Working Papers

01/2018

Data organisation and process design based on functional modularity for a standard production process

E. Esteban, M. Novás, S. Saldaña, D. Salgado, L. Sanguiao

First draft: June 2018

This draft: June 2018

Data organisation and process design based on functional modularity for a standard production process

Abstract

We propose to use the principles of functional modularity to cope with the essential complexity of statistical production processes. Moving up in the direction of international statistical production standards (GSBPM and GSIM), data organisation and process design under a combination of object-oriented and functional computing paradigms are proposed. The former comprises a standardised key-value pair abstract data model where keys are constructed by means of the structural statistical metadata of the production system. The latter makes a profuse usage of the principles of functional modularity (modularity, data abstraction, hierarchy, and layering) to design production steps. We provide a proof of concept focusing upon an optimization approach to selective editing applied to real survey data in standard production conditions at Statistics Spain (INE). Several R packages have been prototyped implementing these ideas. We also share diverse aspects raising from the practicalities of the implementation.

Keywords

Production Architecture, Key-value Pair Data Model, Standardisation, Functional Modularity, Process Design

Authors and Affiliations

E. Esteban, M. Novás, S. Saldaña, D. Salgado, L. Sanguiao

Depto. Metodología y Desarrollo de la Producción Estadística

Instituto Nacional de Estadística

Data organisation and process design based on functional modularity for a standard production process

E. Esteban, M. Novás, S. Saldaña, D. Salgado, L. Sanguiao

June 7, 2018

Abstract

We propose to use the principles of functional modularity to cope with the essential complexity of statistical production processes. Moving up in the direction of international statistical production standards (GSBPM and GSIM), data organisation and process design under a combination of object-oriented and functional computing paradigms are proposed. The former comprises a standardised key-value pair abstract data model where keys are constructed by means of the structural statistical metadata of the production system. The latter makes a profuse usage of the principles of functional modularity (modularity, data abstraction, hierarchy, and layering) to design production steps. We provide a proof of concept focusing upon an optimization approach to selective editing applied to real survey data in standard production conditions at Statistics Spain (INE). Several R packages have been prototyped implementing these ideas. We also share diverse aspects raising from the practicalities of the implementation.

Keywords: Production Architecture, Key-value Pair Data Model, Standardisation, Functional Modularity, Process Design

1 Introduction

The modernisation and industrialisation of official statistical production has been in the centre of the international and national activity in Official Statistics basically since the turn of the century, with the creation of the High-Level Group for the Modernisation of Official Statistics by the Bureau of the Conference of European Statisticians being a noticeable landmark (HLG-MOS, 2017).

Indeed, this group was born with a clear strategic vision (HLG-MOS, 2011) to streamline the statistical production by means of "different and better processes and methods tuned to delivering our products at minimal cost with greater flexibility and in cooperation between institutions" so that these "new and better products and services [are produced] more tuned to the way the world is operating today". Many outputs have been produced by the different groups operating under the umbrella of the HLG-MOS ranging from the establishment of diverse production standards (like the Generic Statistical Business Process Model –GSBPM, the Generic Statistical Information Model –GSIM, the Common Statistical Production Architecture –CSPA, or the Generic Activity Model for Statistical Organizations– GAMSO) over the promotion and development of streamlined statistical methods (e.g. UNECE (2017a)) to capabilities and communication aspects (UNECE, 2017b).

More recently, within the realm of the European Statistical System (ESS hereafter), the future of European Official Statistics is strategically envisaged by the so-called ESS Vision 2020 (Eurostat, 2014a) and its implementation portfolio in key projects such as those focused upon the European System of Business Registers –ESBRs, the Common EU Data Validation Policy –VALIDATION, the Shared Services for European Statistics –SERV, and the Digital Dissemination and Communication –DIGICOM, to name a few (Eurostat, 2014b).

All these initiatives pose a challenge for statistical offices in their attempt to modernise their production, especially regarding the adoption of these new standards and practices: this is to be accomplished under the high pressure of product release calendars within the traditional stove-pipe production model and a decreasing amount of budgeted resources.

In this work we want to present the ongoing efforts at Statistics Spain (INE) to bring a concrete plan for the modernisation of (a part of) the statistical production process into reality. Our rationale is that an official statistical production system constitutes a clear example of a human-generated complex system. We claim that to cope with this complexity, like with the design of computer systems, the principles of functional modularity are also of great value. These principles must fully integrate statistical production metadata, statistical methodology, and computer software design. It is common practice to see the application of these principles in

the construction of software for the production of official statistics, but this is not enough. We claim that these principles must be applied *to integrate fully together these three aspects of statistical production*, otherwise we would be failing at coping with the complexity of the process. To illustrate our proposal we show how we have developed a set of R packages to make a proof of concept already applied in normal production conditions of several Short-Term Business Statistics (STS) at Statistics Spain (INE).

Our proposal is based upon two complementary elements. Firstly, for our data architecture we make use of a key-value pair structure in which keys are composed making a profuse usage of the system of structural metadata. Secondly, closely following GSBPM's and GSIM's principles, for our statistical process architecture we make use of the functional and object-oriented paradigms to incorporate modularity into the statistical methods. As we shall illustrate with the R packages, this paves the way for a natural posterior implementation in software tools. Our central message is thus *to bring modularity by design into the statistical process and the mathematical methodology itself and not just into the construction of computer tools.*

The paper is organised as follows. In section 2 we set up the generic approach taking us from complexity as an essential trait of statistical production systems to the principles of functional modularity to cope with it. In section 3 we argue that the international statistical production standards themselves implicitly suggest the use of a combination of the object-oriented and functional paradigms as the basis to build an information architecture. In section 4 we detail the abstract data model which we propose to use as the central element of our proposed data organisation. Complementarily, in section 5 we explain our proposed process design illustrating with an example in statistical data editing the application of modularity principles upon a very concrete statistical methodological approach to selective editing. In section 6 we share diverse aspects regarding the implementation of this proposal, including the software tools development. We close with conclusions and future prospects in section 7.

2 Generic approach: from complexity to functional modularity

The need for modernisation and industrialisation of official statistical production can be immediately argued from the very concept of *complex system*. The key features of a complex system are (Saltzer and Kaashoek, 2009) (i) a large number of components, (ii) a large number of interconnections between these components, (iii) many irregularities in these interconnections since the lack of regularity is indeed the rule rather than the exception, (iv) a long description of the system and its related management (so-called Kolmogorov complexity), and (v) a team of designers, implementers, and/or maintainers to handle the system. It is evident that an official statistical production system is indeed a clear example of a human-generated complex system.

This conclusion can be illustrated and motivated with a simple superficial description of the production of diverse statistical operations at a statistical office. Let us just consider the execution phases of the process. Data collection needs to be carried out in different data collection modes (CAPI, CATI, CAWI, EDI...) upon a number of statistical units, either business units or households or people, usually in the range of tens of thousands for each survey in a mid-sized country like Spain. This is to be multiplied by the number of variables (either data and metadata) associated to each unit. These data must be duly entered into the system, edited, treated, validated, and curated to produce the corresponding microdata sets. They are further processed to produce the aggregated outputs with the appropriate statistical methods and finally treated for disclosure control and also, if necessary, for seasonality and calendar effects adjustment before the due dissemination. Each production step and data and metadata element in the process is interconnected to some other element. For example, a change of a parameter in a validation rule during collection will need to be followed by a post-capture data editing revision and adjusted aggregation procedure (e.g. in variance estimation). Indeed, the interconnections between all elements cannot be described according to a given regularity thus making explicit the so-called *water-bed effect*: a slight modification of a process step may bring strong consequences in another process step. In the current setting of the statistical process at production offices, the description of how to produce the statistics for a given survey is not only necessarily long showing the imbricate set of process steps but also hardly standardised: two members of the production staff of two different surveys can rarely be interchanged to carry out even the same tasks in the process despite the common standard mathematical procedures underlying the whole estimation. Moreover, the number of actors in the process to be coordinated not only for a given statistical operation but especially for the set of surveys conducted at an office (not to mention a whole national or European statistical system) is very high, introducing evident management challenges.

In our view, the conception of official statistical production as the combination of statistics and complexity lies at the core of the need for the industrialisation of the statistical production process: not only do you need to use sound statistical methodology but you are also required to cope with this complexity in order to have an

efficient production process. Traditionally, in our view, official statistics have been produced in an artisan way in which each survey was independently designed and executed. Moreover, in extreme cases not only have been data and process architectures in different surveys in the same office diverse (occasionally even incompatible) but also within the same survey different agents have made use of unconnected architectures rendering the management of the whole process virtually impossible. Up to present times this so-called stove-pipe production model has been extensively followed.

On a more quantitative footing, the inefficiency of this stove-pipe approach can be also justified by the complex nature of the production system itself. As a complex system, it is subjected to the so-called square law of computation (Weinberg, 2011) (see also Saltzer and Kaashoek (2009)), which in our case can be expressed in terms of resources vs. number of requirements upon the system.

A simplified description of how to detect and correct errors in a process step can illustrate and motivate this law. A process step is basically a collection of both sequential and concurrent production tasks to accomplish a given objective within the process. We can easily assume that the potential number of errors is proportional to the size of the production step (i.e. to the number of tasks) and that they can occur randomly throughout the step. In principle, in a non-modular approach an error is detected after executing the process step, which is then fixed. The process step is then executed again to detect new errors. If the time to find an error is assumed proportional to the execution time, the total amount of time to clean the process step will be proportional to the number of errors times the necessary cleaning time per error, but the latter is proportional to the number of errors itself. Thus, the total amount of time will be quadratic in the number of errors. This argument shows how a naive sequential approach to production becomes unmanageable due to the complexity of the system.

Under this square law, it is clear that an increase in the number of requirements upon the system (because of the non-stopping demand on Official Statistics, e.g. new legal regulations, more disaggregated information...) will produce a quadratic increase in the demand of resources, which is unattainable. Complexity must be coped with to face these challenges. The need for modernisation derives from the complexity of the global statistical production process.

Now, the bottom line of our proposal: we believe that the common principles of computer system design jointly known as *functional modularity* (Saltzer and Kaashoek, 2009) are of great utility in designing and implementing an efficient official statistical production process. Let us remind that functional modularity comprises four elements, namely modularity, data abstraction, hierarchy, and layering. These principles should not only be applied to the development of computer tools: *it is the process itself which must be designed in these lines by conjugating statistical metadata, statistical methodology, and software design.*

Modularity is already at the very heart of production standards (such as the GSBPM – see next section) where the production chain is broken down into different subprocesses. However, modularity per se does not help us cope with complexity, we need data abstraction by which modules are designed and implemented independently of each other except for their interconnecting interface. Statistical processes must be designed independently of each other so that only initial inputs and final outputs do uniquely enter into play in the chained execution of a given set of processes. The details about the execution of each subprocess must be transparent in the whole process.

Layering and hierarchy are principles by which modules are designed and implemented to minimize the number of interconnections among their components seeking optimal efficiency. In our proposal these principles will be translated into organizing both data and process architectures into four layers. A bottom layer for the statistical methodology (purely mathematical in many but not all cases); a second layer for the finest-grained production tasks upon which more complex activities can be composed (third layer). Finally, a top layer to orchestrate the whole process with these elements will complete the process design. We insist on the idea that this structure *must be applied to the statistical processes themselves conjugating metadata, mathematics and software design*, not just to the construction of computer tools.

3 From metadata to architecture

The starting point to concretise our proposal into data organisation and process design is the interrelationship between the GSBPM and GSIM standards. The GSBPM is an international production standard modelling the statistical production chain in 8 phases, each one divided in different production subprocesses. This standard focuses upon production activities. Complementarily, the GSIM is another international production standard providing a model for the information objects in the production process. The inspiring interrelationship between

both standards is represented in figure 1 already originally appearing both in the GSBPM (UNECE, 2013a) and in the GSIM (UNECE, 2013b).

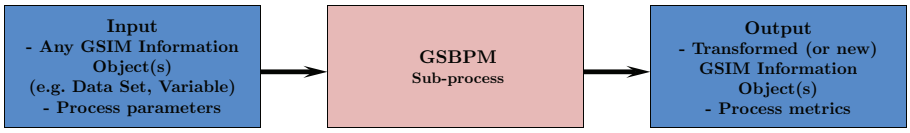

Figure 1: Interrelationship between GSBPM and GSIM standards (taken from UNECE (2013a)).

There is also an implicit reference to this interrelationship appearing in the name of the GSBPM level-2 subprocesses (*Design collection*, *Test production system*, *Calculate aggregates*...) with the clear structure *action + information object*. If several transformations matching figure 1 are concatenated, where the output of a step is the input of the next one, and if each transformation is associated to each input object, we indeed have the conception of a statistical production process as a sequence of objects defined through their attributes (GSIM-like information objects) and transformed according to their methods (GSBPM-like production tasks).

Our proposal suggests a step forward in this direction by profusely using the principles of functional modularity to substantiate this general view of the combination of both GSBPM and GSIM. Notice that these standards do not make any explicit mention to these principles, however their spirit is there. Similarly, in the international DDI standard (DDI, 2018) a modular scheme for the successive transformations upon both data and metadata sets is provided. Here, we also include under the same modular view these data and metadata.

To implement this dual data-process view under the principles of functional modularity we firstly need to provide a data organisation scheme to deal with information objects in a standard way. Indeed, the proposed scheme must be valid for all kinds of statistics (social surveys, business statistics, statistics based on administrative registers, etc.). In the next section we present an abstract data model based on key-value pairs in this sense. Indeed, we will define an object class for representing data in any kind of statistical data processing subprocess.

Complementarily, a process design scheme needs also to be provided. We understand that every "unit of statistical production information" is defined through a set of attributes (GSIM-like part) and a collection of statistical transformations (GSBPM-like part). In other words, they are *objects* (Booch et al., 2007). Furthermore, these objects can be thought of as constituting a sequence of transient transformations also combining data and metadata. This enables traceability and auditability of the whole process.

Indeed, this is extremely evocative of well-known computing models (van Roy and Haridi, 2004): the object-oriented and functional paradigms. Indeed, by making each transformation depend only on its object input they will become stateless, i.e. depending on no previous production step (state, in rigour[1]). This is a natural way of implementing referential transparency, i.e. a property by which the procedure can be replaced with its corresponding value without changing the behaviour and the result of the whole process. As a consequence, executing a referentially transparent subprocess will always provide the same value for the same input arguments irrespective of the rest of the process. This is the functional paradigm. As for the object-oriented paradigm, we concentrate on its advantages to model complex objects and on its characteristics regarding transformations. Thus, transformations are conceived under the functional paradigm and objects are understood and modelled following the object-oriented paradigm.

However, we need to be more concrete about how to combine these paradigms in statistical processes. Let us focus on the recommendations of the METIS group through their informal task force on metadata flows (ITFMF, 2013), in particular, to document each production task by different elements, namely (i) input data, (ii) input parameter, (iii) throughput, (iv) output, and (v) process metric. Indeed, these recommendations are closely followed in the Generic Statistical Data Editing Models (UNECE, 2015). In the present work we will leave out the fifth element about the metric. We propose the following structure for every data processing production task. We conceive every data processing production task as a transforming action upon a data set

[1] A cautious reader may immediately argue whether those steps involving (pseudo)random number generation arise as an exception to this stateless sequence of transient transformations. In full rigour, one can consider the random number generation seed as an internal state of the transformation. However, in the spirit of those statistical methods involving random simulation, we can accept that two processes providing *statistically* similar results can be considered identical under the data organisation and process design we defend here even despite numerical dissimilarities.

under a set of parameters producing a new data set or a new parameter set. We represent this as

$$\mathsf{OutputData},\ \mathsf{OutputParameters} := \mathsf{Action}(\mathsf{InputData},\ \mathsf{InputParameters})$$

It must be remarked that the distinction between data and parameter is somewhat arbitrary since it depends on the semantic context of the concrete computation. For example, in Predict(InputData, PredictParameters) we are computing predicted values for those data in the object InputData according to those parameters specified in the object PredictParameters, e.g. an ARIMA time series model ARIMA(p, d, q). Previously, we would need to compute the degrees p, d, and q. These can be computed similarly by PredictParameters := ComputeDegrees(PredictParameters, DegreeParameters), where an initialized parameter object PredictParameters is updated with the computed degrees and where DegreeParameters specifies the parameters needed to compute p, d, and q. Notice how in this second computation PredictParameters acts as an input data object.

This distinction about data and parameters can be also discussed in other common settings in standard production conditions. For instance, when joining two data sets we can consider both data sets as elements of a more complex InputData object and the join resulting from the parameters specified in the corresponding InputParameters object (inner, outer...). In the same vein, adding new records to an existing data set can be also modelled through a complex InputData object with an appropriate InputParameters object. Depending on the traceability and auditability provided to the whole system, the transient transformations can be further conveniently stored specifying timestamps, usernames...

All in all, functional modularity principles can be used to implement this combination of paradigms by setting up a hierarchy of layers going up from (i) the statistical methodology, over its implementation in (ii) low-level procedures (possibly assembled in libraries) and (iii) high-level procedures thereof, to (iv) a process-orchestrating layer working as a user interface.

Notice how this organisation in layers meets also different traditional profiles within statistical offices. The statistical methodology is under mathematicians' and methodologists' responsibility, possibly also with the collaboration of domain experts. This layer focuses on the more abstract and mathematical part of the production system. The second layer implements the methodology as low-level software procedures. It falls under developers' and programmers' responsibility, possibly with the collaboration of programming-skilled methodologists. This layer still keeps a certain degree of abstraction. It is in the third layer where concrete applications and production activities take form by means of statisticians' and survey managers' responsibility, possibly with the aid of developers. In this layer, the collection of standard low-level procedures adapts to the concrete needs of each statistical program. Finally, a process orchestrator working as user interface for ease of the human-computer interaction can be additionally put into place. This ease of use allows the management to optimize the production resources by potentially assigning tasks to non-specialists following previously specified protocols.

In the next sections we shall illustrate with concrete surveys conducted at Statistics Spain how this information architecture has been partially deployed for the statistical data editing phase. Our first step has been to propose a common data structure for all survey and administrative data sets (thus either InputData or OutputData) based upon a standardised abstract data model for any kind of statistics. This is detailed in section 4.

Next, we have implemented the optimization-based selective editing techniques formerly developed at Statistics Spain (Arbués et al., 2013) following these principles. This boils down to designing and programming Actions together with different sets of InputParameters (also OutputParameters). We undertake this in section 5.

4 Data organisation

We will use the Spanish Retail Trade Survey and Service Sector Indicators Survey monthly conducted at Statistics Spain to illustrate the application of this approach. These are short-term business statistics. Data are collected through paper questionnaires, telephone, fax, email, and CAWI modes. Statistical units are selected according to a stratified simple random sampling design. Target aggregates are mainly Laspeyres indices of both turnover and number of employees, possibly broken down into economic sector code and type of employment contracts, respectively.

In the preceding framework, our first task is to define an abstract data model for all statistical operations. The immediate goals of this model have been the versatility among all kinds of survey or administrative data and a fast and easy deployment in the implementation.

The model essentially consists of a key-value pair data model in which the key is composed by making use of the structural statistical metadata of the production system. We must distinguish between the data model for storing data in a corporative internal repository (the key is not parsed) and the data model for processing (the key is parsed). For manageability and rapid deployment reasons, in the current implementation the information is stored in plain text files, as explained below. These files are not modified once written. Updated information, if any, is included as a new file (with updated key in the name of the new file; see below). Concurrency issues and many other data architecture details are not considered relevant at this point.

The central element in the data model is the composition of the key for each single datum in the global production system at the office scale (or the whole statistical system scale). The key is composed of the following components:

(i) An alphanumerical code to identify the survey/statistical program.
This alphanumerical code is taken directly from the Spanish National Statistical Plan where each survey/statistical program is univocally identified. This code makes reference to the concrete statistics where this value is generated, processed, and used.

(ii) An alphanumerical code to identify the time period of reference (coincident with the time period of the corresponding statistics).
An ad-hoc simplified syntax has been put into place to denote the different reference time periods for all statistical operations according to the following table:

Time Period	**Code**
Month	MM, MR
Trimester	TT, TR
Semester	SS, SR
Year	AA, AR

The second character denotes whether it is an ordinary data set or a duplicated data set containing statistical units from the rotated sample. This is especially used in short-term business statistics making use of chain-linked Laspeyres indices with rotating panels.

(iii) An identifier to indicate whether they are raw or (partially) edited microdata, paradata, identification data...
The different codes are:

Data File Type	**Code**
Finally Validated Values	FF
Partially Edited Values	FD
Raw Values	FG
ParaData	FP
Identification Variable Values	FI
Edit Rules (Longitudinal phase)	FL
Edit Rules (Cross-sectional phase)	FT

(iv) A version number either with the prefix P for provisional or D for definitive values.

(v) An identifier for the statistical variable.
This identifier is taken from the system of structural metadata so that each concept measured with a statistical operation in the whole statistical production system is identified with a standard name. For example, the concept of "turnover" is measured in different surveys (industry, retail trade, service sector...) and the same identifier *Turnover* is used in every survey. Subtleties in this statistical variable arising from its concrete usage in a survey is further specified using qualifiers (see immediately below).

(vi) A set of qualifiers specifying different attributes (statistical unit ID, geographical code, economic activity code...).
Qualifiers are variables further specifying the semantic content of each value. Although from the strictly computer point of view all qualifiers play the same role, this is not the case from the statistical standpoint. There exist basically two types of qualifiers, namely, those allowing us to identify the statistical units and the rest of them. The latter can be further divided into two categories. Firstly, as in the example below, there exist qualifiers amounting to codes of standard classifications such as the NACE, PRODCOM,

COICOP.... To the extent feasible, at Statistics Spain (INE) international standard classifications are in use in agreement with the ESS. In parallel, not all qualifiers of this type can be found in standard classifications. In these cases, the metadata unit, in agreement with domain experts, is putting into place a collection of internal standard classifications for these qualifiers. For example, the number of employees in a business unit is an extensively requested variable usually broken down according to diverse criteria: by type of contract, by professional situation, by type of remuneration. These have given rise to respective classifications with their own codes, which are used as qualifiers in the corresponding key. Secondly, there exist qualifiers not possibly being understood as part of a classification. For example, the economic activity code of a business unit can possibly change because of a change in its business activity, so that this variable in the population frame should be modified after receiving the updated information during the field work. A qualifier (say, IsMod) denoting whether we are referring to the former value (IsMod=0) or the modified value (IsMod=1) must be introduced. This self-evident qualifier value is not part of a classification. More specific qualifiers can always be used according to the specific process under execution. For example, in statistical data editing qualifiers in terms of population, measurement time, measured unit, and measured element can be properly defined, coded, and used as qualifiers (van der Loo, 2015).

The following simplified example clarifies the meaning of these components. Let us consider the validated value of the turnover for a business unit (statistical unit ID 289409300MM) in the Retail Trade Survey (code E30103) in the reference time period of January 2016 in the region of Castilla-La Mancha (geographical code 08) in the economic sector of trade of food and beverages (NACE Rev.2 code 47.11). This value pertains to the first definitive data set for this time period. This is visually depicted in figure 2. Notice that some qualifiers are missing in this simplified example as structural metadata defining the variable type (integer value expressed in euros).

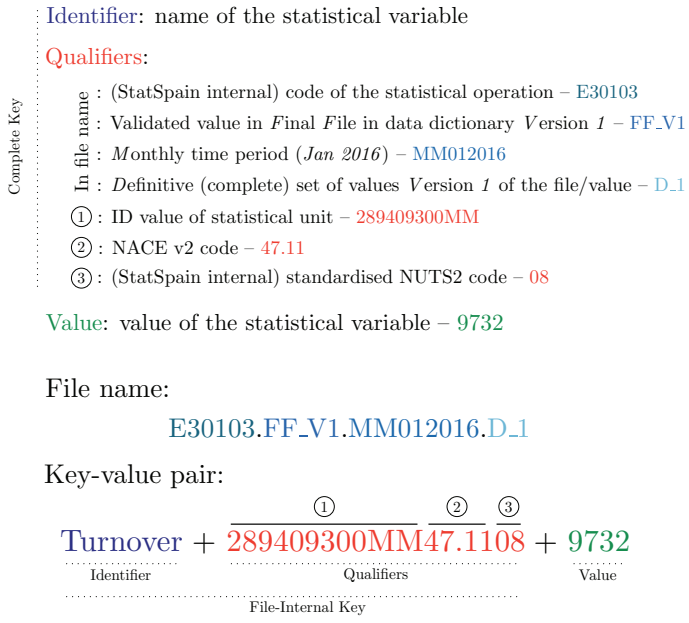

Figure 2: Example of a key-value pair with a key composed of structural statistical metadata.

As stated above, in the current implementation data are stored in files, each one identified by the statistical operation code, the type of data (finally validated data, raw data, paradata...), the reference time period, and the definitive or provisional character of the data within the production process. In other words, the common part of the key for a data set is encoded in the name of the corresponding file where the rest of the key and the values are stored. In each file each line will keep the standardised identifier and the rest of qualifiers together for each value (e.g. Turnover@@289409300MM47.1108@@9732 in our example). Other implementations are also possible.

A data dictionary is also configured and stored containing the specifications of each statistical variable: name, description, data type –numeric or alphanumeric–, maximal length –in terms of number of characters–, qualifiers, corresponding domain-used variable names, range of values and some other technical information for data collection applications. This dictionary allows the user to parse the key to instantiate objects according to a business logic class for all data processing tasks, which is indeed a data frame where the parsed key components are assigned in respective columns together with the corresponding value column. In this way data are

tidy in the sense of Wickham (2014) for further processing with standardised transformations[2]. This business logic class consists essentially of the data frame and the data dictionary. Data transformations are applied upon this class of objects returning updated objects of the same class.

Immediate benefits are obtained after adopting such a data organisation. Firstly, since every data of every survey/statistical program can be managed in this way, a unique data architecture can be adopted for the whole production system in the office. This is a first crucial step towards the suppression of the stove-pipe production model, paving the way for a more efficient architecture. Having a common data architecture allows us to build standardised applications valid for all surveys thus impinging on the rationalization of resources.

Secondly, these data specifications can be adapted to many actual circumstances in daily production. Let us consider for instance the case in which the economic activity code in the example changes along the process because the business unit has changed its activity. The example depicted here is oversimplified for ease of illustration. In practice the metadata system has dozens of standard classifications for qualifiers (always international when possible) to parameterise each single datum along the process. In particular, we have four classifications aiming at pinpointing (i) the process stage in which the value is generated (design, collection, processing, dissemination... or a subprocess thereof), (ii) the element of the process which the value is related to (frame population, sample, questionnaire... or a sub-element thereof), (iii) the role of the related actor in the process (statistical unit, interviewer, editing clerk...), and (iv) the type of value (dichotomic variable, excluding variable, percentage...). The evolution of the value along the process can be followed using these qualifiers. The metadata unit has put in place and is maintaining over 70 classifications and growing as more statistical programs incorporate this architecture. Many classifications are very specific for a given statistical domain but many others refer to common features to a large number of surveys.

Thirdly, the use of metadata in composing the keys to identify data values paves the way for achieving a standardised production system. In this way every single datum in the whole production process is parameterised using, so to say, a common system of coordinates. In contrast to the dangerously common opinion of only conceiving metadata as a cumbersome documenting tool independent of production tasks and effective only after production has been executed (so-called *passive metadata* according to Lundell (2013)), this data organisation makes use of the metadata system from the very beginning in which data are generated and providing an interface between data and the user (*active metadata* according to the same author). Notice how this active role of metadata is key in the sequence of transient transformations along the production process. Every independent transformation upon a given dataset must be implemented depending only on the input data and input parameters, i.e. on the data and metadata contents which transform according to the parameters. If metadata are erroneous, the interface between data and the user is lost, and the process (as a sequence of transformations) cannot be executed.

5 Process design

The design of the process architecture according to the principles set out in section 3 is much more complex than the design of the data architecture. To begin with, a standard class of parameters (InputParameter) for all possible statistical methods (Action) is virtually impossible since there exists a vast number of different statistical techniques. Thus, we will illustrate the application of the functional modularity principles with the concrete example of the optimization approach to selective editing developed at Statistics Spain (INE) (see Arbués et al. (2013)).

The division in layers begins by considering the statistical methodology at the bottom of the hierarchy. We will not go deep into the mathematical details and shall focus on the implementation of a very concrete formula to assign local (item) scores to each statistical unit.

The core of selective editing techniques is based on the assignment of a score to each variable to be edited for each statistical unit providing thus a measure of the degree of suspicion of it containing an influential measurement error. The heuristic approach (de Waal et al., 2011) recommends choosing local (item) score functions such as $s_k = \omega_k \cdot |y_k - \hat{y}_k|$, where ω_k stands for the sampling weight of unit k and y_k, $\hat{y}_k$ denote the reported and predicted (expected) values of the variable y under editing, respectively. The main methodological content of the optimization approach firstly consists of modelling the measurement errors $\epsilon_k = y_k - y_k^{(0)}$ ($y^{(0)}$ denoting the true value) for each unit and computing their first- and second-order moments M_{kl} for each pair of

[2]Tidy data mean Codd's 3rd normal form so that (i) each variable forms a column, (ii) each observation forms a row, and (iii) each type of observational unit forms a table (see Wickham (2014)).

statistical units k and l (business units in our example) and each variable y (turnover and number of employees in our example). These are given by analytical expressions (Arbués et al., 2013):

$$M_{kk} = \sqrt{\frac{2}{\pi}} \cdot \omega_k \cdot \hat{\nu}_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; -\frac{(y_k - \hat{y}_k)^2}{2\hat{\nu}_k^2}\right) \cdot \zeta_k \left(\frac{y_k - \hat{y}_k}{\hat{\nu}_k}\right), \tag{1}$$
$$M_{kl} = 0, \quad k \neq l,$$

for the loss function $L(a, b) = |a - b|$ and

$$m_k = \omega_k \cdot \hat{\nu}_k \cdot \frac{\hat{\sigma}_k^2}{\hat{\sigma}_k^2 + \hat{\nu}_k^2} \cdot \left(\frac{y_k - \hat{y}_k}{\hat{\nu}_k}\right) \cdot \zeta_k \left(\frac{y_k - \hat{y}_k}{\hat{\nu}_k}\right), \tag{2}$$
$$M_{kk} = \omega_k^2 \cdot \hat{\nu}_k^2 \cdot \left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}_k^2 + \hat{\nu}_k^2}\right)^2 \cdot \left[\frac{\hat{\sigma}_k^2 + \hat{\nu}_k^2}{\hat{\sigma}_k^2} + \left(\frac{y_k - \hat{y}_k}{\hat{\nu}_k}\right)^2\right] \cdot \zeta_k \left(\frac{y_k - \hat{y}_k}{\hat{\nu}_k}\right),$$
$$M_{kl} = m_k \cdot m_l, \quad k \neq l,$$

for the loss function $L(a, b) = (a - b)^2$, where in both cases

$$\zeta_k(x) = \frac{1}{1 + \frac{1-\hat{p}_k}{\hat{p}_k}\left(\frac{\hat{\nu}_k^2}{\hat{\sigma}_k^2 + \hat{\nu}_k^2}\right)^{-1/2} \exp\left(-\frac{1}{2}\frac{\hat{\sigma}_k^2}{\hat{\sigma}_k^2 + \hat{\nu}_k^2}x^2\right)}.$$

Exact details about the derivation of expressions (1) and (2) are given by Arbués et al. (2013). In the first case, when $|\frac{y_k - \hat{y}_k}{\hat{\nu}_k}| \to \infty$, $M_{kk} \to \omega_k |y_k - \hat{y}_k|$, which is the usual expression in the heuristic approach (de Waal et al., 2011) (in this case M_{kk} can be viewed as item scores, indeed). Thus, formulae (1) and (2) can be understood as a rigorous generalization of the traditional approach to selective editing by using statistical models for the measurement errors. Now the scores depend also on the other parameters like the probability of reporting an erroneous value and the variability of these errors reported in the past. As a matter of fact, statistical models for the measurement error are behind the diverse parameters in these expressions:

- ω_k denotes the sampling (design) weight of unit k;
- y_k denotes the raw (reported) value of variable y for unit k as collected in the questionnaire;
- $\hat{y}_k$ and $\hat{\nu}_k$ denote the predicted value and its prediction standard deviation for variable y and unit k;
- ${}_1F_1(x; y; z)$ stands for the confluent hypergeometric function of the first kind (Pearson et al., 2017), which arises from the choice of the loss function in the underlying optimization problem;
- $\hat{p}_k$ denotes the estimated probability of measurement error for variable y and unit k, i.e. $p_k = \mathbb{P}\left(y_k \neq y_k^{(0)}\right)$, where $y_k^{(0)}$ stands for the true value of variable y;
- $\hat{\sigma}_k$ denotes the estimated standard deviation for the observed measurement error $\epsilon_k = y_k - y_k^{(0)}$.

These quantities can be computed for the whole population or by population cells (e.g. determined by economic sector or geographical region or both).

Now we consider the second and third layers in which the statistical methodology is implemented in finer- and coarser-grained production tasks. From the methodology it is clear that the error moments can be written as functions of diverse parameters $M_{kl} = M_{kl}(y_k, \hat{y}_k, \hat{\nu}_k, \hat{\sigma}_k, \hat{p}_k, \omega_k)$. Now the question arises regarding how to organise this computation in a modular way.

It is at this point where functional modularity and statistical methodology must be precisely combined. From a strictly computational point of view, there is no distinction between the parameters y_k, $\hat{y}_k$, $\hat{\nu}_k$, $\hat{\sigma}_k$, $\hat{p}_k$, ω_k. However, from the statistical point of view this distinction is fundamental to allow for the system to grow and evolve in the future in an efficient way. Raw values y_k are taken directly from the data collection stage. Independent modules will take care of the computation of $\hat{y}_k$ and $\hat{\nu}_k$ (prediction module), of $\hat{\sigma}_k$ (observation error estimation module), of $\hat{p}_k$ (error probability estimation module), and of ω_k (sampling design module). The computation of these parameters will be completely independent of each another and each one will depend exclusively on its input arguments. They will interact with each other only through their final computed values

so that the computation is transparent.

This organisation in modules is justified by the underlying statistical knowledge. First, there exist many prediction methods potentially applied to obtain both $\hat{y}_k$ and $\hat{\nu}_k$. If new methods need to be added into the system, this will be possibly undertaken without affecting the rest of the computation. This same observation is valid for the rest of modules. Notice that this is a simple example in which we are computing a single value with an analytical formula with just 6 arguments. The consequences of a poor modular organisation from the methodological point of view along the whole production system may produce the opposite effect. This is why functional modularity and statistical methodology must be precisely combined in the design of the production system.

Now each module in turn makes use of these same principles so that different methodological aspects of the computation are considered independently. For example, due to missing values or some other reason predicted values cannot be computed for all statistical units and they must be imputed. An independent module for imputation is thus constructed taking care of this task independently of any other and embedded in the former computation. The architecture is again the same:

$$\mathsf{ImputedObject} := \mathsf{Impute}(\mathsf{InputObject}, \mathsf{ImputationParameters}).$$

The whole computation is then constructed as follows. Firstly the Action element specifying the concrete production task will be denoted by ComputeErrorMoment and it will implement either formulae (1) or (2) depending on its arguments.

As InputData we set all elements in expressions (1) and (2), namely (i) the values of the target variables y (turnover and number of employees in our example), (ii) some other auxiliary variables (e.g. those determining different population domains; economic classification NACE codes and Spanish geographical classification codes in our example), and (iii) the model parameters $\theta_k = (\hat{y}_k, \hat{\nu}_k, \hat{\sigma}_k, \hat{p}_k, \omega_k)$ for each variable y and each unit k. These are indeed the parameters for the continuous variable observation-prediction model (Arbués et al., 2013). We will call this InputData data set contObsPredModParam and it is given the key-value pair structure described in the preceding section. These parameters (hence the object contObsPredModParam) must be computed with their respective modules:

- The predicted values $\hat{y}_k$ and their standard deviations $\hat{\nu}_k$ are computed by initializing the object contObsPredModParam and defining an abstract class PredictionParam for the input parameter. The computation is undergone by updating the object contObsPredModParam:

  $$\mathsf{contObsPredModParam} := \mathsf{ComputePred}(\mathsf{contObsPredModParam}, \mathsf{PredictionParam}).$$

  The concrete statistical method used to compute $\hat{y}_k, \hat{\nu}_k$ is specified by defining a concrete subclass of PredictionParam. In our example, we have defined four time series models (random walks with regular, seasonal, and regular/seasonal differences and automatic ARIMA models), out of which that with lowest $\hat{\nu}_k$ is automatically selected. Any alternative choice (e.g. with machine learning techniques) could be easily implemented by defining the corresponding subclass. Hierarchy and layering principles are applied by internally constructing routines upon the key-value pair data structure in terms of simpler data structures such as vectors. In addition, imputation routines can be embedded in the design of these classes and methods as an attribute of PredictionParam.

- The estimated standard deviation $\hat{\sigma}_k$ of observation errors are computed in the same way:

  $$\mathsf{contObsPredModParam} := \mathsf{ComputeObsErrorSTD}(\mathsf{contObsPredModParam}, \mathsf{ObsErrorSTDParam}).$$

  In this case, another abstract class ObsErrorSTDParam has been defined, whose concrete subclasses determine the statistical method to use for the estimation. In our example, we have defined a subclass thereof to estimate σ_k by maximum likelihood using the historical double sets of raw and validated data. As before, imputation routines can also be embedded in the design of these classes and methods as an attribute of ObsErrorSTDParam.

- The estimated error probabilities $\hat{p}_k$ are also computed in the same way:

  $$\mathsf{contObsPredModParam} := \mathsf{ComputeErrorProb}(\mathsf{contObsPredModParam}, \mathsf{ErrorProbParam}).$$

  In this case, an abstract class ErrorProbParam is defined, whose concrete subclasses determine the statistical method to use for the estimation. In our example, we have defined a subclass thereof to estimate p_k by

maximum likelihood using the historical double sets of raw and edited data. Again, as before, imputation routines can also be embedded in the design of these classes and methods as an attribute of ErrorProbParam.

- The sampling weights ω_k are usually computed at an earlier stage of the production process so that we must only take them from some other data set of the survey at stake. In any other case, if explicitly needed for the editing phase, the computation of the sampling weights can be undertaken along similar lines.

Next, as parameters InputParameter in our error moments computation we essentially need to specify the loss function $L(\cdot, \cdot)$. We will denote this object by ErrorMomentParam.

Finally, the output object OutputData will be denoted by ErrorMoments and is basically an array of error moment matrices $[M_{kl}^{(q)}]$ per population cell (q denotes the turnover and the number of employees in our example). In this way we already have the content of each object in the expression

$$\mathsf{ErrorMoments} \coloneqq \mathsf{ComputeErrorMoment}(\mathsf{contObsPredModParam}, \mathsf{ErrorMomentParam}).$$

The whole computation at the third (scripting) layer is thus executed just by calling something like

$$\begin{aligned}
\mathsf{DD} &\coloneqq \mathsf{readFile}(\mathsf{DataDictionaryFile}) \\
\mathsf{contObsPredModParam} &\coloneqq \mathsf{buildcontObsPredModParam}(\mathsf{DD}) \\
\mathsf{PredictionParam} &\coloneqq \mathsf{buildPredictionParam}(\dots) \\
\mathsf{contObsPredModParam} &\coloneqq \mathsf{ComputePred}(\mathsf{contObsPredModParam}, \mathsf{PredictionParam}) \\
\mathsf{ObsErrorSTDParam} &\coloneqq \mathsf{buildObsErrorSTDParam}(\dots) \\
\mathsf{contObsPredModParam} &\coloneqq \mathsf{ComputePred}(\mathsf{contObsPredModParam}, \mathsf{ObsErrorSTDParam}) \\
\mathsf{ErrorProbParam} &\coloneqq \mathsf{buildErrorProbParam}(\dots) \\
\mathsf{contObsPredModParam} &\coloneqq \mathsf{ComputePred}(\mathsf{contObsPredModParam}, \mathsf{ErrorProbParam}) \\
\mathsf{SamplingWParam} &\coloneqq \mathsf{buildSamplingWParam}(\dots) \\
\mathsf{contObsPredModParam} &\coloneqq \mathsf{ComputePred}(\mathsf{contObsPredModParam}, \mathsf{SamplingWParam}) \\
\mathsf{ErrorMoments} &\coloneqq \mathsf{ComputeErrorMoment}(\mathsf{contObsPredModParam}, \mathsf{ErrorMomentParam})
\end{aligned}$$

In the construction of the diverse parameters objects the same hierarchical scheme can be followed (including e.g. the imputation routines). Notice also the far-reaching consequences for the organisation of the work and the production process at a statistical office. Firstly, survey managers and domain experts can work at a scripting level with high-level functions such as ComputePred, ComputeObsErrorSTD, and ComputeErrorProb above. This does not demand extensive IT skills and they can concentrate on the adapted use of these tools to their concrete survey. Indeed, the modularity allows them to seamlessly combine and choose diverse alternatives to compute the parameters and the error moments according to the characteristics of the statistical operation. On the other hand, developers and methodologists (data scientists, ideally) can work at a lower level implementing new statistical methods as new subclasses and overloaded methods. Needless to say, the communication between both layers for an optimal design of classes and methods is recommended. Notice however that both the naming conventions and the division in modules (both function and libraries) derives directly from the application of the foregoing principles being the statistical methodology behind the borders (interfaces) between the different modules. This paves the way for an easy application of standard good practices in software development, but supported by a strong mathematical basis. At the current development and implementation of our proposal, we can only offer an empirical view on this particular production stage (editing), but if these principles are to be applied throughout the process, the different functional modules should similarly interface with one another thus coping with complexity.

Secondly, this architecture favours software evolution and ease of maintenance over code preservation (Booch et al., 2007). Legacy code is recognised as a heavy ballast in the modernisation of statistical production. We are not providing solutions for the existing legacy code, but this architecture philosophy helps a great deal in not producing legacy code. The code can evolve according to new needs detected in the statistical programs by defining new subclasses and methods while at the same time the produced code is easily maintained because execution statements as above hardly change.

Thirdly, since statistical methods are implemented with an abstraction of concrete statistical operations, the same code at the lower level and highly similar at the scripting level is valid for different surveys. This allows us to optimally manage resources among statistical operations since the methodology and the computer tools

are standardised.

Fourthly, we would like to comment on the granularity of the services and the computer tools. In our example above, by starting with formulae (1) and (2), we also want to suggest that it is indeed the statistical methodology which should determine the degree of granularity of computer tools implementing the different methods. In the modular design, the statistical methods themselves should determine the natural borders among modules (hence also their interconnecting interfaces). Furthermore, the internal components of each module should also be structured according to the statistical methodology. In our example above, the reader can observe how each parameter entering into formulas (1) and (2) is dealt with using an independent method upon the object contObsPredModParam because each parameter can be computed/estimated choosing an adequate statistical method. Should new methodological proposals appear for a concrete computation, these can be easily incorporated without affecting the rest of software routines (like e.g. imputation routines).

Finally, we would like to underline how the scripting philosophy fits perfectly well in the GSDEMs as a processing step in which input statistical data and input metadata, process details, and transformed statistical data and output metadata are clearly expressed (UNECE, 2015). Although we have not yet undertaken the use of this process architecture to deal with process metrics, we are fully convinced that these monitoring parameters can also be computed along similar lines. This can be possibly carried out by complementing every computation or transformation upon an input dataset with a chosen set of indicators in the output monitoring the transformation.

To end, we must mention that along with the foregoing technical, mathematical difficulties, a highly relevant element in the practical implementation of this proposal arises in the staff reaction to changes in the production system. In the current stage of prototyping in production in a few statistical operations, the role of survey managers has been identified as key since in our current production model they take the decision on each survey. The gap between statisticians and computer scientists (and their traditional skills) also stands up as an aspect which needs further work.

6 Implementation: a proof of concept

The principles of functional modularity have been applied by designing and developing independent software packages for concrete aspects of these data organisation and process design. There are many aspects in the implementation worth sharing in order to be acquainted with the interplay between theoretical proposals and the practicalities arising within an already ongoing production system in a statistical office.

Firstly, since both object-oriented and functional paradigms are in the core of the proposal, the natural choice for a programming language is a language supporting these paradigms in a natural way without syntax quirks and twists. Java, C++, R, Python, Scala and many others are candidates fulfilling this condition. Since the user domain is clearly statistical data processing, another requisite is to feasibly develop trustworthy statistical tools in a very fast way. Finally, a good documenting system of classes, methods, and functions is also desirable allowing us to document data and parameter inputs, output, and throughput of each element (the process statistical metadata). Our choice has been R (R Core Team, 2012; Chambers, 2008).

Secondly, the methodology of software development has also been carefully decided. Instead of the more classical waterfall model (see e.g. Palmquist et al. (2013)), we have used a spiral approach (Boehm, 1988). Thus, instead of compiling specifications, designing, coding, and testing in a linear way, we have incrementally agreed on a first round of specifications, made a first design implemented on a first version of several R packages, and constructed a first version of the repository with key-value data files for 3 different short-term business statistics surveys. In this first round, the physical layer (the files themselves), the programming layer (classes, methods, and functions: the R packages), and the scripting layer were constructed. In a second round, apart from bugs and flaws in some functions detected in the testing phase, an important redesign was detected to be necessary in the classes and methods implementation[3]. This affected the second layer and interestingly enough it did not affect the scripting layer. Along this line of work, we pursue the production of constantly evolving pieces of software which can adapt fast and straightforwardly to the needs and changes of production. Again, this change of philosophy is at odds with the traditional culture in a statistical office and a formidable exercise of management for its implementation at the whole office scale is required. For example, the idea that computer

[3]The technical reason was that, for performance reasons to handle these key-value pair data sets, our packages heavily depend on the package data.table (Dowle and Srinivasan, 2016). Formerly we used the S4 system of classes and methods and the method dispatch, which suspends the lazy evaluation, is thus incompatible with the data.table syntax. We migrated all key-value data packages to the system S3.

tools built in this way are not finished and ready for use in production may be dangerous since it may lead to reject the methodology because the tools are immature. These more agile methodologies also allow us to make a more rational use of scarce resources, since the development is incremental. In our view, a change of mindset to conceive software as constantly evolving instead of as a closed definitive tool is necessary to industrialise and modernise the statistical production.

Thirdly, as a by-product of the preceding methodology, the communication between domain experts and survey managers, on the one hand, and developers and methodologists, on the other hand, must be clearly stressed. Although the architecture makes the work of both profiles independent by defining programming and scripting layers, the optimal design of the system will be obtained when the communication between both parts during the development stage is maximal. Again, we face a management challenge possibly impinging on organisational aspects of the whole statistical office (does it make sense to differentiate between statistical methodology and statistical software development departments?).

Fourthly, computer skills of the different actors must be taken into account. Two further actions have been taken in this sense to deploy the preceding architecture at Statistics Spain. On the one hand, the file containing the data dictionary is indeed an XML file for machine-readability. This technology is not usually part of the usual computer skills of domain experts and survey managers. Thus, to build this file we requested these statisticians to record the specifications of each statistical variable of their survey in an Excel file with a pre-specified structure. Excel files, although limited when dealing with some data structures, are easily handled. Then we programmed a specific function building the data dictionary file automatically from this Excel file.

Fifthly, the statistical computing system used as a standard at Statistics Spain is SAS and following this institutional policy, computing routines used by survey managers and domain experts must be written in SAS and not in other languages such as R, Python, Scala... Thus, the fourth layer working as a user interface has been developed as extremely simplified SAS macros executing the former R scripts in batch so that the interaction between the user and the architecture takes place only in SAS (so far this has been only accomplished to feed and read from the repository; the selective editing routines are executed directly by data collection staff in simplified R scripts). Although the functionality of the system is currently severely reduced and the rigidity increases, the ease of use is noticeable since the user only needs to specify a few very generic parameters.

Finally, the collection of packages in constant evolution at diverse degrees of maturity can be found in GitHub (Esteban et al., 2017a,b,c,d,e,f,g,h,i,j,k,l; Sanguiao, 2017). The architecture behind these packages closely follows the statistical methodology of the optimization approach to selective editing. Thus, it is difficult to give a precise description of what each package does without entering into mathematical content. A summarised description of what each package does can be found in Esteban et al. (2017m). It is important to point out that this division into many different packages focusing upon concrete aspects of the statistical process should not be read just as an example of good practices in programming, but as a consequence of the identification of functional modules according to the underlying statistical methodology.

7 Conclusion and future prospects

The main conclusion from this work is that recognising an official statistical production system as a human-generated complex system, the principles of functional modularity can be used to cope with this complexity to design both data and process architectures adapted and adaptable to the evolving needs of statistical production. Moving up a step forward in the direction of international standards we can combine the object-oriented and functional paradigms to define functional modules for the different production tasks whose borders and interacting interfaces are naturally determined by means of the underlying statistical methodology. These principles drive us genuinely to a set of layers from the statistical methodology, over its implementation in lower- and higher-level production tasks and steps to a top orchestrating user interface.

The data organisation revolves essentially around a key-value pair data model where keys are composed of statistical metadata. The process architecture implements transformations over information objects combining thus both paradigms. These architectures, in our view, bring relevant benefits for an efficient production system. They provide due roles for the different professional profiles in a statistical office, favour the evolution of software thus adapting to new needs, give rise to a complete global parametrisation of every single datum along the process, and bring forth standardization into the production tools among surveys and statistical programs of different nature.

Some of the elements presented in preceding sections are attached to the concrete production system at

Statistics Spain. It is advisable then to recognise those exportable elements to other offices. Regarding the data architecture, the core element is the use of metadata to identify values. The key-value pair structure could be substituted by alternative data models such as the SDMX or DDI. Nonetheless, in a deeper stage of analysis, performance issues (among others) should be taken into account to make a choice. In our case, we can process monthly data sets of around 2 million lines and about 15 qualifiers (around 28000 business units) to construct their corresponding traditional data matrices in less than 2 seconds. Regarding the process architecture, the core elements are (i) the application of functional modularity to statistical methods to produce modular computations respecting the natural borders in Statistics, (ii) the layers organizing the production tasks at different degrees of modularity, (iii) the use of object-oriented modelling for the information objects (both data and parameters), and (iv) the use of the functional paradigm to carry out the chained transformations upon these information objects. All other implementation details can be adapted to concrete circumstances.

Our proof of concept, nonetheless, reveals relevant challenges ahead. To be more efficient, an agile software development methodology should be preferred over more static methodologies. Also, the existing gap between methodologists/statisticians and computer scientists/developers must be urgently bridged. All this pushes us to increase the communication standards within the office among the different actors (methodologists, computer scientists, domain experts, survey managers, business managers...). This a remarkable management exercise.

In this line, being stakeholders and members of the ESS, alignment with international initiatives is strategic. Thus in future revisions and developments this alignment with CSPA services and European standards will be taken into account and pursued. Previously, technical requisites to be CSPA-compliant and to reach shareability of computer application must be agreed by the international community (see e.g. the 2017 meeting report of UNECE (2017a)).

Acknowledgments

J.M. Bercebal, from Statistics Spain's IT Department, made the original proposal to use a key-value pair data structure as a standard data structure for all statistical operations. A.-I. Sánchez-Luengo and her team is responsible for the use of Statistics Spain's structural metadata system, whose input has been essential. All tests and prototypes for this architecture have been undergone and constructed using real survey data from the Spanish Retail Trade Survey, the Spanish Service Sector Indicator Survey, and the Spanish Transport Passenger Statistic. E. Sánchez-Núñez and E. Rosa-Pérez and their teams are sincerely acknowledged for their data provision and subject matter expert input collaboration. We also acknowledge the invaluable comments and suggestions by two editors and four anonymous referees.

References

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

Boehm B. (1988). A spiral model of software development and enhancement. *IEEE Computer* **21**(5), 61–72.

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

Chambers, J.M. (2008). *Software for data analysis*. Springer.

DDI Alliance (2018). Data Documentation Initiative (2018). Available at `https://www.ddialliance.org/`.

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

Dowle, M. and Srinivasan, A. (2016). *data.table: Extension of 'data.frame'*. Available at `https://CRAN.R-project.org/package=data.table`. R package version 1.10.0.

Esteban, E., Saldaña, S., and Salgado, D. (2017a). *RepoTime: Implementation of a notation for time intervals*. Available at `https://github.com/david-salgado/RepoTime`. R package version 0.2.2.

Esteban, E., Saldaña, S., and Salgado, D. (2017b). *StQ: Tools to manage metadata-incorporated key-value pair datasets*. Available at `https://github.com/david-salgado/StQ`. R package version 0.4.34.

Esteban, E., Saldaña, S., and Salgado, D. (2017c). *RepoReadWrite: Read and write files from/to the microdata repository*. Available at `https://github.com/david-salgado/RepoReadWrite`. R package version 0.4.5.

Esteban, E., Saldaña, S., and Salgado, D. (2017d). *RepoUtils: Implementation of tools to map and work with repositories*. Available at `https://github.com/david-salgado/RepoUtils`. R package version 0.1.2.

Esteban, E., Saldaña, S., and Salgado, D. (2017e). *contObsPredModelParam: Class and methods for the parameters of a continuous observation- prediction model*. Available at `https://github.com/david-salgado/contObsPredModelParam`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017f). *StQPrediction: Define S4 classes and methods to make predictions*. Available at `https://github.com/david-salgado/StQPrediction`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017g). *StQImputation: Classes and methods to implement different imputation methods upon StQ objects*. Available at `https://github.com/david-salgado/StQImputation`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017h). *SelEditErrorMoment: Compute the conditional measurement error moments under the optimization approach to selective editing*. Available at `https://github.com/david-salgado/SelEditErrorMoment`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017i). *SelEditFunctions: Functions for selective editing*. Available at `https://github.com/david-salgado/SelEditFunctions`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017j). *SelEditUnitPriorit: Classes and methods to implement unit prioritization*. Available at `https://github.com/david-salgado/SelEditUnitPriorit`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017k). *TSPred: Point and std prediction of time series*. Available at `https://github.com/elisa-esteban/TSPred`. R package version 0.2.5.

Esteban, E., Saldaña, S., and Salgado, D. (2017l). *BestTSPred: Construction of objects of class BestTSPredParam*. Available at `https://github.com/elisa-esteban/BestTSPred`. R package version 0.0.1.

Esteban, E., Saldaña, S., and Salgado, D. (2017m). Software implementation of optimization-based selective editing techniques at Statistics Spain (INE). UNECE Work Session on Statistical Data Editing. The Hague, 24-26 April 2017.

Eurostat (2014a). ESS Vision 2020. Available at `http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020`.

Eurostat (2014b). Vision 2020 Implementation Portfolio. Available at `http://ec.europa.eu/eurostat/web/ess/about-us/ess-vision-2020/implementation-portfolio`.

HLG-MOS (2011). High-Level Group for the Modernisation of Official Statistics. Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics. *Conference of European Statisticians Geneva. 59th Plenary Session*. 14-16 June, 2011. Working Paper 1. Available at `https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2011/1.e.pdf`.

HLG-MOS (2017). High-Level Group for the Modernisation of Official Statistics. UNECE Statistics Wikis. Available at `http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics`.

Informal Task Force on Metadata Flows (2013). Metadata flows in the GSBPM. *Work Session on Statistical Metadata*. Geneva, 6–8 May, 2013. Working Paper 22.

Lundell, L.-G. (2013). Framework of metadata requirements and roles in the S-DWH. ESSnet on microdata linking and data warehousing in production of business statistics. Deliverable 1.1. Available at `https://ec.europa.eu/eurostat/cros/content/dwh-sga2-wp1-11-metadata-framework-statistical-data-warehousing-v112-final_en`.

Palmquist, M.S., Lapham, M.A., Miller, S., Chick, T., and Ozkaya, I. (2013). Parallel worlds: agile and waterfall differences and similarities. Technical Note CMU/SEI-2013-TN-021. Software Engineering Institute. Carnegie Mellon University. Available at `http://repository.cmu.edu/cgi/viewcontent.cgi?article=1761&context=sei`.

Pearson, J.W., Olver, S., and Porter, M.A. (2017). Numerical methods for the computation of the confluent and Gauss hypergeometric functions. *Numerical Algorithms* **74**, 821–866.

R Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Available at `http://www.R-project.org`.

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

Sanguiao, L. (2017). *Transformation of Standard Questionnaires*. Available at `https://github.com/Luis-Sanguiao/StQT`. R package version 0.1.0.9000.

UNECE (2013a). Generic Statistical Business Process Model. Version 5.0. Available at `http://www1.unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model`.

UNECE (2013b). Generic Statistical Information Model. Version 1.1. Available at `https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model`.

UNECE (2015). Generic Statistical Data Editing Models. Version 1.0. Available at `https://statswiki.unece.org/display/kbase/GSDEMs`.

UNECE (2017a). Statistical Data Editing Work Sessions. Available at `http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing`.

UNECE (2017b). Capabilities and Communication Group. Available at `http://www1.unece.org/stat/platform/display/MCOOFE/Capabilities+and+Communication+Group%3A+Home`.

van der Loo, M. (2015). A formal typology of data validation functions. UNECE Work Session on Statistical Data Editing. Budapest, 14-16 September 2015.

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

Weinberg, G.M. (2011). *An introduction to general systems thinking*. Weinberg and Weinberg.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software* **29**(10), 1–23.