A modern vision

of

official statistical production[1]

[1]This work has been awarded the 2016 edition of the INE, Eduardo García España Prize.

A modern vision of official statistical production

David Salgado
Dept. Methodology and Development of Statistical Production
Statistics Spain (INE)
Paseo de la Castellana, 183
28046 Madrid (Spain)

Abstraction is real, probably more real than nature.

J. Albers.

No sólo le costaba comprender que el símbolo genérico "perro" abarcara tantos individuos dispares de diversos tamaños y diversa forma; le molestaba que el perro de las tres y catorce (visto de perfil) tuviera el mismo nombre que el perro de las tres y cuarto (visto de frente).

J.L. Borges en *Funes, el memorioso*.

Abstract

This work is devoted to defend the claim that the modernisation and industrialisation of official statistical production needs a unified combination of statistics and computer science in its very principles. We illustrate our vision with concrete proposals under current implementation at Statistics Spain. Following a bottom-up approach we give a precise formulation of the estimation problem in a finite population, which by using functional modularity principles has allowed us to propose a methodological classification of level-3 production tasks within the Generic Statistical Business Process Model. Additionally, in the same spirit we show our attempts to industrialise the statistical data editing phase by carefully combining rigorous statistical methodology proposals with a light-weight object-oriented software implementation. Finally, we argue that the new sources of information for official statistics will underline the need for this unified combination.

Contents

[1] As of uploading this work to the web, the actual data model implemented has evolved to a more efficient version according to joint work with S. Saldaña, E. Esteban, J.M. Bercebal and M.R. Díaz-Santos. Details will be published elsewhere.

1 Introduction

During the past two decades official statistical production has been undergoing an internationally driven process of modernisation and industrialisation. In this regard the most distinguishable initiative is constituted by the activities of the UNECE High-Level Group of Modernisation of Official Statistics (HLG-MOS) [1] through all their committees. In particular, we recognize as landmarks the Generic Statistical Business Process Model (GSBPM) [2], the Generic Statistical Information Model (GSIM) [3], and the Common Statistical Production Architecture (CSPA) [4]. Others are the new Generic Activity Model for Statistical Organizations (GAMSO) [5], the sets of Generic Statistical Data Editing Models (GSDEMs) [6], and the Big Data project [7].

Some complementary initiatives within the European Statistical System (ESS) are the recommended practices for editing and imputation in cross-sectional business surveys (EDIMBUS manual) [8] (being the seed for the GSDEMs), the European Statistics Code of Practice [9] with the related Quality Assurance Framework of the ESS (ESS QAF) [10], the Euro-SDMX Metadata Structure (ESMS) [11] for the dissemination of reference metadata, and the Validation and Transformation Language (VTL) [12] as a standard language to express data editing validation rules, to name a few.

All these are measures to industrialise the statistical production process proposing standard tools for the many aspects of this process. By and large, all of them follow a top-down approach by which generic proposals are made not taking into account specific methodological details of the production. As an immediate positive consequence, statistical offices can find a straightforward adaptation of these standards to their particular processes and statistical production is more undemandingly comparable in the international realm and thus susceptible of standardisation to a certain degree.

In this paper we want to offer a complementary vision of the statistical production based on a bottom-up approach by taking into account methodological aspects of the process. We want to underline that the **unified combination of statistics and computer science** must be in the **core** of these attempts to industrialise official statistical production. The following vision is intended to illustrate how this combination allows us to firmly root modernising initiatives upon very concrete statistical and computer science principles.

Recognizing a noticeable distinction between scientific and technological approaches to data analysis [13], our spirit is to select and to fuse both kinds of principles to propose a firm theoretical basis of official statistical production which allows us to identify scientific core elements of the process detaching them from those other elements subjected to business organization and purely technological decisions. Rather than advocating an alternative to the former initiatives, this is intended to offer a complementary view hopefully providing an even stronger defense of the adoption of these international measures.

Most of the work empirically illustrating our vision has been and are being done in collaboration with many other colleagues at Statistics Spain, whom are duly credited along the paper citing their concrete contributions. Those aspects of this work originally proposed by the author, especially those not yet published in journals nor in international or national meetings, have been detailed in different appendices. This is intended to let the central part of the paper show our main message: **the modernisation and industrialisation of official statistical production needs a unified combination of statistics and computer science in its very principles**. More technical details intended to rigorously illustrate our vision are thus relegated to the appendices for clarity's sake of the discourse.

The paper is organised as follows. In section 2 we include a precise formulation of the estimation problem in a finite population. In section 3 we argue about the complexity of

official statistical production systems, motivating thus the usage of computer system design principles to cope with it. These principles are briefly introduced and used for classifying production tasks in section 4, which also illustrates its application to the process metadata standard developed at Statistics Spain. In section 5 we include our proposals to industrialise the statistical data editing phase currently under implementation at Statistics Spain. These proposals illustrate the combination of statistical and computer science principles. In section 6 we argue that the new sources of information for official statistics will clearly demand this combination. In section 7 we include some closing conclusions. All technical details of novel proposals not yet published elsewhere have been relegated to the appendices A, B, C, D, and E in order to make the message as clear as possible.

2 A detailed definition of the estimation problem in a finite population

In a statistical office there exists a number of estimation problems with diverse characteristics to be solved and disseminated as official statistical products. The most prominent and most mathematically grounded one is the estimation of a given quantity of interest for a given finite population of people, enterprises, establishments, etc. We shall call this the estimation problem in a finite population, which we shall define in a detailed form in this section.

But this is not the only one. National accounts are also usually under the responsibility of statistical offices. Demographic figures are typically within national statistical plans too and computed with their own methodology. There are also a number of statistics compiled not as an estimation but as a thorough counting exercise (causes of death, ...). In addition, occasionally some statistical offices are in charge of election registers, official municipal population registers, and other administrative registers.

Herein we will concentrate only on the estimation problem in a finite population. We will provide a detailed definition of this problem with three nested layers of complexity. The analysis carried out throughout this work revolves exclusively around this problem. The problem is formulated in abstract terms so as to embrace as many situations in practice as possible.

2.1 Layer 1: the point and interval estimation problem in a finite population

This is esentially the problem dealt with in survey sampling textbooks [14–16]. Paradoxically, they do not usually include a fully-fledged detailed definition, although most of them do share common core elements.

Definition 2.1 (**Point and interval estimation problem in a finite population**)**.** Let U be a finite population of known size N, represented by frame populations $U_{F_1}, \ldots, U_{F_M}$ of size $N_{F_1}, \ldots, N_{F_M}$, respectively, whose units u_k are associated to object and auxiliary variables $(\mathbf{y}_k, \mathbf{x}_k)$, respectively. Let $\mathbf{F}_U = \mathbf{f}(\mathbf{y}_1, \ldots, \mathbf{y}_N; \mathbf{x}_1, \ldots, \mathbf{x}_N)$ be population aggregates and $c : \mathcal{P}\left(\cup_{m=1}^{M} U_{F_m}\right) \to \mathbf{R}$ be a cost function. Let $c_0 > 0$. The objective of the point and interval estimation problem in a finite population is to provide accurate point and interval estimations of the population aggregates $\mathbf{F}_U$ in a partition $\mathcal{I}$ (and usually refinements thereof) of the population $U = \cup_{i \in \mathcal{I}} U_i$ by carrying out an (imperfect) measurement process of the object variables $\mathbf{y}$ over a sample $s \subset \cup_{m=1}^{M} U_{F_m}$ of units under the restriction $c(s) \leq c_0$. The estimates to be disseminated must satisfy agreed statistical disclosure control requirements. □

This definition portrays a number of elements which deserve further attention for its rel-

evance in the construction of the problem. Firstly by *finite population*[2] U we understand a set of a **finite** and **known** number N of **identifiable** units u_k. This concept makes the difference with that of estimation in classical inference problems[3]. Notice the three key features [17]: finite size, known size, and unit identifiability. These properties allow us to denote the population units just by the simple label k.

The concept of population of analysis is an abstract concept, especially in comparison with that of frame population. A *frame population* U_{F_m} is a physical or electronic enumeration of units satisfying those criteria to be considered elements of the finite population U. In practice, most statistical operations make use of one single frame population, but it is perfectly possible to make use of several of them for a given population of analysis [18]. The generation and maintenance of up-to-date frame populations is a key activity in a statistical office with specific methodological procedures.

Variables associated to each unit k alternatively belong to two main categories. Object variables $\mathbf{y}$ are those whose values are object of the analysis and they are only known after executing the measurement process over the sampled units. On the contrary, auxiliary variable $\mathbf{x}$ have values known for all frame population units before the measurement process is executed. It must be clear that in practice (i) it is possible to measure again these variables $\mathbf{x}$ when measuring the variables $\mathbf{y}$ and (ii) it is possible that for some auxiliary variables we only know their aggregate values $\sum_{k\in U_D} \mathbf{x}_k$ for different domains $U_D \subset U$.

The *population aggregates* $\mathbf{F}_U$ are functions of the object and/or auxiliary variables. In practice they are usually function of totals $\mathbf{F}_U = \mathbf{f}\left(\sum_{k\in U} \mathbf{y}_k; \sum_{k\in U} \mathbf{x}_k\right)$. The usual macroeconomic indices lie within this category. It is important to notice that they are never parameters of a probability distribution, as in classical inference problems [19], hence the term *aggregate* instead of *parameter*. Moreover, we underline the multivariate character of this estimation problem: we are estimating several population aggregates, not just one (hence the boldface in $\mathbf{F}_U$ and $\mathbf{f}$).

Estimations must be *point* and *interval* estimations. This obliges the statistical office to produce both estimates for population aggregates and a measure of their precision. Indeed, in statistical rigour, this measure is as important as the population aggregate estimates themselves (anyone can produce an estimate, only a few can make it accurate).

Although we explicitly mention the *sample* s in definition 2.1, notice that we do not claim that the sample s must be selected according to a probabilistic sampling design $p(\cdot)$. As a matter of fact, in practice many statistical operations do not use probabilistic designs but *cut-off* sampling designs. In any case, the sample selection is a central element of the solution to this problem.

Let us underline an essential difference with the estimation problem in classic inference. In the estimation problem in a finite population **there is no element of probability theory** in its definition; in classical inference the concept of population itself inherently incorporates the notion of probability distribution (see e.g. [19]). This difference was already pointed out in [20]. Definition 2.1 does not contain an implicit approach to sample selection, which can range from probabilistic sampling designs to model-based techniques.

So far, we have commented on usual elements of production. But definition 2.1 also include unconventional ones. First, we make explicit the *multivariate* character of the estimation problem in the sense that estimates are required not only for the whole population U, but usually for a great number of domains $U_d \subset U$ generally in a nested hierarchy. This is usually the basis for a further implicit requirement of numerical consistency of all estimates,

[2]Often also *population of analysis*, *object population*, or simply *population*.
[3]Sometimes also misleadingly referred to as *estimation in an infinite population*.

which is behind the multipurpose property of sampling weights [21]. In practice, there exist two types of domains: planned and unplanned [22]. The former are basically population partitions known beforehand and taken into account in the design phase of the solution to the problem. The latter are domain estimates about which users demand information about. This occasionally drives the statistical office towards the problem of small domain estimation [23].

The second novel element is the *measurement process* of the object variables. This explicit mention wants to underline that, given the imperfect nature of this process, the consideration of non-sampling errors is absolutely essential to reach a satisfactory solution [24]. As a matter of fact, this imperfection is the common link among this kind of errors. Non-response is but the impossibility to carry out the measurement process. Frame coverage errors arise due to an imperfect measurement process over auxiliary variables in the frame population. Measurement errors (i.e. informed values differing from true values $\mathbf{y}_k^{obs} \neq \mathbf{y}_k^0$) are the direct consequence of the imperfect measurement over object variables.

The measurement process is the core of one of the most resource-consuming and management-demanding production phases, namely data collection. Also, some theoretical digressions arise when considering the operative consequences of data collection. For instance, complementary notions of statistical units appear when considering those population entities (people, households, enterprises, establishments, etc.) providing the values of the object variables (the so-called *observation* units in contrast with the *analysis* units). As an illustration notice the difference in price statistics between product prices themselves and the establishments informing about these prices.

The third novel element in definition 2.1 is *cost*. Cost must be understood in a generalized sense involving financial and human resources, response burden, timeliness, punctuality, ... The inclusion of cost in the definition of the problem, in our view, is fundamental not only from a practical standpoint but especially from the purely mathematical point of view. Without cost the mathematical solution is trivial: make a census[4]. Thus, without cost, the mathematical problem is meaningless. Cost arises in different phases of the solution to the problem: sample selection, statistical data editing, etc.

Finally, the last novel element in definition 2.1 is the *statistical disclosure control*, which must be dealt with in both the legally normative and purely methodological sides. This inclusion arises from the relevance of statistical dissemination in official statistics. Probably, in other statistical production realms, this requirement is not necessary, but in official statistical production is fundamental.

2.2 Layer 2: longitudinal extension of the problem

Definition 2.1 does not include the time dimension of the problem. However, in practice, populations evolve with time and statistical offices must monitor this evolution.

In this second layer of the complete definition we explicitly include the time variable t in the former elements of definition 2.1. A priori this does not seem difficult but it involves certain complexities needing explicit formulation not only in practice but also theoretically.

When the time variable t is introduced, problems regarding time scales arise. To begin with, if we consider the finite population U_t as it evolves in time, it is clear that estimates are needed not for particular instances of time t_i but for extended periods of time. Let us consider Labour Force Surveys (LFSs) as an illustration. Estimates are usually referred

[4]Notice that a census is a mathematical solution, never a statistical nor practical solution. Indeed taking into account accuracy as the most primitive form of cost, a census can be more inaccurate than a sample [25], not to mention other cost issues such as timeliness, cost-effectiveness, response burden, ...

to natural year terms or months, but within these periods, the population of analysis evolves.

How should estimates referred to extended time periods be understood? E.g. how should an unemployment rate along a term be understood? Firstly we recognize two relevant time scales associated to problem 2.1. On the one hand, we denote by τ_U the time scale associated to the natural evolution of the population of analysis U_t. On the other hand, we denote by τ_D the time scale given by the time period of dissemination (the reference time period).

In terms of these time scales we can formalise *coarse-graining* procedures of the concepts of population and of variable. What should we understand by the population $U_{[t_i,t_{i+1}]}$ for an extended time period $[t_i, t_{i+1}]$ (say, a term or a month)? What should we understand in the case of variable values $y_{k[t_i,t_{i+1}]}$ for each unit k?

By coarse-graining procedure we mean a rigorous mathematical formulation by which we obtain a synthetic population $U_{[t_i,t_{i+1}]}$ out of all populations U_t for each time instant $t \in [t_i, t_{i+1}]$. Similarly for variable values $y_{k[t_i,t_{i+1}]}$. With abuse of notation we denote both cases as $U_{[t_i,t_{i+1}]} = g_{[t_i,t_{i+1}]}(U_t)$ and $y_{k[t_i,t_{i+1}]} = g_{[t_i,t_{i+1}]}(y_{kt})$.

To ease the notation, it is reasonable to expect that practical coarse-graining procedures do not depend on both initial and final time instances, but only on the duration of the interval $\tau = t_{i+1} - t_i$, i.e. on the time scale τ of the period. Thus we denote by g_τ the coarse-graining procedure over the scale τ. For producing estimates we will typically need populations and variable values coarse-grained over the dissemination time scale τ_D.

Let us consider some particular cases.

- Case $\tau_D \approx \tau_U$.

  In this case the dissemination time scale coincides with the natural time scale of change of the population. An illustrative example is constituted by monthly business statistics where the business population can be considered constant in one-month periods[5]. The coarse-graining procedure g_{τ_D} is trivial in this case, since $U_{[t_i,t_i+\tau_D]} = U_{t^*}$ where $t^* \in [t_i, t_i + \tau_D]$. Trivially, the size of the coarse-grained population will be given by $N_{[t_i,t_i+\tau_D]} = N_{t^*}$.

  This coarse-graining procedure must also be similarly applied upon the frame populations $U_{F_m t}$ to produce the coarse-grained version $U_{F_m[t_i,t_i+\tau_D]} = U_{F_m t^*}$ and $N_{F_m[t_i,t_i+\tau_D]} = N_{F t^*}$.

  For variables, since the population does not change over the coarse-grained period, their coarse-grained values are trivially given by $y_{k[t_i,t_i+\tau_D]} = y_{kt^*}$, i.e. their constant values over the period.

- Case $\tau_D > \tau_U$.

  This represents the situation in which the population changes faster than the dissemination time scale. For example, in European LFSs the population of unemployed workers changes weekly[6] ($\tau_U = 1$ week) whereas the dissemination is made in a three-month time period ($\tau_D = 1$ term).

  In this case the coarse-graining procedure g_{τ_D} must formalize which units k in every population U_t, $t \in [t_i, t_i + \tau_D]$, constitute the coarse-grained population $U_{[t_i,t_i+\tau_D]}$.

[5] This is open to debate. There may be cases in which the business does change in less than a monthly scale. For the present discussion, for illustrative purposes we are assuming that businesses are "frozen" during each month.

[6] It is weekly because of the very definition itself of employment [26].

The size of $U_{[t_i,t_i+\tau_D]}$ can be expressed in abstract terms as $N_{[t_i,t_i+\tau_D]} = \int_{\tau_D} N_t d\mu\,(N_t)$ and can be straightforwardly computed as the cardinal of $U_{[t_i,t_i+\tau_D]}$.

These coarse-graining procedures must also be similarly applied upon the frame populations $U_{F_m t}$ to produce their coarse-grained counterparts.

For variables, now the population do change over the coarse-grained period, and the coarse-grained values must then be given in abstract terms by $y_{k[t_i,t_i+\tau_D]} = \int_{\tau_D} y_{kt} d\mu(y_{kt})$, where μ denotes an abstract measure expressing how values are synthetised for the whole time period.

This point deserves a further illustration. Let us consider a typical LFS where the variable "employed" $y_{kt} \in \{0,1\}$, $t \in \{\text{week}_j\}$ has a natural time scale of $\tau_U = 1$ week (reference time period). Aggregate estimates are disseminated in a time scale of $\tau_D =$ 1 term. The coarse-grained version of the variable y_{kt} is trivially given by

$$y_{k[t_i,t_{i+1}]} = \sum_{j \in \text{weeks}[t_i,t_{i+1}]} y_{kt_j}.$$

- Case $\tau_D < \tau_U$

  This case is meaningless.

These are not the only time scales associated to the problem which has to be taken into account. We can also consider the time scale τ_M associated to the measurement process upon the statistical units. Let us think of a statistical operation monthly disseminated ($\tau_D = 1$ month) but with a week-long questionnaire ($\tau_M = 1$ week). Let y_k denote the variable "number of open days" for establishment k. Notice that $\tau_M < \tau_D$. What is the (coarse-grained) number of open days for the whole month for each unit k, which indeed is to be used to produce the estimates of interest, when the unit only responds for a week-long period within that month? A first crude choice could be to impute the same pattern all over the month (D days) [27]:

$$y_{k[t_i,t_i+\tau_D]} = \frac{D}{7} \times y_{k\text{week}}.$$

Alternatively we can try a more sophisticated model

$$y_{kt} = y_{k\text{week}} + y^*_{k\text{week1}} + y^*_{k\text{week2}} + y^*_{k\text{week3}},$$

where $y^*_{k\text{weekj}}$ are imputed values for the rest of weeks j within the referenced month not included in the response. These imputed values can be the result of some elaborated statistical models.

All this is condensedly expressed in the compact notation $y_{k[t_i,t_i+\tau_D]} = g_{\tau_D}(y_{kt})$. Once the time dimension is introduced through coarse-graining procedures, a more general definition of the estimation problem in a finite population can be formulated in terms of the coarse-grained elements. Henceforth we shall denote the coarse-grained dissemination periods by t_j, $j = 1, \ldots, J$.

Definition 2.2 (**Longitudinal extension of the estimation problem in a finite population**). Let $\{U_{t_j}\}_{j=1,\ldots,J}$ be a family of finite populations, with natural evolution time scale τ_U, coarse-grained over the dissemination time scale τ_D, with sizes N_{t_j}, respectively, and represented by coarse-grained frame populations $U_{F_1 t_j}$, $\ldots$, $U_{F_M t_j}$ of sizes $N_{F_1 t_j}, \ldots, N_{F_M t_j}$, whose units k are associated with coarse-grained object and auxiliary variable values $(\mathbf{y}_{kt_j}, \mathbf{x}_{kt_j})$, respectively. Let $F_{U_{t_j}} = \mathbf{f}(\mathbf{y}_{1t_j}, \ldots, \mathbf{y}_{Nt_j}; \mathbf{x}_{1t_j}, \ldots, \mathbf{x}_{Nt_j})$ be coarse-grained population aggregates and

$$c_{t_J} : \mathcal{P}\left(\cup_{m=1}^{M} U_{F_m t_J}\right) \to \mathbf{R}$$

be a cost function. Let $\mathbf{c}_0 > 0$. The objective of the point and interval estimation problem in a finite population over time (aka longitudinal extension of the problem) is to provide accurate point and interval estimations of the population aggregates $\mathbf{F}_{U_{t_j}}$ in a partition $\mathcal{I}$ (and usually refinements thereof) of the populations $U_{t_j} = \cup_{i \in \mathcal{I}} U_{it_j}$ by carrying out an (imperfect) measurement process of the object variables $\mathbf{y}_{t_J}$ over a sample $s \subset \cup_{m=1}^{M} \tilde{U}_{F_m t_J}$ of units under the restrictions $c_{t_J}(s) \leq c_0$. The estimates to be disseminated must satisfy agreed statistical disclosure control requirements. □

Notice that this extended definition is similar to definition 2.1 but using coarse-grained elements over the time scale τ_D.

We are assuming that the last time period t_J is the time period under which the current estimates are to be produced. Thus, definition 2.2 establishes that estimations must be computed for all periods $j \leq J$, which implicitly introduces the need to generate and coherently maintain time series of estimates and their coefficients of variation (through introducing late responses and/or using link coefficients, backcasting techniques, index rebasing procedures, ... when necessary). Furthermore, it is usual to disseminate these time series both seasonally and calendar-effect adjusted.

2.3 Layer 3: cross-sectional extension of the problem

A statistical office never produces one single statistics. On the contrary, a number of statistical operations must be produced and disseminated simultaneously for different domains according to a national or international statistical Act. Many of them share common elements of the problem (households, establishments, enterprises, etc.), not to mention human and material resources.

It is natural (even compulsory) to request that resources for production and dissemination must be shared among all statistical operations within a statistical office. Furthermore, in the international realm, there already exist important efforts to share resources across different statistical offices themselves.

This suggests that a new layer of complexity is needed in the definition of the estimation problem:

Definition 2.3 (**Cross-sectional extension of the estimation problem in a finite population**). Let $\mathcal{P}_q$, $q = 1, \ldots, Q$ be longitudinal estimation problems in a finite population, respectively. Let C be a global cost function defined over the set of statistical operations $\{\mathcal{P}_q\}_{q=1,\ldots,Q}$. Let $C_0 > 0$. The objective of the point and interval estimation problem in a finite population across several statistical operations (aka cross-sectional extension of the problem) is either to provide accurate point and interval estimations for each problem $\mathcal{P}_q$ under the restriction $C(\mathcal{P}_1, \ldots, \mathcal{P}_Q) \leq C_0$ or to minimize the global cost $C(\mathcal{P}_1, \ldots, \mathcal{P}_Q)$ provided that all statistical operations $\mathcal{P}_q$ produce accurate point and interval estimations. The estimates to be disseminated across all statistical operations must satisfy agreed statistical disclosure control requirements. □

Again, cost must be understood in a generalized sense involving financial and human resources, response burden, timeliness, punctuality, ... As an illustrative consequence, let us consider response burden, which introduces a particular methodological problem: the sample selection coordination problem (see e.g. [28]).

But consequences are not only methodological. The cross-sectional character introduces very relevant restrictions in the field work (especially in the data collection phase). Important management science issues arise when dealing with the planning, coordination, and usage of both human and material resources, since they have to be reused in different statistical operations.

Notice that the problem has been formulated with two different possible objectives, i.e. either given a fixed budget we concentrate on producing and disseminating accurate estimates or we seek to minimize the production and dissemination global cost provided that estimates are accurate enough. This is a top management decision intricately related with the financial and fund-raising capacity of the organization.

Finally we underline that the problem is not necessarily to be applied as a whole to all the statistical production in an office, which can be divided in sets of operations upon which the most adequate coarse of action in each case shall be taken.

2.4 The problem of official statistical production

As stated, the preceding definitions only focus on the estimation problem in a finite population. We believe that the rest of problems to be solved in a statistical office must be provided with similar definitions and put them all together to formulate the problem of global official statistical production.

In the rest of this work we shall make use only of the above definitions, thus the resulting analysis will be necessarily partial at the global production scale.

3 On the complexity of official statistical production

When one first steps into a statistical office as an official statistician it is struck by the usual statement about how different statistical production and textbook academic statistics are. Apart from other factors, in our view, this is due to the lack of experimental science background in the academic skills of most official statistics professionals, especially mathematicians, statisticians, economists, and econometricians.

Nonetheless, they are completely right. We have a good distance from statistical formulas on a piece of paper or even in the form of simple software prototypes to their industrial usage in a production chain of, say, hundreds of statistical operations. But this should not be surprising. We claim that the missing element in this gap is *complexity*: **a statistical production system is a complex system**.

Although a definitive scientific definition of complexity is extremely difficult, we can easily recognize in statistical production systems several features defining a complex system [29]:

- **Large number of components**: a statistical operation can range from a few tens of object and auxiliary variables for short-term statistics to a few hundreds in the case of structural surveys, not mentioning the several dozens of estimates to be produced accross a great number of population domains or paradata generated during the execution of the process. This must be multiplied by the number of statistical operations under production in a statistical office.
- **Large number of interconnections**: needless to say, most of the different production tasks are intricately interconnected in such a way that a variation in a given task can have unexpected *waterbed* effects in another tasks [30].
- **Many irregularities**: as survey conductors and domain experts rightfully underlined when referring to their own statistical operations, each survey portraits specific characteristics somehow singling them out from the rest. No clear-cut regularity allowing production designers to pose universal rules can be cross-sectionally identified among all statistical operations.
- **A long description**: as a further complication to the preceding feature, protocols, rules, guidelines, or instructions to accomplish the diverse production tasks cannot be described in a homogeneous fashion. A lot of exceptions are indeed the rule.

- **A team of designers, implementers, or maintainers**: not only is it that a high number of different professional profiles ranging from IT experts to statisticians are needed to produce official statistics, but also do they need intensive coordination and communication.

Based on these features, different simple models can be used to justify the so-called *square law of computation* [31], which can be adequately represented by figure 1, in terms of resources and complexity.

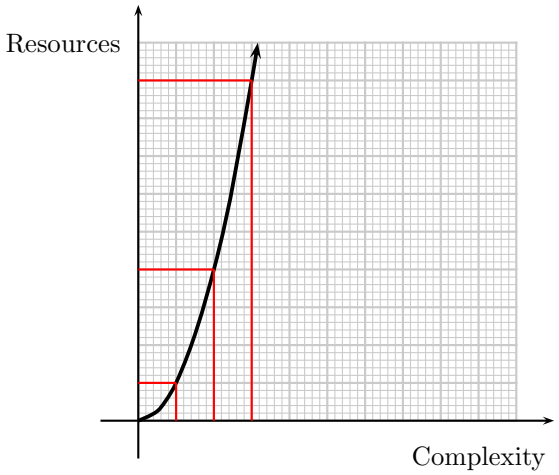

Figure 1: Square law of computation in terms of complexity and resources

A simple model justifying this behaviour can be borrowed from computer software engineering [29]. Just consider the usual conception of the execution of the statistical production process as the downstream sequence of sample selection, data collection, statistical data editing, aggregate and variance estimation, and dissemination. Let N_T denote the number of tasks involved in the execution of this process. Assume that the number of errors to be corrected N_e in the whole process is proportional to the number of involved tasks $N_e \propto N_T$ (the higher the number of tasks, the higher the number of errors to be corrected) and that they are randomly distributed accross the process. Survey conductors execute the process and notice an error, then find and fix the error. Let us also assume that the time t_e it takes to find an error is proportional to the number of involved tasks $t_e \propto N_T$. Then the total time needed to find and fix all errors T_e will be roughly given by

$$\begin{aligned} T_e &\propto N_e \times t_e \\ &\propto N_T^2. \end{aligned}$$

The bottom line of this law is the fact that for each increasing unit of complexity (in arbitrary units; e.g. a new breakdown of estimates, a change of normative regulations, ... a new task in the process, in definitive) requires an increasing amount of resources (see figure 1). It is not **only** a matter of how much resources and how many tasks are to be executed, but also of how they are **related through the production model**.

Indeed, should the production model be kept under the former downstream premises, the increasing demand of information upon statistical offices will eventually collapse the production. Notice that the quadratic behaviour is somehow arbitrary for our argument and any increasing convex curve deduced from those features will support our claim. Complexity is the ultimate reason why resources hypothetically enough to accomplish a given task are not sufficient when that task is part of the production system.

4 A methodological classification of production tasks

4.1 A high-level proposal

Our approach to deal with the complexity of statistical production bears on the reflection that, *since the fabric of statistical production is information, we should use some principles of computer system design to cope with this complexity.* In particular, we claim that *functional modularity* (modularity + abstraction) together with *hierarchy* and *layering* [29] arise as useful principles to structure statistical production in a way to cope with its inherent complexity.

To illustrate how these principles tackle complexity let us elaborate on the preceding simple model underlying the quadratic law (see also [29]). If the production process is divided into M ***independent*** modules so that each error can be immediately categorized within one of the modules, then the total time needed to find and fix all errors T_e can be somehow ameliorated. Assuming for simplicity that each module involves roughly $\frac{N_T}{M}$ tasks, we can write

$$\begin{aligned} T_e &\propto (T_e \text{ for each module }) \times M \\ &\propto \left(\frac{N_T}{M}\right)^2 \times M \\ &\propto \frac{N_T^2}{M}. \end{aligned}$$

Thus, two such modules would entail a reduction up to one half of the original time, three such modules would entail a reduction up to one third, and so on. Notice that the key element is the independent character of the modules so that any error can be straightforwardly classified within one of them. This is achieved not only with modularity *per se* but designing modules with appropriate interfaces, i.e. following the principle of abstraction. This is functional modularity.

Thus, the design of functional modules is key, which implies identifying adequate interfaces, i.e. natural boundaries among the different parts of the production process. We claim that the methodology[7] of statistical production must be the guiding principle to design these modules.

We make the following proposal concentrating upon the definition of the estimation problem in a finite population formulated in section 2. All tasks in the production process will be parameterised according to (i) its methodological group and (ii) its implementation state. We identify the following methodological groups motivated by the preceding formulation of the estimation problem in a finite population:

1. **Problem construction**.- This group embraces all tasks devoted to set up precise identifications of each concrete element of the estimation problem as formulated in its definition, especially the population of analysis (equivalently, the population units), the variables, the population aggregates precisely expressed in terms of the variables, and the dissemination domains, but also the frame population(s), the accuracy restrictions, the limiting costs, and the statistical disclosure control restrictions. Details can be numerous boiling down to legal and budget issues. The contact with stakeholders is compulsory to accomplish an adequate problem construction. The key ingredient in this group is the domain expert knowledge.

2. **Frame construction**.- This group embraces all tasks devoted to the construction and maintenance of population frames (see e.g. [32]). This is usually carried out by collecting

[7] Methodology must not be understood in the usual restricted sense referring to statistical theory (see section 5) but embracing all aspects of statistical production, i.e. both mathematical-statistical and computer science aspects.

a high number of administrative registers of diverse nature (population, citizenship, taxes, ...). Record linkage techniques and thorough data editing (different in objectives to statistical data editing –see below) are the core statistical tools used in this group.

3. **Sample selection**.- This group involves all statistical tools devoted to select the sample, from probabilistic sampling theory [16] to cut-off sampling techniques [33] and other more specific methods such as capture-recapture methods [34].

4. **Data collection**.- This group embraces all tasks devoted to measure the values of object (and possibly some auxiliary) variables [35]. Notice the essential difference of this measurement process with that in the frame construction group. It is important in this stage to design, generate, and feed an adequate system of data collection paradata.

5. **Statistical data editing**.- This group involves all tasks devoted to detect and treat errors in the measurement process, i.e. frame coverage errors, non-response and measurement errors [24]. It embraces both error detection and error treatment with interactive, automatic, selective, and macro editing techniques, as well as imputation techniques [36].

6. **Aggregate and accuracy estimation**.- This group consists of the construction of estimators according to the sample selection and the edited (possibly imputed) variable values [37]. The estimation involves not only population aggregates estimation but also that of their accuracy (typically through coefficients of variation) [38]. Also variation rates in a yearly, quarterly, monthly, ... basis are of usual interest. The main statistical methods include re-weighting techniques and variance estimation methods, but in the case of unplanned domains it also embraces small domain estimation and related techniques [23].

7. **Time series construction**.- This group embraces all tasks devoted to produce up-to-date time series of different estimates. In most cases this involves adding the new estimates to the times series, but from time to time backcasting techniques due to standard classification changes [39] or linking coefficients because of index rebasing must be used. Seasonal and calendar-effects adjustments [40] are also included here.

8. **Statistical dissemination**.- This group embraces all tasks devoted to make resulting estimates reach users. This involves statistical data communication techniques and data visualization methods. In the case of official statistics, techniques for statistical disclosure control [41] are of special relevance.

9. **Process configuration**.- This group embraces all tasks devoted to establish the correct workflows and to prepare and maintain overall human and material resources.

10. **IT architecture configuration**.- This group contains all tasks devoted to the design, development, execution and monitoring of the configuration of the IT architecture supporting all preceding modules.

The second coordinate to categorize a production task is to classify its implementation state as (i) design, (ii) build, (iii) execute, and (iv) monitor (see also [2]). Thus, any production task can be categorized in any of the cells of table 1. We include in this table some possible subgroups for illustration's sake (not complete).

Notice that this is just a high-level classification, since each group can be further decomposed into more detailed subgroups. For example, in the statistical data editing group we can recognize error detection and error treatment tasks, which although intricately related in the execution stage, they are clearly different in the design phase. These subgroups can be further decomposed down to finer methodological methods (e.g. interactive, automatic, selective, and macro editing techniques or editing and imputation strategy workflow design, execution, and monitoring).

	Design	Build	Execute	Monitor
Problem construction				
Frame construction				
Construction by record linkage				
Thorough data editing				
Sample selection				
Probabilistic sample selection				
Non-probabilistic sample selection				
Data collection				
Statistical data editing				
Error detection				
Error treatment				
Estimation				
Aggregate estimation				
Variation rate estimation				
Accuracy estimation				
Times series construction				
Backcasting				
Linking				
Seasonal and calendar-effect adjustment				
Statistical dissemination				
Statistical disclosure control				
Data visualization				
Communication to stakeholders				
Process configuration				
IT architecture configuration				
Databases configuration				

Table 1: Methodological classification of statistical production tasks.

A suspicious reader may enquire about the relationship with the GSBPM, which is the internationally accepted business production model. As stated in section 1, the proposed classification is not an alternative but a complement to the usage of the GSBPM. Table 1 can be viewed as an assistance to identify production tasks within the GSBPM.

For example, all tasks in the group *1. Problem construction* can be categorized in the GSBPM group *1. Specify needs*. Also, all tasks in the design state of any methodological group must be assigned to some of the GSBPM level-2 subprocesses within the GSBPM level-1 Design phase. Similarly for (i) those in the build state with respect to the GSBPM level-1 Build phase, (ii) those in the execution state with respect to GSBPM level-1 Collect, Process, Analyse, and Disseminate phases, and (iii) those in the monitoring state with respect to GSBPM level-1 Evaluate phase.

Occasionally we find GSBPM's methodological sizing somehow hard to recognize and thus categorizing specific production tasks also hard to assign to level-2 subprocesses. The purely methodological classification in table 1 can assist in this goal. Next subsection illustrates this with the specific example of the process metadata standard for the statistical production at Statistics Spain [42].

4.2 Application to the GSBPM-based process metadata standard at Statistics Spain

Recently, Statistics Spain has produced a GSBPM-based standard for the process metadata of its statistical production [42]. The standard is fully based upon the GSBPM incorporating

	Design	Build	Execute	Monitor
Problem construction	1.m.n	3.2.10, 3.2.15	2.1.n, 2.2.n, 2.5.7, 5.6.1	6.5.1, 8.1.n, 8.2.n
Frame construction	2.4.1	3.2.1, 3.5.1	4.1.1	6.5.1, 8.1.n, 8.2.n
Sample selection	2.4.2-2.4.5	3.2.2-3.2.3, 3.5.1	4.1.2	6.5.1, 8.1.n, 8.2.n
Data collection	2.3.n, 2.5.1, 2.5.10-2.5.11	3.1.n, 3.2.4, 3.5.2	4.2.n, 4.3.n, 4.4.n, 5.2.1	5.2.2, 6.5.1, 8.1.n, 8.2.n
Statistical data editing	2.5.2-2.5.4, 2.5.10-2.5.11, 2.5.12	3.2.5-3.2.7, 3.5.3	5.3.1-5.3.2, 6.2.1-6.2.2	5.3.3, 6.5.1, 8.1.n, 8.2.n
Estimation	2.5.8-2.5.11	3.2.11-3.2.15, 3.5.3	5.5.1-5.5.2, 5.7.1-5.7.2, 6.1.3	5.7.3, 6.5.1, 8.1.n, 8.2.n
Times series construction	2.5.13-2.5.14	3.2.16-3.2.17, 3.5.3	6.1.1-6.1.2	6.5.1, 8.1.n, 8.2.n
Statistical dissemination	2.5.15	3.2.18, 3.3.n, 3.5.4	6.1.4-6.1.5, 6.4.n, 7.1.1, 7.2.n, 7.3.n, 7.4.n, 7.5.n	6.5.1, 8.1.n, 8.2.n
Process configuration	2.6.1-2.6.3		3.4.n, 3.5.n, 3.6.n, 3.7.n	6.5.1, 8.1.n, 8.2.n, 8.3.n
IT architecture configuration	2.5.5-2.5.6, 2.6.4-2.6.6	3.2.8-3.2.9, 3.2.19-3.2.20	5.1.n, 5.8.1-5.8.2, 7.1.2-7.1.3	6.5.1, 8.1.n, 8.2.n

Table 2: Methodological classification of tasks of the Spanish GSBPM-based process metadata standard.

the use of the formerly mentioned computer science principles to develop a third level of statistical production tasks within this international business process model. The adaptation of the former computer science principles to the construction of this standard has been done in collaboration with A.I. Sánchez-Luengo, S. Lorenzo, and M.A. Martínez-Vidal and implemented by a large internal working group at Statistics Spain embracing domain experts and experts from data quality, data collection, statistical dissemination, survey sampling, and IT.

Here we offer a complementary view not included in the standard explicitly showing the underlying classification of tasks under the above group categorization. The standard was produced striving for functional modularity in the production process. These underlying methodologically classified functional modules were the basis for identifying GSBPM level-3 subprocessess adapted to Statistics Spain's needs. In table 2 we show the categorization of each task in the standard (not included in its formulation).

5 Industrialising the statistical data editing phase

We will choose the statistical data editing module to illustrate our proposals to industrialise the production process. We want to underline once more the interrelation between both statistical and computer science standpoints: a full methodological proposal must embrace all these aspects.

5.1 Statistical methodology

Statistical data editing is one of the most resource-consuming phases in official statistical production [43]. This drove methodologists some decades ago to develop selective editing techniques [44–48] whose main objective is to rationalize resources and to detect influential errors upon which interactive editing is concentrated in contrast to non-influential errors which can be edited automatically.

The bottom line of these techniques is the assignment of an *item score* $s_k^{(q)}$ to each object variable $y^{(q)}$ for each measured unit k. This score is computed by means of the so-called *local score functions* $s_k^{(q)} = s\left(y_k^{(q,obs)}, \hat{y}_k^{(q)}; \mathbf{x}_k\right)$ upon the observed values $y_k^{(q,obs)}$, the anticipated values $\hat{y}_k^{(q)}$, and possibly some auxiliary variables $\mathbf{x}_k$ (sampling weights, etc.). All item scores are then used in the so-called *global score function* $S_k = S\left(s_k^{(1)}, \ldots, s_k^{(Q)}\right)$ to produce the *unit score* S_k. Those units with unit scores S_k above a chosen threshold t_k will be interactively edited; the rest will be edited by automatic editing techniques.

The choice of anticipated values, local score functions, global score function, and threshold values is made in a purely heuristic basis [36], no rigorous connection to statistical principles is made. However, let us remind one of the virtue to be pursued by official statistics producers: "[...] it seems desirable, to the extent feasible, to avoid estimates or inferences that need to be defended as judgments of the analysts conducting the survey" [49]. Our contribution tries to follow this direction by putting selective editing techniques upon a firm theoretical basis and, in this formalization, to pave the way to extend its application to qualitative and semicontinuous variables.

We will provide a high-level view of the statistical methodology, so that the combination of this methodology and its software implementation can be more straightforwardly appreciated. The starting point of the statistical methodology is the formulation of two general principles [51, 52]:

(i) Editing must **minimise** the amount of resources deployed to **interactive tasks**.

(ii) Data **quality** must be **ensured**.

This drives us to a generic optimization problem. This problem boils down to two extreme cases, namely (i) the longitudinal case, in which only object and auxiliary information of the current time period in course is available for each unit separately (input editing) and (ii) the cross-sectional case, in which all object and auxiliary information of this time period is available for every unit (output editing), since data collection for the whole sample has been accomplished.

5.1.1 The longitudinal case

The longitudinal case reduces to the resolution of a stochastic optimization problem, formerly proposed in [50] not taking into account the sampling design, assuming stationarity for the measurement error generation, and using the principles of duality, interchangeability, and sample average approximation (see [50]). This optimization problem made use of general optimization routines in languages not adopted for production conditions at Statistics Spain. In [52] this approach was generalized in terms of a generic loss function L. To be specific, we will concentrate on the absolute value loss function $L(a, b) = |a - b|$, thus providing a direct connection with the traditional heuristic approach of score functions. The resulting unit score for each unit k in the original longitudinal case was given by

$$S_k = \begin{cases} 1 & \text{if } \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} \leq 1, \\ 0 & \text{if } \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} > 1, \end{cases} \tag{1}$$

where $M_{kk}^{(q)}$ are so-called conditional measurement error moments playing the role of item scores and λ_q^* are Lagrange multipliers associated to the dual optimization problem [52]. The error moments are given by

$$M_{kk}^{(q)} = \mathbb{E}_m \left[\omega_k \cdot \left| \left(y_k^{(q,obs)} - y_k^{(q,0)} \right) \right| \left| \mathbf{Z}_k^{cross} \right. \right],$$

where m stands for the statistical model of the measurement error $\epsilon_k^{(q)} = y_k^{(q,obs)} - y_k^{(q,0)}$, ω_k is the (design) sampling weight, and $\mathbf{Z}_k^{cross}$ denotes the cross-sectional available information about unit k.

It is important to notice that no heuristic choice of anticipated values $\hat{y}_k^{(q)}$, local and global score functions $s(\cdot)$ and $S(\cdot)$, and threshold values t_k is necessary. Everything arises naturally from the choice of a statistical model and a loss function. This is purely statistical language. So far, these are ideas already contained in [52].

We are currently working on a streamlined version of the longitudinal case adapted to standard production conditions [53]:

- The unit score value (1) involves only **linear** global score functions. It is more or less straightforward to generalize this approach to obtain arbitrary Minkowskian global score functions [54]. See appendix A. Unit score values are then given by

$$S_k^{(r)} = \begin{cases} 1 & \text{if } \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)r} \leq 1, \\ 0 & \text{if } \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)r} > 1, \end{cases} \tag{2}$$

  where $r = 1, 2, 3, \ldots$. Notice that $r = 1$ recovers the former case.

- We propose an alternative way to find the Lagrange multipliers (see appendix B). The most controversial assumption in the original derivation in [50] is the stationarity of the measurement error generation mechanism. In the original approach it is assumed that measurement errors in each time period are indeed realizations of the same random variable (the measurement error). This is too crude an assumption for realistic applications and we have dropped it by solving an alternative associated optimization problem. In appendix B we include the theoretical proposal which is currently under analysis using actual survey data. Results will be published elsewhere when the proposal will thoroughly be tested with survey data [53].

Nonetheless, in production we have proposed, analysed with actual survey data, made an experience pilot with one concrete statistical operation, and implemented in standard production conditions for other surveys an alternative approach not using optimization for the longitudinal case [55, 56]. This alternative approach makes use of time series modelling to produce a validation interval for each variable for each statistical unit, and for each time period. It is computationally demanding but effective in the reduction of recontact rates (we obtain up to one half of the original recontact rates and just using simple time series models for predicting each value in our current experience).

In any case, the standard format decided to express these validation intervals, either from score functions or time series modelling techniques, amounts to associating to each statistical unit k (i) an object variable name VarName_k upon which to apply the edit, (ii) a condition $C_k(\mathbf{y}_k, \mathbf{x}_k)$ determining whether to apply or not the edit, (iii) the lower bound b_k of the validation interval, and (iv) the upper bound of the validation interval u_k. For example, the edit with VarName_k = Turnover, $C_k = \{\omega_k > \bar{\omega}\}$, and validation interval $I_k = [b_k, u_k]$ amounts to applying for each unit k to its object variable turnover the edit by firstly checking if the sampling weight ω_k is greater that a fixed amount $\bar{\omega}$ and then, in the positive case, checking if the reported value y_k^{obs} lies within the interval I_k or not (thus flagging the variable). This expression of edits allows us to make a standardized software implementation independent of the nature of the variables, of the concrete computational procedure, and of the statistical operation.

5.1.2 The cross-sectional case

The cross-sectional case arises when object and auxiliary information about all measured units concerning their reported values is available for the current time period in course. This case reduces to the resolution of a combinatorial optimization problem [57, 52] whose solution [58] provides the selection of units to be edited interactively:

$$\begin{array}{ll} [P_{co}\left(\boldsymbol{\eta}, \Omega\right)] & \max \ \mathbf{1}^T \mathbf{r} \\ & \text{s.t.} \quad \mathbf{r}^T M^{(q)} \mathbf{r} \leq \eta_q, \quad q = 1, 2, ..., Q, \\ & \mathbf{r} \in \Omega \subset [0,1]^{\times n}, \end{array}$$

The selection of units will ultimately depend on how tight we decide to be with the presence of non-influential measurement errors in the estimates of each aggregate $Y^{(q)}$, i.e. on the bounds η_q. These bounds can be chosen according to accepted a priori coefficient of variations or as measures of errors relative to estimates in preceding periods.

Nonetheless, we argue that for the output editing field work it is better to have a prioritization of units rather than a selection of units [52]. Should a number of units be selected, either the selection may be short and thus resources are underused, or the selection is too large and resources are not enough. With a prioritization, resources can be more easily optimised.

In [52] an algorithm was proposed for the prioritization based on successive solutions $i = 0, 1, \ldots, n$ of the combinatorial optimization problem $P_{co}\left(\boldsymbol{\eta}_i, \Omega_i\right)$ upon particular choices of the bounds $\boldsymbol{\eta}_i$. In [56] an argument was given for the equivalence between this prioritization algorithm and the score function approach making use of the greedy algorithm to solve P_{co} proposed in [58]. Here, in appendix C we prove this equivalence under any algorithm.

Therefore, we have been able to set a purely statistical non-heuristic basis for local score functions s as measurement error (conditional) moments $s_k = M_{kk} = \mathbb{E}_m \left[\omega_k \cdot \left|y_k^{obs} - y_k^0\right| \middle| \mathbf{Z}_k^{cross}\right]$ for each variable y within a measurement error model m conditional on the available cross-sectional information for that unit. Complementarily, we have also proved that a global score function S amounts to choosing a prioritization of units according to a sequence of successively tighter bounds upon the allowed non-influential measurement errors in the aggregate estimates.

Finally, the error moments M_{kk} are computed according to so-called observation-prediction models [52]. For example, for continuous variables the observation-prediction model given by

$$y_k^{obs} = y_k^{(0)} + \epsilon_k^{obs} \qquad\qquad y_k^{(0)} = \hat{y}_k + \epsilon_k^{pred},$$

where $y_k^{(0)}$ are true values and $\hat{y}_k$ are predicted values according to an auxiliary model m^*. The model satisfies the following specifications:

1. $\epsilon_k^{obs} = \delta_k^{obs} \cdot e_k$;
2. $e_k \simeq Be(p_k)$, where $p_k \in (0, 1)$;
3. $\left(\epsilon_k^{pred}, \delta_k^{obs}\right) \simeq N\left(\mathbf{0}, \begin{pmatrix} \nu_k^2 & \rho_k \sigma_k \nu_k \\ \rho_k \sigma_k \nu_k & \sigma_k^2 \end{pmatrix}\right)$.
4. ϵ_k^{pred}, δ_k^{obs}, and e_k are mutually independent of $\mathbf{Z}_k$.
5. e_k is independent of ϵ_k^{pred} and δ_k^{obs}.

Under these assumptions, the error moments are given in closed form by:

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

$$\xrightarrow{\frac{|y_k^{obs} - \hat{y}_k|}{\nu_k} \to \infty} \quad \omega_k |y_k^{obs} - \hat{y}_k|, \tag{3b}$$

where we have denoted

$$\zeta_k(x) = \frac{1}{1 + \frac{1-p_k}{p_k}\left(\frac{\nu_k^2}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{\sigma_k^2 + 2\rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}x^2\right)}.$$

Notice that the usual heuristic local score function [36] given by (3b) is obtained in the limit case where the prediction error is negligible. The predicted values $\hat{y}_k$ are computed and the model parameters $\nu_k, \sigma_k, \rho_k, p_k$ are estimated using the historic set of raw and edited values y_k^{obs}, y_k^{ed}.

Notice that this model is not the only option which can be used, although it is very general. In section 5.4 we explore a theoretical proposal of an observation-prediction model for a binary variable.

5.2 Software implementation[8]

As stated, a methodological proposal for statistical production is definitely incomplete if computer science considerations about its implementation are not dealt with and even tested and/or prototyped with real (not simulated) data.

To implement the preceding theoretical proposals the first requirement is to have access to the historic set of raw and edited values of the statistical operation at stake. The larger this set, the higher the number of theoretical (thus also practical) possibilities to compute and estimate the involved quantities ($\hat{y}_k$, ν_k, σ_k, ρ_k, p_k). For example, should we have at least 36 months of data in a monthly survey we could possibly make use of ARIMA modelling techniques to produce predictions and estimates of these quantities (see e.g. [56]).

Furthermore, if these techniques are to be applied in a standard way on many surveys at a statistical office, then it is necessary that the information system of the statistical office provides access to all these historic sets of raw and edited data, preferentially in a standard format.

This was not the case of Statistics Spain and a repository of files with this content was immediately put in place. The original proposal currently used in production was made by J.M. Bercebal. The repository is essentially made up of different files: one per reference time period and editing degree (raw, input-edited, completely edited) and one per survey containing a dictionary of data (variable name, type, length, role in the repository, description). All file names are standardised.

All data files are given a key-value pair structure [59] in which the key is compound of the values of different qualifiers (statistical units ID variables, statistical variable geographical qualifiers, ...). Every variable composing this key is recorded in its corresponding column in the data file, so that each file contains as many columns as the total number of variables

[8] As of uploading this work to the web, the actual data model implemented has evolved to a more efficient version according to joint work with S. Saldaña, E. Esteban, J.M. Bercebal and M.R. Díaz-Santos. Details will be published elsewhere.

composing the key and another column for the values.

This structure is very adequate for querying and processing data because it is very close to Codd's 3rd normal form [60]. However, when storing a large number of both microdata and paradata, this proposal needs a step further to prevent files from growing in their number of columns thus betraying the original philosophy of the key-value pair structure. We have recently made the following proposal by taking advantage of the dictionary of data. We propose to compose the key under the syntactic rule roughly expressed by

$$VarName : [VarNameValue]\_[QualName1] : [QualName1Value]\_\ldots\_[QualNameN] : [QualNameNValue], \tag{4}$$

where as many N qualifiers as necessary are to be included independently for each variable.

Now only two columns are necessary for each data file: one for the compound key and one for the data value. The dictionary of data allows us to build a lexer [61] to go from the alphanumeric compound key expressed by (4) to the column-based compound key suitable for processing and querying.

This proposal has been implemented by designing S4 classes in R according to the class diagram in figure 2. Their attributes and methods are depicted in figure 8 in appendix D. The coding of these classes has been carried out by E. Esteban, S. Saldaña, P. García-Segador, and the present author in the form of R packages which are publicly available at [62–64].

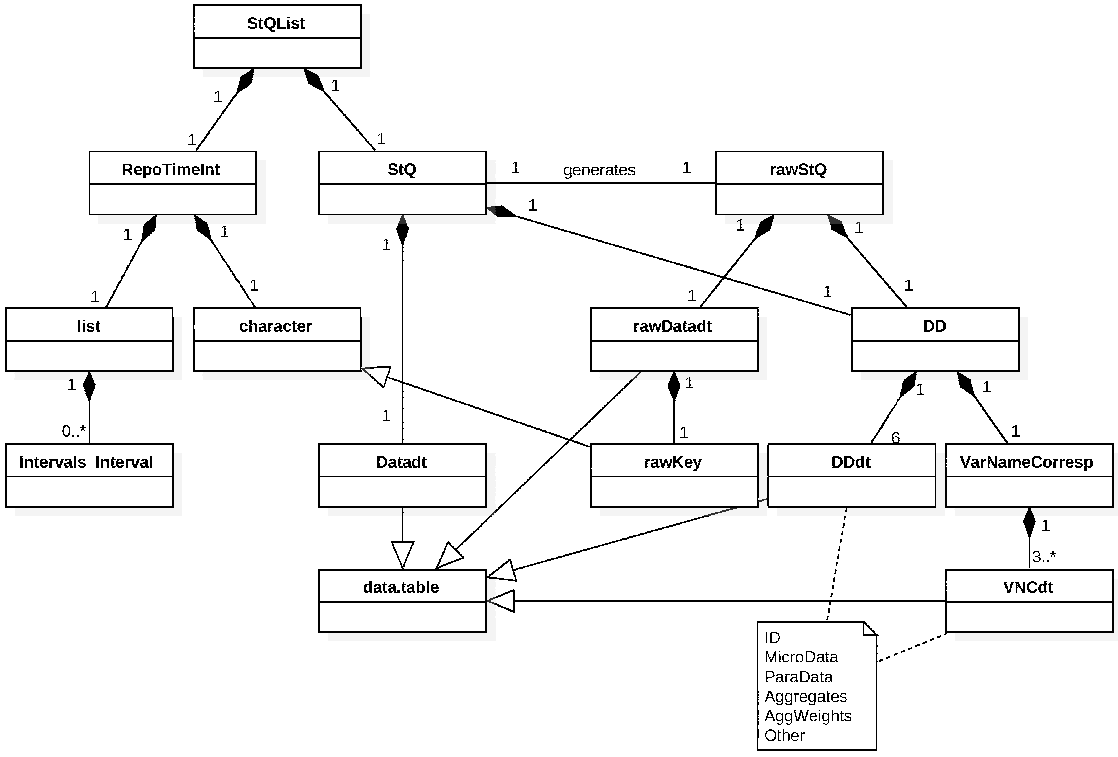

Figure 2: UML class diagram for the data model implementing in R the repository key-value pair structure.

We do not proceed to discuss the many details and design decisions behind this implementation (like the choice of `data.table` [65] as the basic class, the design of getters and setters, etc.), which will be thoroughly discussed elsewhere. Notice that the data model

depicted in figure 2 is valid for any statistical operation. Nonetheless we want to underline again the combination of computer science and statistical skills as a necessary ingredient in the design and development of an industrialised official statistical production system: are these classes (or a generalization thereof independent of any language) to be designed by traditional computer scientists or by traditional statisticians?

A thoughtful view of the interrelation (see figure 3) between business process model (BPM) and information model (IM) standards [2] invites us to think of the design of an official statistical production system as a formidable exercise of object-oriented analysis and design [66]. The argument can be posed graphically. Redraw figure 3 as in figure 4.

Is this but a sequence of objects (i.e. attributes + methods)? Is it not a natural invitation to design the statistical production process under OOA/D principles?

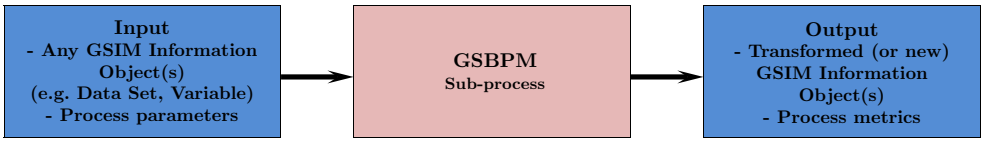

Figure 3: GSBPM and GSIM interrelation.

The traditional academic background of statisticians embrace neither these techniques nor associated modelling languages such as UML [67] or BPMN [68]. Nor does the traditional academic background of computer scientists embrace statistical techniques for official statistical production. The design of an efficient statistical production process must arise as a combination of both.

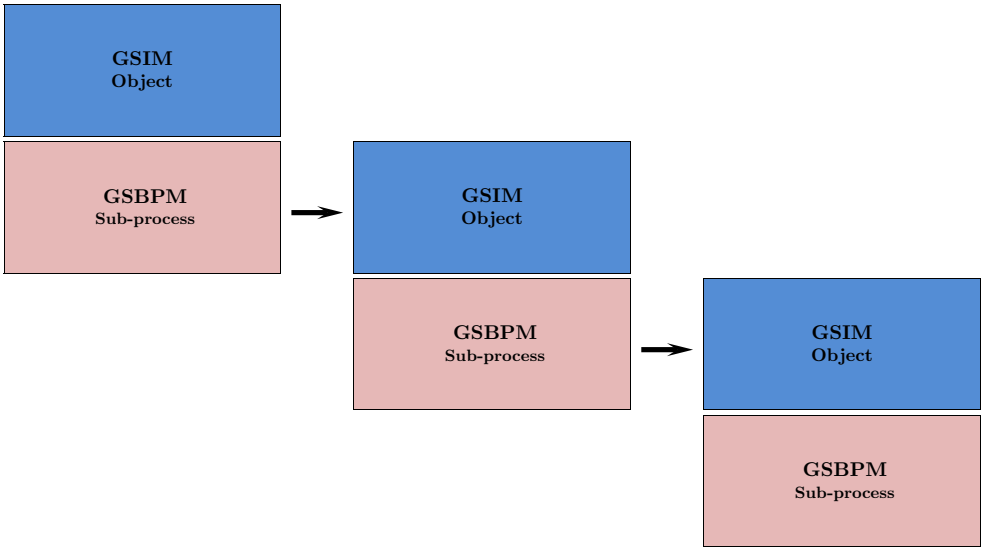

Figure 4: GSBPM and GSIM interrelation as an object.

Once an IT architecture for the repository was put in place (see figure 8 in appendix D for details), the next step already in course is to design prototype S4 classes[9] for the implementation of the former selective editing techniques independently both of the observation-prediction model for measurement errors (hence for both qualitative, quantitative, and semi-continuous variables) and of the time series methods to compute validation intervals. This

[9] As a matter of fact, they are already in production; we are currently streamlining these classes.

will hopefully allow us to make use of R packages already developed in the field of survey data editing [69, 70]. Notice even further the high relevance of the proposed combination of statistics and computer science to efficiently integrate both the statistical and computer-science aspects of this implementation.

5.3 Statistical data editing workflow

An example where the unified combination of statistics and computer science skills is clearly visible is the organization of the workflow within an editing and imputation strategy. The original EDIMBUS strategy [8] for business surveys is depicted in figure 5. More elaborated variants of this idea can be found in [6].

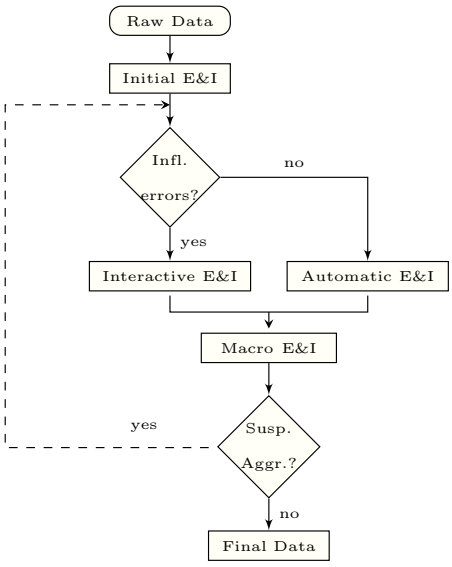

Figure 5: EDIMBUS strategy workflow [8].

We have translated this diagram into a UML sequence diagram explicitly tackling some business process design issues. Our proposal is represented in figure 6.

The key novelties are schematically:

- The strategy is expressed in terms of (reusable) production functions [72] understood as precise formulations of sequenced production steps possibly needing input parameters (despite the notation, not to be confused with mathematical functions).
- The phase of editing during data collection is explicitly included in the strategy (function `EditColl`).
- The distinction between functions applied upon single statistical units or upon the whole sample of statistical units is made explicit through the use of fork and join diagram elements.

We do not want to discuss methodological issues and notational details of this diagram (see [71]). Instead we want to underline the benefits from integrating statistical and computer science knowledge. It is clear that both diagrams 5 and 6 are clearly motivated from the statistical business process knowledge. A modelling language like UML or any other alternative can be simply viewed as an extra point of rigour in this sense.

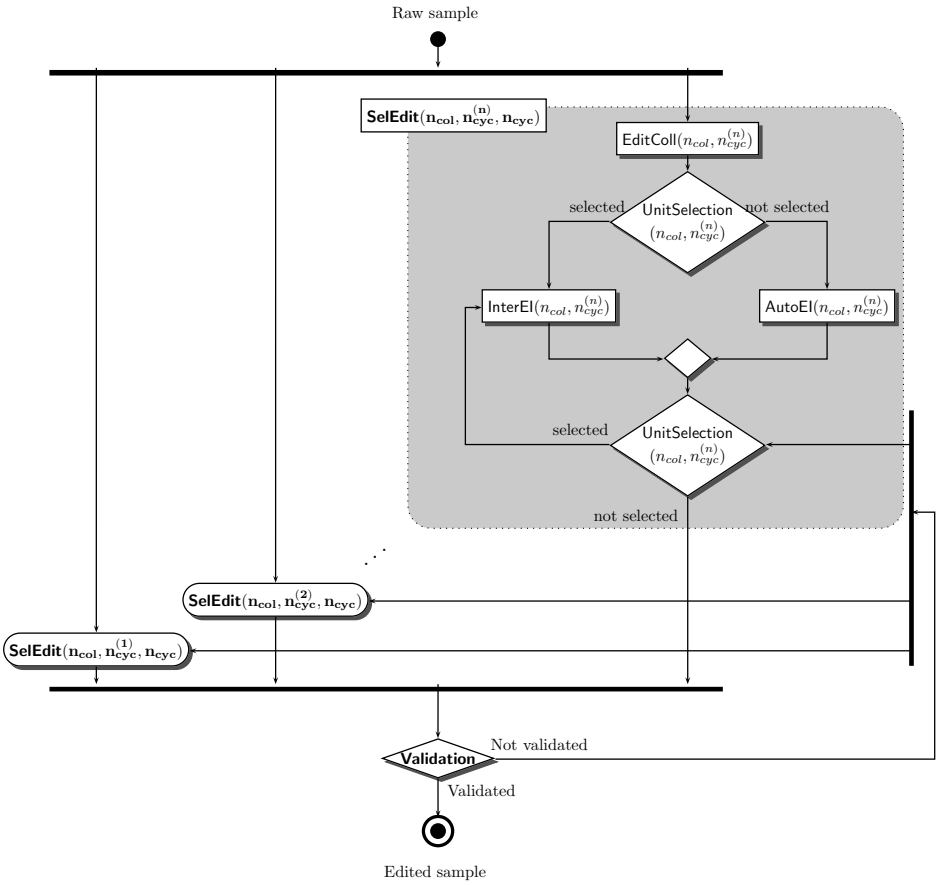

Figure 6: Extended EDIMBUS strategy workflow [71].

However we find it more interesting to call official statisticians' attention upon the fact that in the effort to express a business process in a rigorous standard modelling language like UML or BPMN or any other, the language elements themselves force us to think of many details of the process. This is a clear benefit from the usage of these computer science tools.

By the way, in the same reasoning line expressed regarding the construction of class diagrams, are these sequence diagrams to be designed by traditional computer scientists or by traditional statisticians? Notice that, against that misunderstood reduction of computer science to a matter of coding, no single line of computer code is written when designing these diagrams. Also, no computer scientist without methodological knowledge about the business process can design them.

5.4 An observation-prediction model for a binary variable

The final illustration regarding the modernisation of the editing phase of official statistical production comes from putting together again both computer science and statistical tools to define a variant of the `UnitSelection` function in diagram 6 when the variables are qualitative.

The argument reads as follows. Once the editing workflow has been formalized and standardized using the preceding computer science modelling tools, if the statistical content of

the production functions in the strategy can be extended to embrace not only quantitative but also qualitative and semicontinuous variables, the strategy will be of wider application under the same structure.

Let us concentrate on binary variables $y^{(q)} \in \{0,1\}$, $q = 1, \ldots, Q$. The longitudinal case is meaningless in these conditions. Thus we will concentrate on the cross-sectional case. The combinatorial optimization problem is still $P_{co}(\boldsymbol{\eta}, \Omega)$, where now

$$M^{(q)} = \text{diag}\left(\omega_k \cdot \mathbb{E}_m\left[\left|y_k^{(q,obs)} - y_k^{(q,0)}\right| \left|\mathbf{Z}_k^{cross}\right.\right]\right).$$

Therefore, the key point is the computation of the conditional moments

$$\mathbb{E}_m\left[\left|y_k^{(q,obs)} - y_k^{(q,0)}\right| \left|\mathbf{Z}_k^{cross}\right.\right]$$

under an adequate observation-prediction model. Emulating the usual manual editing tasks, as available information $\mathbf{Z}_k^{cross}$, apart from the set of reported values $y_k^{(q,obs)}$, we assume the existence of continuous variables $\mathbf{x}_k$ for each statistical unit k. The observation-prediction model for the binary variables y_k^{obs}, y_k^0 (we drop out the superscript (q) for ease of notation) assumes the following probabilities:

$$\begin{aligned}
\mathbb{P}\left(y_k^0 = 1|\mathbf{x}\right) &= \pi_k(\mathbf{x}), \qquad \mathbb{P}\left(y_k^0 = 0|\mathbf{x}\right) = 1 - \pi_k(\mathbf{x}),\\
\mathbb{P}\left(y_k^{obs} = 1|y_k^0 = 0, \mathbf{x}\right) &= q_k(\mathbf{x}), \qquad \mathbb{P}\left(y_k^{obs} = 0|y_k^0 = 0, \mathbf{x}\right) = 1 - q_k(\mathbf{x}),\\
\mathbb{P}\left(y_k^{obs} = 1|y_k^0 = 1, \mathbf{x}\right) &= 1 - p_k(\mathbf{x}), \qquad \mathbb{P}\left(y_k^{obs} = 0|y_k^0 = 1, \mathbf{x}\right) = p_k(\mathbf{x}),
\end{aligned}$$

where the probabilities $\pi_k(\mathbf{x}), p_k(\mathbf{x}), q_k(\mathbf{x}) \in [0,1]$ for all $k \in s$ play the same role as the parameters p_k, ν_k, $\hat{y}_k$, ν_k, ρ_k in the continuous variable observation-prediction model.

By Bayes' theorem we can write:

$$\begin{aligned}
\mathbb{P}\left(y_k^0 = 1|y_k^{obs} = 0, \mathbf{x}\right) &= P_k(\mathbf{x}), \qquad \mathbb{P}\left(y_k^0 = 0|y_k^{obs} = 0, \mathbf{x}\right) = 1 - P_k(\mathbf{x}),\\
\mathbb{P}\left(y_k^0 = 1|y_k^{obs} = 1, \mathbf{x}\right) &= 1 - Q_k(\mathbf{x}), \qquad \mathbb{P}\left(y_k^0 = 0|y_k^{obs} = 1, \mathbf{x}\right) = Q_k(\mathbf{x}),
\end{aligned}$$

where

$$\begin{aligned}
P_k(\mathbf{x}) &= \frac{\pi_k(\mathbf{x}) p_k(\mathbf{x})}{\pi_k(\mathbf{x}) p_k(\mathbf{x}) + (1 - \pi_k(\mathbf{x}))(1 - q_k(\mathbf{x}))},\\
Q_k(\mathbf{x}) &= \frac{q_k(\mathbf{x})(1 - \pi_k(\mathbf{x}))}{q_k(\mathbf{x})(1 - \pi_k(\mathbf{x})) + (1 - p_k(\mathbf{x}))\pi_k(\mathbf{x})}.
\end{aligned}$$

After elementary manipulations we have:

$$M_{kk} = \omega_k \cdot \mathbb{E}_m\left[\left|y_k^{obs} - y_k^0\right| \left|y_k^{obs}, \mathbf{x}\right.\right] = \begin{cases} \omega_k \cdot P_k(\mathbf{x}) & \text{if } y_k^{obs} = 0,\\ \omega_k \cdot Q_k(\mathbf{x}) & \text{if } y_k^{obs} = 1, \end{cases}$$

The model probabilities π_k, p_k, q_k can be estimated by logistic models with $\mathbf{x}$ as exogenous variables. For example, let us partition the population into domains $U_d \subset U$ in which the units show a similar error generation mechanism. Then the error probabilities $\pi_k(\mathbf{x})$, $p_k(\mathbf{x})$, and $q_k(\mathbf{x})$ can be estimated by formulating logistic models

$$\begin{aligned}
\text{logit}\,(\pi) &= \alpha_{d0} + \boldsymbol{\alpha}_d^T \mathbf{x} + \epsilon_d,\\
\text{logit}\,(p) &= \beta_{d0} + \boldsymbol{\beta}_d^T \mathbf{x} + \tilde{\epsilon}_d,\\
\text{logit}\,(q) &= \gamma_{d0} + \boldsymbol{\gamma}_d^T \mathbf{x} + \tilde{\tilde{\epsilon}}_d,
\end{aligned}$$

within each domain U_d so that $\hat{p}_k(\mathbf{x}) = \hat{p}_d(\mathbf{x}_k)$ and $\hat{q}_k(\mathbf{x}) = \hat{q}_d(\mathbf{x}_k)$ for all $k \in U_d$.

Thus, following the same general arguments as in the quantitative variables case, we may view these error moments as item scores for each binary variable $y^{(q)}$:

$$s_k^{(q)} = s\left(y_k^{(q,obs)}; \pi_k(\mathbf{x}), p_k(\mathbf{x}), q_k(\mathbf{x})\right) = \begin{cases} \omega_k \cdot P_k^{(q)}(\mathbf{x}) & \text{if } y_k^{(q,obs)} = 0, \\ \omega_k \cdot Q_k^{(q)}(\mathbf{x}) & \text{if } y_k^{(q,obs)} = 1. \end{cases} \tag{6}$$

Standardization is thus achieved both in the methodological and in the production implementation realms.

This approach to selectively edit a survey with categorical variables is currently under exploration. It needs a deep search of adequate models for the parameter probabilities in terms of available continuous variables and thorough testing with data from a real survey. See appendix E for a simple exploration model with data from the Spanish National Health Survey.

6 New sources of information for official statistics

Radermacher, in the opening address of the Conference NTTS 2015 [73], portrayed in a glimpse a history of official statistics in four periods:

- **XIX century**.- Official statistics production feeds from censuses and complete enumeration of data directly from the populations.
- **1900-1970**.- Sampling is introduced, macroeconomic statistics is born, and deepening and widening of survey methodology is produced along different lines. Data come directly from selected statistical units.
- **1970-2010**.- Information and communication technologies are introduced and become widespread. Methodologies adapt to these new technologies. Administrative registers begin to be used as source of information. Data protection and privacy begins to be an issue. Data still come directly from selected statistical units, either in electronic or physical collection modes.
- **2010-onwards**.- Data are generated from machine to machine interaction. Data deluge increases exponentially. Methodologies need further adjustment.

Dillman [74] already predicted somehow the decreasing availability of data directly from statistical units due to the widespread generalization of self-management of most social and human activities (self-managed luggage check-in at airports, automatic public transport tickets dispensers, ...). All these data are increasingly stored and processed by machines and are finally culminating into huge amount of data everywhere generated, stored, processed, and analyzed solely by machines. It is the era of Big Data.

However, in the spirit of Dillman's analysis, it is not size that matters. The key point is the digital footprint of human activity [75]. In this sense, both administrative registers and Big Data can be understood under the same umbrella: it is the trace of human activity in an information system, namely the information system of the State Public Administration and of corporations, respectively.

Both types of information systems pose different problems for their use in official statistics. In the case of administrative registers, data access is not really a major issue inasmuch as an agreement among different parts of the Public Administration is enough. However, for Big Data, digital data are mostly owned by private corporations. Data extraction and delivery in this case cannot be considered the same as filling a traditional questionnaire, however

complex it may be, either physical or electronic.

Regarding statistical production methodology, in both cases adjustments to the source of information are needed. In the case of administrative registers, variables are not designed and collected for statistical purposes. In the case of Big Data, this is also so, but in addition a strong change in the statistical methodology is entailed: design-based inference apparently will not be enough and model-based techniques will be ineludibly necessary, especially those Bayesian-based machine learning techniques [76].

All in all, this new scenario poses a high risk for official statistics. A public-private partnership [77] is in due order for public statistical agencies to have access to that huge amount of information. Shouldn't these agencies have access to this information, this will ineludibly be statistically exploited by someone else (the corporations themselves?) and the agencies may become gradually irrelevant.

All this digital information, either huge, large, big, or small, will be impossible to exploit unless a unified combination of statistics and computer science rules the production.

7 Some conclusions

We argue that the modernisation and industrialisation of official statistical production requires a unified combination of statistics and computer science in its very principles. To defend this idea we have detailed some proposals, many of them already applied in production in some Spanish statistical operations, where this combination is clearly shown.

The industrialisation of official statistical production needs a modern system of metadata, in particular of process metadata. The standard for these metadata at Statistics Spain takes GSBPM level-2 production subprocesses as a starting point and develops a third level in which production tasks are described according to their input(s), output(s), throughput, documentation, tools, and responsible unit(s). This has been accomplished following computer system design principles such as functional modularity, hierarchy, and layering. Among other things, these make us establish natural borders between tasks. Thus, under this view, only a combination of computer science and statistical principles will allow us to design an efficient statistical process in which complexity is properly dealt with.

In our proposal to modernise the statistical data editing phase, we have combined both skills in different ways. Proceeding top-down, the workflow of production tasks has been expressed using a UML sequence diagram, which makes us think of and give details not contained in more informal diagrams. One of these new elements are production functions, which if properly designed according to functional modularity principles, can be substituted for streamlined versions without affecting the rest of the editing and imputation strategy.

The most prominent example of such a functionally modular production function is UnitSelection, which selects each statistical unit for interactive editing or automatic editing (due to the presence or absence of influential errors). We have provided a non-heuristic statistical basis for the use of score functions, which arises as the application of so-called observation-prediction statistical models. The use of rigorous statistical language allows to define new UnitSelection functions applicable to binary, categorical, and semicontinuous variables provided that we model the measurement errors.

All this statistical methodology demands an adequate IT architecture to standardize its use in actual production conditions. Firstly, since model parameters are estimated making use of historical sets of raw and edited data values, we need access to these data upon

which these methodological proposals are to be applied. We have programmed a light-weight object-oriented implementation of a versatile key-value pair structure for a repository of files using a high-level language such as R. This has allowed us to go from design to production in a very short period of time. Everything is reduced to feeding this repository.

Once the access to historic double sets of raw and edited data values is solved, the implementation of the statistical methodology for selecting units with influential errors is undertaken by designing new classes valid for any statistical operation under consideration. An initial proposal is already in production with good results (see figure 7). Currently we are working in streamlined versions allowing us to also deal with the cases of categorical and semicontinuous variables. Notice that this requires our defended combination and will hopefully drive us to a standardization of this production phase at a larger scale.

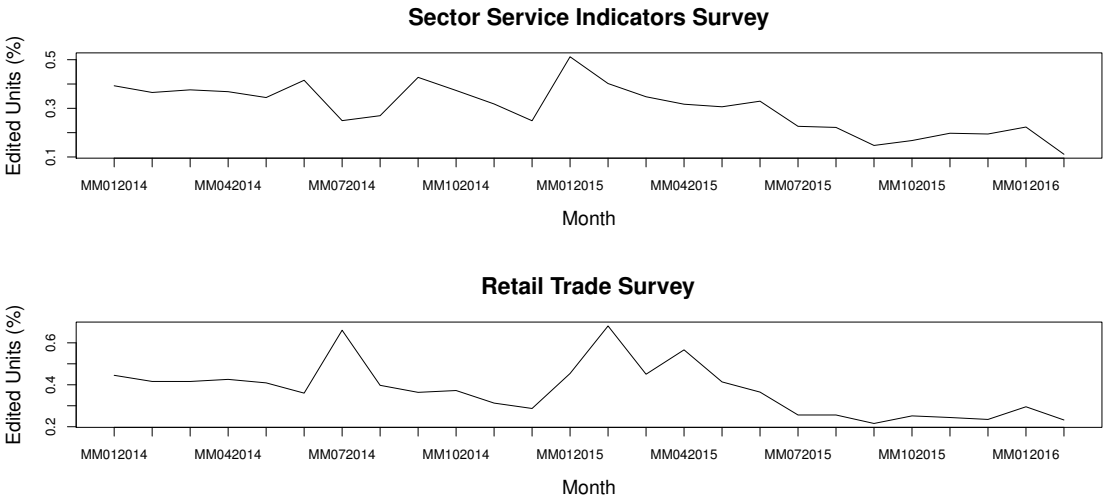

Figure 7: Recontact rates in the Spanish Retail Trade Survey and the Sector Service Indicators Survey after applying the optimization approach to selective editing. For the Industrial Turnover and New Orders Received Surveys the recontact rate was also roughly halved on average from its original 55% after applying this proposal starting in January, 2013 (not included in the figure). The initial anomalies in the figure reflect the difficulties met in the implementation during the first months. Special thanks to E. Sánchez-Núñez and her teams and staff at Statistics Spain's provincial delegations for their patience and perseverance in the implementation process (parameter fine-tuning, bug fixing, migration from SAS to R, ...).

However as key as we find this combination of computer science and statistics to modernise official statistical production, we do not claim that this combination *per se* will automatically industrialise a statistical office. Nor do we claim that this unified combination in the professional profile of new young official statisticians will do so.

The process of modernisation and industrialisation must be undergone in current production conditions, whatever they are, and implemented by those (often overloaded) actors whose professional profiles needs this adjustment. Skills will surely be faster acquired provided suitable training programmes are put in place. Moreover, this modernisation must be accomplished without affecting the release of official statistical products mostly ruled by legal regulations. This is an extraordinary challenge for the top management thus requiring a detailed and carefully designed action plan.

A Derivation of Minkowskian global score functions

To derive the generalized stochastic optimization problem, first we notice that for loss functions $L^{(r)}(a,b) = |a-b|^r$, the quantities $\mathbb{E}_{pm}^{\frac{1}{r}}\left[L^{(r)}\left(\hat{Y}^{(*,q)}(\mathbf{R}), Y^{(q)}\right)|\mathbf{Z}\right]$ satisfy the triangle inequality so that $\mathbb{E}_{m}^{\frac{1}{r}}\left[L^{(r)}\left(\hat{Y}^{(*,q)}(\mathbf{R}), \hat{Y}^{(q,0)}\right)|\mathbf{Z}\right] \leq \eta_q$ also entails the upper-bounding of $\mathbb{E}_{pm}^{\frac{1}{r}}\left[L^{(r)}\left(\hat{Y}^{(*,q)}(\mathbf{R}), Y^{(q)}\right)|\mathbf{Z}\right]$ provided that $\mathbb{E}_{pm}^{\frac{1}{r}}\left[L^{(r)}\left(\hat{Y}^{(q,0)}, Y^{(q)}\right)\Big|\mathbf{Z}\right]$ is also bounded (by the sampling design).

For the generalized restrictions we can write

$$\begin{aligned}\mathbb{E}_m\left[L^{(r)}\left(\hat{Y}^{(*,q)}(\mathbf{R}), \hat{Y}^{(q,0)}\right)|\mathbf{Z}\right] &\leq \mathbb{E}_m\left[\sum_{k\in s} R_k\left|\omega_{ks}\epsilon_k^{(q)}\right|^r|\mathbf{Z}\right] \\ &= \mathbb{E}_m\left[\mathbf{R}^T\text{diag}\left(\Delta^{(q)r}\right)|\mathbf{Z}\right],\end{aligned}$$

where $\epsilon_k^{(q)} = y_k^{(q)} - y_k^{(q,0)}$ denotes the measurement error and $\Delta^{(q)}$ stands for the diagonal matrix with entries $|\omega_{ks}\epsilon_k^{(q)}|$.

Besides, the longitudinal case is to be applied as a form of input editing, so that no cross-sectional information will be available when resolving about unit k, except for the information regarding unit k itself. Thus we can write $R_k \in \mathcal{S}(\mathbf{Z}_k^{cross})$ to indicate that the selection of unit k is a function of the available information (cross-sectional included, i.e. the values in the questionnaire). Then we can write

$$\begin{aligned}\mathbb{E}_m\left[\mathbf{R}^T\text{diag}\left(\Delta^{(q)r}\right)|\mathbf{Z}\right] &= \mathbb{E}_m\left[\mathbb{E}_m\left[\mathbf{R}^T\text{diag}\left(\Delta^{(q)r}\right)|\mathbf{Z}^{cross}\right]|\mathbf{Z}\right] \\ &= \mathbb{E}_m\left[\mathbf{R}^T\text{diag}\left(M^{(q,r)}\right)|\mathbf{Z}\right],\end{aligned}$$

where $M^{(q,r)} = \mathbb{E}_m\left[\Delta^{(q)r}\Big|\mathbf{Z}_k^{cross}\right]$. Now, to impose an upper bound on this conditional expectation for arbitrary values of r is too restrictive. Thus we will focus on the looser expression (by means of Lyapunov's inequality):

$$\mathbb{E}_{pm}\left[\mathbf{R}^T\text{diag}\left(M^{(q,1)r}\right)|\mathbf{Z}\right] \leq \mathbb{E}_{pm}\left[\mathbf{R}^T\text{diag}\left(M^{(q,r)}\right)|\mathbf{Z}\right].$$

The generalized stochastic optimization problem is given by:

$$\begin{aligned}[P_{st}^{(r)}] \qquad & \max\ \mathbb{E}_m\left[\mathbf{1}^T\mathbf{R}|\mathbf{Z}\right] \\ \text{s.t.}\quad & \mathbb{E}_m\left[\mathbf{R}^T\text{diag}\left(M^{(q,1)r}\right)|\mathbf{Z}\right] \leq \eta_q, \quad q = 1,2,...,Q, \\ & R_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s.\end{aligned}$$

The solution to this problem is *mutatis mutandis* similar to the original case (see eq. (2)):

$$R_k = \begin{cases} 1 & \text{if } \sum_{q=1}^{Q}\lambda_q^* M_{kk}^{(q,1)r} \leq 1, \\ 0 & \text{if } \sum_{q=1}^{Q}\lambda_q^* M_{kk}^{(q,1)r} > 1, \end{cases}$$

where λ_q^* are the optimal Lagrange multipliers of the associated dual problem [50].

B Alternative theoretical proposal to find the Lagrange multipliers

The original approach in [50] proceeds to convexify the stochastic optimization problem by setting a general form for the selection strategy vector given by

$$R_k = \begin{cases} 1 & \text{if } \xi_k \leq Q_k, \\ 0 & \text{otherwise,} \end{cases}$$

where ξ_k are continuous uniform random variables $U(0,1)$ and $Q_k \in [0,1]$ are continuous random variables independent from ξ_k. The generalized stochastic optimization problem reads then

$$\begin{aligned} [P_{st}^{(r)}] \qquad & \max \ \mathbb{E}_m \left[\mathbf{1}^T \mathbf{Q} | \mathbf{Z} \right] \\ \text{s.t.} \quad & \mathbb{E}_m \left[\mathbf{Q}^T \text{diag} \left(M^{(q,1)r} \right) | \mathbf{Z} \right] \leq \eta_q, \quad q = 1, 2, ..., Q, \\ & Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s. \end{aligned}$$

The solution to the generalized stochastic optimization problem $P_{st}^{(r)}$ is reached using the duality theorem (see e.g. [78]). Let the Lagrangian

$$\mathcal{L}^{(r)}(\mathbf{Q}, \boldsymbol{\lambda}) = \mathbb{E}_m \left[\mathbf{1}^T \mathbf{Q} - \sum_{q=1}^{Q} \lambda_q \left(\mathbf{Q}^T \text{diag}(M^{(q,1)r}) - \eta_q \right) | \mathbf{Z} \right].$$

Let $\tilde{P}_{st}^{(r)}(\boldsymbol{\lambda})$ denote the associated maximization problem

$$\begin{aligned} [\tilde{P}_{st}^{(r)}(\boldsymbol{\lambda})] \qquad & \max \mathcal{L}^{(r)}(\mathbf{Q}, \boldsymbol{\lambda}) \\ & Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s. \end{aligned}$$

Then $\mathbf{Q}(\boldsymbol{\lambda}^*)$ is an optimal solution to problem $P_{st}^{(r)}$, when $\boldsymbol{\lambda}^*$ are the optimal Lagrange multipliers of the dual problem given by

$$[D^{(r)}] \qquad \min_{\boldsymbol{\lambda} \geq \mathbf{0}} \max_{Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross})} \mathcal{L}^{(r)}(\mathbf{Q}, \boldsymbol{\lambda})$$

Now problem $D^{(r)}$ is solved approximately in [50] by applying the sample average approximation and the principle of interchangeability. To support the former, stationarity in the measurement error generation must be assumed. This allows us to view the loss matrices $M_t^{(q,1)}$ in each time period t as realizations of the same random variables. Thus they write

$$\begin{aligned} \max_{Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s} \mathbb{E}_m \left[\mathbf{1}^T \mathbf{Q} - \sum_{q=1}^{Q} \lambda_q \left(\mathbf{Q}^T \text{diag}(M^{(q,1)r}) - \eta_q \right) | \mathbf{Z} \right] \quad \approx \\ \frac{1}{T} \sum_{t=1}^{T} \max_{\mathbf{q}_t \in [0,1]} \left[\mathbf{1}^T \mathbf{q}_t - \sum_{q=1}^{Q} \lambda_q (\mathbf{q}_t^T \text{diag}(M_t^{(q,1)r}) - \eta_q) \right]. \end{aligned}$$

They solve this numerical optimization problem using standard maximization routines [50]. An alternative approach to reach a solution to problem $D^{(r)}$ is to avoid the assumption about stationarity and to use predictions for the error moments $M^{(q)r}$, denoted by $\widehat{M^{(q)r}}$. Thus we propose to write

$$\max_{Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s} \mathbb{E}_m \left[\mathbf{1}^T \mathbf{Q} - \sum_{q=1}^{Q} \lambda_q \left(\mathbf{Q}^T \operatorname{diag}(M^{(q,1)r}) - \eta_q \right) | \mathbf{Z} \right] =$$

$$\max_{Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s} \left[\mathbf{1}^T \mathbb{E}_m \left[\mathbf{Q} | \mathbf{Z} \right] - \sum_{q=1}^{Q} \lambda_q \left(\mathbb{E}_m \left[\mathbf{Q}^T | \mathbf{Z} \right] \operatorname{diag}(\widehat{M^{(q,1)r}}) - \eta_q \right) + \right.$$

$$\left. \sum_{q=1}^{Q} \lambda_q \mathbb{E}_m \left[\mathbf{Q}^T \operatorname{diag} \left(\widehat{M^{(q,1)r}} - M^{(q,1)r} \right) | \mathbf{Z} \right] \right]. \tag{7}$$

Now, by the Cauchy-Schwarz inequality we can write

$$\begin{aligned} \mathbb{E}_m \left[Q_k \left(\widehat{M_{kk}^{(q,1)r}} - M_{kk}^{(q,1)r} \right) | \mathbf{Z} \right] &\leq \mathbb{E}_m \left[\left| Q_k \left(\widehat{M_{kk}^{(q,1)r}} - M_{kk}^{(q,1)r} \right) \right| \left| \mathbf{Z} \right. \right] \\ &\leq \mathbb{E}_m^{\frac{1}{2}} \left[Q_k^2 | \mathbf{Z} \right] \cdot \mathbb{E}_m^{\frac{1}{2}} \left[\left(M_{kk}^{(q,1)r} - \widehat{M_{kk}^{(q,1)r}} \right)^2 | \mathbf{Z} \right] \\ &\leq \sigma_m \left[M_{kk}^{(q,1)r} | \mathbf{Z} \right] \end{aligned}$$

Instead of problem (7) we shall concentrate upon the looser problem given by

$$\max_{Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s} \left[\mathbf{1}^T \mathbb{E}_m \left[\mathbf{Q} | \mathbf{Z} \right] - \sum_{q=1}^{Q} \lambda_q \left(\mathbb{E}_m \left[\mathbf{Q}^T | \mathbf{Z} \right] \operatorname{diag}(\widehat{M^{(q,1)r}}) - \left(\eta_q + \sum_{k \in s} \sigma_m \left[M_{kk}^{(q,1)r} | \mathbf{Z} \right] \right) \right) \right] \tag{8}$$

Notice that the more precise the predictions $\widehat{M^{(q,1)r}}$, the lower the conditional standard deviations $\sigma_m \left[M_{kk}^{(q,1)r} | \mathbf{Z} \right]$, hence the closer problem (8) is to problem (7). Regretfully, we do not know the distribution of $M^{(q,1)r}$, thus nor do we know $\sigma_m \left[M_{kk}^{(q,1)r} | \mathbf{Z} \right]$. So, instead of problem (8), we will estimate each σ_m and concentrate upon the problem given by

$$\max_{Q_k \in \mathcal{S}(\mathbf{Z}_k^{cross}), k \in s} \left[\mathbf{1}^T \mathbb{E}_m \left[\mathbf{Q} | \mathbf{Z} \right] - \sum_{q=1}^{Q} \lambda_q \left(\mathbb{E}_m \left[\mathbf{Q}^T | \mathbf{Z} \right] \operatorname{diag}(\widehat{M^{(q,1)r}}) - \tilde{\eta}_q^{(r)} \right) \right],$$

where we denote $\tilde{\eta}_q^{(r)} = \eta_q + \sum_{k \in s} \widehat{\sigma}_m \left[M_{kk}^{(q,1)r} | \mathbf{Z} \right]$.

Now this maximization problem has the trivial solution $\mathbf{q}^{(r)} = \mathbf{q}^{(r)}(\boldsymbol{\lambda}, \widehat{\mathbf{M}^{(\cdot,1)}})$ given by

$$q_k^{(r)}(\boldsymbol{\lambda}, \widehat{\mathbf{M}^{(\cdot,1)}}) = \begin{cases} 1 & \text{if } \sum_{q=1}^{Q} \lambda_q \widehat{M_{kk}^{(q,1)r}} \leq 1, \\ 0 & \text{if } \sum_{q=1}^{Q} \lambda_q \widehat{M_{kk}^{(q,1)r}} > 1. \end{cases}$$

Let

$$L^{(r)}(\boldsymbol{\lambda}) = \sum_{k \in s} q_k^{(r)}(\boldsymbol{\lambda}, \widehat{\mathbf{M}^{(\cdot,1)}}) + \sum_{q=1}^{Q} \lambda_q \tilde{\eta}_q^{(r)} - \sum_{q=1}^{Q} \sum_{k \in s} \lambda_q \widehat{M_{kk}^{(q,1)r}} q_k^{(r)}(\boldsymbol{\lambda}, \widehat{\mathbf{M}^{(\cdot,1)}}),$$

then the optimal Lagrange multipliers $\boldsymbol{\lambda}^*$ will be given by the solution of the relaxed problem expressed as $\min_{\boldsymbol{\lambda} \geq \mathbf{0}} L^{(r)}(\boldsymbol{\lambda})$.

It can be proved [53] that the function $L^{(r)}$ is continuous in $\{\boldsymbol{\lambda} \geq \mathbf{0}\}$ but non-differentiable in $\bigcup_{k \in s} H_k^{(r)+}$, where $H_k^{(r)+}$ denotes the intersection $H_k^{(r)+} = H_k^{(r)} \bigcap \{\boldsymbol{\lambda} \geq \mathbf{0}\}$, $H_k^{(r)}$ being

the hyperplane $H_k^{(r)} = \{\boldsymbol{\lambda} \in \mathbb{R}^Q : \sum_{q=1}^{Q} \lambda_q \widehat{M_{kk}^{(q,1)r}} = 1\}$. Thus the objective function $L^{(r)}(\boldsymbol{\lambda})$ is also continuous in $\{\boldsymbol{\lambda} \geq \mathbf{0}\}$ and non-differentiable in $\bigcup_{t=1}^{T} \bigcup_{k \in s} H_k^{(r)+}$. Furthermore, it can also be proved that $L^{(r)}$ is coercive (see e.g. [79]). Thus, by Weierstrass' theorem the objective function is lower-bounded and attains its minimum in its domain of definition.

This suggests to use a direct search optimization method (see e.g. [80]), which does not make use of derivatives to find the minimum. In particular, we are currently considering a *compass search* method. Let $\{\mathbf{e}_q\}_{q=1,\ldots,Q}$ denote the canonical basis in $\mathbb{R}^Q$ and $f = L^{(r)}$. The algorithm is depicted in Algorithm 1 below.

Algorithm 1 Compass search algorithm to find the minimum of $L^{(r)}$

1: $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}_0$ ▷ Initial guess solution
2: $s \leftarrow s_0$ ▷ Initial step length
3: **for all** q=1,...,Q **do**
4: $\quad \boldsymbol{\lambda}_q \leftarrow \boldsymbol{\lambda} + s \cdot \mathbf{e}_q$
5: **end for**
6: $\boldsymbol{\lambda}^* \leftarrow \mathsf{argmin}_{q=1,\ldots,Q} f(\boldsymbol{\lambda}_q)$
7: **if** $f(\boldsymbol{\lambda}^*) < f(\boldsymbol{\lambda})$ AND $\left\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}\right\| > \epsilon$ **then** ▷ Stopping criterion
8: $\quad s \leftarrow s/2$
9: $\quad \boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}^*$
10: $\quad$ RETURN TO STEP 3
11: **end if**
12: **return** $\boldsymbol{\lambda}^*$

The parameters of the algorithm are the initial guess solution $\boldsymbol{\lambda}_0$, the initial step length s_0, and the accepted tolerance ϵ. Details regarding their choice are discussed elsewhere [53].

It is important to underline that the evaluation of this approach will be done according to the selection efficiency measure introduced in [52] and applied upon real survey data. Any other numerical assessment, although necessary for rigour's sake, must be subjected to the final goal of efficiently selecting units with influential errors.

C The unit prioritization algorithm and its equivalence with the score function approach

Firstly, we sketch the prioritization algorithm. Then we will prove the equivalence with the traditional score function approach.

Algorithm 2 Unit prioritization from a sequence of combinatorial optimization problems

1: $\Omega \leftarrow \Omega^{(0)} \equiv \{1, \ldots, n\}$ ▷ Initial empty selection
2: Initialize $\mathbf{s} \in \mathbb{N}^{\times n}$
3: **for** i=1,...,n **do**
4: $\quad \eta_q \leftarrow \eta_q^{(i)}$ so that $\mathbf{r}^{(i)} = \text{solve}\left(P_{co}\left(\boldsymbol{\eta}^{(i)}, \Omega^{(i-1)}\right)\right)$ satisfies $\sum_{k \in s} r_k = i$
5: $\quad \Omega^{(i)} \leftarrow \Omega^{(i-1)} - \{i\}$
6: $\quad s_i \leftarrow k \in [1, n] : r_k^{(i)} \neq r_k^{(i-1)}$
7: $\quad$ **return** $\boldsymbol{s}$
8: **end for**

The main idea behind the prioritization of units using the combinatorial optimization problem $P_{co}(\boldsymbol{\eta}, \Omega)$ is the construction of an adequate sequence of upper bounds $\boldsymbol{\eta}$. We start

by setting $\Omega^{(0)} = \{1, \ldots, n\}$. Then we reduce the upper bounds to $\boldsymbol{\eta}^{(1)}$ so that the solution $\mathbf{r}^{(1)}$ to the new problem $P_{co}(\boldsymbol{\eta}^{(1)}, \Omega^{(0)})$ is $r_k = 1$ for all k, except for one $r_{k_1} = 0$. Unit k_1 has been selected. We set $\Omega^{(2)} = \Omega^{(1)} - \{k_1\}$ and again we reduce the upper bounds to $\boldsymbol{\eta}^{(2)}$. This procedure is repeated n times. The prioritization is given by the sequence of indices $\mathbf{s} = (k_1, k_2, \ldots, k_n)$ indicating which new unit has been selected in each iteration.

Now, conjugating both this algorithm with an appropriate sequence of bounds $\{\boldsymbol{\eta}^{(i)}\}_{i=1,\ldots,n}$ we can justify the traditional role of global score functions.

Proposition C.1. Let

$$\eta_q^{(i+1)} = \sum_{k \in \Omega^{(i)}} M_{kk}^{(q)r} - \max_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right),$$

where the global score function S is such that

$$k_{i+1} = \text{argmax}_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$$

in each resolution $i+1$ of the problem $P_{co}\left(\boldsymbol{\eta}^{(i+1)}, \Omega^{(i)}\right)$. Then the prioritization of units is given by the descending order of the values of $S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$.

Proof. Note that in iteration $i+1$ of the prioritization algorithm the restrictions of the problem read

$$\mathbf{r}^{(i+1)T} \mathbf{M}^{(q)r} \mathbf{r}^{(i+1)} - \eta_q^{(i+1)} \leq 0 \quad q = 1, \ldots, Q. \tag{9}$$

Now to solve the problem in this iteration we substitute

$$\eta_q^{(i+1)} = \sum_{k \in \Omega^{(i)}} M_{kk}^{(q)r} - \max_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$$

into equation (9) to arrive at

$$\mathbf{r}^{(i+1)T} \mathbf{M}^{(q)r} \mathbf{r}^{(i+1)} - \mathbf{r}^{(i)T} \mathbf{M}^{(q)} \mathbf{r}^{(i)} + \max_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right) \leq 0, \quad q = 1, \ldots, Q. \tag{10}$$

Now the solution $\mathbf{r}^{(i+1)}$ found with the recipe

$$k_{i+1} = \text{argmax}_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$$

is precisely $\mathbf{r}^{(i)}$ with $r_{k_{i+1}} = 0$, since equation (10) reduces to

$$M_{k_{i+1} k_{i+1}}^{(q)r} \geq \max_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$$

for all q. Thus the new selected unit k_{i+1} makes $\mathbf{r}^{(i+1)}$ satisfy every restriction.

Finally the ordering given by $k_{i+1} = \text{argmax}_{k \in \Omega^{(i)}} S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$, $\Omega^{(i+1)} = \Omega^{(i)} - \{k_i\}$ is exactly the decreasing order of the values $S\left(M_{kk}^{(1)r}, \ldots, M_{kk}^{(Q)r}\right)$. □

To sum up, if we think of $M_{kk}^{(q)}$ as local scores and of S as a global score function, we are reproducing the traditional approach. This entails the following interpretation of global score functions: they translate our priority about the gradual accuracy of the different aggregate estimates $\hat{Y}^{(q)}$ upon editing execution into a unit prioritization.

D S4 classes and methods for the repository data model

The S4 classes and methods in language R implementing the key-value pair structure for the repository is detailed in the following collection of class diagrams (see figure 8).

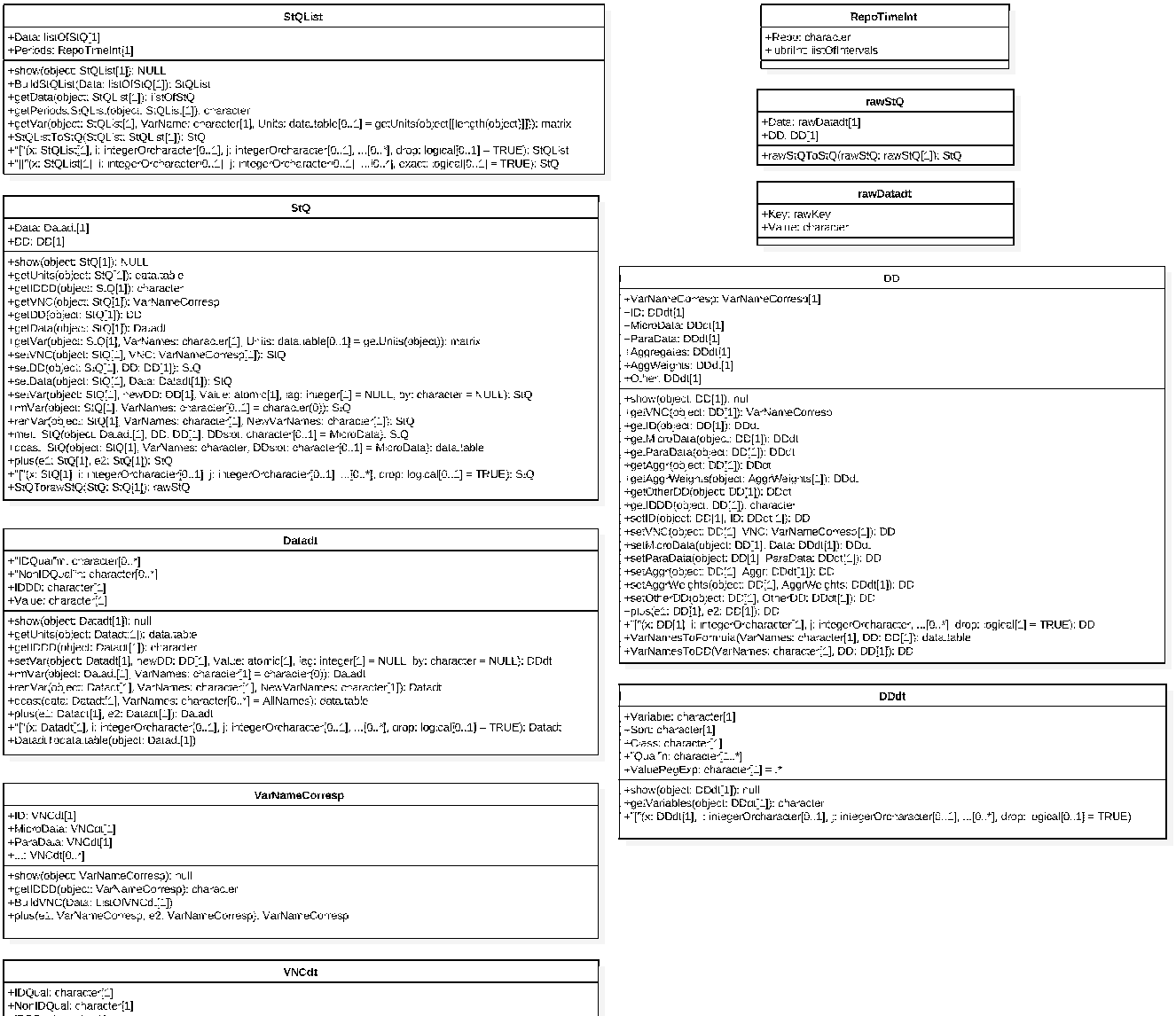

Figure 8: UML class diagrams with their attributes and methods.

E A simple observation-prediction model for a binary variable

We include an illustrative example of our currently on-going research on the optimization approach to selective editing of binary variables. We give details of an initial exploratory example using data from the Spanish National Health Survey[10].

[10]We thank M.R. González-García for providing raw and edited data sets and insight into the meaning of the variables.

We will focus on a binary variable indicating membership to a domain $U_d \subset U$. In particular, we concentrate on so-called social class membership, where by "social class" it is understood a grouping of CNO-11[11] codes. We define $y \in \{0, 1\}$ as the binary variable indicating whether a person belongs to social class 1 or not. Social class 1 comprises a priori those occupations[12] with higher expected salaries.

A prioritization of units for the editing field work according to the measurement error of the binary variable y is impossible using traditional item score functions $s(\cdot)$ (as e.g. $s(y_k; \omega_k) = \omega_k |y_k^{obs} - \hat{y}_k|$, especially useless in self-weighting samples). In practice, this kind of variable is edited by its semantic content and expert judgment. In the case of "social class 1" editing clerks explore the relationship of this variable with, among others, the variable "salary" included in the questionnaire searching for inconsistencies. Highest salaries are expected when $y = 1$. Editing is completely interactive.

Using a double set of raw and edited values we train the logistic models (5a), (5a), (5a), where $\mathbf{x} = (x_1, x_2)^T$ are the ratio between reported salary and number of consumption units in the household $x_1 = \frac{\text{Salary}}{\text{ConsumpUnits}}$ and this ratio weighted by the sampling weight $x_2 = \text{SampWeight} \cdot \frac{\text{Salary}}{\text{ConsumpUnits}}$. The parameters π_k, p_k, q_k are then estimated by $\hat{\pi}_k(\mathbf{x}) = \hat{\pi}(\mathbf{x}_k)$, $\hat{p}_k(\mathbf{x}) = \hat{p}(\mathbf{x}_k)$, $\hat{q}_k(\mathbf{x}) = \hat{q}(\mathbf{x}_k)$, where the predicted values $\hat{\pi}$, $\hat{p}$, $\hat{q}$ are computed according to the trained models and the values $\mathbf{x}_k$ are taken from the test data set. To be precise, in order to prevent high nonresponse rates when asking the detailed salary, a censored variable in terms of intervals is asked. Therefore, the center of this censored interval is used as Salary in the preceding variables x_1, x_2. Thus, its discriminating power is rather limited indeed since only 10 intervals are queried about.

The item scores s_k for variable y are then given by (6) with $\pi_k(\mathbf{x}) = \hat{\pi}(\mathbf{x}_k)$, $p_k(\mathbf{x}) = \hat{p}(\mathbf{x}_k)$, $q_k(\mathbf{x}) = \hat{q}(\mathbf{x}_k)$. We produce a prioritization of units according to these values s_k and investigate its quality by using the editing efficiency measure introduced in [52] (see figure 9).

Despite the high simplicity of the models and the low discriminating power of the available continuous variables x_1, x_2, the prioritization is more efficient than a random sorting of the units. Although this is a simple exploratory example, it is promising since it encourages us to explore more sophisticated models closer to more realistic conditions.

The current lines of research are:

- As an initial exploration, we use half of the total double set as training set and the other half as test set (halves chosen randomly). Our plan is to reproduce actual production conditions by dividing the data set according to the progression of data collection, use progressively these as training sets and the on-going new data set chunk (usually 2 or 3 weeks) as test set.
- Binary variables are to be generalized to categorical variables, so multinomial logistic models instead of binomial logistic models are to be explored.
- A systematic analysis of all available continuous variables as exogeneous variables in the models is to be accomplished.
- Editing both continuous and categorical variables in the same strategy is to be explored.

[11] CNO-11 stands for *Clasificación Nacional de Ocupaciones 2011* [81], which is the Spanish version of the International Standard Classification of Occupations (ISCO-08) [82].

[12] CNO-11 codes: 111, 112, 121, 131, 132, 211, 213, 214, 215, 221, 223, 241, 242, 243, 244, 245, 251, 259, 261, 262, 265, 271, 281, 282, 291, 292, 283.

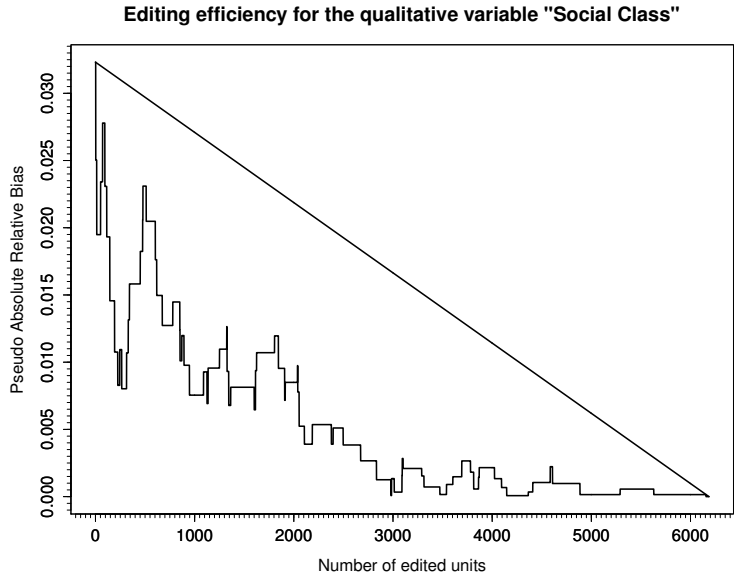

Figure 9: Editing efficiency measure for prioritizing units to edit the binary variable "Social Class 1".

References

[1] High-Level Group for the Modernisation of Statistical Production. Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics. *Conference of European Statisticians*. Geneva, 14-16 June, 2011.

[2] UNECE. Generic Statistical Business Process Model v5.0. UNECE, 2013. Available at http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model (accessed on May 19, 2016).

[3] UNECE. Generic Statistical Information Model v1.1. UNECE, 2013. Available at http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model (accessed on May 19, 2016).

[4] UNECE. Common Statistical Production Arquitecture v1.5. UNECE, 2013. Available at http://www1.unece.org/stat/platform/display/CSPA/Common+Statistical+Production+Architecture+Home (accessed on May 19, 2016).

[5] UNECE. Generic Activity Model for Statistical Organizations v1.0. UNECE, 2015. Available at http://www1.unece.org/stat/platform/display/GAMSO/GAMSO+Home (accessed on May 19, 2016).

[6] UNECE. Generic Statistical Data Editing Models v1.0. UNECE, 2015. Available at http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=117774163 (accessed on May 19, 2016).

[7] UNECE. Big Data in Official Statistics. UNECE, 2015. Available at http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics (accessed on May 19, 2016).

[8] ISTAT, CBS, SFSO, and EUROSTAT. *Recommended practices for editing and imputation in cross-sectional business surveys (EDIMBUS manual).* Eurostat, 2007. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf (accessed on May 19, 2016).

[9] European Statistical System. European Code of Practice (rev. ed.). Eurostat, 2011. Available at http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955 (accessed on May 19, 2016).

[10] European Statistical System. Quality Assurance Framework of the European Statistical System. Eurostat, 2012. Available at http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646 (accessed on May 19, 2016).

[11] European Statistical System. Euro-SDMX Metadata Structure (ESMS). Eurostat, 2014. Available at http://ec.europa.eu/eurostat/data/metadata/metadata-structure (accessed on May 19, 2016).

[12] European Statistical System. New Validation and Transformation Language (VTL). 2014. Available at http://ec.europa.eu/eurostat/web/ess/-/new-validation-and-transformation-language-vt-1 (accessed on May 19, 2016).

[13] L. Breiman. Statistical modeling: the two cultures. *Statistical Science*, **16**, 199–231 (2001).

[14] M.H. Hansen, W.N. Hurwitz, and W.G. Madow. *Sample survey: methods and theory (7th ed.)*. Wiley, 1966.

[15] W.G. Cochran. *Sampling Techniques (3rd ed.)*. Wiley, 1977.

[16] C.-E. Särndal, B. Swensson, and J.H. Wretman. *Model assisted survey sampling*. Springer, 1992.

[17] C.-M. Cassel, C.-E. Särndal, and J.H. Wretman. *Foundations of Inference in Survey Sampling*. Wiley, 1977.

[18] S. Lohr. Multiple-frame surveys, in D. Pfeffermann and C.R. Rao (eds.) *Sample Surveys: Design, Mehods, and Applications*. Handbook of Statistics **29A**, pages 71–88, Elsevier, 2009.

[19] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury Press, 2002.

[20] W.E. Deming. *Some theory of sampling*. Wiley, 1950.

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

[22] J.N.K. Rao. *Small area estimation*. Wiley, 2003.

[23] J.N.K. Rao and I. Molina. *Small area estimation (2nd ed.)*. Wiley, 2015.

[24] J.T. Lessler and W.D. Kalsbeek. *Nonsampling error in surveys*. Wiley, 1992.

[25] F. Yates. *Sampling methods for censuses and surveys (3rd ed.)*. Charles Griffins, 1965.

[26] Eurostat. EU labour force survey - methodology. Eurostat, 2016. Available at http://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey_-_methodology (accessed on May 19, 2016).

[27] E Rosa-Pérez. Improving the statistical process in the hotel occupancy survey. *To be presented at Q2016 - European Conference on Quality in Official Statistics*, Madrid, June 1-3, 2016.

[28] D. Nedyalkova, L. Qualité, and Y. Tillé. Sampling procedures for coordinating stratified samples: methods based on microstrata. *International Statistical Review* **76**, 368–386 (2008).

[29] J.H. Saltzer and M.F. Kaashoek. *Principles of computer system design: an introduction*. MIT Press, 2009.

[30] Wikipedia. Waterbed theory. Available at https://en.wikipedia.org/wiki/Waterbed_theory (accessed on May 19, 2016).

[31] G.M. Weinberg. *An introduction to general systems thinking*. Weinberg and Weinberg, 2011.

[32] K. Washington, J. Burton, and R. Detlefsen. Frame construction and sample maintenance for current economic surveys. *Proc. ASA Section on Survey Research Methods*, 2010.

[33] D. Haziza, G. Chauvet, and J.-C. Deville. A note on sampling and estimation in the presence of cut-off sampling. *Australian and New Zealand Journal of Statistics* **52**, 303 –319 (2010).

[34] J.D. Nichols. Capture-recapture models. *Bioscience* **42**, 94–102 (1992).

[35] R.M. Groves. *Survey errors and survey costs*. Wiley, 1989.

[36] T. de Waal, J. Pannekoek, and S. Scholtus. *Handbook of statistical data editing and imputation*. Wiley, 2011.

[37] C.-E. Särndal and S. Lundström. *Estimation in Surveys with Nonresponse*. Wiley, 2005.

[38] K. Wolter. *Introduction to variance estimation (2nd ed.)*. Springer, 2007.

[39] United Nations Statistical Division. *Backcasting Handbook*. United Nations, 2013. Available at http://unstats.un.org/unsd/class/intercop/training/ece13/ac258-Bk4-e.PDF (accessed on May 19, 2016).

[40] European Statistical System. ESS guidelines on seasonal adjustment (2015 edition). Eurostat, 2015. Available at http://ec.europa.eu/eurostat/documents/3859598/6830795/KS-GQ-15-001-EN-N.pdf (accessed on May 19, 2016).

[41] A. Hundepool and J. Domingo-Ferrer and L. Franconi and S. Giessing and E.S. Nordholt and K. Spicer and P.P. de Wolf. Statistical disclosure control. Wiley, 2012.

[42] INE. Estándar de documentación de procesos de producción de operaciones estadísticas del INE (in Spanish). INE, 2015. Available at http://www.ine.es/clasifi/estandar_procesos.pdf (accessed on May 19, 2016).

[43] Federal Committee on Statistical Methodology. Data Editing in Federal Statistical Agencies. 1990. Available at http://fcsm.sites.usa.gov/files/2014/04/spwp18.pdf (accessed on May 19, 2016).

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

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

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

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

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

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

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

[51] I. Arbués, P. Revilla, and D. Salgado. Optimization as a theoretical framework to selective editing. *UNECE Work Session on Statistical Data Editing*, WP1, 2012.

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

[53] A. Salvador and D. Salgado. A generalization of the stochastic optimization approach to selective editing. *In preparation*.

[54] D. Hedlin. Local and global score functions in selective editing. *UN/ECE Work Session on Statistical Data Editing*, WP 31, 2008.

[55] R. López-Ureña, M. Mancebo, S. Rama, and D. Salgado. 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*, pages 1–9, 2013.

[56] R. López-Ureña, M. Mancebo, S. Rama, and D. Salgado. Application of the optimization approach to selective editing in the Spanish Industrial Turnover Index and Industrial New Orders Received Index survey. *INE Spain Working Paper 04/14*, pages 1–20, 2014.

[57] D. Salgado. Exploiting auxiliary information: selective editing as a combinatorial optimization problem. *INE Spain Working Paper 04/11*, pages 1–38, 2011.

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

[59] Wikipedia. Attribute-value pair. Available at https://en.wikipedia.org/wiki/Attribute%E2%80%93value_pair (accessed on May 19, 2016).

[60] H. Wickham. Tidy data. *Journal of Statistical Software* **29**, 1–23 (2014).

[61] Wikipedia. Lexical analysis. Available at https://en.wikipedia.org/wiki/Lexical_analysis (accessed on May 19, 2016).

[62] E. Esteban, P. García-Segador, S. Saldaña, and D. Salgado. *RepoTime: Implementation of a notation for time intervals*, 2016. R package version 0.1.0. Available at https://github.com/david-salgado/RepoTime (accessed on May 19, 2016).

[63] E. Esteban, P. García-Segador, S. Saldaña, and D. Salgado. *StQ: Construction of Standard Questionnaires*, 2016. R package version 0.1.0. Available at https://github.com/david-salgado/StQ (accessed on May 19, 2016).

[64] E. Esteban, P. García-Segador, S. Saldaña, and D. Salgado. *RepoReadWrite: Reading and writing files of the microdata repository*, 2016. *Package RepoReadWrite*. Available at https://github.com/david-salgado/RepoReadWrite (accessed on May 19, 2016).

[65] M Dowle, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan. *data.table: Extension of Data.frame*, 2015. R package version 1.9.6.

[66] G. Booch, R.A. Maksimchuk, M.W. Eagle, B.J. Young, J. Conallen, and K.A. Houston. *Object-oriented analysis and design with applications (3rd ed.)*. Addison-Wesley, 2007.

[67] OMG. OMG Unified Modeling Language (UML) v2.5, 2015. Available at http://www.omg.org/spec/UML/2.5/PDF/ (accessed on May 19, 2016).

[68] OMG. Business Process Model and Notation v2.0, 2011. Available at http://www.omg.org/spec/BPMN/2.0/PDF/ (accessed on May 19, 2016).

[69] M. van der Loo and E. de Jonge. *validate: Data Validation Infrastructure*, 2016. R package version 0.1.4.

[70] U. Guarnera and M.T. Buglielli. *SeleMix: Selective Editing via Mixture models*, 2013. R package version 0.9.1.

[71] S. Rama and D. Salgado. Standardising the editing phase at Statistics Spain: a little step beyond EDIMBUS. *INE Spain Working Paper 05/2014*, 2014.

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

[73] W. Radermacher. *Statistics 4.0*. Eurostat, 2015. Opening address at the Conference NTTS2015, Brussels, March, 2015.

[74] D.A. Dillman. Navigating the rapids of change: some observations on survey methodology in the early 21st century. *Public Opinion Quaterly* **66**, 473–494 (2002).

[75] F. Reis. Big Data in Official Statistics - progress at EU level. Talk presented at the Conference Webdatanet 2015, Salamanca. Accessed on 23 June, 2015.

[76] K.P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2013.

[77] N. Robin, T. Klein, and J. Jütting. Public-private partnerships for statistics: lessons learned, future steps. *OECD Development Co-operation Working Papers* Issue 27, 1–39 (2016).

[78] L.A. Wolsey. *Integer Programming*. Wiley, 1998.

[79] K. Lange. *Numerical analysis for statisticians*. Springer, 2010.

[80] T.G. Kolda, R.M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review* **45**, 385–482 (2003).

[81] INE. *Clasificación Nacional de Ocupaciones 2011*. INE, 2011. Available at http://www.ine.es/jaxi/menu.do?type=pcaxis&path=/t40/cno11&file=inebase (accessed on May 19, 2016).

[82] ILO. *International Standard Classfication of Occupations*. International Labour Organization, 2008. Available at http://www.ilo.org/public/spanish/bureau/stat/isco (accessed on May 19, 2016).