Trade Indices and, as of this writing, it is currently being implemented in the Industrial Price survey and the Export and Import Price Indices for Industrial Products survey.

The paper is organized as follows. In section 2 we describe the proposed parametrisation of E&I strategies and its motivation. In section 3 we describe our proposal to extend the generic EDIMBUS strategy, including the editing during collection phase and the decomposition in terms of parameterised editing functions. In section 4 we concentrate on a standardised description of E&I strategies and the unit selection functions which we are using to streamline the editing phase at Statistics Spain. We close with some conclusions in section 5.

## 2 Parametrisation of the E&I strategy

In order to neatly choose a specific function to be executed at each step of the E&I strategy we firstly need to parameterise the latter. As in other phases of the whole statistical production process we recognise the available information as a key tool to execute the different tasks. By available information we mean any sort of information which can be used to optimise the execution of a task. For example, when editing data of a given unit in a given survey, it will embrace the information about both that particular unit and all related units at present and past time periods in that or other surveys or related directories. We propose to parameterise the E&I strategies around this notion of available information.

To this end we shall consider three dimensions of the available information: (i) longitudinal, (ii) cross-sectional and (iii) multivariate. By longitudinal we mean the value of variables for each unit in previous time periods. This implicitly assumes that the survey is periodical. In other cases we always have the possibility to consider a certain level of aggregation into larger domains. By cross-sectional we refer to the information stemming out from the whole sample at the current period. Finally, by multivariate we signify the information arising from the multidimensional character of the survey (usually several variables are investigated with the same questionnaire).

With this in mind, we introduce a first parameter $n_{col} = 0, 1, \ldots, n$ which denotes the number of questionnaires collected up to the time point of execution of the function under considera-

tion. Notice that this parameter gives an indirect measure of the time passed from the starting instant of the strategy to the execution of the current function. However, more often than not a questionnaire undergoes a second editing cycle (Granquist, 1997), thus we introduce a second parameter $n_{cyc}^{(k)} = 0, 1, 2, \ldots$ for each unit $k \in \{1, \ldots, n\}$. This parameter denotes the number of editing cycles which unit $k$ has undergone. Parallelly we shall denote by $n_{cyc}$ the number of cycles that the whole sample has undergone. In other terms, $n_{cyc} = m$ if $n_{cyc}^{(k)} \geq m$ for all $k \in s$.

Observe that again all parameters $n_{cyc}^{(k)}$ and $n_{cyc}$ give an indirect measure of the execution time of the whole strategy up to this point. It is clear that the lesser the parameter values are, the more efficient the strategy will be provided that the accuracy is guaranteed.

With this set of parameters, editing functions upon questionnaire $k$ will be specified by $\mathsf{FunctionName}(n_{col}, n_{cyc}^{(k)}, n_{cyc})$ whereas those upon the whole sample $s$ will be denoted by $\mathsf{FunctionName}(n_{col}, n_{cyc})$. A whole E&I strategy will be completely described when each editing function FunctionName is clearly specified for each possible set of values of the parameters $n_{col}$, $n_{cyc}^{(k)}$, $n_{cyc}$.

Indirectly this entails the creation of a library of editing functions among which to choose the specific function at each step. A useful taxonomy for this library is already given by Pannekoek et al. (2013). However, instead of specific functions, we will focus on categories of particular functions. For example, instead of using the name of a particular algorithm of automatic data editing, we shall use the generic function AutoEI to embrace any of these functions. In practice, when describing a particular E&I strategy of a given survey we ultimately need to specify the concrete editing function to use. In the next section we propose the categories of editing functions which we shall use in describing the generic strategy.

## 3 An extended generic E&I strategy

The starting point of our proposal is the generic EDIMBUS strategy itself. This generic strategy is given in some more detail by de Waal et al. (2011). However for our purposes we retain the original description (see figure 1). This strategy is an extraordinary tool to organize the editing phase, but to introduce the proposed extension some points deserve a deeper analysis.

Firstly, as the EDIMBUS manual states itself, the original strategy is intended only for the post data capture stage. In our opinion, a streamlined version of E&I strategies should also comprise input editing, that is, editing performed during data collection, including all previous necessary activities to carry it out. This points towards the integration of production functions already underlined by e.g. Pierzchala (1990). Thus we shall also consider editing during collection.

Secondly, the decomposition of the strategy depicted in figure 1 revolves around the historical modalities of data editing (de Waal, 2009), namely, interactive, automatic, selective and macro editing. However, for the adoption of production functions, this decomposition is not detailed enough, since it does not specify if the tasks regard an individual questionnaire or the

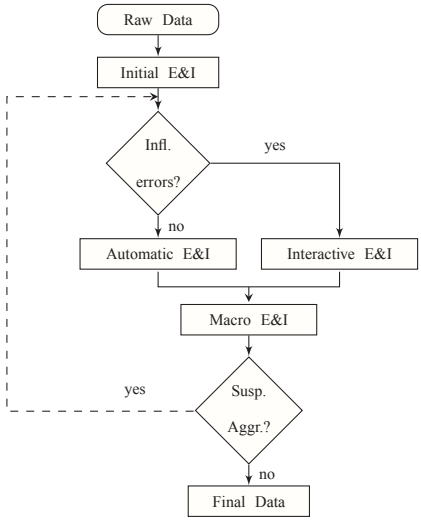


Figure 1: Original generic EDIMBUS strategy (EDIMBUS, 2007).

whole sample or a subset thereof. Thus in the extended version we will specify whether a particular function is to be applied upon an individual unit or upon a set of units.

Thirdly, being important as they are, we claim that the sheer historical modalities are not fully appropriate as categories of editing activities within a strategy. Instead, we propose the following categories, very close to those already contained in the EDIMBUS strategy:

1. Editing during collection, which we will denote by the generic function EditColl. This family of functions refers to any modality of editing during the collection of data. A first natural classification of these functions can be undertaken by distinguishing them through the collection mode, i.e. CAPI, CATI, CAWI, mail, etc. or using any layered classification of these modes (see e.g. Biemer and Lyberg (2003)).

2. Interactive editing and imputation, which we will denote by the generic function InterEI. It embraces all those editing and imputation functions in the interactive editing modality (de Waal, 2009; de Waal et al., 2011). A further decomposition is hard given the varieties of course of action under this modality.

3. Automatic editing and imputation, which we will denote by the generic function AutoEI. It comprises all those editing and imputation functions in the automatic editing modality (de Waal, 2009; de Waal et al., 2011). Again, the increasing varieties of algorithms to perform these tasks (de Waal et al., 2011) indicates a high degree of complexity in the classification of this family.

4. Selection of units, which we will denote by the generic function UnitSelection. It refers to those tasks committed to decide whether a unit is considered influential or not. Equivalently, in more general terms, this function will determine whether the unit will enter into the interactive editing modality or into the automatic editing modality. We shall describe in detail three of these functions in section 4.

5. Validation, which we will denote by the generic function Validation. This family encompasses all tasks dedicated to validate the data set as a valid final microdata set to be passed to the next production phase.

There exist subtle differences between these generic functions and those derived from the historical modalities in the original EDIMBUS strategy. To begin with, we have dropped out the initial E&I stage which has been substituted by the editing during collection phase. Some of the considerations regarding the initial E&I step (EDIMBUS, 2007) are to be embedded in this new phase. Furthermore, the design of these functions should include those principles and guidelines already appearing in the literature (Nichols et al., 2005).

Next, in our opinion the distinction between micro selective editing and macro editing must be blurred, since they can be considered as the initial and final steps of the same editing process which evolves according to the number of questionnaires already collected. Both the statistical methodology and the data technology will bring in how the progression of the data collection can be exploited for a more efficient data input editing. Notice how this again points towards the integration of production functions. As a consequence of this blurring, we have introduced a generic function whose main result is the selection of a given unit to be subjected to interactive or automatic editing. The specific function to apply in any case is given by the parameters $n_{col}$, $n_{cyc}^{(k)}$, $n_{cyc}$ introduced in the preceding section. In consonance, we have also dropped out the macro E&I phase.

Finally, the validation stage is equivalent to the determination of whether the computed aggregates are suspicious or not, thus validating the microdata set. Notice that this generic function can be viewed somewhat as a unit selection function to be applied on the whole sample, instead of on an individual unit.

With all these considerations we only need to give the flow of functions to determine the extended version of the strategy. This is depicted in figure 2.

The starting point is the selected sample with the raw data. Then a sequence of functions identified with the name SelEdit is applied upon each individual unit. It comprises several steps. Firstly, at the same time as data are collected, they are subjected to editing during collection. This is denoted by the generic function EditColl. Notice that this must be an ingredient of an integrated production function involving indeed the data entry into the system, the interview and the input editing itself. The specific choice of function depends on the characteristics of the survey. Should the collection mode preclude the editing during collection, this function would reduce to the null function (no task involved).

Next, once the data of the questionnaire $k$ are within the system, it must be decided whether the unit $k$ is considered influential or not. We will detail instances of this function in the next section. If the unit is selected, then it undergoes interactive editing and imputation indicated under the function InterEI. The particular choice will depend on the number $n_{col}$ of questionnaires already collected and the number $n_{cyc}^{(k)}$ of editing cycles which unit $k$ has already undergone. The function comprises any possible course of action as respondent recontact, survey data imputation, etc.

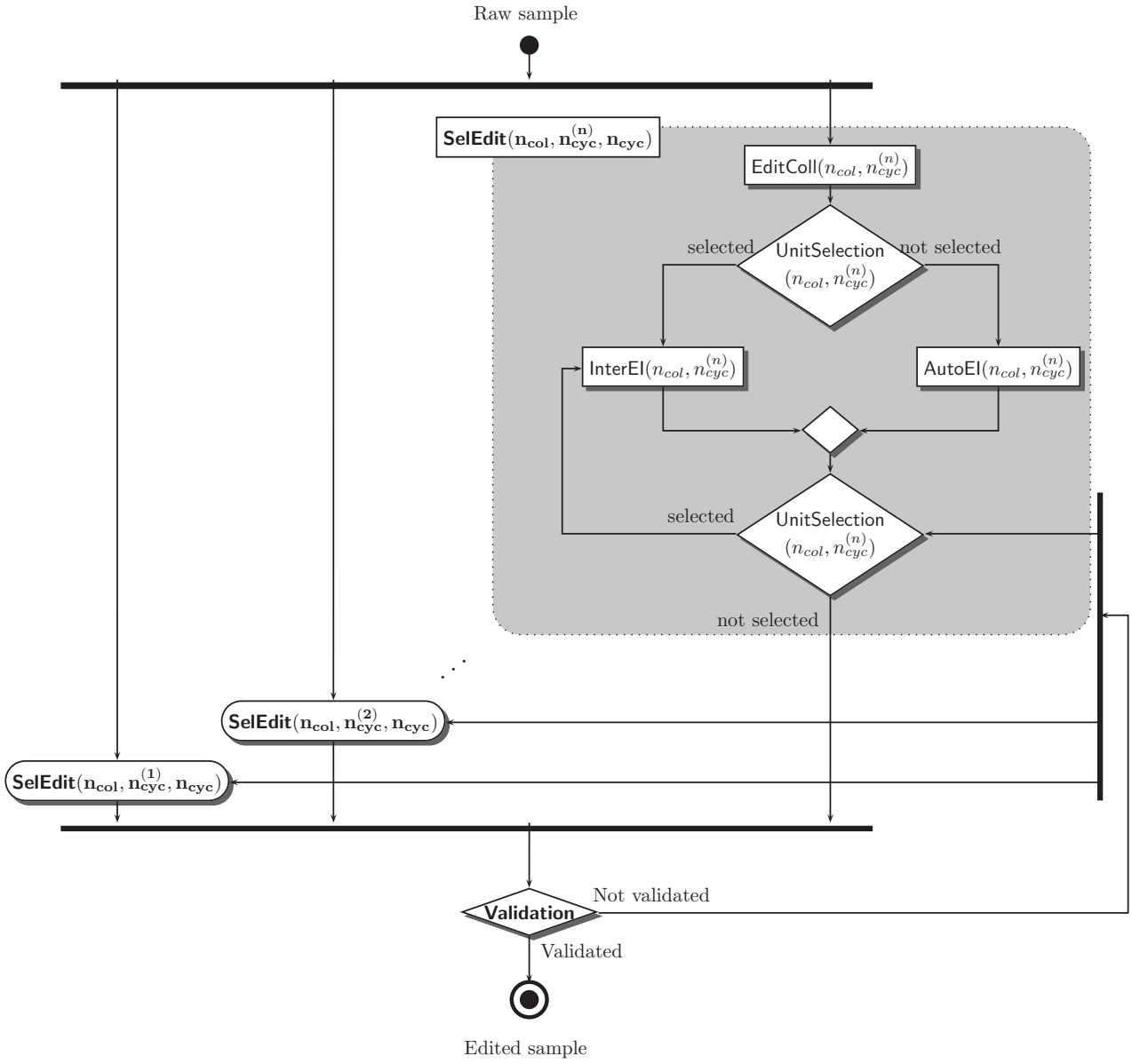


Figure 2: Proposed extended generic strategy.

If the unit is not selected, then it undergoes automatic editing. Notice that the corresponding function AutoEI is applied upon each unit $k$. However, the application of the function can also make use of all sort of available information such as historical data, collected data for the present realization of the survey and administrative data. This means that, if desirable, the application of the function can be deferred until the optimal amount of information is available.

The next step is the application of a UnitSelection function on unit $k$. The difference with the previous application is the available information which can be exploited to determine if unit $k$ is considered influential or not. In particular, this second application is conducted when the cross-sectional information about the complete sample is available. If the unit is selected, then it must undergo interactive editing, now with a different parametrisation. On the contrary, if the unit is not selected, then it must join the rest of units of the sample so that it is subjected to the Validation function.

If the complete sample if validated as a result of the application of the function Validation, we are done and we have the final validated microdata set. On the contrary, if the sample is not validated, then each unit $k$ must undergo a UnitSelection function to decide whether it must enter into interactive editing or not. This sequence of steps is repeated thus incrementing the number of editing cycles until the sample is finally validated.

# 4 Standardising E&I strategies at Statistics Spain

As described above, the extended generic E&I strategy depicted in figure 2 only contains top-level editing functions. This is but a very first step to standardise the data editing phase: a complete library of specific functions remains to be elaborated so that they can be neatly identified and chosen for every step of the execution of the strategy in actual production conditions for every possible survey. Furthermore, this library should comply with some (preferrably internationally agreed) standards.

Clearly, this is a huge exercise (which nonetheless should begin with the very first step). At Statistics Spain, under the need for streamlining the production process and optimising resources, we have concentrated on the UnitSelection functions. Upon the basis of the experience pilot described by López-Ureña et al. (2014), we have defined three (families of) unit selection functions based on the optimization approach to selective editing (Arbués et al., 2013) and the so-called *interval-distance* edit (López-Ureña et al., 2013, 2014).

The whole approach pursues to unify the selection of units under the same theoretical framework: an optimization problem derived from general principles to execute the data editing phase (Arbués et al., 2013). Under this approach the selection of units can be viewed as the result of a resource-consumption minimization problem posed according to the progression of the data collection process. So far, the methodology (Arbués et al., 2013) has proven the input and output editing to be possibly conceived as the initial and final steps, respectively, of the editing process according to the completion of the data collection.

On the one hand, when no cross-sectional information is actually used to perform the selection of units (i.e., in input editing conditions), we are under the stochastic optimization approach by which a unit $k$ is selected if

$$S_k = \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} > 1$$

where $\lambda_q^*$ are Lagrange multipliers obtained by solving an auxiliary optimization problem $P^*$ and $M_{kk}^{(q)}$ are conditional moments of the measurement error associated to the estimator $\hat{Y}^{(q)}$, $q = 1, \ldots, Q$ (Arbués et al., 2012, 2013). Thus, this stochastic optimization approach produces global score functions with an implicit threshold value. We detail this family of functions in section 4.1. Apart from this linear form, more possible combinations of error moments are under study (Salvador and Salgado, 2014).

The auxiliary optimization problem $P^*$ can only be solved so far using a general optimization routine (Arbués et al., 2012), which is not appropriate for our current production conditions. Thus more research is being conducted in order to propose a specific numeric algorithm to solve this problem so that it can be implemented both in SAS and R (Salvador and Salgado, 2014).

In the meantime, both for the pilot experience described by López-Ureña et al. (2013, 2014) and for the short-term business statistics whose E&I strategies are, as of this paper writing, under the redesign programme at Statistics Spain, a so-called *interval-distance* edit is being used instead. It comprises (i) a validation interval $I_k^{(qt)} = [l_k^{(qt)}, u_k^{(qt)}]$ for each variable $y^{(q)}$ under editing control, each unit $k \in s$ and the current time period $t$, (ii) a distance type $d_k = 1, 2, 3$ determining a corresponding distance function $d_k(I_k^{(qt)}, y_k^{(qt,obs)})$ between the interval $I_k^{(qt)}$ and the variable value $y_k^{(qt,obs)}$ collected in the questionnaire, and (iii) a threshold value $t_k^{(qt)}$. If $d_k(I_k^{(qt)}, y_k^{(qt,obs)}) > t_k^{(qt)}$ for any $q = 1, \ldots, Q$, then unit $k$ is selected. In practice, the original intervals $I_k$ can be redefined to absorb both the distance type and the threshold so that only a final validation interval $\overline{I_k}$ to check whether $y_k^{obs} \in \overline{I_k}$ or not is actually needed. In section 4.2 we detail interval-distance functions based on time series.

On the other hand, when the cross-sectional information is actually used from the complete sample (i.e., in output editing conditions), we are under the combinatorial optimization approach by which a unit $k$ is selected if the solution $\mathbf{r}^*$ of a derived combinatorial problem is such that $r_k^* = 0$ (Arbués et al., 2013). However, for actual production conditions, a unit prioritization instead of a unit selection solution may seem more adequate (see Arbués et al. (2013) for a wider discussion). Out of this prioritization the resource availability in the realization of the survey will produce an actual selection of units. In section 4.3 we summarise the selection of units based on the combinatorial optimization approach.

In all the three cases we will follow the recommendations of the Informal Task Force on Metadata Flows (2013) to describe the functions. By and large, we will specify the input data, the input parameters, the output, a metric of the process and a description of the process. This is intended to conform jointly the GSBPM and GSIM standards. Notice that we have introduced a slight generalization since the output can be either a data set (output data) or another function

(output function). This distinction is also valid for the input parameters, which can be data or functions.

Once the statistical methodology is decided, the standardisation of the E&I strategies at Statistics Spain has been initiated with a standard way to express the whole strategy and each of its conforming edit rules. The fundamental goal is to express the E&I strategy so that each unit (statisticians, IT personnel, data collection and data editing clerks, ...) involved in its implementation, execution and maintenance can clearly recognise their corresponding elements of information to play their role in the statistical production process. Additionally, E&I strategies expressed in a standard way and conveniently disseminated contribute to a higher degree of transparency.

As pointed out by Pannekoek et al. (2013), an E&I strategy can be specified by setting out the collection of edits which conform it. Taking this as a starting point, we have begun by describing each edit in a hierarchical fashion. At the top level, we include the following elements:

- Edit name. Each edit must be assigned an identifier so that it can be clearly identified within the whole strategy. In the following, we will denote this edit name by $\mathsf{e}$. However notice that by and large each edit $\mathsf{e}$ will correspond to a (possibly derived) variable $y^{(q)}$ under editing control: $q \leftrightarrow \mathsf{e}$.

- Data collection mode. Possibly there could be defined edits to be applied and/or with parameters depending on the data collection mode, thus each edit should also be specified with the data collection mode upon which it must be applied.

- Edit character. An edit must be hard or soft (de Waal et al., 2011) (however see also Scholtus (2013)), thus this should also be specified.

- Edit message. An active edit will trim a message either for the respondent or the data collection/editing clerk depending on the collection mode. This message must be composed following the general guidelines corresponding to each collection mode (see e.g. Nichols et al. (2005)).

- Edit parameters. Each edit is defined according to a set of parameters. To standardise this set of parameters we have slightly generalised the if-then edit form by de Waal (2005) (see also de Waal et al. (2011)). Accordingly, we set that if unit $k$ satisfies the condition $C_k(\mathbf{y},\mathbf{z})$ imposed on the questionnaire variables $\mathbf{y}$ and/or the auxiliary variables $\mathbf{z}$ (e.g. directory variables or sampling design variables), then the edit will be active if the (possibly derived) variable $y_k^{(qt,obs)}$ lies outside the corresponding validation interval $\overline{I}_k^{(qt)}$, that is, if $y_k^{(q)} < \overline{l}_k^{qt}$ or $y_k^{(q)} > \overline{u}_k^{(qt)}$.

  This standardised expression of edits is valid not only for interval-distance edits but also for more traditional types of edits such as balance edits (e.g. $y_k^{(1)} + y_k^{(2)} = y_k^{(3)} \Rightarrow y_k^{(*)} = y_k^{(1)} + y_k^{(2)} - y_k^{(3)} \in [0,0] = \overline{I}_k^{(*)}$) or range edits ($y_k > r \Rightarrow y_k \in [r,\infty)$), to pose a couple of examples.

This is only the top level and further development for a complete normalised specification of an E&I strategy is necessary. For example, it is advisable to develop a controlled vocabulary or a thesaurus of terms so that namespaces for the edit identifiers could be built in a standardised way. The same remark is valid either for the data collection mode (which requires an updateable classification of data collection modes) and for the edit character (hard, soft and any possible new option (Scholtus, 2013)).

So far, only this top level has been partially developed at Statistics Spain, which however allows the different units to identify their labour. IT personnel can recognise the ingredients to program the data collection software; data collection and data editing clerks have information enough either for the validation of the questionnaire values or for the possible respondent recontact; statisticians conducting the survey can use the parameters together with both responded and validated values to perform any kind of further analysis as well as monitoring and fine-tuning the strategy, and finally methodologists can refine the determination of the parameters of each edit without any further change on this scheme.

### 4.1 The unit selection functions UnitSel.StOpt

These are the functions obtained under the stochastic optimization approach to selective editing (Arbués et al., 2012, 2013). For their application in production we only need the optimal Lagrange multipliers $\lambda_q^*$ for each variable $y^{(q)}$ under editing control. Thus, once the values $y_k^{(q,obs)}$ have been collected, we can compute the measurement error conditional moments $M_{kk}^{(q)}$ and construct the global score $S_k = \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)}$. Finally if $S_k > 1$, then unit $k$ is selected and vice versa.

From the preceding paragraph we can easily deduce the descriptive elements of these functions. Firstly, the input data for each unit $k$ is the collected questionnaire with the specific variable values $y_k^{(q,obs)}$ under editing control. If these variables are derived from primary variables, then we assume that the input data contain also the derived values, that is, we compute them if necessary.

Secondly, as input parameters we need both the Lagrange multipliers $\lambda_q^*$ (which are common for all units $k$, but change from one time period to the next) and the function to compute the conditional moments $M_{kk}^{(q)}$ out of the collected values $y_k^{(q,obs)}$ and any other parameters. The computation of both the Lagrange multipliers and the conditional moments is implemented as independent functions. In the former case, only its output data is used in UnitSel.StOpt as input data parameters. In the latter case the moment computation function itself is used as an input function parameter in UnitSel.StOpt. We proceed this way to reach *referential transparency* (see e.g. van Roy and Haridi (2004)) so that if a change in the computation of $\lambda_q^*$ and/or $M_{kk}^{(q)}$ is introduced, the higher-level function UnitSel.StOpt does not need to be modified and just will incorporate the improved procedures. In particular, we are actually using the definition of $M_{kk}^{(q)}$ derived from the observation-prediction model for continuous variables under the linear loss function (see Arbués et al. (2013)). In turn, as already pointed out, the computation of the Lagrange multipliers is still under study (Salvador and Salgado, 2014).

Thirdly, as output data we obtain a binary variable $I_k^{\mathsf{e}} \in \{0,1\}$ flagging whether the unit $k$ has been selected or not. Fourthly, as a metric of the process we keep the weighted local scores $s_k^{(q)} = \lambda_q^* \cdot M_{kk}^{(q)}$ and the global score $S_k$.

Finally as a description of the process we can specify the following steps:

1. Compute the moments $M_{kk}^{(q)}$ out of the input data and using the input function parameter.
2. Compute the weighted local scores $s_k^{(q)} = \lambda_q^* \cdot M_{kk}^{(q)}$ out of the previously computed moments and the input data parameters.
3. Compute the global score $S_k = \sum_{q=1}^{Q} s_k^{(q)}$.
4. If $S_k > 1$, then we set $I_k^{\mathsf{e}} = 1$; otherwise, we set $I_k^{\mathsf{e}} = 0$.

## 4.2 Interval-distance functions based on time series functions

While a numeric algorithm for the computation of the optimal Lagrange multipliers is under construction, we make use of interval-distance edits based on time series. The main idea is to exploit historical data of the survey using time series techniques to assign an interval $I_k^{(q)}$, a distance type $d_k^{(q)}$ and a threshold value $t_k^{(q)}$ to each unit $k$ and the variable $y^{(q)}$ under control. Then, if $d_k^{(q)}(y_k^{(q,obs)}, I_k^{(q)}) > t_k^{(q)}$, unit $k$ is selected.

As input data for each unit $k$ we take the collected value $y_k^{(q,obs)}$ of the specific variable under editing control. If this variable is derived from primary variables, then we assume that the input data contain also the derived value, that is, we compute it if necessary.

Secondly, as input data parameters we need the interval $I_k^{(q)}$, the distance type $d_k^{(q)}$ and the threshold value $t_k^{(q)}$. Again, the computation of these quantities is implemented as independent functions to reach referential transparency. See López-Ureña et al. (2014) for a concrete example of computation of these values based on time series. As input function parameter we make use of the geometric distance $\mathsf{d}_2$ between a point value and an interval.

Thirdly, as output data we obtain a binary variable $I_k^{\mathsf{e}} \in \{0,1\}$ flagging whether the unit $k$ has been selected or not. Fourthly, as a metric of the process we keep the synthetic interval $\vec{I}_k^{(q)}$ (see below) and the geometric distance $d_k^{\mathsf{e}}$ between the value $y_k^{(q,obs)}$ and the synthetic interval $\vec{I}_k^{(q)}$.

Finally as a description of the process we can specify the following steps:

1. Compute the synthetic interval $\vec{I}_k^{(q)}$ out of the input data parameters as follows.
   - If $d_k^{(q)} = 1$, then $\vec{I}_k^{(q)} = I_k^{(q)}$. In these cases, the thresholds are irrelevant.
   - If $d_k^{(q)} = 2$, then $\vec{I}_k^{(q)} = [l_k^{(q)} - t_k^{(q)}, u_k^{(q)} + t_k^{(q)}]$.
   - If $d_k^{(q)} = 3$, then $\vec{I}_k^{(q)} = [l_k^{(q)} - t_k^{(q)} \cdot (u_k^{(q)} - l_k^{(q)}), u_k^{(q)} + t_k^{(q)} \cdot (u_k^{(q)} - l_k^{(q)})]$.

2. Compute the geometric distance $d_k^{\mathsf{e}} = \mathsf{d}_2(y_k^{(q,obs)}, \overline{I}_k^{(q)})$ for the variable $y^{(q)}$ in the input data using the input function parameter $\mathsf{d}_2$.

3. If $d_k^{(q)} > 0$, then we set $I_k^{\mathsf{e}} = 1$; otherwise, we set $I_k^{\mathsf{e}} = 0$.

### 4.3 The unit selection functions UnitSel.ComOpt

Finally, the selection of units exploiting the cross-sectional information under the combinatorial optimization approach is implemented in the UnitSel.ComOpt functions. As pointed out above, in practice we find it more convenient to use a prioritization of units rather than a direct selection (see Arbués et al. (2013) for a wider discussion). This prioritization is carried out in each disjoint domain $s_i$ of the sample. The partition $s = \cup_i s_i$ is chosen as part of the design of the E&I strategy, thus it will work as an input in the execution of the function.

Once we have the prioritization of units within the domain $s_i \ni k$, given the number of units to edit $N_i^{ed}$ within this domain, unit $k$ will be selected if its position $p_{ki}$ in the prioritization satisfies $p_{ki} \leq N_i^{ed}$. The computation of $N_i^{ed}$ and the prioritization of units are implemented as independent functions to achieve referential transparency. The former function implements an allocation algorithm such as that depicted by López-Ureña et al. (2014). The latter implements the combinatorial optimization approach, thus it contains the specification of the conditional moment computation and of an infeasibility function (see López-Ureña et al. (2014) for details).

Thus as input data we take the identifying label $k$ of the unit. As input parameters we take the prioritization of units and the number of units to edit $N_i^{ed}$ within the domain $s_i \ni k$. As output data we obtain a binary variable $I_k^{\mathsf{e}} \in \{0,1\}$ flagging whether the unit $k$ has been selected or not. As a metric of the process we keep the ordinal position $p_{ki}$ and $N_i^{ed}$.

Finally as a description of the process we can specify the following steps:

1. If $p_{ki} \leq N_i^{ed}$, then we set $I_k^{\mathsf{e}} = 1$; otherwise, we set $I_k^{\mathsf{e}} = 0$.

## 5 Conclusions

Upon the need for streamlining the statistical production process we draw as main conclusions from our experience in the standardisation of E&I strategies across surveys at Statistics Spain the following.

Firstly, the notion of statistical production function seems to us crucial to bring together the top-down view of standards such as the GSBPM and the GSIM and the day-to-day tasks conforming this process. This notion does not only provide a neat structure for the production process but also, under an appropriate hierarchy of functions, makes it possible to bring in referential transparency to the organization of the production process at a statistical office. This reinforces the concept of reusability of the production functions, thus optimising them.

Secondly, the different statistical production processes across different statistical offices should follow the same international standards so that the preceding notion of statistical production function becomes even more useful. Regarding the editing phase within the realm of the European Statistical System, the generic EDIMBUS E&I strategy stands as the first landmark for business statistics.

However, this generic strategy does not make use of the notion of statistical production function and does not include the phase of editing during data collection. We have proposed a slightly extended generic E&I strategy which includes both (see figure 2). Besides, this extended generic strategy identifies each production function by parameterising the whole strategy through the number $n_{col}$ of already collected questionnaires, the number $n_{cyc}^{(k)}$ of editing cycles which unit $k$ has already undergone and the number $n_{cyc}$ of editing cycles which the whole sample has already undergone. This generic strategy also makes explicit whether each function is applied on a single unit or on the whole sample.

Also, the EDIMBUS strategy was proposed for business statistics, i.e. to edit and impute quantitative variables. Our proposed extension is also valid for household surveys provided that adequate unit selection functions are constructed. This would imply that selective and macro editing could be applied also to qualitative variables. In this line of thoughts, the optimization approach to selective editing (Arbués et al., 2013), which makes use of statistical models, appears as a versatile tool. By the current analysis under work we can claim that the use of logistic models offers this theoretical possibility. However more research is needed to adapt this theoretical possibility to actual production conditions.

Next, the preceding proposed generic strategy is only the very first step in the complete standardisation of the editing phase. Most work remains to be done. Partially this is intended to be so. We have taken just a few minimal decisions regarding the top-level functions. Proceeding like this will allow us to adapt our production process to future international standards about the definition of statistical production functions and their flow (see Informal Task Force on Metadata Flows (2013) for a first proposal to conjugate both the GSBPM and the GSIM).

As a consequence of this whole approach, the need appears to construct a library of reusable and exchangeable production functions for the different steps of the statistical production process and of its editing phase, in particular. In this sense, at Statistics Spain we have concentrated upon the construction of unit selection functions, so that the resource consumption is optimized. Three of these functions for quantitative variables have been presented in this paper at their top descriptive level (methodological details can be found in (Arbués et al., 2013; López-Ureña et al., 2013, 2014)).

From a more general viewpoint, we are aware of the increasing degree of integration of functions in the statistical production process. Thus, despite the fact that we have provided just a high-level description of the generic extended strategy and of the unit selection functions, there remains to integrate them into a higher level which includes other aspects of the statistical production as the data entry into the system and the design of the data collection instrument. For example, the generic function EditColl in figure 2 is taking into account only the editing aspects of the production process but it is equally important how this function is implemented

in conjunction with the data entry into the system. Thus, a higher-level function is required to aggregate and describe all these complementary aspects.

In the other extreme, the generic extended strategy in figure 2 is posed as a model. That is, in more realistic conditions it could be necessary to introduce modifications in the strategy to accommodate the peculiarities of the actual production process. For example, consider a short-term business statistics with a tight dissemination calendar. Consider a situation in which a highly influential unit responds late with very short time for the survey conductor to apply the same sequence of functions after EditColl as in the rest of units. This suggests the introduction of a new parameter regarding the calendar time so that the appropriate choice of functions for each unit at each instant of time can be taken. In this sense, the generic extended strategy is proposed with versatility enough as to accommodate this sort of contingencies.

To wrap up, we support the view that the standardisation of the statistical production process and in particular of the editing phase is more efficiently undergone using the notion of statistical production function, which in the appropriate workflow instantiates an E&I strategy. Besides, the integration of different production functions points towards the inclusion of editing during collection in the design these strategies. We propose an extension of the generic EDIMBUS strategy to include this phase and its formulation in terms of production functions. Awaiting an international standard for the expression of statistical production functions, we have developed just top-level descriptions and, so far, we have concentrated on unit selection functions to optimize resource consumption. More work is under way at Statistics Spain to consolidate this approach.

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

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

Camstra, A. and Renssen, R. Standard process steps based on standard methods as part of the business architecture. In Camstra, A. and Renssen, R. (eds.), *Proc. 58th World Statistical Congress* 2011, Session STS044, pp. 3061–3070.

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

de Waal, T. (2009). Statistical data editing. In Pfefferman, D. and Rao, C.R. (eds.), *Sample Surveys: Design, Methods and Applications*, pp. 187–214. North Holland, Amsterdam.

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

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.

Granquist, L. (1997). On the current best methods document: edit efficiciently. *UNECE Work Session on Statistical Data Editing*, WP**30**.

López-Ureña, R., Mancebo, M., Rama, S., and Salgado, D. 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 April, 2013.

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

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.

Informal Task Force on Metadata Flows (2013). Metadata flows in the GSBPM. *UNECE Work Session on Statistical Metadata*, WP**22**.

Pannekoek, J., Scholtus, S., and der Loo, M. V. (2013). Automated and manual data editing: a view on process design and methodology. *Journal of Official Statistics* **29**, 511–537.

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

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

Scholtus, S. (2013). Automatic editing with hard and soft edits. *Survey Methodology* **39**, 59–89.

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

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

van Roy, P. and Haridi, S. (2004). *Concepts, Techniques, and Models of Computer Programming*. MIT Press.