Standard process steps based on standard methods as part of the business architecture[1]

Camstra, Astrea
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11
6412 EX, Heerlen, The Netherlands
E-mail: acma@cbs.nl

Renssen, Robbert
Statistics Netherlands, Division of Methodology and Quality
CBS-weg 11
6412 EX, Heerlen, The Netherlands
E-mail: rrnn@cbs.nl

Abstract

Over the past five years, Statistics Netherlands has been working on an ambitious program to redesign the statistical process. The general ideas of this program are represented in a comprehensive enterprise architecture. More recently, the architecture has been complemented by a series of standard methods and standard tools that should facilitate the design of the production process. Before standard methods or tools can be readily applied in the production of statistics, the statistical processes themselves need to be standardized to a certain extent. For this purpose, a conceptual business model for processing statistical data was developed. An important concept in this model is the standard process step. Standard process steps correspond to applications of statistical functions which can be implemented as business services. Statistical functions are usually based on standard statistical methods. By identifying these standard steps and providing guidelines for their use, the model aims to close the gap between the high-level view taken in the business architecture and designing statistical processes in practice. This paper briefly discusses the model and its concepts. The model will be illustrated by applying it to the field of data-editing.

1 Introduction

Statistical agencies are under constant pressure to improve efficiency and reduce reporting burdens, in particular for businesses. In addition, they are facing demands to maintain high quality standards, to enhance flexibility, and to focus more on rapidly changing user needs and product innovation. To meet these contrasting objectives, Statistics Netherlands has been working on an ambitious redesign program (see Braaksma 2009, for an overview). The general ideas of this program are embedded in comprehensive enterprise architecture (e.g., Huigen et al. 2009).

Recently, the architecture has been complemented by a series of standard methods and a set of standard tools that can be used to further streamline the core production process. The methodology series is a catalogue of approved statistical methods presently used at Statistics Netherlands, and is meant to inform and assist statisticians in applying the correct methods. Eventually, all statistical processes should only (re)use methods included in the series. The continuing high IT maintenance costs make it necessary to limit and standardize the diversity of tools and applications as much as possible. A study evaluating the current tools in

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands

statistics production (Renssen, Wings & Paulussen 2008) has resulted in a preliminary list of 18 preferred tools for the domain of statistical data processing, the two most important criteria for making the shortlist being the possibilities of the tool to deal with metadata and its ability to separate design from implementation.

Before standard methods or tools can be applied in the production of statistics, the statistical processes also need to be (partly) standardized. Renssen et al. (2009) give some initial ideas for the standardization of statistical processes. In Renssen (2010) these ideas are extended to a general conceptual business model for data processing. An important concept in this model is the so-called standard process step. Standard process steps are statistical functions (automated as generic components or building blocks) that are applied in a statistical process for a specific statistical purpose. Underlying these components are often statistical methods. There may be several different methods that can achieve the same functionality, for example hot deck and nearest neighbour are two methods used for imputation. Alternatively, a specific method can be applied for different functions. A regression method for example, can be used for imputing data values or for estimating population totals. The idea is to design processes in terms of (conceptual) standard process steps which then serve as the link between methods and tools. That is, rather than implementing methods directly the standard process steps are implemented in statistical processes using a limited number of standard tools. By creating a repository of standard process steps, the transparency and flexibility of the design process will increase and the reuse of building blocks will be facilitated.

Section 2 describes the relation of the model of Renssen (2010) to the business architecture. The model and its main concepts are summarized in Section 3. In Section 4 the model is applied to the field of data-editing. Based on an overall data-editing strategy, standard data-editing methods are further unravelled in a limited number of elementary functions and process steps. These building blocks can then be combined to form processes. Finally, Section 5 gives some conclusions and an overview of future research.

2 Relation to the Business Architecture

The business architecture outlines the ideal design of the statistical process. Often this takes the form of a value chain of activities that are administered to a statistical data set from data collection through data dissemination. Every activity in the chain adds value to the data being processed.

As an example we take the Generic Statistical Business Process Model (Vale, 2009), an architectural framework that is adopted by many statistical offices. The GSBPM distinguishes nine stages in the statistical process including “Collect”, “Process” and “Analyse”, and each stage in turn is divided into a number of subprocesses or activities.

- Collection stage: select sample, setup collection, run collection
- Processing stage: integrate data, classify and code, validate and edit, impute, derive variables and units, calculate weights and aggregate.
- Analysing stage: validate outputs, scrutinize and/or explain outputs, apply disclosure control

The (statistical) activities have a number of features (e.g. input, output functionality, owner) that are specific to the statistical process being considered.

This kind of classification is usually methodologically oriented, i.e., the subjects covered are mainly those methodologists have traditionally focussed on. Not surprisingly therefore, most of these activities also correspond to topics in the methodology series of Statistics Netherlands. The methods underlying the building blocks can be complex (e.g., estimate) but also simple (e.g., recode) or even trivial.

If we look at actual redesign projects carried out at Statistics Netherlands, the process descriptions frequently show a lot of variation. Although the GSBPM activities can often be more or less distinguished, it is not always clear what an activity entails and what it does exactly. Furthermore, the application of certain (complex) methods can lead to a series of other preparatory activities. For example, for "macro-editing” it may be necessary to match reference data first, in order to be able to match the data, a recoding step may be required etc. Sometimes the "preparatory" activities are implicitly assumed, in other cases explicitly

modelled. Also, some processing activities are not always recognized as applications of specific statistical methods (aggregation may be seen as a special form of estimation with weights equal to one). In addition, the process descriptions often contain several physical activities (e.g., retrieve files, select variables, and transform format) that are usually associated with the tools used in the process. Because processes are described in this relatively unstandardized manner, they often appear more varied and complex than suggested by an architectural framework and they provide insufficient insight into the possibilities of reusing parts of the process.

The conceptual business model (Renssen, 2010b) aims to give more structure to process design by defining standard process steps and by establishing explicit relationships with statistical methods and statistical goals. By identifying these standard process steps and providing guidance on their use, the model also seeks to bridge the gap between the high-level view taken in a business architecture and designing statistical processes in practice.

3 The conceptual model for data processing

The concept of a standard process step does not stand alone but only becomes meaningful in the context of our model for designing statistical processes. The model basically has three parts, 1) the operationalization of concepts, 2) the identification of statistical functions, and 3) the application of these functions in statistical processes. These three topics will be discussed in Sections 3.1 through 3.3.

3.1 Operationalization of concepts

The design of the statistical process begins with determining the needs for statistical information. These information needs must then be translated into statistical products. A statistical product is a realization of a well-defined subset of the information, and is uniquely determined by 1) the population of units being reported on, 2) variable definitions for the aspects or properties that are described and 3) the date or period to which the information relates (Renssen, 2010). The following specifications, for example, may apply for unemployment figures: the population consists of Dutch residents aged 15-65 years, “unemployment” is defined as “working less than 12 hours per week and actively looking for (more) work” and the reference period is a calendar month.

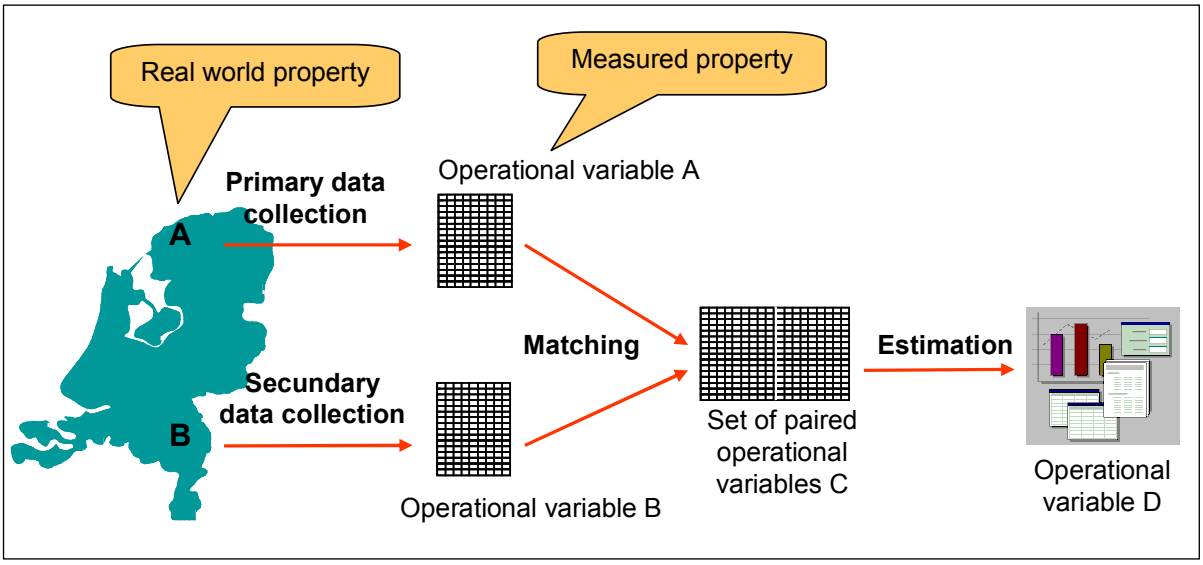
The property "unemployment" is called a conceptual attribute variable. To measure a conceptual attribute variable, it must first be translated into an operational attribute variable. An operational definition of a conceptual variable describes the actions required to measure the concept, including classification and measurement scales. For example, unemployment can be operationalized by means of one or more questions in a survey. An important difference between a conceptual and an operational attribute variable is that the former relates to a real-world property and the latter to a measured property. This also means that the value of an operational variable can be missing or erroneous. Finally, the physical or technical representation of the data results in a physical attribute variable (e.g. gender with outcomes 0,1,99 representing the categories male, female and unknown). The three variable levels correspond to different stages in the design process.

In the practice of producing statistics, the operationalization of a statistical concept is generally a multi-step process. In Figure 1 a very simple example is given. To publish figures on the total number of unemployed by age group (D), the property "unemployment" (A) is measured in a sample using the Labour Force Survey, while person characteristics such as age (B) are obtained from the population register. These measured operational variables are combined into a new measurement C, from which D is finally estimated. Due to data collection errors, sampling errors and/or matching errors the resulting estimate may differ from the conceptual total that had to be estimated.

Frequently, there are several (sets of) operationalizations to achieve the same statistical target (e.g., measurement of D). However, statistical strategies can limit the possible choices. These strategies usually

stem from policies for efficiently managing the statistical production processes and often favour a particular solution. The general strategy for data collection at Statistics Netherlands states that every effort must be made to reduce statistical reporting, meaning that primary data collection can only be conducted when there is no alternative (secondary) data source. Hence, in Figure 1 two data sources are shown instead of collecting all data in the survey.

Figure 1: Operationalization of a variable in several steps



3.2 Statistical functions

The series of operationalizations in Figure 1 already shows the outlines of the design of the statistical process. The next stage in the design is to elaborate these steps further. This means identifying the set of functions needed for each step to realize the end product from one or more input products. In Figure 1 the second step uses a matching function to obtain C from A and B. In the third step an estimation function is used. Matching and Estimation are examples of statistical functions. These statistical functions are in turn related to statistical methods. The estimation function, for example, is often based on a regression method.
Performing a particular step may require a number of preparatory actions. For matching two data sets it might be necessary to derive and/or encode a set of variables to obtain a unique key variable. This is modeled as applying successively a derivation function, a coding function and a matching function. To obtain a successful matching result several iterations of the matching function may also be needed, using different key variables in each iteration. A detailed description of the example of matching two data sets is given by Renssen and Camstra (2011).

As mentioned earlier, measurements of real world properties may contain errors or missing data. In addition to measuring the properties themselves, it is therefore important to provide information about the quality of these measurements. Corresponding to the levels of attribute variables levels, the model distinguishes between conceptual, operational and physical quality indicators. Furthermore, we consider quality functions as specific statistical functions that measure quality according to some statistical method or subject matter knowledge.

After identifying the required functions, the input, output and method of each function application need to be exactly specified. In Section 4 we will take a closer look at statistical functions, especially those involved in data editing.

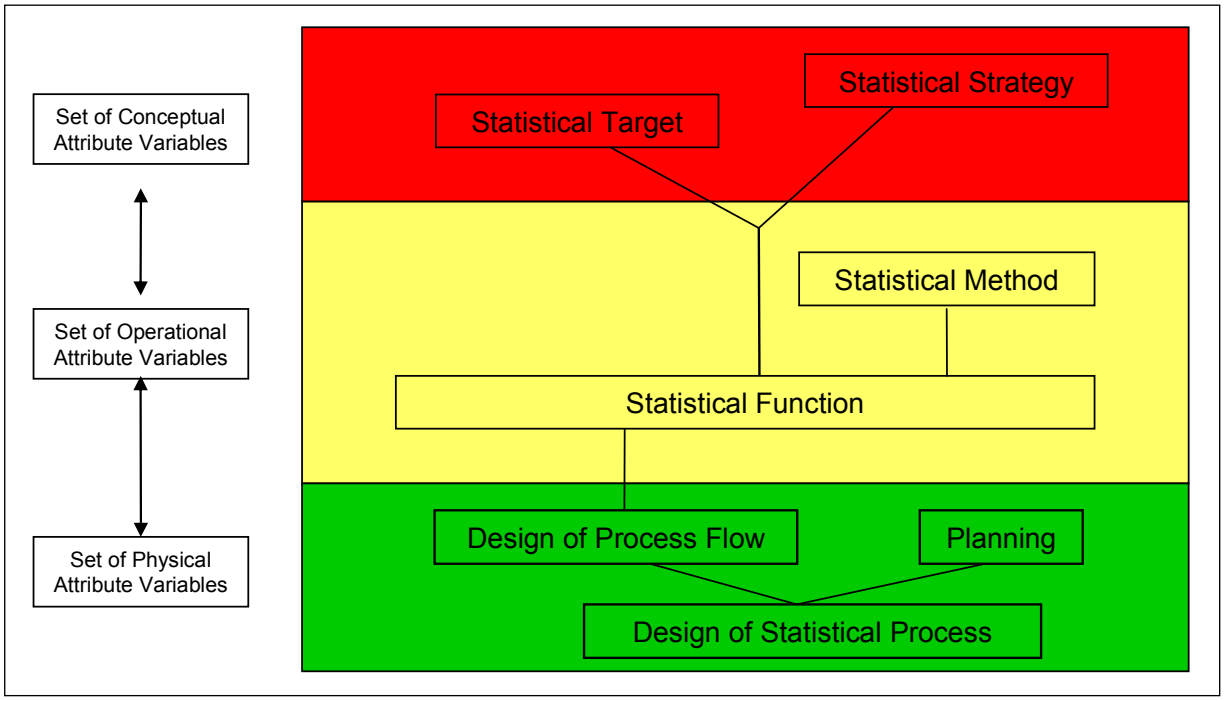
3.3 Designing the statistical process

The total collection of specified statistical functions and quality functions form the basis for the design of the process flow. The intended statistical (sub)goals of the specified functions should be realized in the process flow, but the (physical) implementation still leaves several choices.

In our model each function basically corresponds to an elementary standard process step in the process flow, or more precisely, the elementary process step is an application of a function. The order in which the functions are carried out is described in the process flow taking into account the dependencies between functions. The (same) matching function can sometimes be applied for multiple purposes (e.g. to delineate the population or for data editing). One possible choice is to apply this function only once, for example at the beginning of the process. It may also be useful to combine certain functions. For example, data validation and correction functions are often used in a fixed combination. We speak of a composite standard process step when a combination of functions is used for one or more (sub) goals.

Sometimes the application of a method in a function is so trivial that the function is no longer recognizable as such in the process flow but implicitly assumed. For example, estimating population totals based on register data can be regarded as the (trivial) application of a weight function with weights equal to one. However, in modelling the functions it is often useful to specify the trivial functions as well, just to show that a choice for a trivial method was made. Moreover, the process flow contains non-statistical functions such as data transformations and decision functions that indicate when to start and end the application of a function. The main concepts of the model are shown schematically in Figure 2.

Figure 2: Designing a statistical process



4 The model applied to data editing

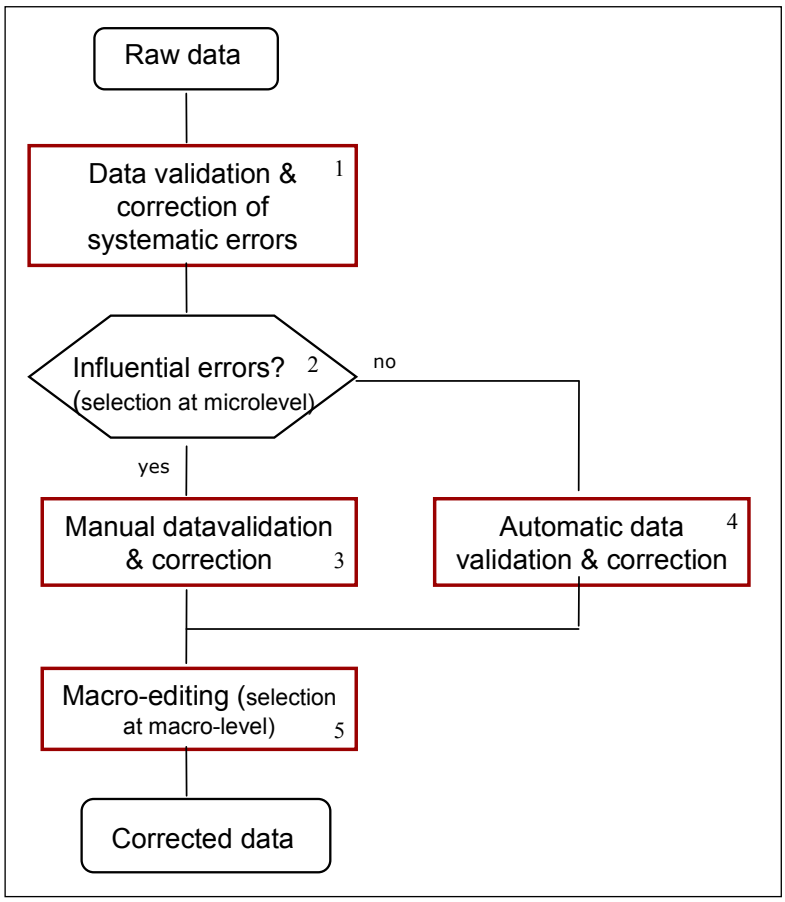
In this section we will apply the main concepts of the model described above to the processing domain of data editing. Section 4.1 discusses the strategy for data editing used at Statistics Netherlands, Section 4.2 explains the concept of a statistical function, while in Section 4.3 an example of a very simple data editing process is given.

4.1 Strategy for statistical data editing

The subject "Data Validation and Correction" of the methodology series (Hoogland et al, 2010) discusses the data editing techniques most frequently used at Statistics Netherlands. The authors describe an overall strategy for the data validation and correction process as shown in Figure 3. The specific way this strategy is applied for different statistical processes may vary and not all steps have to be completed.

In the first step of the data validation and correction process "obvious" systematic errors are detected and corrected (1). An example of a systematic error is a 'thousand-error', i.e., a value that is wrong by a factor of 1000. If the systematic errors are known they can be easily corrected using deductive methods. The next step is to manually check and correct the data, also referred to as interactive data editing (2). Given the general data editing instructions, the subject matter specialist determines which data are wrong and how they should be corrected. Corrections are usually made on the basis of expert knowledge, possibly supplemented with reference data. Since manually checking and correcting data is costly and time consuming, interactive editing is often restricted to influential errors that cannot be reliably solved automatically. This is called selective data editing (3). This technique assigns scores to data values indicating the expected impact on publication figures if these values were to be manually corrected. High scores have a high priority to be examined interactively. The remaining less important errors can subsequently be edited automatically (4). A set of data values of a unit are checked against a predefined set of edit rules and the wrong values(s) are automatically located. At Statistics Netherlands, error localisation methods are frequently based on the Felligi-Holt paradigm. In the last step, provisional publication figures are estimated and compared with historical data or external data sources. If the estimated figures are implausible, the underlying micro-data are analysed further and corrected if necessary. This process of macro-detection and micro-correction is called macro-editing (5). Macro-editing can be interpreted as a form of selective editing where the selection of influential errors is made through the population estimates.

Figure 3: Data-editing strategy



4.2 Statistical functions involved in data editing

In the different techniques covered by the overall editing strategy described in the previous section, a number of basic data editing operations or functions can be distinguished. We define the following four editing functions.

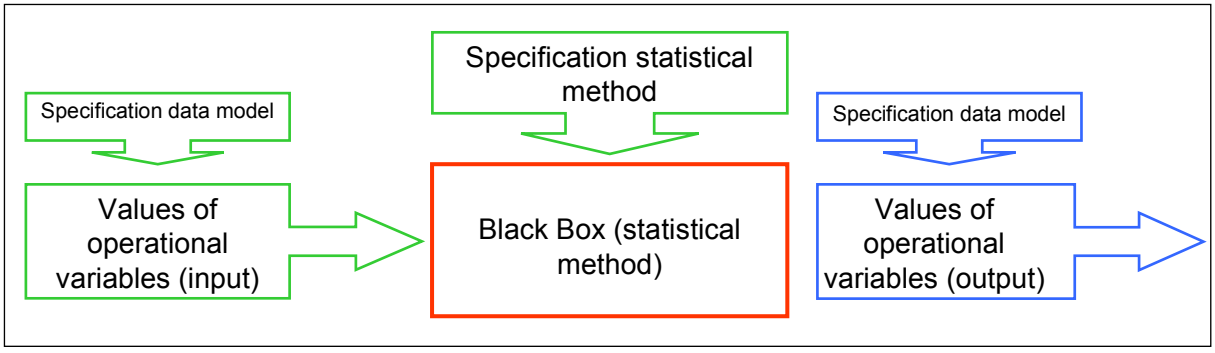
- Data validation function: checks variables for errors or inconsistencies.
- Error localisation function: determines, in case of inconsistencies, which variable is 'wrong' and needs to be corrected.
- Score function: identifies influential observations, i.e., observations that have a substantial impact on publication figures.
- Correction function: corrects the (located) errors or inconsistencies.

The validation, localisation and correction functions are all used in each of the editing techniques, although these functions are not always recognized as separate steps. For example, upon detecting a systematic error, it is automatically located and it can instantly be corrected by deduction. The score function stands somewhat apart and is mainly meant to divide the data into two data streams, one that needs to be inspected further and one that can be handled automatically. The score function operates at both micro and macro level.

The general structure of a function is taken from Renssen (2010), where a function is represented as a black box with input and output as depicted in Figure 4. The input consists of values of (operational) attribute variables and/ or quality indicators. The data model of the input contains the descriptions of the variables including the relevant population and reference period. Similarly, the output consists of new or improved values of operational attribute variables and/or quality indicators and the data model for the output comprises the variables. The method specification describes how the output is obtained from the input and thus establishes the internal process of the function.

The design condition for a statistical function implies that the variables that will be used according to the method specification must be present in de data model of the input, or in other words, the method specification must be consistent with the input data model. The data condition means that the input data must also be of sufficient quality.

Figure 4: A statistical function represented as a black box



As an example we take a closer look at the data validation function (see also Renssen and Camstra, 2011). We make a distinction between the nature of a function, i.e. its functionality and the types of input and output that are associated with it (see Figure 4), and the use of the function in different contexts.

For the data validation function the input consists of the values on a set of N variables, $a_1,\ldots,a_N$ with domains $D_1,\ldots,D_N$. The values are denoted by a. The variables are incorporated in the data model of the input. The edit rules, or edits for short, constitute the method specification. There are different ways to specify the edits. For categorical (discrete) variables or combinations of categorical and numerical variables IF-THEN statements are frequently used. An example is, IF *age* < 18 THEN *driver's license* = no. For numerical variables linear edits are often specified. Examples of edit rules for detecting thousand errors are, for

example *Turnover*(t) > 300 *x Turnover*(t-1) or *Turnover*(t) > 100 *x* stratum median *Turnover*(t-1). Note that these edits contain variables from a previous period (t-1).

If there are Q edits, the output for each unit consists of the values on Q quality indicators indicating whether the edit has been violated or not. The quality indicators make up the data model of the output. The validation function has as a design condition that the variables involved in the edits are part of the data model of the input. If the present data are compared to data from previous periods (t-1) or to data from registers these should be part of the input.

The specification of the edits follows from knowledge about the concept being measured or about the relationship between concepts. They can also result from knowledge about frequently made errors (systematic errors). Typical situations in which the data validation function is used are 1) to check the validity of values with respect to the domain of a variable, e.g., the variable gender can only assume the values male or female, 2) to check (conceptual) relations between two or more variables, e.g., 'men cannot be pregnant' and 3) macro checks which can be regarded as a special kind of relation edit.

We conclude this section with a few general observations about functions.

- A distinction is made between statistical and non-statistical functions. A statistical function changes an operational variable, while a non-statistical function only changes a physical variable (leaving its operational counterpart unaltered). A transformation function changing the format of the data is an example of a non-statistical function. When designing a process both types of functions are needed, but only statistical functions 'improve' or add value to the statistical data. This is also one of the reasons why process descriptions in practice differ from high level architectural models.
- The internal process of a statistical function to process the input into output is, often based on a statistical method. As mentioned in the introduction, there is often more than one method for the same function. This means that there may be several methodological implementations and hence standard process steps with respect to the same function.
- Statistical functions can be used for different statistical purposes. The function itself, however, does not have knowledge about a specific use and can be used at several points in the statistical process. An estimation function may be used to efficiently validate the micro-data in a macro-editing editing process, or to estimate population totals for publication. For both purposes the estimation function could have the same specification e.g., the same specification of data models and method, but that is not necessary.

4.3 A simple data editing process

In Figure 5 an example of a very simple data editing process is given. The goal is to obtain a set or file of edited data which, by design, contains the original values a (optional), the values on two quality indicators, $q1$ and $q2$, and the edited data values a'. First the raw data are checked against an edit rule, resulting in $q1$ to indicate whether the edit has been passed or failed. Subsequently the error is located, resulting in $q2$, which points out which of the variables in a is wrong, and finally the located error is corrected, yielding a'.

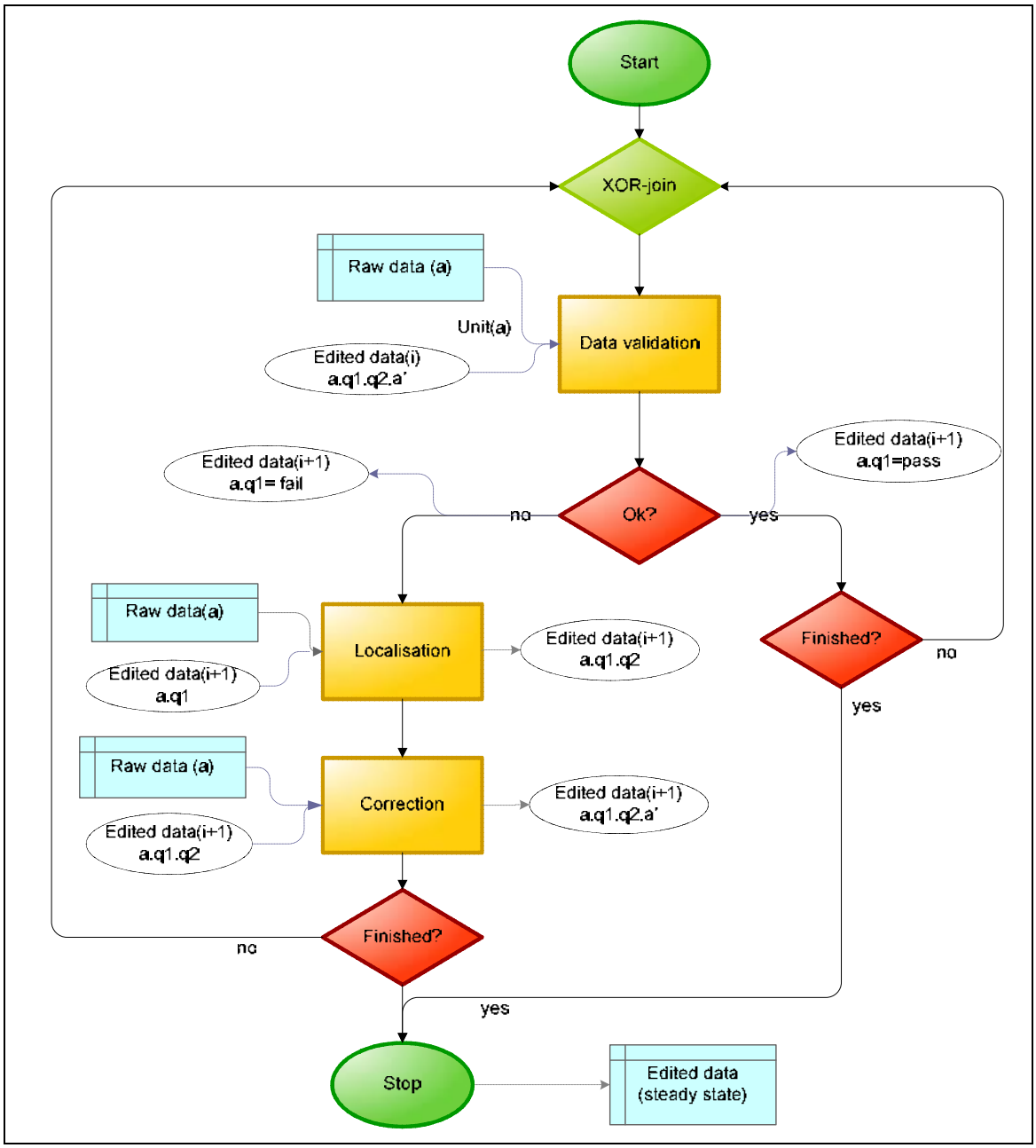
The process is modelled as event driven and the data for the units ('records') are handled one by one. The process is started by a trigger, for example a questionnaire for unit i has been received. When the edit is passed the record is written to a 'running file' of edited data ($a' = a$) and the process waits to receive the next questionnaire $i+1$. When the edit is violated, the record goes through the localisation and correction process, and at the end the corrected data is written to the running file (a'). When the data of all the units are processed or when a certain amount of time has elapsed the process stops (end trigger) and the data can be stored as a 'steady state'. Steady states are a main feature of the architecture of Statistics Netherlands (Braaksma, 2009) and contain data and metadata in an explicitly described state of processing with predefined quality.

Figure 5 shows the difference between a process and the application of a function. Even for a simple process several design decisions have to be made. Usually the data are checked against a (large) number of

edit rules at once. The method specification for the data validation function then consists of a bundle of (similar) rules that can be versioned together.

Applications of the individual functions correspond in principle to elementary standard process steps that can be implemented separately with different tools. In automatic data editing, for example, different tools may be required for carrying out a large set of checks and for locating errors using complicated mathematical algorithms. Alternatively, when the three functions are applied in the same fixed combination, the whole process of Figure 5 can be considered as a composite standard process step.

Figure 5: Process flow of a simple data-editing process (event driven)



5 Future research

The research on standardization of statistical processes currently focuses on developing statistical functions, starting with the functions involved in data editing. Besides giving detailed descriptions of the elementary data editing functions, we look at how these functions can be applied in statistical processes and what is the best way to define standard process steps. In the next stage the remaining functions will be considered. At present, we have defined about 15 functions that are most frequently used in statistical processes (see Renssen and Camstra, 2011). In another line of research the conceptual model discussed in this paper is applied to complete statistical production chains. The process of the Short Term Statistics was modelled using the framework both at the functional level and the process level. This has shown that our approach could be useful for standardizing the designs and descriptions of statistical processes, as well as providing insight into the relationship with the methodology series.

The purpose of the present research is primarily descriptive and we are not directly concerned with the physical implementation of standard process steps and the relation to the standard toolbox. This will be addressed at a later stage in collaboration with the IT-architects. The ultimate goal is to build a repository of standard process steps or 'building blocks', which form the basis for designing statistical production processes.

REFERENCES

Braaksma. B. (2009). Redesigning a statistical institute: The Dutch case. In: Proceedings of MSP2009, workshop on Modernisation of Statistics Production 2009.

Hoogland, J., Van der Loo, M., Pannekoek, J., en Scholtus, S. (2010). Methodology series, theme Data validation and correction. Internal report (in Dutch), Statistics Netherlands, The Hague.

Huigen, R., Bredero, R., Dekker, W. and Renssen, R. (2009). Statistics Netherlands Architecture; Business and Information model. Statistics Netherlands discussion paper 09018.

Renssen, R., Morren, M., Camstra, A. and Gelsema, T. (2009). Standard processes. Discussion paper 10013, Statistics Netherlands, The Hague.

Renssen, R. (2010). Basic principles of a conceptual business model for data processing. Internal report (in Dutch) Statistics Netherlands.

Renssen, R. and Camstra, A. (2011a). The data validation function. Internal report (in Dutch), Statistics Netherlands, Heerlen.

Renssen, R. & Camstra, A. (2011b). Standard process steps in statistics. Meeting on the Management of Statistical Information systems (MSIS 2011).

Renssen, R. Wings. J. and Paulussen, R. (2008). Processes, methods and tools. Internal report (in Dutch), Statistics Netherlands.

Vale, S. (2009). Generic Statistical Business Process Model, Version 4.0, Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS). Eurostat, Luxembourg.