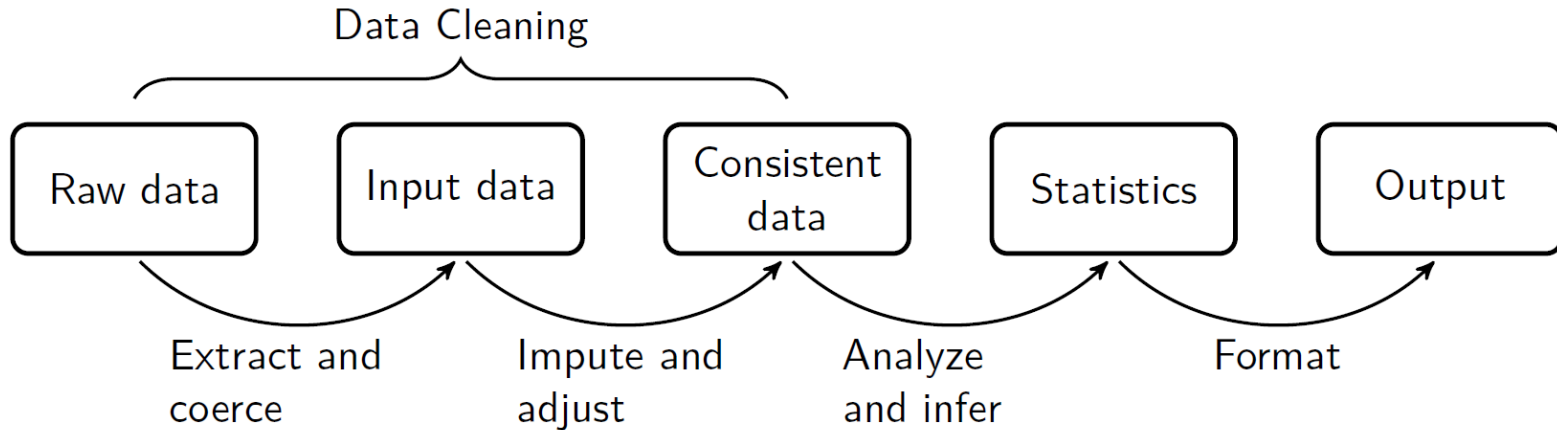


Data validation in national and international context: where are we and where are we going?

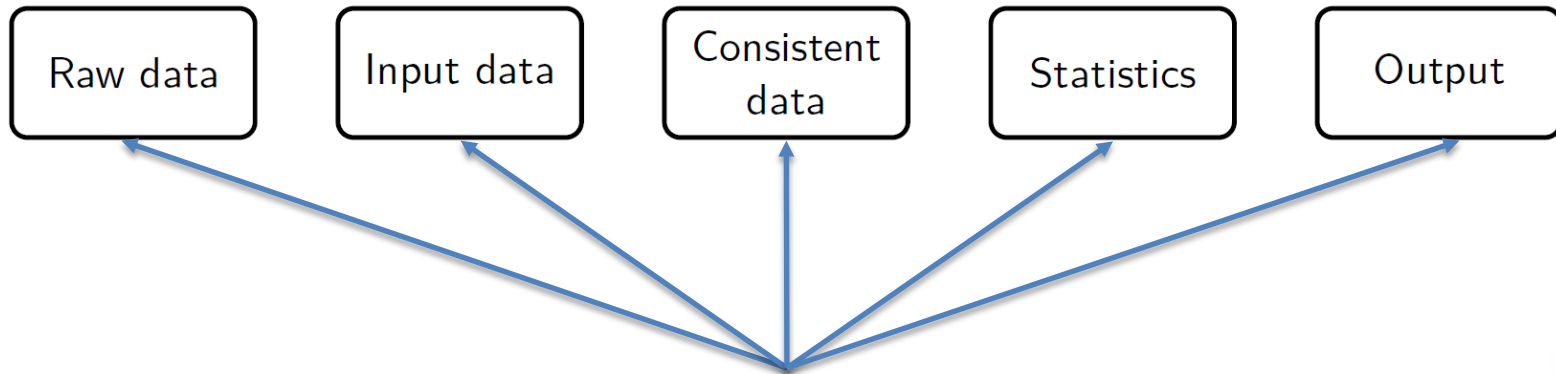
Mark van der Loo
Statistics Netherlands
EMOS webinar 2021-03-16



Statistical production



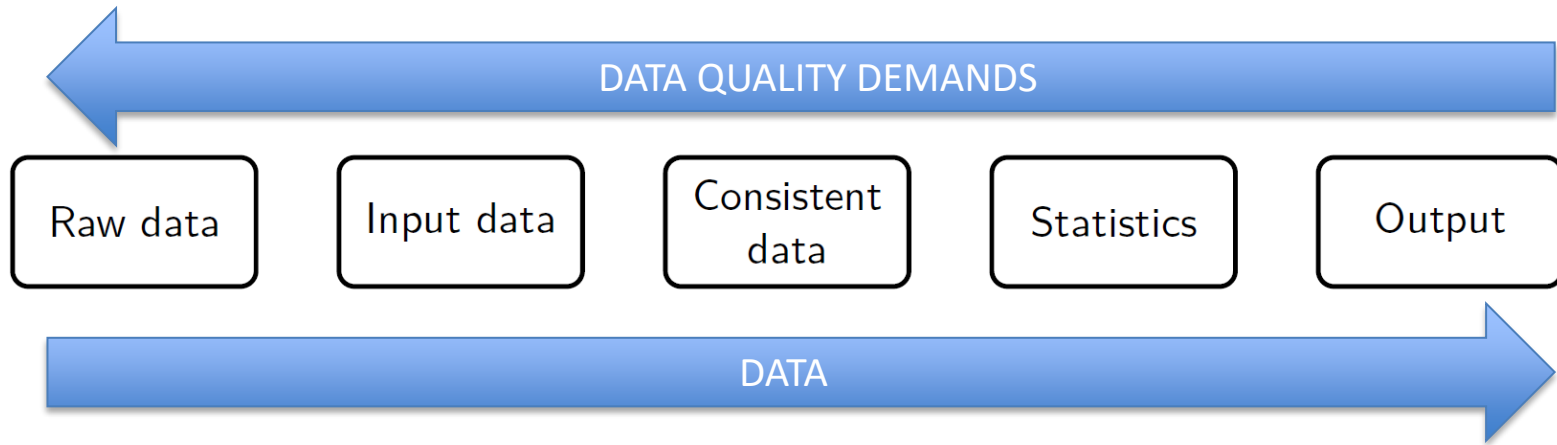
Statistical production



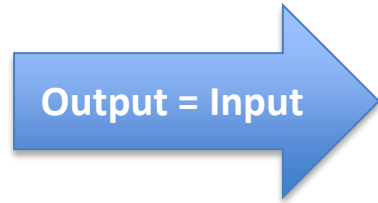
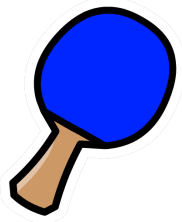
Statistical Products: data with a guaranteed level of quality



Statistical production

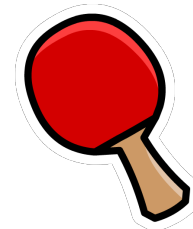


Validation in international context



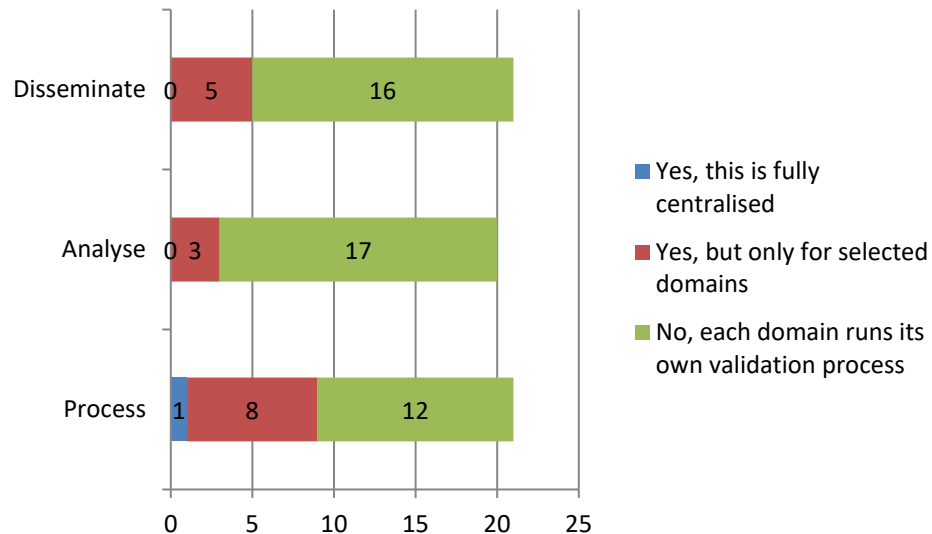
42

eurostat 



Data validation in NSIs (2015)

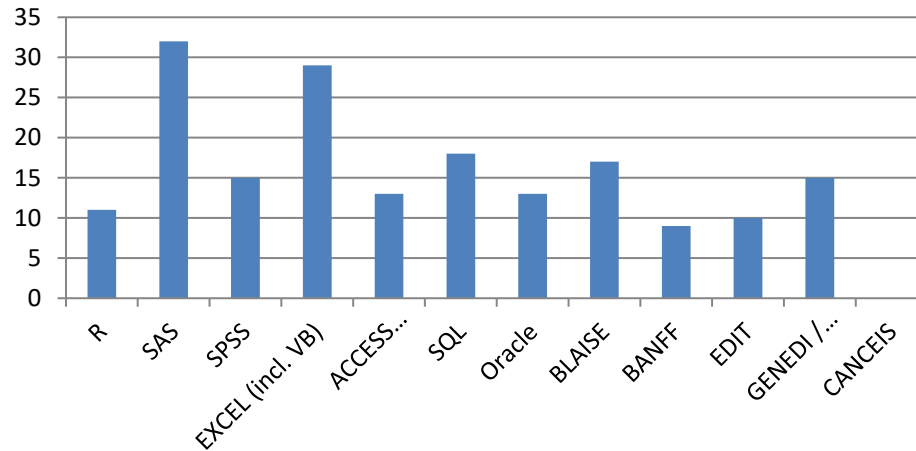
Do you use central validation services?

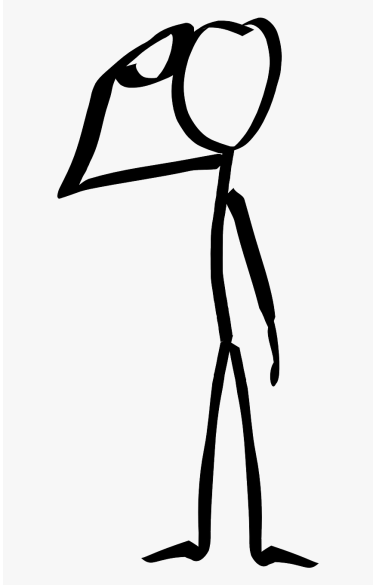


Source: ESSnet validatFoundation (2015)

Data validation in NSIs (2015)

Which tools do you use?





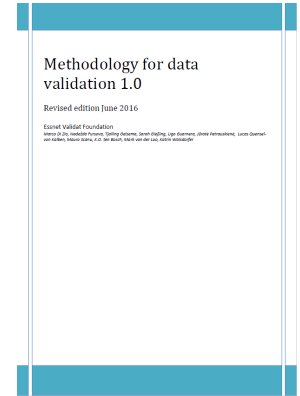
Questions...



Towards a common understanding: definition, principles, and methodology



Definition of data validation



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%?

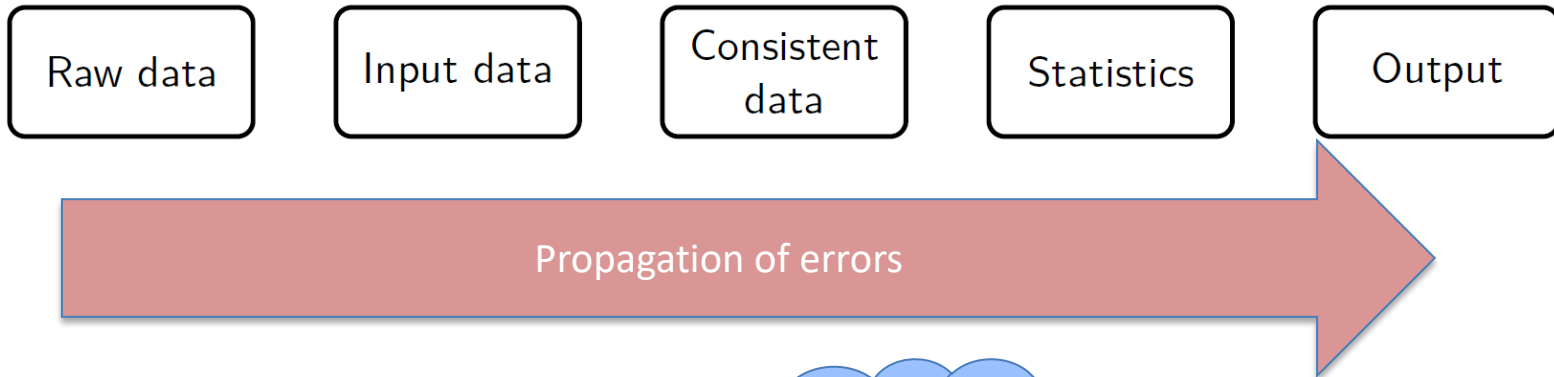


Data Validation Principles

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



The sooner the better



As soon as possible, but not earlier than that



Trust but verify

Доверяй, но проверяй



WELL-DOCUMENTED AND APPROPRIATELY COMMUNICATED VALIDATION RULES

(ed)+A_3_1_3 ("Jointly owned agricultural land with other exploitation status")
 (gender=male of female) and 18 <= age < 99
 // If the person only finished primary or lower secondary school, then his/her field of education is not applicable. // If Q12 then Q54 + Q14 <= 140 hours/ If person worked in
 <= 20.05.2011 or AQ1348 is NULL
 Rules are obtained by combining elementary ones. These checks are used in order to perform macro-editing. If the data is out of the boundaries the firm is controlled manually.
 (((Grund.SurveyYear*12)+akt_maaned)-((B2STAAR*12)+B2TMND))<=12) THEN Startetindenfor1aar := Ja ELSE Startetindenfor1aar := Nej ENDIF"
 months of the year when the entity was active <= 12
 Agricultural Area"" = ""Total Utilised Agricultural Area (see Chapters 8, 9, 10 and 11)""
 Agricultural products, price more than 0 can be entered only when particular product is in season. In out-of-season periods, the price is set equal to 0. Attention is paid to those representative product price changes which are 3%
 stocks at the begin of the year >0 and stocks at the end of the year >0 when turnover>0"
 IF ttyp_leibkonnas='1' and A22='2' and phys_ehak=A23A_ then A23A_-'2' ttyp_leibkonnas='1' and synhlaeg <= 1 and a21='1' and a22='1' then a27='-2' ""
 Utilised Agricultural Area = arable land + meadows and pastures + permanent crops + kitchen gardens
 ""Income on sales inland (code AOP256)"" + ""Income on sales abroad (code AOP257)"" = ""Total income on sales (code AOP258)""
 ""Total assets (code AOP060)"" = ""Total liabilities (code AOP107)""
 /01/1900<=Birth date<= date of the reference week"
 <=120<=Actually hours worked<=168"
 A_2 ""Organic pigs"" &%C_4 ""Total pigs""
 S1[100073]>NULL, potom R1S1[100074]>NULL Z SJ Apeficikovala v domacnosti v otazke AA-slo 100073 ale nebyla ziskovala v davky na vaskum a v voj v otazke AA-slo 100074
 Plausibility checks: Manager is younger than 16 years or older than 18 years with major occupation ""retired"", ""pupils/student"", ""unemployed/seeking employment"" or ""military or civilian"
 e IF (K12<YEAR(GebDat)+14 and K12<RF and K12<DK) then Signal W9cc=DK ""(K12=empty) ""Sie waren zum Zeitpunkt des Abschlusses jÃhrlich 14 Jahre. Stimmt die Angabe?"" (Finished education
 if W8=R02 and (((VQ.W9aFB*100/W9a)>=150) and VQ.W9aFB<DK and VQ.W9aFB<0 and VQ.W9aFB<empty) then Signal W9cc=DK ""Der Wohnungsaufand ist jetzt um Ãber 50% von
 (((D3d=R01) or (D3d=R02) or (D3d=R03)) and (((D108>6000) OR ((D108>9999) and (D108>9200)))) then Signal (D3d=empty) ""Sie sind Landwirt haben aber einen anderen P
 franzug=Ja and FBAnzahl=Nein) then Anzahl:=NeuPers+Anzahl endif (Follow Up interview and no changes prior household members, than Number is sum of old sur
 en check (D7b<=month(Do1+(0,0,7))) ""Datum liegt in der Zukunft"" (Date is in the future) (dber08>=2300 & dber08<=2359 & (dbers=2 | dber
 OR W9cc=DK) ""So hohe Heizkosten?"" (Heating costs too high) (b1fst>9 & (bfst>1) b1fst=9
 age<empty then A1:=R02 endif (younger than
 s in price indic
 to be equal to the
 protection (code
 lication for PL

e and ap
 re: e.g. ch
 ve: e.g. Th
 onnal

Validation rules are data



CREATE



READ



UPDATE



DELETE

C

R

U

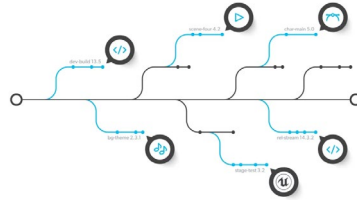
D



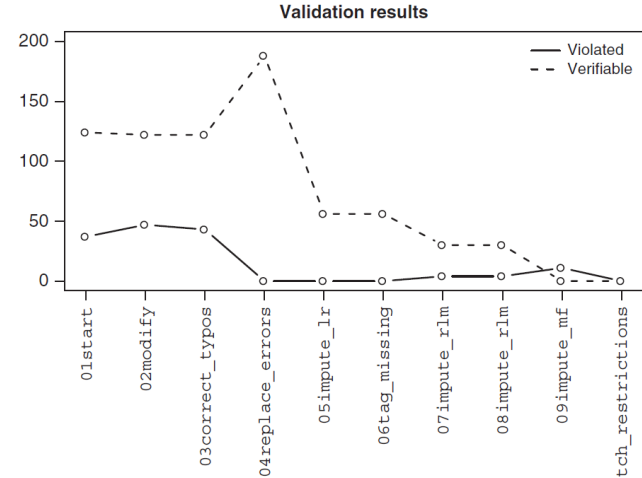
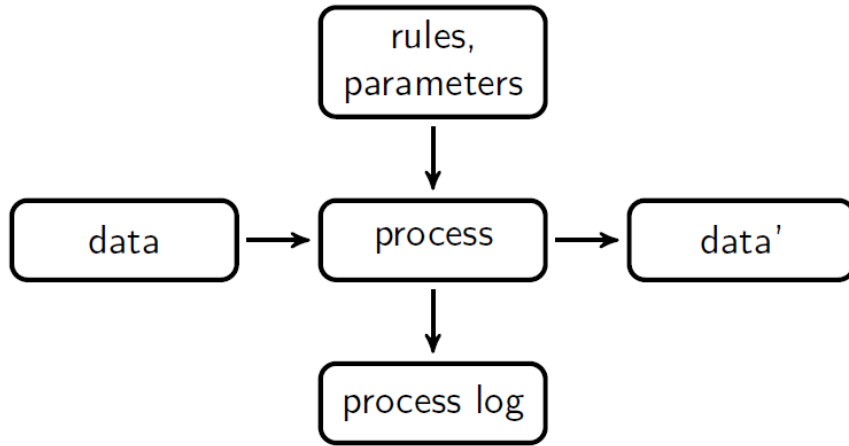
Validation rules are source code



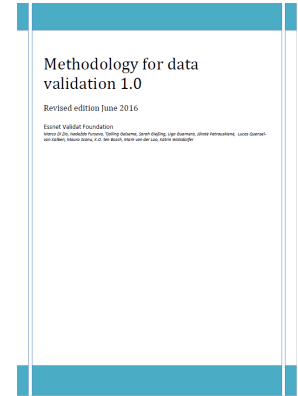
- Version control
- Review
 - (still) valid?
 - (still) relevant?
 - Understandable?
 - Redundant? / Contradictions?
- Naming & documentation



Validation rules are process parameters



Definition of data validation -- revisited



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

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



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 0% in a quarter. Here, the 0% 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 in most high 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^{1–8}. 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/stat.1144

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

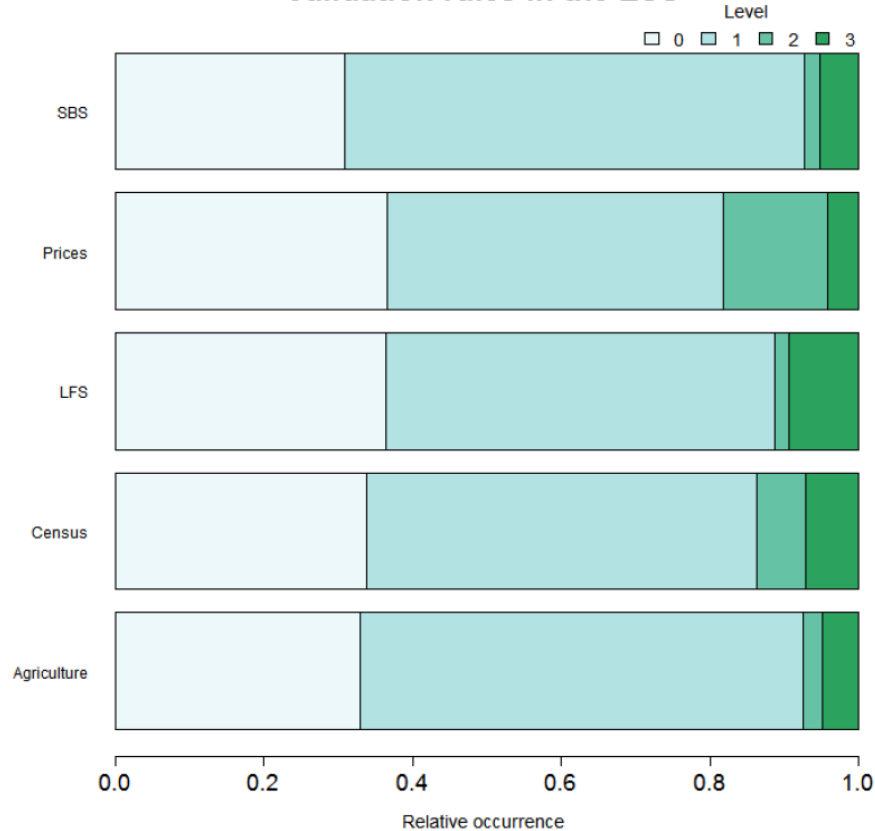
Complexity level = number of m 's assigned to a rule.



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

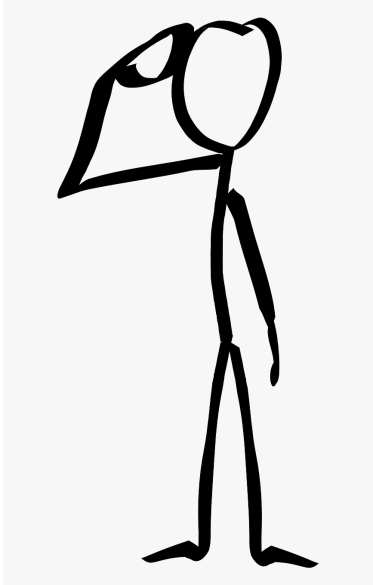


Validation rules in the ESS



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





Questions...



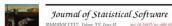
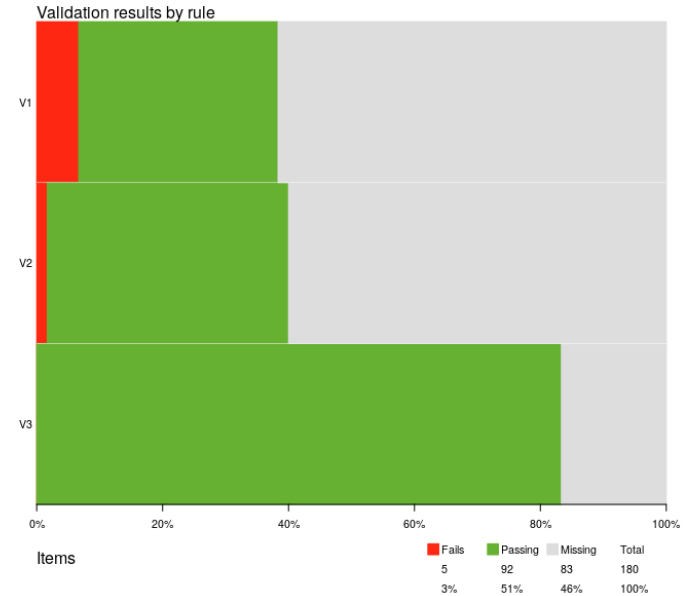
Tools, and looking forward



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)
```



Data Validation Infrastructure for R

Mart P.J. van der Loo
Ede de Jonge

1. Introduction

The Data Validation Infrastructure for R (DVI) is a software package that provides a framework for data validation in R. It is designed to be used by data analysts and researchers who need to ensure the quality of their data before performing any analysis. The DVI provides a set of tools and functions that can be used to check for errors and inconsistencies in data. It also provides a way to generate reports and visualizations that can be used to communicate the results of the validation process.

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
```

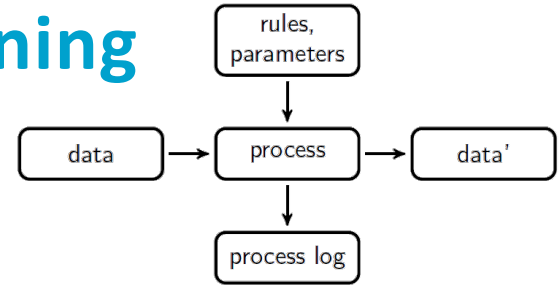
```
>
```

```
> rules <- validator(turnover - cost == profit)
```

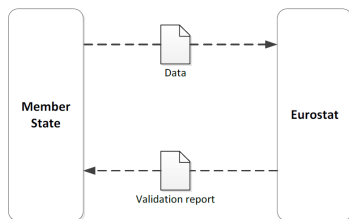
```
>
```

```
> deductive::impute_lr(SBS, rules)
```

```
  cost profit turnover  
1   10      5       15
```

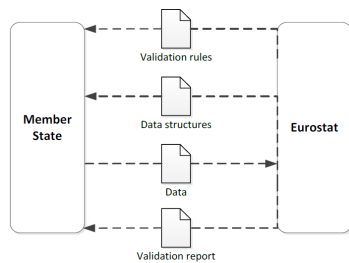


Future of data validation in the ESS



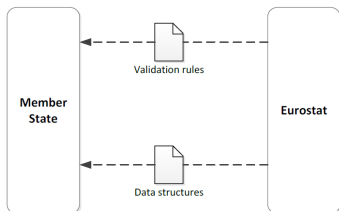
Use ESTAT validation service

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



Facts that matter