Extending data validation with standardised metadata from SDMX registries

Olav ten Bosch, Mark van der Loo
Statistics Netherlands

9th International Conference The Use of R in Official Statistics, ***uRos2021***
November 2021

Contents

- International data validation
- SDMX registries
- Data cleaning with R
- Connecting R-validate to SDMX
- Wrap up

International data validation (1)

Problem:

- ***Invalid*** data in international data reporting may lead to ***costly retransmissions*** or ***reprocessing*** of statistics, ‘data ping pong’

Solution:

- Agree on ***rules*** (statistical working groups)
- ***Validate*** data against these rules ***on both sides***

International data validation (2)

Results from earlier projects:

- Handbook on validation

> *Data Validation is an activity verifying whether or not a combination of values is a member of a set of acceptable combinations.*

- Validation principles

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*

- "Main types of rules"(Eurostat):
  minimal set of rules covering most of the validation needs in ESS

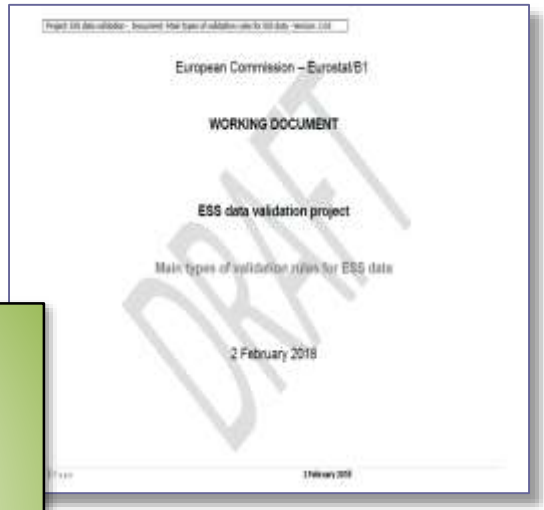
https://ec.europa.eu/eurostat/cros/content/validation-ess

International data validation (3)

'*Main types of validation rules*'

- FDT: FielD Type
- FDL: FielD Length
- FDM: FielD is Manatory or empty
- COV: COdes are Valid
- RWD: Records are Without Duplicate id
- REP: Records Expected are Provided
- RTS: Records are all present for Time Se
- RNR: Records' Number is in a Range
- COC: COdes are Consistent
- VIR: Values are In a Range
- VCO: Values are COnsistent
- VAD: Valueas for Aggregates are consistent with Details
- VSA: Values for Seasonally Adjusted data are plausible

ValidatFOSS2:
Can we easily use internationally agreed (SDMX) metadata to automatically derive such rules for data validation?


European Commission – Eurostat/B1

WORKING DOCUMENT

ESS data validation project

Main types of validation rules for ESS data

2 February 2018

DRAFT

Eurostat, 2018

ValidatFOSS1: rules implemented in R package:
https://github.com/SNStatComp/GenericValidationRules

What is SDMX?

- SDMX: ***S***tatistical ***D***ata and ***M***etadata E***x***change
- SDMX consortium (2001):
  - BIS, ECB, EuroStat, IMF, OECD, UN, WorldBank
  - www.sdmx.org
- Open (ISO 17369:2013) standard for the exchange of statistical data
- UN 2008: SDMX chosen for International data exchange in ESS
- SDMX versions: 1.0, 2.0, 2.1; 3.0 (released 2021)
- Generic information model for (multidimensional) statistical data
- Central metadata management

SDMX registries (1)

registry.sdmx.org

webgate.ec.europa.eu/sdmxregistry

sdmxcentral.imf.org/overview.html

sdmx.data.unicef.org

Other & Internal SDMX registries

SDMX registries (2)

Codelists

SDMX	Id	Name	State
SDMX	CL_AGE	Age	Final
SDMX	CL_AREA	Reference area code list	Final
SDMX	CL_CIVIL_STATUS	Civil (or Marital) Status	Final
SDMX	CL_COFOG_1999	Classification of the Functions o...	Final
SDMX	CL_COICOP_1999	Classification of Individual Cons...	Final
SDMX	CL_CONF_STATUS	Confidentiality Status	Final
SDMX	CL_COPNI_1999	Classification of the Purposes o...	Final
SDMX	CL_COPP_1999	Classification of the Outlays of ...	Final
SDMX	CL_DECIMALS	Decimals	Final

Viewing: Civil (or Marital) Status [1.0]

Position	Id	Name
1	S	Single person
2	M	Married person
3	W	Widowed person
4	D	Divorced person
5	L	Legally separated person
6	P	Person in registered partnership
7	Q	Person whose registered partnership ended with the death of the partner
8	E	Person whose registered partnership was legally dissolved

- Access to many metadata resources
- concepts, variable definitions, data flows, structures and code lists
- Versioning, ownership

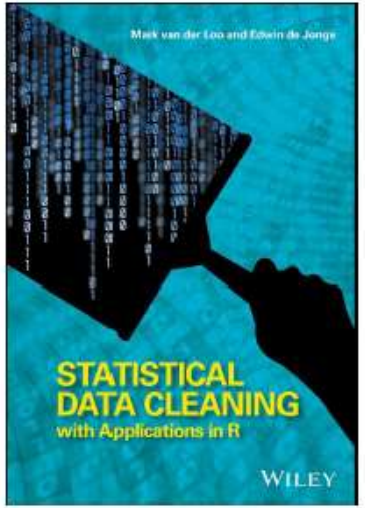
SDMX registries (3)


Other & Internal SDMX registries

- Programmatic access via SDMX 2.1 REST API: https://github.com/sdmx-twg/sdmx-rest
- ☹ Some registry implementations offer slightly ***different*** functionality
- ☺ We found ***one generic access method*** for all registries using rsdmx: cran.r-project.org/package=rsdmx
- Demo notebooks Python & R in: https://github.com/SNStatComp/validatesdmx

Data cleaning with R (1)

MPJ van der Loo and E de Jonge (2018)
Statistical data cleaning with applications in R
John Wiley & Sons, NY.



- R data cleaning ecosystem
  - ***validate***: check data based on validation rules
  - ***dcmodify***: change data based on 'if-this-then-that' rules
  - ***errorlocate***: locate errors based on validation rules and mark them for correction
  - ***simputation***: many different imputation methods
  - ***rspa***: adapt numerical records to fit (in)equality restrictions
  - ***deductive***: solve errors based on control rules
  - ***validatetools***: find inconsistencies and redundancies

Data cleaning with R (2)

Rules

```
# Range limits:
Age >= 0
Age <= 120
Working_hours >= 0
Working_hours <= 100

# Some checks between variables:
if (Married > 0) Age > 18
if (Working_hours > 0) Employed > 0

#Such a rule depends on country legislation:
if (Age > 65) Working_hours = 0

# ID must be unique
any(duplicated(ID)) == FALSE
```

Data

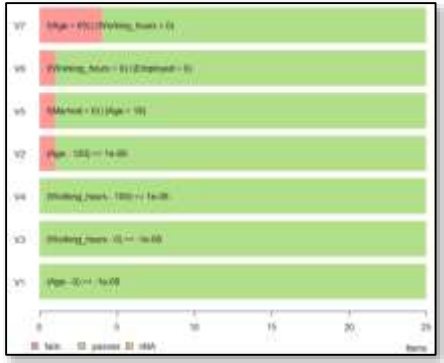
ID	Age	Marital status	Status in employment	Working hours per week
1	36	0	1	40
2	40	1	1	40
3	25	0	0	0
4	31	0	1	20
5	62	1	1	43
6	55	1	1	41
7	34	1	1	40

confront

Summary

```
> summary(validation)
  name items passes fails nNA error warning                                expression
1   V1    25     25     0   0 FALSE   FALSE                          (Age - 0) >= -1e-08
2   V2    25     24     1   0 FALSE   FALSE                        (Age - 120) <= 1e-08
3   V3    25     25     0   0 FALSE   FALSE              (Working_hours - 0) >= -1e-08
4   V4    25     25     0   0 FALSE   FALSE            (Working_hours - 100) <= 1e-08
5   V5    25     24     1   0 FALSE   FALSE                 !(Married > 0) | (Age > 18)
6   V6    25     24     1   0 FALSE   FALSE !(Working_hours > 0) | (Employed > 0)
7   V7    25     21     4   0 FALSE   FALSE      !(Age > 65) | (Working_hours = 0)
8   V8     1      0     1   0 FALSE   FALSE           any(duplicated(ID)) == FALSE
```

Per rule



Dashboard: data & results

Connecting R-validate to SDMX (1)

R-validate (≥ 1.1.0) now supports:

- Rules based on ***any codelists*** from ***any registry***
- ***Caching*** of registry results within a session
- Convenience functions for global and ESTAT registry

function	what it does
`sdmx_endpoint`	retrieve URL for SDMX endpoint
`sdmx_codelist`	retrieve sdmx codelist
`estat_codelist`	retrieve codelist from Eurostat SDMX registry
`global_codelist`	retrieve codelist from Global SDMX registry
`validator_from_dsd`	derive validation rules from DSD in SDMX registry

- Result can be used in a natural way in R expressions:

```
Activity %in% global_codelist(agency_id="ESTAT", resource_id="CL_ACTIVITY")
```

Connecting R-validate to SDMX (2)

- Deriving all rules from a DSD:

```
# import data
my_data <- read.csv(“mydata.csv”)

# derive all rules from a DSD
rules <- validator_from_dsd(endpoint = sdmx_endpoint("ESTAT")
   , agency_id = "ESTAT", resource_id = "STSALL", version="latest")

# confront data with rules
out <- confront(my_data, rules)

# plot results
plot(out)
```

- Generates multiple codelist rules derived from the ESTAT registry
- Easy to integrate in statistical processes

Wrap-up

- The ESS works on improving ***international data validation*** e.g. handbook, principles, main types of rules
- R has a rich ***data-cleaning ecosystem*** covering many of today's validation needs. Main types of rules implemented.
- R-validate now ***supports checks*** derived from ***any*** version of ***any*** codelists from ***any SDMX registry***
- Functionality documented in new chapter on SDMX in on-line ***cookbook:*** data-cleaning.github.io/validate
- Next steps? Usage? Other SDMX artefacts? Relation VTL?

Questions, ideas, suggestions

Olav ten Bosch	o.tenbosch@cbs.nl	@olavtenbosch
Mark van der Loo	mpj.vanderloo@cbs.nl	@markvdloo

awesomeofficialstatistics.org awesome ☆ Star 184 Fork 48