Implementation of generic data validation methodology for Short Term Statistics

Mark van der Loo[1] and Olav ten Bosch

Statistics Netherlands, PO box 24900 HK, The Hague, the Netherlands

Key words: data cleaning, interoperability, European Statistical System, Free and Open Source Software (FOSS), R.

Introduction and context

When data is transferred from a producer to a consumer, it is important to agree on quality aspects of the transferred data. Cases that are relevant for official (business) statistics include transfer of microdata between businesses and a national statistical institute (NSI), between administrative authorities and NSIs and transfer of statistical data between NSIs and other organizations. Besides data transfer between organizations, data sets may be handed over within organizations between organizational units, for example in cases where data gathering or data publication is executed by a centralized service. For reasons of efficiency, it is important that sender and receiver of the data agree on the state of the data under scrutiny.

Focusing on the European context, data transfer between NSIs and Eurostat, and between other national statistical authorities and Eurostat is a critical part of producing statistics on a European level. Although much of the transfer process has been standardized on a technical level (for example through the EDAMIS[1] and SDMX[2]), it has been recognized by the European Statistical System (ESS) that the way that quality demands are formulated, and the way that failure to meet quality demands are communicated is fragmented across statistical domains and statistical institutes[3]. Lack of clarity in communicating (demands on) data quality has had demonstrably negative effects on efficiency, causing frequent retransmissions between institutes where a single transmission should do. The current record stands at more than 20 retransmissions before acceptance.

To increase efficiency, the ESS has undertaken a series of activities to standardize the way quality demands are formulated and how quality reports are communicated. The main results include a handbook on data validation methodology [4] including a clear definition of the concept of 'data validation', an overview of common validation rules [6], and a machine-readable and language-independent data quality reporting format[22] [2].

This paper focuses on the application of some of these results to measure the quality of Short Term Business Statistics in terms of standardized validation rules, using tools that are published as Free and Open Source Software (FOSS) in the statistical programming language R [8]. In the following paragraphs, first the concept of data validation is made more precise, and next the application to STS is discussed.

---

[1] Corresponding author: m.vanderloo[at]cbs.nl

[2] There are more results including an assessment of the VTL language[5], data validation principles[20] and a data validation business architecture for the ESS [21].

Data validation and data validation tools

In order to agree on how data quality is measured, it is important to agree on the concept of 'data validation' itself. Informally, data validation is the activity where one decides whether or not a particular data set is fit for a given purpose[4, 9, 10, 11]. To formalize this process one usually defines a set of predicates over the variables and metadata of a particular data set such that the value TRUE corresponds to a data set passing a test and FALSE corresponds to a data set failing a test. A data set is deemed valid when it passes all tests. These predicates are usually referred to as data validation rules or edit rules, and in references [9, 10, 11] it is demonstrated that a formal definition allows for deriving a natural classification of validation rules. The classification separates the rules into levels of complexity, where a more complex rule can only be evaluated when a larger variety of information is available.

Once data validation is formally defined, it becomes possible to communicate data quality demands unambiguously and with mathematical precision. It also allows for automation of the task of data validation, separating the definition of quality demands based on domain knowledge from execution of the rules and administrating the results. It also allows for investigating the validation rules themselves, yielding a better grip on the data quality assessment process.

At Statistics Netherlands, the R-based tool `validate' [12] has been built for the purpose of defining, maintaining, applying data validation rules, and analyzing their results. A companion package called 'validatereport' [13] was developed under a European grant to export data validation results to a standardized ESS format. Efforts were also undertaken to facilitate the expression of cross-domain validation rules with validate, in the form of the 'GenericValidationRules' [14] package. Based on these packages, domain-specific validation rules for the ESS have been implemented and are now reusable by the entire ESS [15] (Figure 1). Finally, it is worth mentioning that rules, once defined in 'validate', can also be reused for data editing (data cleaning) purposes in several R-based tools [10]. An overview of these, and many more FOSS tools for official statistics can be found in the so-called awesomelist for official statistics software [16].

Data validation for Short Term Statistics

The way STS data is to be transferred between NSIs and Eurostat is defined in the 'SDMX for Short-Term Business Statistics' guidelines[17]. These guidelines include a set of validation rules that are formulated in informal, human-readable language. For instance, `*Zeroes are not admitted for prices*' and `*No missing observations (gaps) are accepted in time series[…]*'. Observe that although these rules are easy to interpret by humans, they require a possibly complex variety of information to execute. In the first instance one must select price data and test whether they are non-zero. In the second instance one must first isolate a time series for a certain variable, sort it, compute time differences, and confirm that no gaps arise.

Since the rules hold for every STS producer in the ESS, and since the technical transmission format is fixed, it makes sense to automate these rules once and for all in an open and freely available tool. Here, this was done by implementing the rules in the R packages 'validate' and 'GenericValidationRules' . The expression for the first example looks as follows.

```
if (INDICATOR %in% c("IMPZ","PRBB","PREN","PREX","PREZ","PRIN","PRON"))

      OBS_VALUE != 0
```

Here, the condition selects certain indicators that represent prices, and then demands that the observed value is not equal to zero.

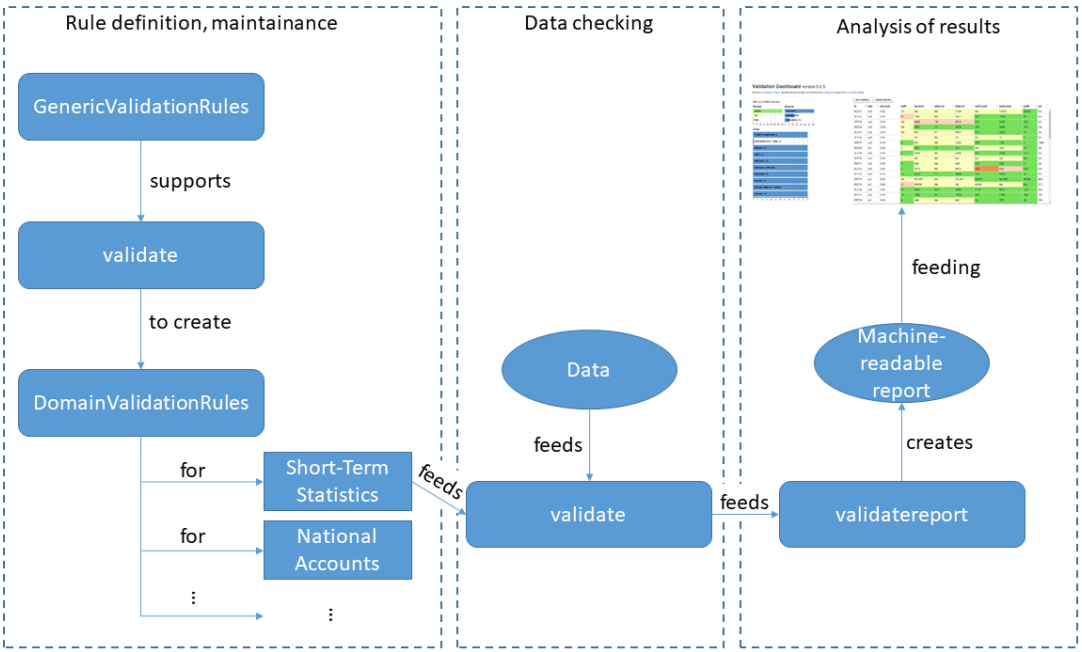

Figure 1. Overview of the generic data validation tools and procedures. The R package 'GenericValidationRules' makes it easier to define complex ESS validation rules for the ESS in 'validate' syntax. This yields a set of reusable Domain-specific data validation rules that can be reused across the ESS for several domains. These rules, together with data feed into validate, yielding validation results. These are fed into `validatereport' to create an ESS generic machine-readable reporting format. This format can for instance be used to feed a web-besed dashboard.

The second rule needs an elaborate algorithm to be computed over generic STS data. However, since the demand to have gapless time series is very common, it is one of 20 generic data validation rules, and a general implementation is available in the form of the RTS function from the GenericValidationRules package. The rule looks as follows on an example data set.

```
RTS(TIME_PERIOD, ftp="2017-Q1",ltp="2019-Q3",FREQ, REF_AREA,
SEASONAL_ADJUST, INDICATOR, ACTIVITY) == TRUE
```

Where 'RTS' is the unique label assigned to the 'gapless time series' rule in [7]. The parameters for 'RTS' have been taken from the same reference.

To facilitate maintenance and understanding of rules, it is possible in 'validate' to endow them with metadata. To be precise, a rule can be labeled with a name, a short description (label) a long description, a time stamp, and so on. So here is what the first rule looks like precisely in the 'validate' YAML-based storage format[18].

```
- expr: RTS(TIME_PERIOD, ftp="2017-Q1",ltp="2019-Q3",FREQ, REF_AREA,
        SEASONAL_ADJUST, INDICATOR, ACTIVITY) == TRUE
  name: "STS02"
  label: "No gaps"
  description: |
    No missing observations (gaps) are accepted in time series,
    sent in one or several files – i.e. files should be sent in
    the chronological order based on the latest observation.
```

Here, 'description' is copied from the human-readable guidelines. This way the generic machine-readable expression is kept together with the human-readable description.

In this proof-of-concept the application of this approach was validated in the following ways. First, generalizability was demonstrated by implementing a set of rules for National Accounts data as well. Second, the rules were confronted with data sets from STS and national accounts. The validation results were then translated into the generic data validation report structure for the ESS. These results were subsequently read into a generic data validation dashboard[19], demonstrating full compatibility from rule definition to result analysis.

Summary and conclusion

Recent proof-of-concepts in the European Statistical System demonstrated the viability of reusing a set of generic validation tools to both express data quality demands and to communicate on their results. In this paper some results of the Short-Term Statistics have been highlighted. The results discussed in this paper are based on freely available and open source tools based on R, facilitating re-use both within the ESS and beyond.

References

[1] EDAMIS: Electronic Data files Administration Information System, https://webgate.ec.europa.eu/edamis/helpcenter/website/tools/ewp/index.htm.

[2] SDMX: Statistical Data and Metadata eXchange, https://sdmx.org.

[3] Eurostat (2015) *Business Case ESS.VIP.BUS VALIDATION (version 1.8)*. pdf on circab.

[4]] M. Di Zio, N. Fursova, T. Gelsema, S. Giessing, U. Guarnera, J. Ptrauskiene, Q.L. Kalben, M. Scanu, K. ten Bosch, M. van der Loo, K. Walsdorfe (2015). Methodology of data validation. *Deliverable No 2 of the ESSnet on data validation.*

[5] T. Gelsema (Editor), Ten Bosch, O., Vignola, L., Di Zio, M. Bianchi, G., Scanu, M. (2015) *A study of VTL.* Deliverable of the ESSnet on Validation. https://ec.europa.eu/eurostat/cros/content/essnet-validation-study-vtl-final_en.

[6] VTL: Validation and Transformation Language, https://sdmx.org/?page_id=5096

[7] V. Tronet (2018) Main types of validation rules for ESS data (version 1.0.3). Eurostat Working document. pdf.

[8] R core team (2019). *R: A Language and Environment for Statistical Computing*. R foundation for statistical computing, Vienna, Austria. https://r-project.org.

[9] MPJ. van der Loo and De Jonge, E. (2019) Data Validation. Wiley StatsRef Online (to be published)

[10] MPJ van der Loo and De Jonge, E (2018) *Statistical data cleaning with applications in R*, John Wiley & Sons.

[11] M. van der Loo (2015) *A formal typology of data validation* UNECE Work Session on Statistical Data Editing (Budapest). https://www.unece.org/stats/documents/2015.09.sde.html.

[12] MPJ van der Loo and De Jonge, E. (2019) *Data validation infrastructure for R. Journal of Statistical Software* (submitted for publication). See https://cran.r-project.org/package=validate.

[13] M. van der Loo (2019). validatereport: report on validation results. https://github.com/data-cleaning/validatereport.

[14] M. van der Loo, Windmeijer, D., Ten Bosch, O. (2019) GenericValidationRules. Generic validation rules for the ESS. https://github.com/SNStatComp/GenericValidationRules.

[15] M. van der Loo, Windmeijer, D., Ten Bosch, O. (2019) DomainValidationRules. Domain-specific validation rules for the ESS. https://github.com/SNStatComp/GenericValidationRules.

[16] Awesomelist of official statistics software: http://www.awesomeofficialstatistics.org. Official statistics software is 'awesome' when it is (1) Free, Open Source, and available for download, and (2) confirmed to be used in the production of official statistics by at least one institute. Software that facilitates access to official statistics is accepted as well, as long as it conforms to (1).

[17] SDMX for Short-Term Business Statistics Guidelines (STS). https://circabc.europa.eu/sd/a/adfce7f9-49ff-47cf-8a47-4224011a8d48/SDMX%20for%20STS_guidelines_170201.pdf

[18] YAML is a json-like structured data standard, supported by many programming languages. See https://yaml.org

[19] O. ten Bosch and Van der Loo, M. (2019). *A generic Shiny/JS dashboard for data validation results*. Use of R in Official Statistics (uRos2019, Bucharest). https://github.com/data-cleaning/ValidatReport.

[20] Principles for Data Validation: https://ec.europa.eu/eurostat/cros/content/principles_en

[21] Validation Scenarios for member states: https://ec.europa.eu/eurostat/cros/content/scenarios-member-states_en

[22] M. van der Loo and Ten Bosch, O (2017). *Design of a Machine-Readable Validation Report Structure*. Deliverable No 2 of the ESSnet ValidatIntegration. https://ec.europa.eu/eurostat/cros/content/essnet-validat-integration_en