Quality assurance from an internationally standardized and generic data validation ecosystem

Olav ten Bosch, Statistics Netherlands, o.tenbosch@cbs.nl

Mark van der Loo, Statistics Netherlands, mpj.vanderloo@cbs.nl

Abstract

One of the challenges in official statistics is monitoring data quality. Data characteristics observed when designing statistics may not hold during production in the long run. If this is not detected in time, this may lead to errors or costly recalculations. Therefore data has to be validated before being used. This holds for data processing within a statistical institute as well as for data exchange among statistical organizations.
On the International level data validation rules have been agreed upon in International statistical working groups. Eurostat performed an analysis on these rules to identify the so-called 'main types of validation rules' that cover the majority of international statistical processes. Some of these rules use metadata described in the International SDMX standard.
On the National level data validation is incorporated in the national statistical production processes. NSIs increasingly use the R programming language in these processes and the R-package validate is a popular tool to validate data. It offers a language for defining validation rules supporting many type of data checks. The software can execute analyses on possibly large datasets, providing statisticians with feedback on the health of their data.
In the ValidatFOSS project a free and open source generic validation ecosystem has been developed that maximally aligns with ESS standards. The architecture is generic and can be used in any statistical domain, on any SDMX compliant registry and for data validation within organizations as well as for international data reporting. An online validation cookbook offers recipes for implementing the most common validation scenarios found in official statistics. In this paper we briefly present the results and philosophize about the role this building block can play to monitor data quality in the new emerging data ecosystem of official statistics.

Keywords: data quality, data validation, ESS validation rules, R, open source, SDMX

1. Introduction

Quality assurance is an important topic in official statistics. Within this broad subject monitoring data quality - especially the challenge to guarantee that consecutive data transfers adhere to minimal quality criteria - is not straightforward. One reason for this is that data quality tends to vary over time. Data characteristics observed when designing a statistical process may not hold forever. Inputs might change, processes can change, systems can change and they may have an effect on the data being produced. If these data flaws are not detected in time, this may lead to costly recalculations or – even worse – undetected errors. This is even more of a problem with new indicators that combine traditional data (surveys, registers) with other, more volatile, new data sources such as transaction

data, web data or spatial data. Therefore data has to be validated before being used in production. This holds for statistical processes within statistical organisations as well as for cross-organisational processes, such as data reporting from National Statistical Offices (NSIs) to international organisations as part of the European Statistical System (ESS).

From 2018, Statistics Netherlands executed a project called ValidatFOSS[1]. The goal is to deliver an open source ecosystem that facilitates data validation in all stages in the statistical process. In chapter 2 we describe some important concepts of international data validation. In chapter 3 we describe the data validation functionality already covered in the R world. Chapter 4 presents the setup of the generic open source data validation ecosystem that meets the international as well as national validation needs. Chapter 7 contains the conclusions and some further reflections.

2. Data validation in international context

To meet the increasing demands on data quality it is necessary to standardise and automate data validation where possible. There have been multiple international projects working on this, resulting in a handbook on validation (Methodology, 2018), a set of data validation principles (Eurostat TF, 2019), a standard validation report (van der Loo, 2017) and a list of common types of rules in the ESS (Eurostat/B1, 2018). As they play a major role in designing and realising a standardised and automated data validation ecosystem, we briefly explain the first validation principle and the ESS main types of rules. In addition we briefly dive into the existence and access to standardised international metadata in SDMX.

3.1. Validation principle 1: the sooner the better

In 2019 an ESS task force wrote down 6 principles for data validation[2]. Principle number one is stated "*the sooner the better*", which is to be interpreted as the vision that the sooner[3] errors are detected in a statistical production chain, the easier and more efficient it is to correct them. Applied to international data reporting this means that NSIs should check their data against international validation rules before sending them to international organisations. Of course, this can only be done if the validation rules are well-defined and understood, which is contained in principle 3. However checking data just before just sending to a data consumer is only a minimum requirement. To

[1] ValidatFOSS: Validation with Free and Open Source Software. ESTAT (GAs) No: 825659 and 882817
[2] For details, see https://ec.europa.eu/eurostat/cros/content/principles_en
[3] While respecting the natural order of checking technical errors first, before content errors.

minimize costs of recalculations every validation rule should be checked as early as possible in the respective statistical process at the NSI and should therefore be integrated into the statistical system where possible. Hence the search for a generic data validation tool that can be integrated in production chains and adheres to international standards.

3.2. ESS main types of rules

An important cornerstone in international data validation was the identification of 20 so-called 'main types of validation rules' that cover the most common validation rules across all statistical domains (Eurostat, 2018). Figure 1 shows these rules with their three letter abbreviation and some other characteristics such as whether they are mandatory, to what validation level[4] they belong, and a severity level (error, warning, for information). They vary from more simple rules on field formats, value ranges, to more complicated rules such as data completeness, consistence between details and their aggregates or plausibility of seasonally adjusted values. Because of the general applicability of these rules, they were felt to be a logical candidate[5] for implementing validation at the side of the statistical offices.

The 20 main types of validation rules in the ESS and their characteristics

Rule type	Mandatory	Default	Validation level						SDMX	Micro data	Severity level		
			0	1	2	3	4	5			E	W	I
(EVA) Envelope is Acceptable	X		X						X	X	X		
(FLF) File Format	X		X						X	X	X		
(FDD) Fields Delimiter	(X)	";"	X						X	X	X		
(DES) Decimals Separator	(X)	"."	X						X	X	X		
(FDT) Field Type	X		X	(X)					(X)	X	X		
(FDL) Field Length	X		X						X	X	X		
(FDM) Field is Mandatory or empty			X	(X)					(X)	X	X	(X)	
(COV) Codes are Valid	(X)			X					(X)	X	X	(X)	
(RWD) Records are Without Duplicates	(X)	Key		X					X	(X)	X		
(REP) Records Expected are Provided				X	X					X	X	(X)	
(RNR) Records' Number is in a Range	X	>=1		X	(X)					X	X	(X)	(X)
(COC) Codes are Consistent				X	X				(X)	X	X	(X)	
(VIR) Values are in Range		>=0		X	X				(X)	X	X	(X)	(X)
(VCO) Values are Consistent				X	X	X	X	X		X	X	(X)	(X)
(VAD) Values for Aggregates are consistent with Details	(X)	=		X	X						X	(X)	(X)
(VNO) Values are Not Outliers				X	X						(X)	X	(X)
(VSA) Values for Seasonally Adjusted data are plausible				X	X						X	(X)	(X)
(RRL) Records Revised are Limited					X					(X)	(X)	X	(X)
(VRT) Values are Revised within a Tolerance level					X					(X)	(X)	X	(X)
(VMP) Values for Mirror data are Plausible						X					(X)	X	(X)

Figure 1: ESS main types of validation rules

[4] These levels are defined in the International handbook on validation.
[5] With the exception of mirro data checks among country data (MVP) which typical Eurostat tasks. However the generic concept of executing mirror checks across multiple data producers is valuable.

3.3. SDMX registries

The table in Figure 1 contains a column 'SDMX[6]' indicating whether the rule depends on metadata specified in this international ISO standard. At the core SDMX has a logical information model describing the key characteristics of statistical data and metadata, which can be applied to any statistical domain. This metadata is defined in an SDMX registry where data producers can download or query the necessary metadata. Alternatively, metadata is distributed in a so-called Data Structure Definition (DSD) file, which is usually an XML format. Both types of modes should result in exactly the same metadata agreements.

SDMX registries can be accessed through a REST API, using a standardized set of parameters. Two important SDMX registries are:

- *Global SDMX Registry:*
  This registry is maintained by the SDMX consortium. As such, it is the top-level central place for ESS-wide metadata. It contains general statistical metadata such as for the Harmonised Index of Consumer Prices (HICP), National Accounts (NA), Environmental accounting (SEEA), Balance of Payments (BOP), and many more.
- *Eurostat SDMX Registry*:
  This registry contains is operated by Eurostat and contains statistical metadata for all other official statistics in the ESS.

A look at the contents of the registries makes it immediately clear that the metadata provided in these registries together define most of the metadata of the ESS. Since some of the main types of rules rely on those definitions it is necessary to take them into consideration in the design of the generic data validation ecosystem.

3. Data validation using R

The elements presented in the previous chapter were designed with international data reporting in mind. The art of producing official statistics on *national* level is slightly different but not unrelated. In national statistical processes there is an increasing use[7] of the R programming language for processing data. One of the reasons for this is that there are quite some readily available R packages

[6] Statistical Data and Metadata eXchange (SDMX), see https://sdmx.org
[7] See for example: http://r-project.ro/conference2021.html

that facilitate the implementation of typical statistical operations such as validation, data cleaning, error localisation, editing, linking, imputation, aggregation, visualisation and dissemination.

Among these, the R-package validate (van der Loo, 2021) is a popular tool for data validation. It offers a language for validation rules supporting any type of data checks, including uni- and multivariate checks, statistical checks, checks on time series, hierarchical aggregation, and more. The software can work on possibly large datasets, providing statisticians with feedback on the health of their data. Validation results are presented graphically and in a machine readable report, which can be used as input for consecutive processes. Common rules can be re-used among multiple validation processes. Figure 2 sketches an example workflow of a validation session. A dataset on person level is 'confronted' with a set of validation rules that should hold. Results are displayed in a summary, bar chart or together with the data in an (experimental) dashboard.

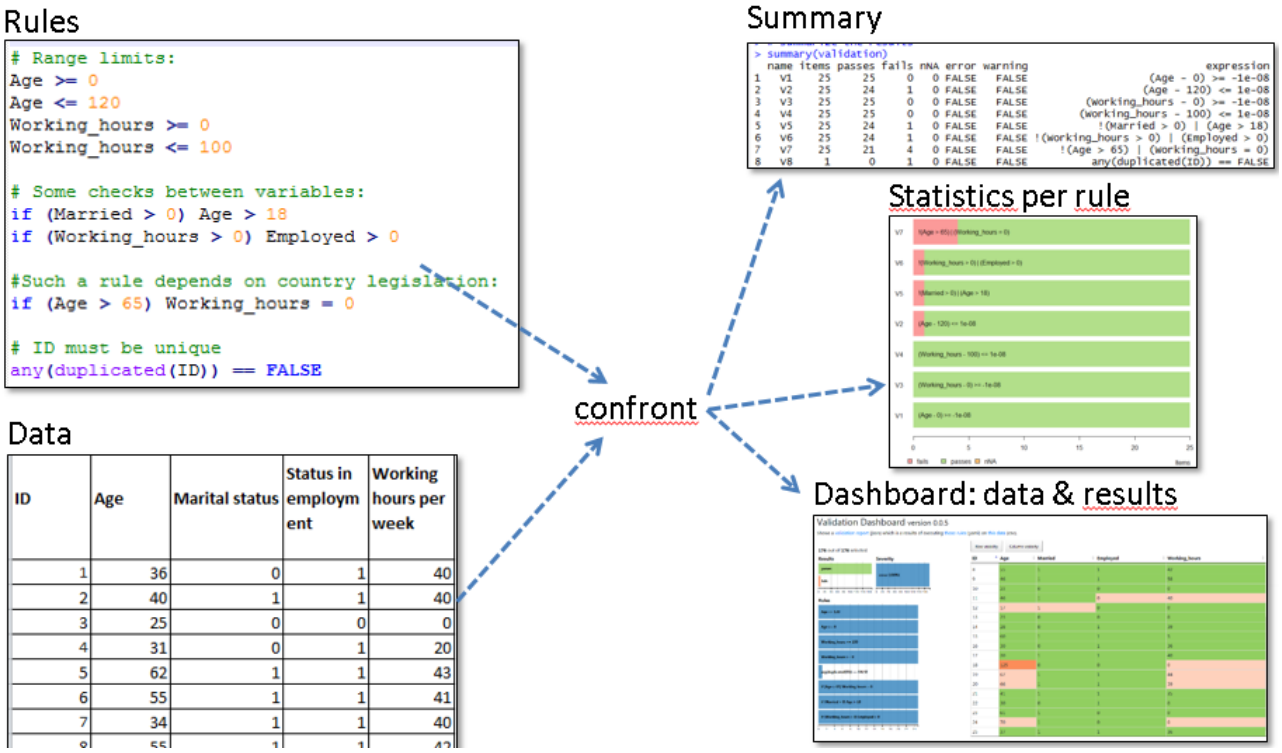

Figure 2: workflow R-package validate

The rule language used in this package is R-based. Many of the rules in this form are easy to read as the checks on range limits and if checks between variables in the example in Figure 2. The fact that any R expression that evaluates into a boolean can be used as a rule definition, makes it possible to also implement more complicated rule expressions. To facilitate the use on larger datasets the R-

package validatedb has been developed. It allows users to check whether records in a database are valid using the same validation rule syntax as in 'validate'.

Within Statistics Netherlands these R-packages are growing in popularity. At time of writing they are known to be used in statistics on healthcare, national accounts, short term statistics, asylum, museums and nautica to name a few. In addition there are other NSIs using them for various purposes, which is fully in-line with the FOSS objectives of the ValidFOSS project.

4. A generic data validation ecosystem

In the ValidatFOSS project all ingredients listed in the previous chapters were combined to provide a free and open source generic validation ecosystem that maximally aligns with ESS standards. The existing R package validate was extended with higher level functions that make it easy to implement generic statistical data validation functionality often found in official statistics as well as the ESS main types of rules.

The results are documented in an online validation cookbook (Loo, 2022) that offers recipes for implementing the most common validation scenarios found in official statistics. Figure 3 shows the data validation typology used in this cookbook. It organises the checks in variable checks, checks on availability and uniqueness, multivariate checks, statistical checks and checks based on SDMX metadata.

On the most detailed level the checks are implemented in higher level validation functions that have a clear use in national validation processes, such as ‘is_unique’ or ‘contains_at_least’. In some simple cases they map directly to base R functions, such as the checks on type or length. The ESS main types of rules were used as an input for the design of this typology. As an example, the validate function ‘hierarchy’ was heavily inspired by the ESS equivalent ‘VAD’ and can thus be used to easily implement such international validation rule. The functions 'in_linear_sequence' and 'is_linear_sequence' can be used to check for gaps or duplicates in numerical or time series as in main type rule ’REP’.

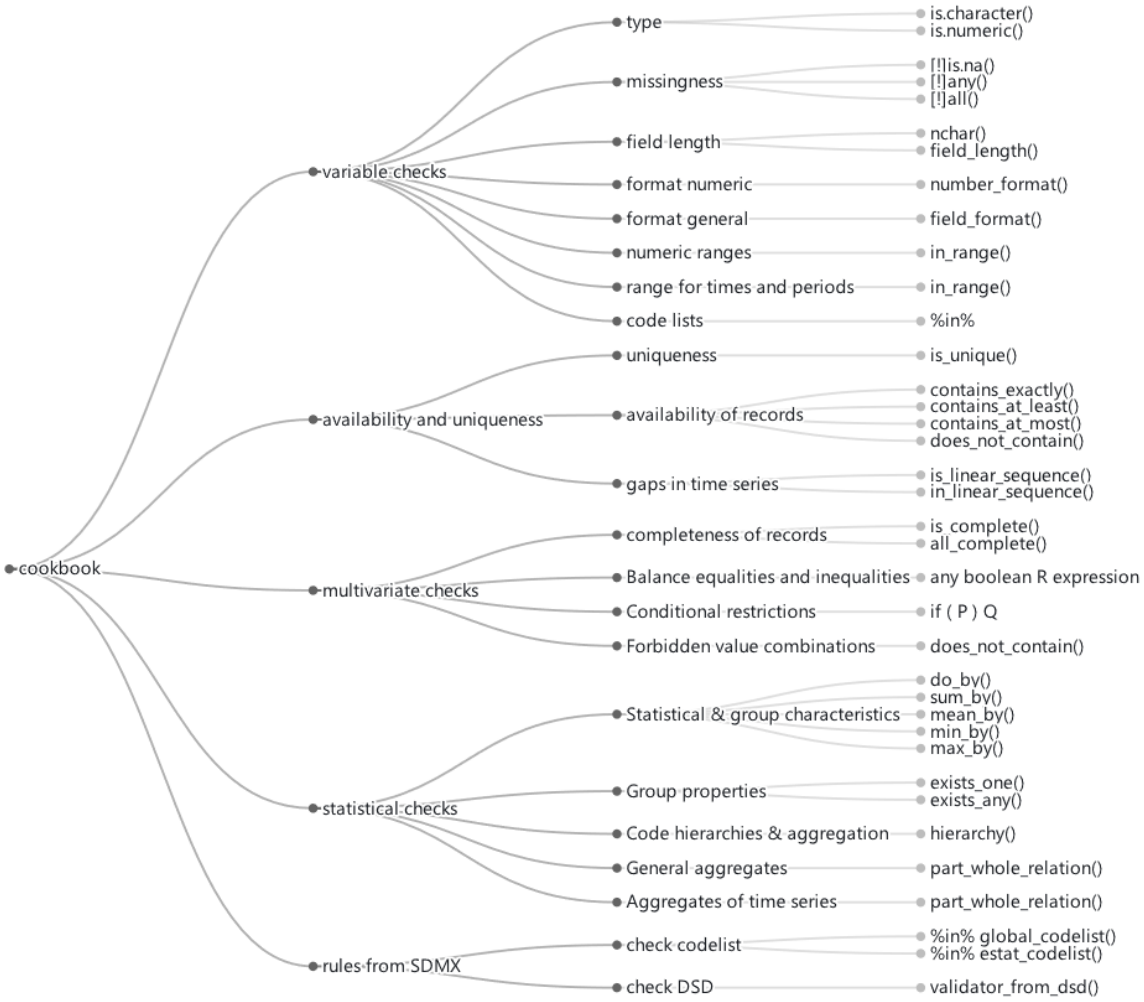

Figure 3: data validation typology from cookbook

The checks on codelists and DSDs retrieve the necessary metadata from the respective SDMX registry via the SDMX 2.1 API. For efficiency results from consecutive SDMX calls are cached within an R session. Because the functions use the standardised SDMX API, organisations that have an in-house SDMX registry to manage their internal metadata can use the same functions to check against their company metadata.

One could ask how this setup relates to other initiatives, such as the implementations of the main types of rules into SQL and the Validation and Transformation Language (VTL)[8]. Although we do not claim to have the total overview we sketched our understanding of the situation in Figure 4. At the highest level of abstraction we find the ESS main types of rules. They have been implemented in the cookbook and in VTL. Cookbook functions offer direct access to SDMX registries and use underlying R packages. The R package validatedb understands cookbook functions and translates them to SQL. Italy and Poland work on translators from VTL to SQL. Portugal implemented some

[8] https://sdmx.org/?page_id=5096

of the main types of rules into SQL directly (Bosch, SDE2020). Both VTL as well as the R language can also implement other high level rules than the main types of rules.

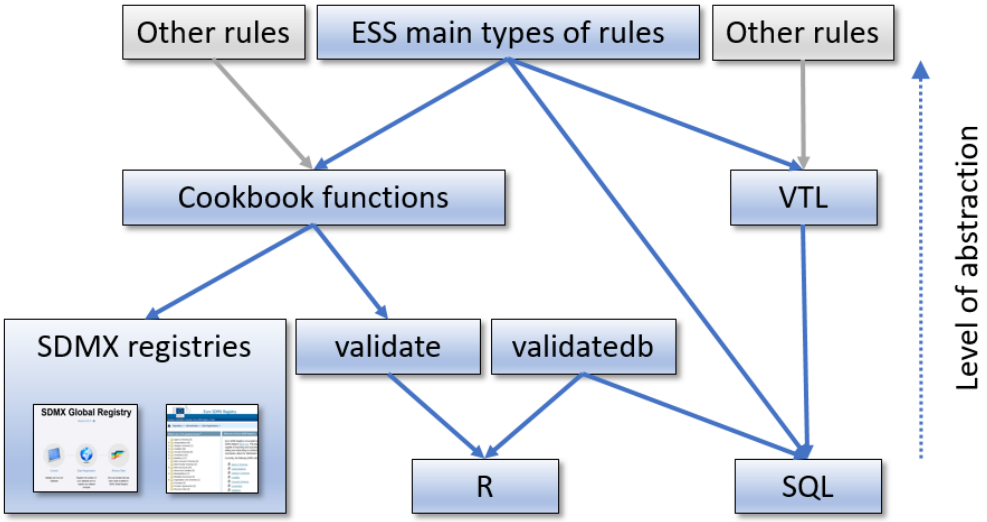

Figure 4: Rule implementations on various levels of abstraction

5. Conclusions and reflections

Data validation – both in national as well as international context - is important for guaranteeing quality of official statistics. It has been recognised that NSIs should ideally check their data as early as possible aligning to the validation principle "the sooner the better". To do this rules must be known beforehand and tools must exist to efficiently and effectively implement such rules in all stages of the statistical process. This holds for International validation rules as well as for rules that hold in a more local context.

In the ValidatFOSS project a free and open source generic validation ecosystem was developed that offers access to a rich set of validation rules often needed in official statistics. Moreover the system makes it easy to implement the International main types of rules as defined by Eurostat. The system connects to SDMX registries to retrieve structured metadata such as code lists and DSDs. An online validation cookbook offers recipes for common validation scenarios found in official statistics. The implementation in R makes it easy to implement data validation into R-based statistical processes. The system is used in a number of statistical domains at Statistics Netherlands and in some other countries, which aligns with the concept of re-using statistical tools as free and open source software in the ESS.

We hope that the work presented in this paper can be used as a basic building block for further improvements on monitoring data quality in the new emerging data ecosystem. New data sources such as web data, data from commercial providers or even administrative data are often more volatile than one might think and certainly more volatile than traditional data sources. Therefore the necessity to validate data before its use will grow. The international rules and the typology presented are a good starting point to face this challenge and could be further extended to new demands. In a datafied society it may even be necessary to develop even more intelligent data validation concepts such as the ability to derive validation checks from existing data streams in sound combination with human knowledge with the ultimate goal to guarantee or further improve quality of statistics.

5. References

Methodology for data validation 2.0 (revised edition 2018), available at https://ec.europa.eu/eurostat/ramon/statmanuals/files/methodology_for_data_validation_v2_0_rev2018.pdf

Eurostat Task Force on validation (2019), Principles for data validation, available at: https://ec.europa.eu/eurostat/cros/content/principles_en

Loo van der, M., Bosch ten, O. (2017) Design of a generic machine-readable validation report structure, version 1.0.0 August 15, 2017, https://ec.europa.eu/eurostat/cros/content/validation-report-structure_en

Eurostat/B1 (2018), Main types of validation rules for ESS data, Summary table available at: https://ec.europa.eu/eurostat/cros/content/02c-main-types-data-validation-rules-and-fictive-domain_en

Bosch ten, O., Loo van der, M., Quaresma S (2020), Implementing main types of International validation rules in national validation processes, UNECE workshop in statistical data editing (SDE), available at https://unece.org/statistics/events/SDE2020

Loo, MPJ van der, Jonge E de, (2021) Data validation infrastructure for R (2021), Journal of Statistical Software 97 1-22. See also https://cran.r-project.org/package=validate

Loo, MPJ van der and ten Bosch, O (2022), The data validation cookbook, https://data-cleaning.github.io/validate