UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Discover the hidden validation rules in your data with 'validatesuggest'

Olav ten Bosch, Edwin de Jonge, Mark van der Loo (Statistics Netherlands, the Netherlands)

o.tenbosch@cbs.nl; e.dejonge@cbs.nl; mpj.vanderloo@cbs.nl

I. Introduction

1. Data validation is an essential element in the challenge to provide high quality official statistics. Survey data, data from administrative sources, data from the web, sensor data, or data from other alternative data sources often are dirty and need to be checked before being used as an input to official statistics. The same holds for the processing and analysis phase. It is good practice to check and validate data at intermediate stages in the statistical production processes varying from processing raw data up to aggregation. It also holds for data to be disseminated: there is a strong demand to set up automated validation processes that execute just before publishing new statistical indicators to the general public.

2. In a wider context, data validation plays an important role in the data exchange among statistical organisations. In the European Statistical System (ESS), national Statistical Institutes (NSIs) report national aggregates to international organisations such as Eurostat. Multiple NSIs and Eurostat have worked on this, resulting in a set of validation principles [1], a methodological handbook on validation [2], a validation architecture and generic software. An important element in the approach is the principle that rules should be divided from the software. Validation rules should ideally be configured and maintained by domain specialists so they need a flexible setup to do this.

3. Crucial to the subject of data validation are the definition of rules of high quality. Although it may seem straightforward at first, in practice it can be a challenge to discover the rules, to specify them with the right precision - not too relaxed but also not to strict - and to maintain them over time. In practice the number of rules tend to grow organically and this may result in a large set of rules to be maintained. In this paper we explore a new approach where validation rules are derived from data. The approach is not meant to replace the knowledge-driven definition of validation rules but as a complementary and as a bootstrapping process. We expect that it can be used as an additional mechanism to discover and maintain validation rules to guarantee high quality statistics on the longer run. It can also be helpful to bootstrap rule definitions, i.e. to create an initial set of rules that can be further explored and adjusted.

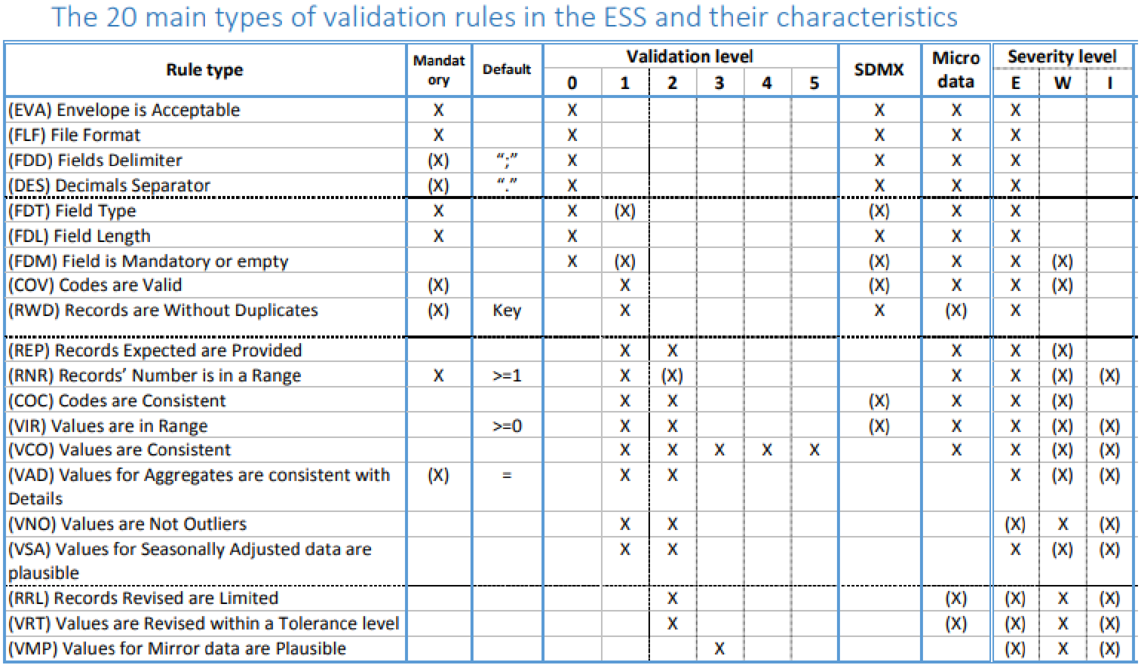
4. In chapter II we briefly introduce the existing open source R-based ecosystem for national and international data validation. In chapter III we explain the concept of deriving validation rules from data. Chapter IV shows an early implementation of the concept in the experimental R package 'validatesuggest'. In chapter V we perform some reflections and come to our conclusions.

II. An ecosystem for national and international data validation

5. The necessity to monitor data quality has been recognized before. It is general knowledge that data characteristics observed when designing statistics may not hold during production in the long run and that, if this is not detected in time, this may lead to errors or costly recalculations. Therefore data to be used in official

statistics 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.

6. On the International level data validation rules have been agreed upon in International statistical working groups. Eurostat developed the so-called ‘main types of validation rules’ [3] that cover the majority of international statistical processes. They are used by Eurostat and some NSI’s to implement international rules in national validation processes [4]. Some of these rules use metadata described in the International SDMX standard. Figure 1 shows these rules with their three letter abbreviation and some other characteristics such as whether they are mandatory, to what validation level they belong, and a severity level (error, warning, for information). They vary from more simple rules on field formats and value ranges to more complicated rules such as on data completeness, consistency between details and aggregates or plausibility checks.



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

7. On the National level data validation has to be incorporated in national statistical production processes. The optimal way to do so might vary per organization, however generally speaking NSIs increasingly use the R programming language in their offices. Hence the open source R-package *validate* [5] is becoming a popular tool for validation. It supports many different validation rules and its supports the principle that rules should be under control of domain specialists. The validation functions supported have recently been extended to make it easy to implement the ESS main types of validation rules and to connect to internationally agreed metadata from SDMX registries [6]. The *validate* software can execute analyses on possibly large datasets, providing statisticians with feedback on the health of their data. An online cookbook [7] explains the use of the tool for the most common validation scenarios found in official statistics. Figure 2 shows the data validation typology used in this cookbook. It organises the validation checks in variable checks, checks on availability and uniqueness, multivariate checks, statistical checks and checks based on SDMX metadata.

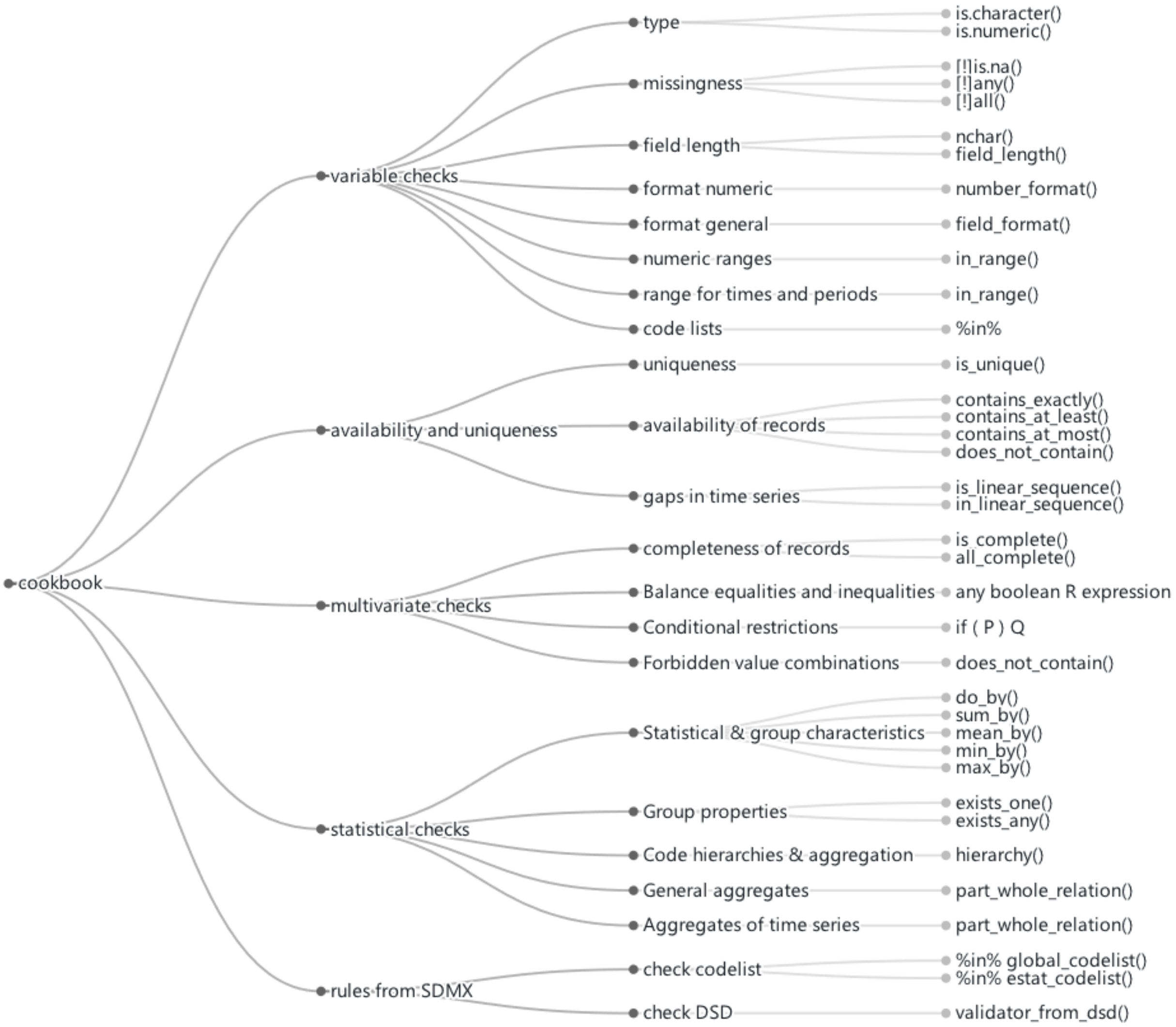


Figure 2: R validate data validation typology

III. A data-driven approach to validation

8. In the data-driven approach we infer rules from data: the knowledge contained in the data suggests the rules. Properties of the data, such as type, range, distribution, correlation can be derived and used as rule suggestions. It is preferable to start with a clean dataset to generate rules, but in practice it also works with (slightly) dirty datasets: the generated rules can make data errors explicit. For example if the age variable in a dataset contains negative values, the generated range rule suggestion will have a negative lower boundary, which is clearly false, but can easily be adjusted: the suggested rule is therefore useful.

9. The more data available and the longer time span it covers, the better the rules inferred will catch the data dynamics over time. Once in a while it might be valuable to re-suggest validation rules on fresh data, somewhat comparable with re-training in Machine learning. Obviously not all types of validation rules can be suggested from data and suggestions are never guaranteed to hold in a wider context. Rules defined by and from domain experts are a valuable basis, but the data-driven approach may be a starting point or extra help for the domain specialist.

10. Because most of the official statistics measure a certain phenomenon over time, the production of official statistics is an iterative and repetitive process by nature. In terms of the Generic Statistical Business Process Model (GSBPM) [8] in the design and build phases the methodological design, the collection, the instruments and the process design are developed. In the collect, process, analyse and disseminate phases, a yearly, quarterly or monthly (or in some cases weekly) production cycle is performed. Over time, these

production cycles generate data that contain an increasing amount of knowledge on the specific statistical process. On the hypothesis that data that passed the production cycle is supposed to be valid, in the data-driven approach even more advanced validation rules can be derived if data is available on multiple production cycles.

11. The data-driven approach can work because we know what type of validation rules make sense in official statistics. We can build on the existing rule ecosystem presented in the previous chapter. Both the definition of the International ESS rules as well as the definition of the rule typology in the cookbook actually cover experience from many years of validation practices in official statistics. We use them to select the most valuable and most common set of rules to suggest.

12. With respect to the ESS main type of rules in Figure 1, rules such as mandatory or range checks seem feasible to detect and they align with the selection of checks from the cookbook. Advanced rules such as revisions and seasonally adjustments are probably difficult to detect without further guidance and metadata from the user on what type of adjustments to expect. An interactive approach where the suggestion system identifies blank spots in metadata and asks for additional information where needed could work here, but we do not further explore such mechanism in this paper.

13. With respect to the R validate typology in Figure 2 rules with the highest chance of correct suggestions are variable checks, availability and uniqueness, and checks from SDMX. Multivariate checks might be more difficult to detect in general, but for a limited number of variables – two or three – it is certainly possible and useful. Statistical checks can be derived, but statistical distributions are typically domain dependent and to be checked for plausibility. Again the additional expert knowledge needed to infer complex rules could be retrieved interactively, but in this first approach we focus on what we can derive from the data without extra interactions.

14. Up to now we described data-driven rule suggestions based on a single dataset. If data from multiple production cycles is available we could extend the concept to infer rules that hold for the common characteristics of that particular stream of datasets as a whole. The consecutive data transmissions, also called periodic data reports, between institutes or between departments within an institute are the input. We see two types of rules that could be discovered that way:

(a) *cross sectional derived rules*: rules that apply to all datasets of all periodic reports and thus have a higher probability of being right then inferred from only a single period. These can be discovered by concatenating the single datasets and executing the rule discovery on the complete set. A refinement could be to give data from more recent time periods a higher preference than older data, from the viewpoint that recent data characteristics are more important than those in older data.

(b) *time series derived rules*: rules that depend on the time period in the data transmission, such as a growth rate. To infer this types of rules we need knowledge on what variable holds the time period and an interpretation of the relationship between the contents of that variable and the target variable(s). In an SDMX context the time period variable can be read from the data structure description (DSD). In other contexts, where this is not contained in the meta data, this could be derived by automatic time variable determination. With respect to the ESS validation rules and the validation typology, rules based on time-period that are candidates to be discovered are 'gaps in time series', 'aggregation of time series'.

15. In all cases rule suggestion should allow for adding tolerances and ideally also for slight data errors.

16. The rule suggestion methods can be framed in a Machine Learning context, both unsupervised as well as supervised. Unsupervised rule suggestion, as proposed here, uses data type, shape, correlation and other properties to derive suggestions for rules that the data must abide to. For supervised both a clean and dirty dataset is needed to be able to learn what records and values are deemed erroneous. Generally speaking supervised machine Learning for error detection and validation seems promising, but an important advantage of using validation rules is that they are explicit, understandable and can be evaluated with ease, which makes it a natural tool for executing and explaining the validation process in official statistics: each invalid value is *explainable*. Rule based validation has therefor an edge over black box machine learning detection methods,

which may do a good job at detecting if a record is (in)valid, but cannot explain what is wrong and should be adjusted. Interesting research in the realm of explainable artificial intelligence (XAI) suggests that rules may be derived from black box models.

IV. Implementation in ‘validatesuggest’

17. The R package *validatesuggest* [9] is an early implementation of the data-driven approach to validation rule discovery. It generates rules in R validate syntax from a supplied dataset. Up to now only a limited number of common types have been implemented, which however could already add considerably to the validation rule maintenance challenge. They are:

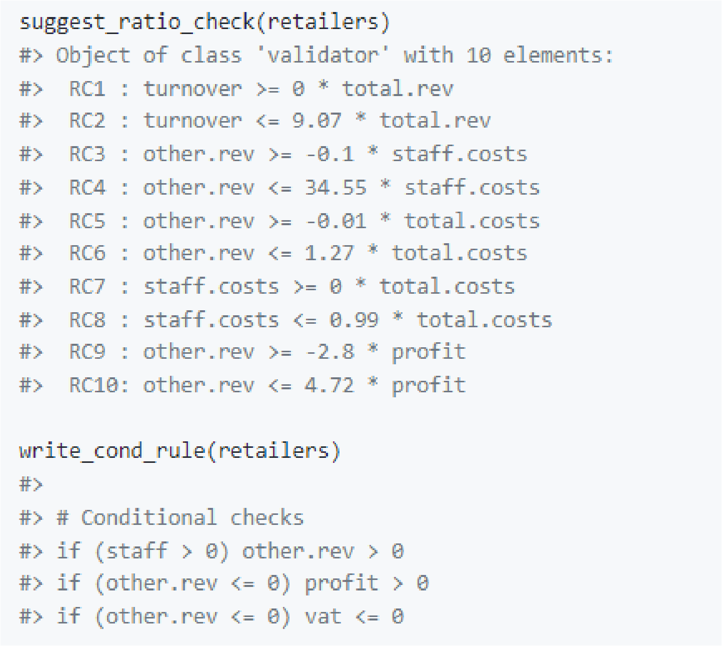
(a) Positivity checks
(b) Range checks
(c) Checks on na: whether a variable may contain nas
(d) Checks on uniqueness
(e) Type checks
(f) Ratio checks:
(g) Discovery of conditional rules

A high level function ‘suggest_all’ is available to infer all supported validation rules in one pass.

The ratio check (f) uses a correlation threshold: only variables that are (enough) correlated are considered.

The discovery of conditional rules (g) implements a unsupervised machine learning technique (association rules), that checks the co-occurrence frequency of values. e.g. if the values “job: retired”, “income_type: pension” co-occur, the rule “if (income_type == pension) job == retired” will be derived. The direction of the causality relation is derived from the occurrence of other values for the respective variable. In this example for “job: retired” there is only one income_type, where for “income_type: pension” there are multiple job values. If no direction can be derived two rules are generated. No supervised methods have been implemented at the moment.

18. Figure 3 shows which ratio checks and conditional rules have been discovered on the retailers dataset which is included in the R validate package. 10 ratio checks were discovered. As can be seen the results of the rule suggestions is presented in a form that is both user-readable as well as suited for interpretation by the R package validate.



```
suggest_ratio_check(retailers)
#> Object of class 'validator' with 10 elements:
#>  RC1 : turnover >= 0 * total.rev
#>  RC2 : turnover <= 9.07 * total.rev
#>  RC3 : other.rev >= -0.1 * staff.costs
#>  RC4 : other.rev <= 34.55 * staff.costs
#>  RC5 : other.rev >= -0.01 * total.costs
#>  RC6 : other.rev <= 1.27 * total.costs
#>  RC7 : staff.costs >= 0 * total.costs
#>  RC8 : staff.costs <= 0.99 * total.costs
#>  RC9 : other.rev >= -2.8 * profit
#>  RC10: other.rev <= 4.72 * profit

write_cond_rule(retailers)
#>
#> # Conditional checks
#> if (staff > 0) other.rev > 0
#> if (other.rev <= 0) profit > 0
#> if (other.rev <= 0) vat <= 0
```

Figure 3: Discovery of ratio checks and conditional checks from retailers dataset

19. We strongly suggest that rules discovered from data presented are *always manually inspected* by a domain specialist before they are potentially used for real-world validation. The specialist may know from experience that certain checks may hold on this particular dataset, but not in general. Also it is good practice to try to minimize validation rules for maintainability. In the example above rule RC1 and RC7 are both redundant with the positivity checks on *turnover* and *staff.costs* that are discovered by the positivity check suggest function. In this case it is probably easy to see and to decide by the specialist. For more complex cases there is the R package *validatetools* [10] that can automatically detect and remove inconsistencies and redundancies in validation rule sets.

20. Several improvements can be made to the validatesuggest package. One line of thinking is to also detect some of the more advanced statistical checks and the international standardized ESS main types of validation rules. In particular validation rules that rely on standardized metadata as defined in SDMX registries could be detected if automatic access to such registries or a DSD is added as a feature. This would be a very practical feature, since it is common practice in the production of official statistic to use and include standardized variables and categories. Another practical addition could be to add interactive rule adjustment. Rule (suggestions) may need to be adapted, it is insightful to see what the impact on (the validity of) the dataset is when adjusting a rule. Finally, implementing or using a supervised method for rule generation would also be a useful addition.

V. Conclusions and reflections

21. In this paper we presented a new way to assist the rule based validation process found in official statistics. It takes existing data as a starting point and suggests validation rules to the statistical expert. Knowledge on common validation rule types from internationally agreed ESS main types of rules and from the validation cookbook was taken as an input. Rule suggestions can be generated from data either to create an initial rule set for a dataset, as well as to assist the domain expert in completing the set of validation rules. In all cases rule suggestion should allow for adding tolerances and ideally also for slight data errors.

22. If data on multiple production cycles is available the concept can be extended to take the time dimension in the data into consideration to infer rules that hold for the common characteristics of that particular stream of datasets as a whole. We make a distinction between ‘cross sectional derived rules’ that apply to all datasets of all periodic reports and ‘time series derived rules’ that depend on the time period in the data transmission, such as for example a growth rate.

23. An early implementation of the concept is available in the experimental R-package ‘validatesuggest’. It supports, positivity checks, range checks, checks on na, checks on uniqueness, type checks, ration checks and discovery of conditional rules in R validate syntax. Extensions could be in the direction of detecting statistical checks, connecting to SDMX metadata found in registries, to add interactivity in the suggestion process or to add supervised methods for rule generation.

24. With respect to the relationship with machine learning approaches, we think that supervised machine learning for error detection and validation seems promising, but an important advantage of using validation rules is that they are explicit, understandable and can be evaluated with ease. This makes it a natural tool for executing and explaining the validation process in official statistics: each invalid value is *explainable*. Rule based validation has therefore an edge over black box machine learning detection methods, which may do a good job at detecting if a record is (in)valid, but cannot explain what is wrong and should be adjusted. Interesting research in the realm of explainable artificial intelligence (XAI) suggests that rules may be derived from black box models.

VI. References

[1] Principles for data validation, https://ec.europa.eu/eurostat/cros/content/principles_en

[2] Methodology for data validation 2.0. Revised Edition 2018. https://ec.europa.eu/eurostat/cros/system/files/ess_handbook_-_methodology_for_data_validation_v2.0_-_rev2018_0.pdf

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

[4] 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

[5] R-package validate: https://cran.r-project.org/package=validate

[6] O. ten Bosch, M. van der Loo, *Quality assurance from an internationally standardized and generic data validation ecosystem*, European Conference on Quality in Official Statistics (Q 2022), Vilnius, June 2022, https://q2022.stat.gov.lt/scientific-information/papers-presentations/session-16

[7] M. van der Loo, O. ten Bosch, The Data Validation Cookbook: http://data-cleaning.github.io/validate/

[8] Unece GSBPM, https://statswiki.unece.org/display/GSBPM

[9] R-package validatesuggest: https://github.com/data-cleaning/validatesuggest

[10] R-package validatools: https://cran.r-project.org/package=validatetools