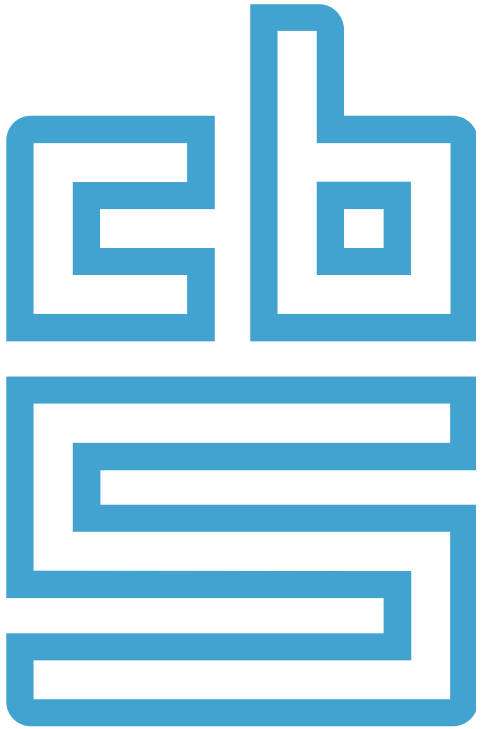
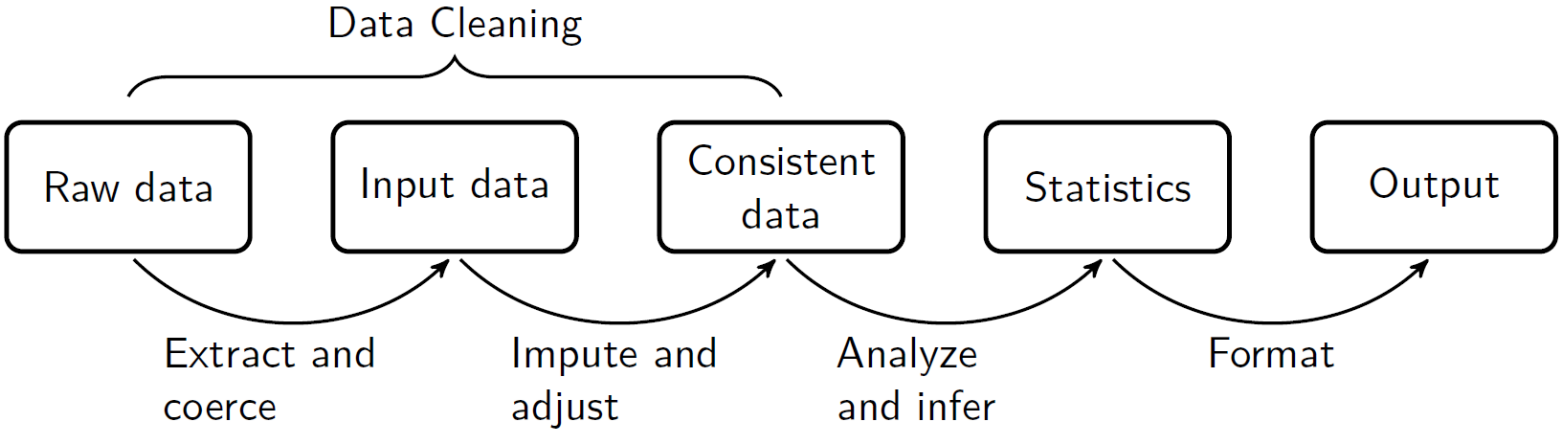


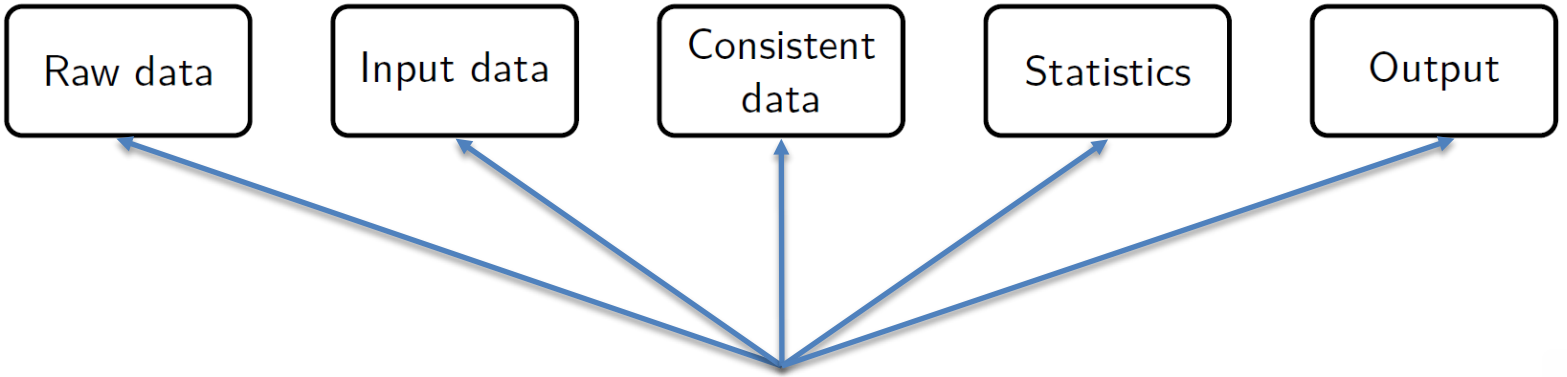
Data validation in national and international context: where are we and where are we going?

Mark van der Loo
Statistics Netherlands
EMOS webinar 2021-03-16

Statistical production

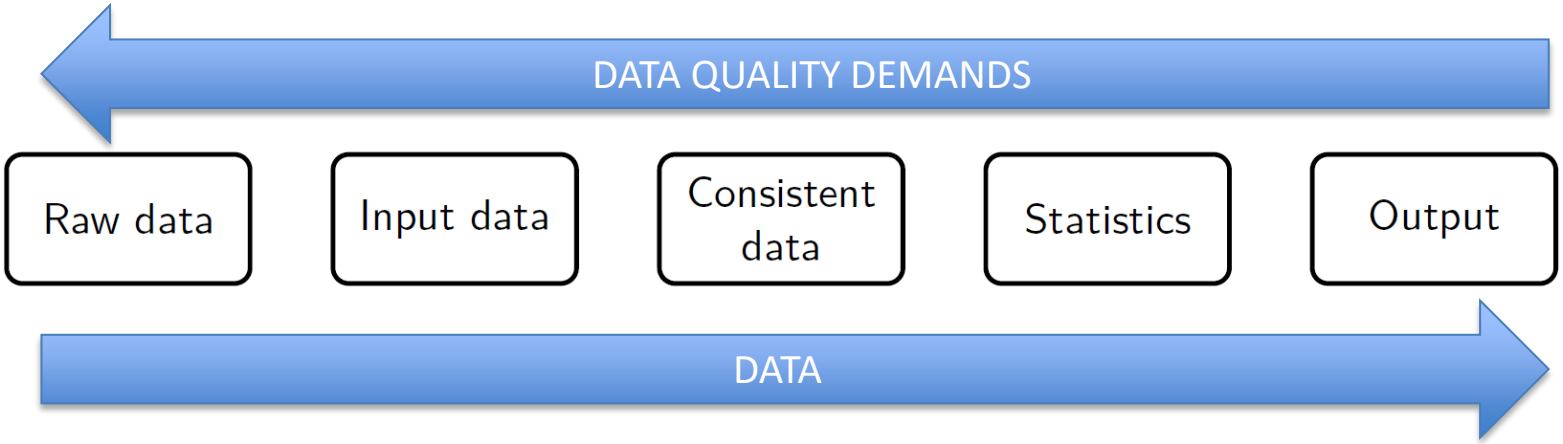


Statistical production



Statistical Products: data with a guaranteed level of quality

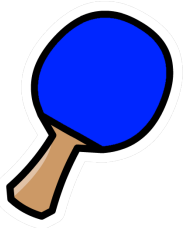
Statistical production



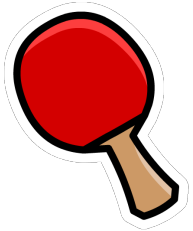
Validation in international context

Output = Input

eurostat

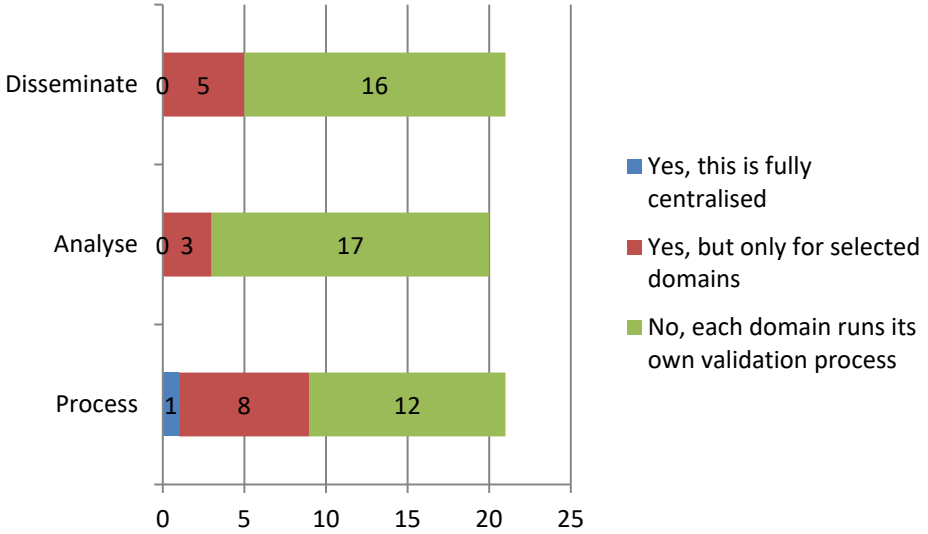


42



Data validation in NSIs (2015)

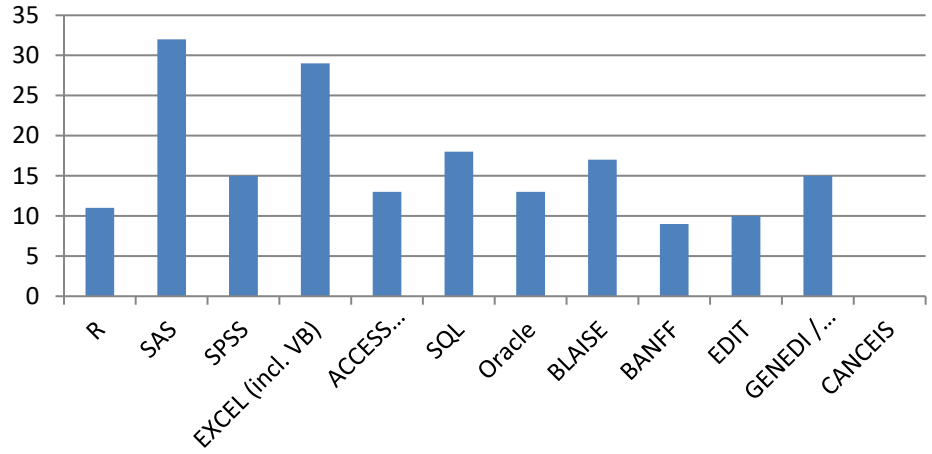
Do you use central validation services?

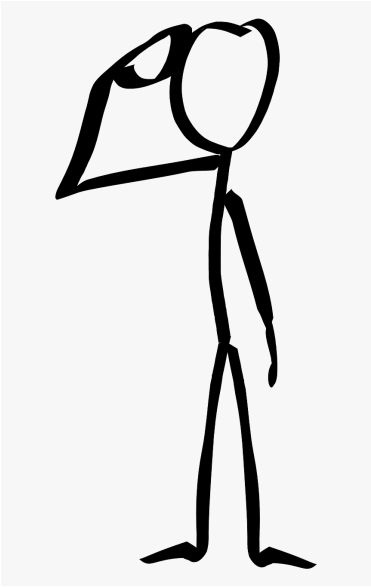


Source: ESSnet validatFoundation (2015)

Data validation in NSIs (2015)

Which tools do you use?

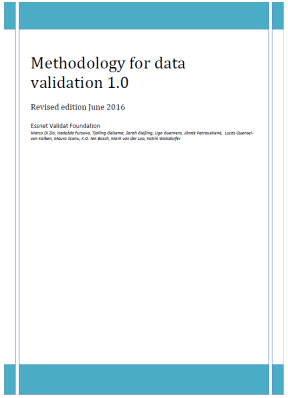




Questions…

Towards a common understanding: definition, principles, and methodology

Definition of data validation


Methodology for data validation 1.0

Revised edition June 2016

Essnet Validat Foundation

An activity in which one verifies whether or not a combination of values is acceptable.

(ESS Handbook on data validation)

- Is Age a positive number?
- Turnover – Costs equals Profit?
- Average profit change less than 10%?

Data Validation Principles

1. The sooner, the better
2. Trust, but verify
3. Well-documented and appropriately communicated validation rules
4. Well-documented and appropriately communicated validation errors
5. Comply or Explain
6. Good enough is the new perfect

Principle 1

The sooner the better

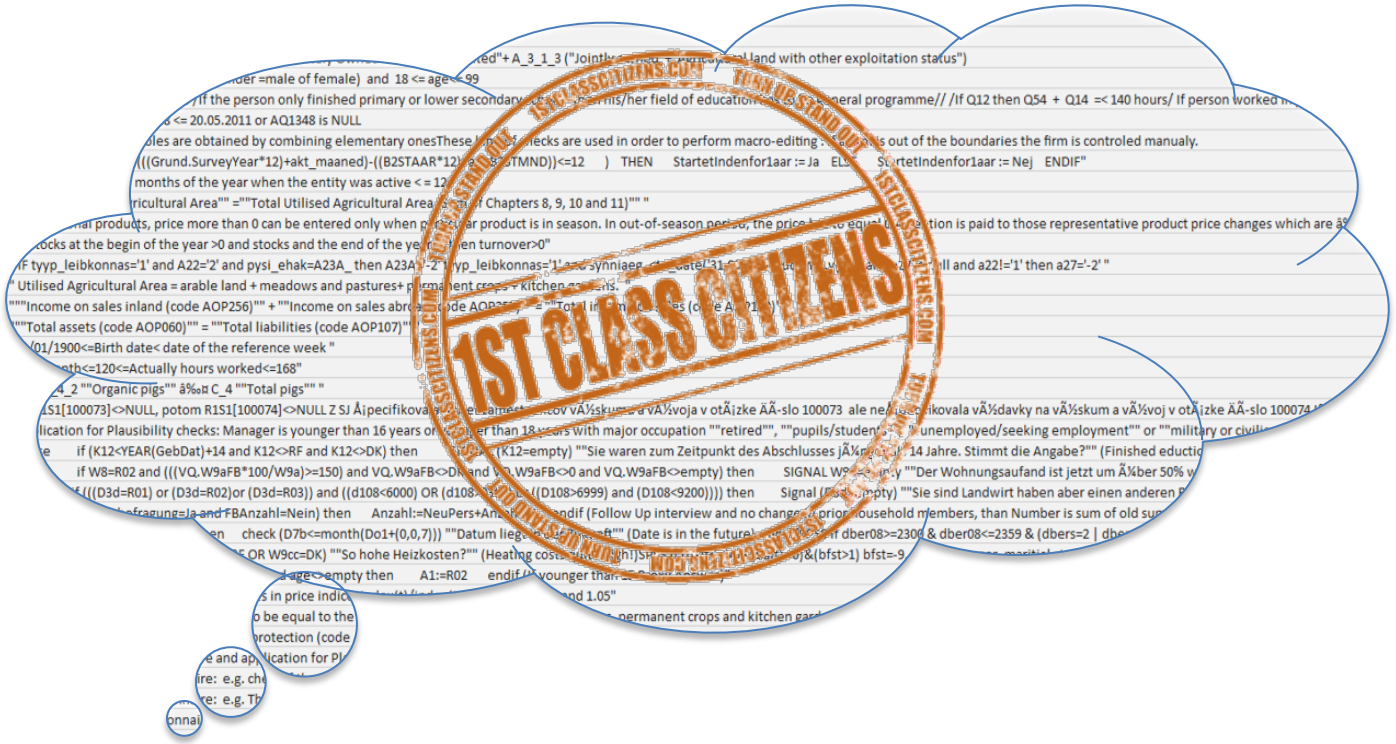
Raw data → Input data → Consistent data → Statistics → Output

Propagation of errors

As soon as possible, but not earlier than that

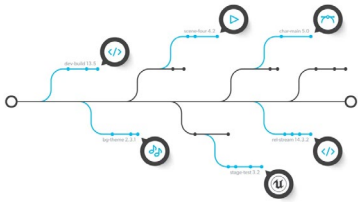
Trust but verify

WELL-DOCUMENTED AND APPROPRIATELY COMMUNICATED VALIDATION RULES



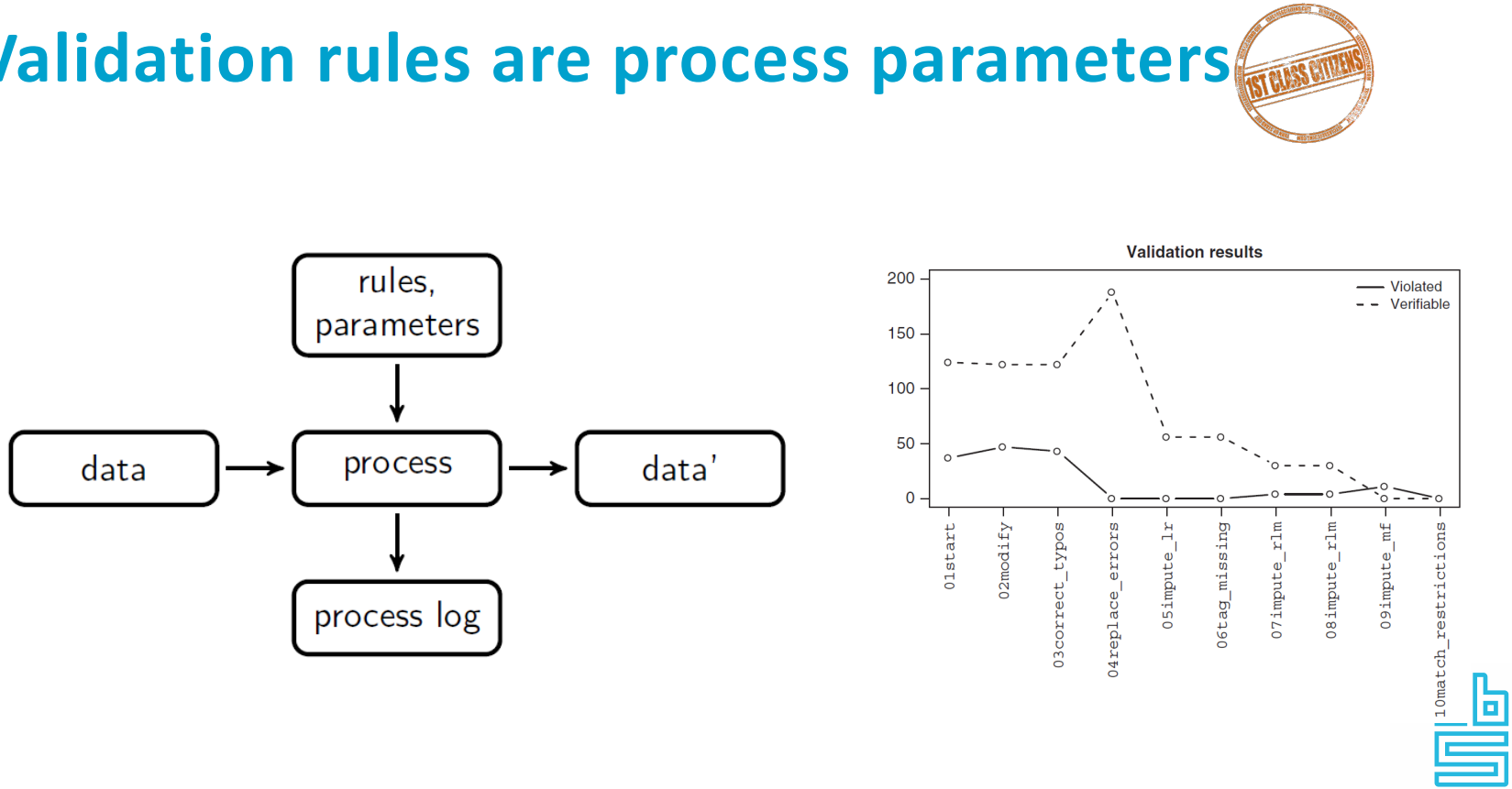
Validation rules are data

Validation rules are source code

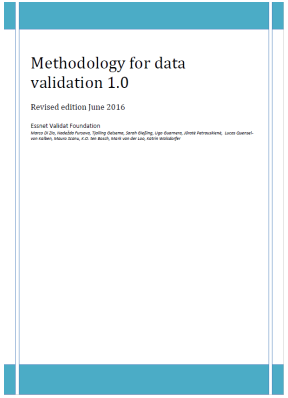


- Version control
- Review
  - (still) valid?
  - (still) relevant?
  - Understandable?
  - Redundant? / Contradictions?
- Naming & documentation

Validation rules are process parameters



Definition of data validation -- revisited


Methodology for data validation 1.0

Revised edition June 2016

Essnet Validat Foundation

An activity in which one verifies whether or not a combination of values is acceptable.

(ESS Handbook on data validation)

- Is *Age* a positive number?
- *Turnover – Costs* equals *Profit*?
- *Average profit* positive?
- *Change in Average Profit Margin* less than 10%?

Data validation theory

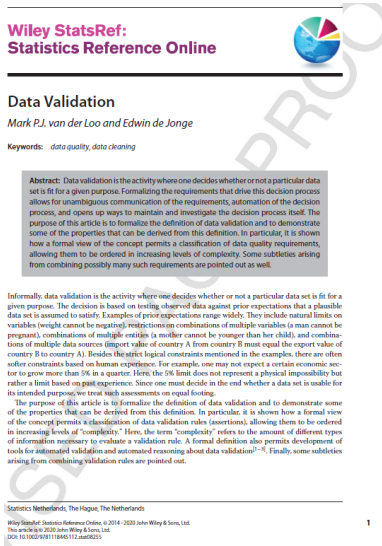
Definition 3 *A* data validation function *is a surjective function*

$$v : D^K \twoheadrightarrow \{False, True\}.$$

Table 1. The 10 possible classes of validation rules, grouped into "validation levels."

	Validation level			
0	1	2	3	4
ssss	sssm ssms smss	ssmm smsm smms	smmm msmm	mmmm

A higher level indicates that a wider variety of information is necessary to evaluate a validation rule.



Wiley StatsRef:
Statistics Reference Online

Data Validation

Mark P.J. van der Loo and Edwin de Jonge

Keywords: *data quality, data cleaning*

Abstract: Data validation is the activity where one decides whether or not a particular data set is fit for a given purpose. Formalizing the requirements that drive this decision process allows for unambiguous communication of the requirements, automation of the decision process, and opens up ways to maintain and investigate the decision process itself. The purpose of this article is to formalize the definition of data validation and to demonstrate some of the properties that can be derived from this definition. In particular, it is shown how a formal view of the concept permits a classification of data quality requirements, allowing them to be ordered in increasing levels of complexity. Some subtleties arising from combining possibly many such requirements are pointed out as well.

Informally, data validation is the activity where one decides whether or not a particular data set is fit for a given purpose. The decision is based on testing observed data against prior expectations that a plausible data set is assumed to satisfy. Examples of prior expectations range widely. They include natural limits on variables (weight cannot be negative), restrictions on combinations of multiple variables (a man cannot be pregnant), combinations of multiple entities (a mother cannot be younger than her child), and combinations of multiple data sources (import value of country A from country B must equal the export value of country B to country A). Besides the strict logical constraints mentioned in the examples, there are often softer constraints based on human experience. For example, one may not expect a certain economic sector to grow more than 5% in a quarter. Here, the 5% limit does not represent a physical impossibility but rather a limit based on past experience. Since one must decide in the end whether a data set is usable for its intended purpose, we treat such assessments on equal footing.

The purpose of this article is to formalize the definition of data validation and to demonstrate some of the properties that can be derived from this definition. In particular, it is shown how a formal view of the concept permits a classification of data validation rules (assertions), allowing them to be ordered in increasing levels of "complexity." Here, the term "complexity" refers to the amount of different types of information necessary to evaluate a validation rule. A formal definition also permits development of tools for automated validation and automated reasoning about data validation[2–9]. Finally, some subtleties arising from combining validation rules are pointed out.

Statistics Netherlands, The Hague, The Netherlands

Wiley StatsRef: Statistics Reference Online, © 2014 - 2020 John Wiley & Sons, Ltd.
This article is © 2020 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat08255

MPJ van der Loo and E de Jonge (2019)
Wiley StatsRef Online
arxiv.org/abs/2012.12028

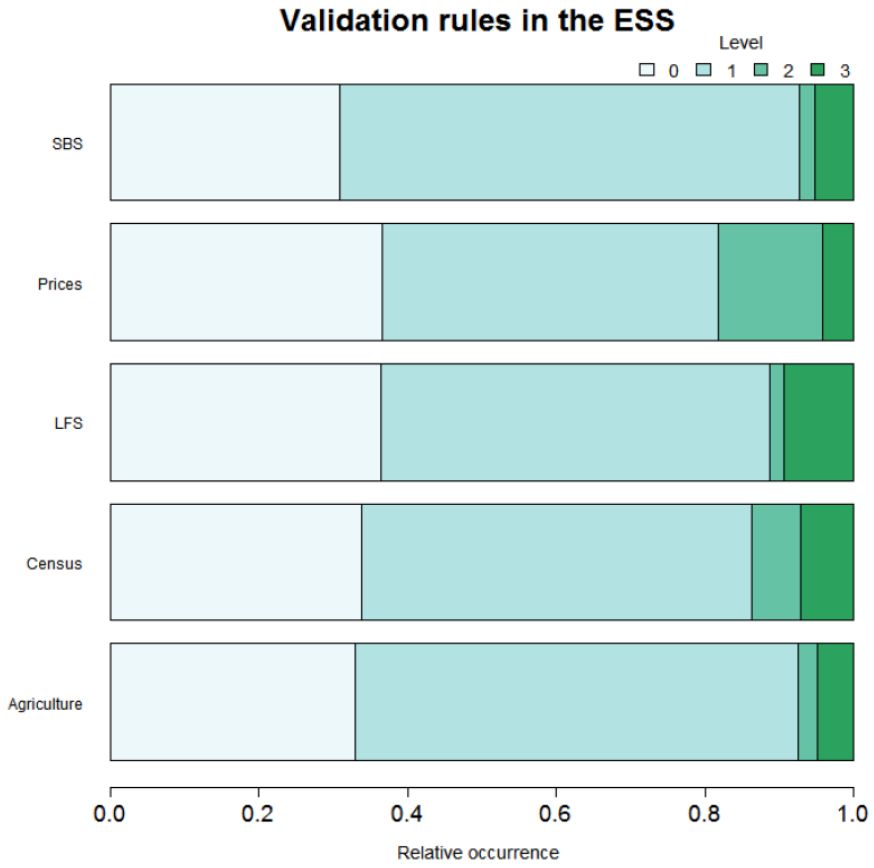
Do I need more than one…

- Entity type?
- Time point or period?
- Population unit?
- Variable?

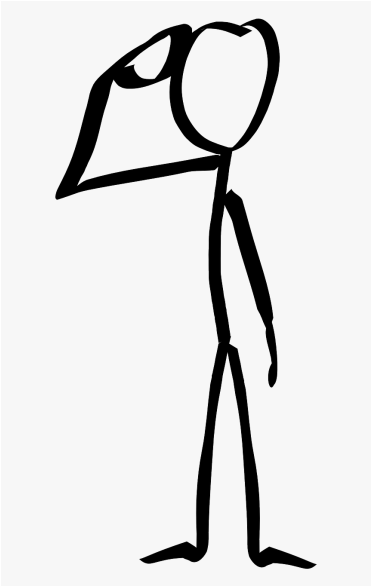
If '**no**' assign an *s*
If '**yes**' assign an *m*

Complexity level = number of *m*'s assigned to a rule.

Rule	Level
Is *Age* a positive number?	0
Turnover minus *Costs* equals *Profit*?	1
Average profit positive?	1
Change in Average Profit Margin less than 10%?	3



Analyses of ~1300 rules accross 5 statistical domains in 28 EU member states.



Questions…

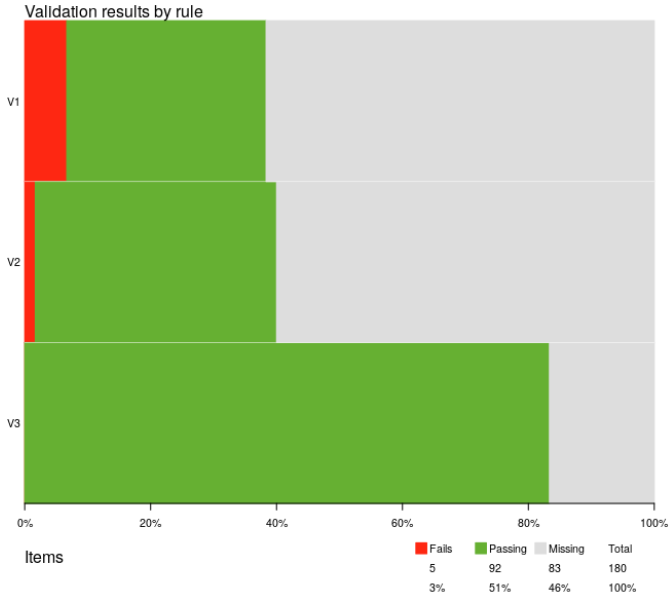
Tools, and looking forward

Implementation: R package validate


```
rules <- validator(
  turnover + other.rev == total.rev
  , other.rev >= 0
  , if (staff > 0) staff.costs > 0
)

out <- confront(retailers, rules)

plot(out)
```



Journal of Statistical Software

Data Validation Infrastructure for R

Mark P.J. van der Loo
Statistics Netherlands

Edwin de Jonge
Statistics Netherlands

Abstract

Keywords: data checking, data quality, data cleaning, R.

1. Introduction

MPJ van der Loo and E de Jonge (2020)
Data validation infrastructure for R. J. Stat. Soft (accepted)
arxiv.org/abs/1912.09759

Validate: rules are source code with metadata


```
- expr: any(FREQ == meta$FREQ & INDICATOR == meta$INDICATOR & TIME_PERIOD == meta$PERIOD)
  name: "STS01"
  label: "Correct series"
  description: |
    The indicators, the periodicity and the last observation
    period of at least one time series must be the same as in
    the identification in the EDAMIS flow.
```

Rule taken from STS transmission guidelines.
Source: https://github.com/SNStatComp/DomainValidationRules

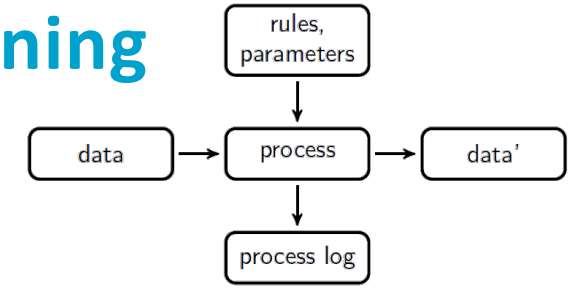
Validate: rules are data


```
> library(validate)
> rule1 <- validator( x >= 0)
> rule2 <- validator( y >= 0, z >= 0)
> allrules <- rule1 + rule2
> allrules[1:2]
Object of class 'validator' with 2 elements:
 V1   : x >= 0
 V1.1: y >= 0
```

Validate: rules can be investigated

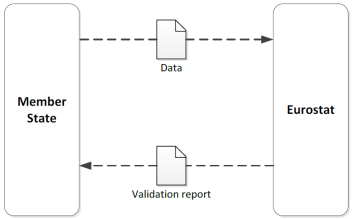
```
> rules <- validator(x > y, y > x)
> variables(rules)
[1] "x" "y"
>
> validatetools::detect_infeasible_rules(rules)
[1] "V1"
>
```

Validate: use rules for data cleaning



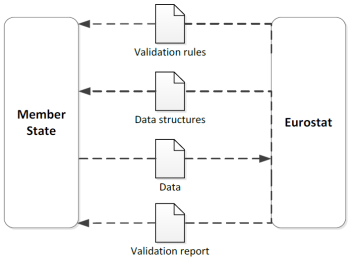

```
> SBS
  cost profit turnover
1   10     NA       15
>
> rules <- validator(turnover - cost == profit)
>
> deductive::impute_lr(SBS, rules)
  cost profit turnover
1   10      5       15
```

Future of data validation in the ESS



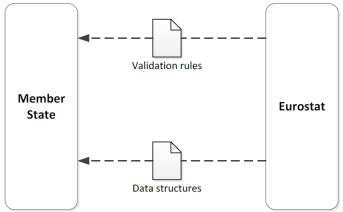
Use ESTAT validation service

OR



Download rules, install validation service locally

OR

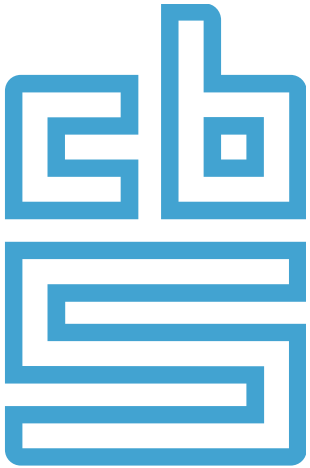


Download rules, run on your own software (e.g. validate)

Validation rules are agreed upon by domain working groups

Thank you for your attention

- The Data Validation Cookbook (data-cleaning.github.io)
- Data Validation - Overview | CROS (europa.eu)
- ESS Handbook - Methodology for data validation v1.1 - Rev2018 | CROS (europa.eu)
- [2012.12028] Data Validation (arxiv.org)
- [1912.09759] Data Validation Infrastructure for R (arxiv.org)



Facts that matter