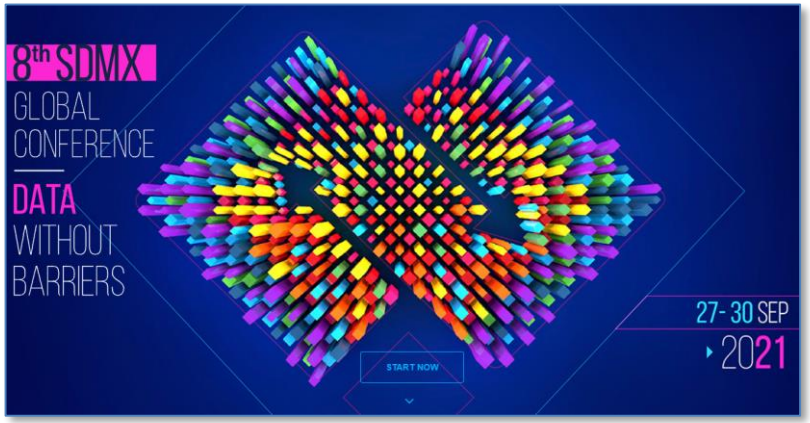


Validation in R using metadata from SDMX registries

Olav ten Bosch, Mark van der Loo
Statistics Netherlands

8th SDMX Global Conference, 27-30 September 2021

Contents

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

International data validation (1)

- ***Invalid*** data may lead to ***costly*** retransmissions or reprocessing (data ping pong)
- To guarantee overall data ***quality*** and ***efficiency***, the European Statistical System (ESS) is moving towards more harmonised validation activities
- International validation rules are agreed in domain specific ***statistical working groups***
- Data producer (NSIs) and data consumers (international organisations) ***validate*** data against the ***same rules***

International data validation (2)

ESSnet Validat Foundation 2015-2016

ESSnet Validat Integration, 2017 (DE,

- Handbook on validation
- Validation principles
- Main types of rules

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*

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

ValidatFOSS & ValidatFOSS2, 2018-2021

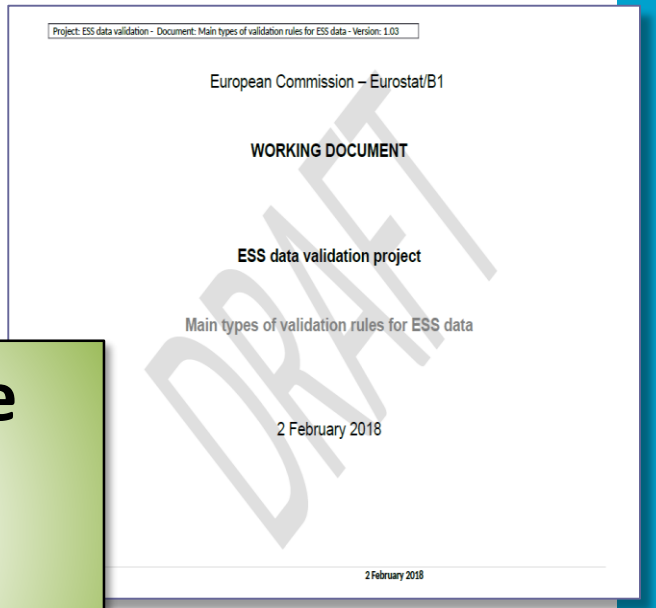
- Implementing validation in Free and open source software (R)

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-keys
- REP: Records Expected are Provided
- RTS: Records are all present for Time Series
- 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

Can we easily use internationally agreed metadata for data validation?


Project: ESS data validation - Document: Main types of validation rules for ESS data - Version: 1.03

European Commission – Eurostat/B1

WORKING DOCUMENT

ESS data validation project

Main types of validation rules for ESS data

2 February 2018

DRAFT

2 February 2018

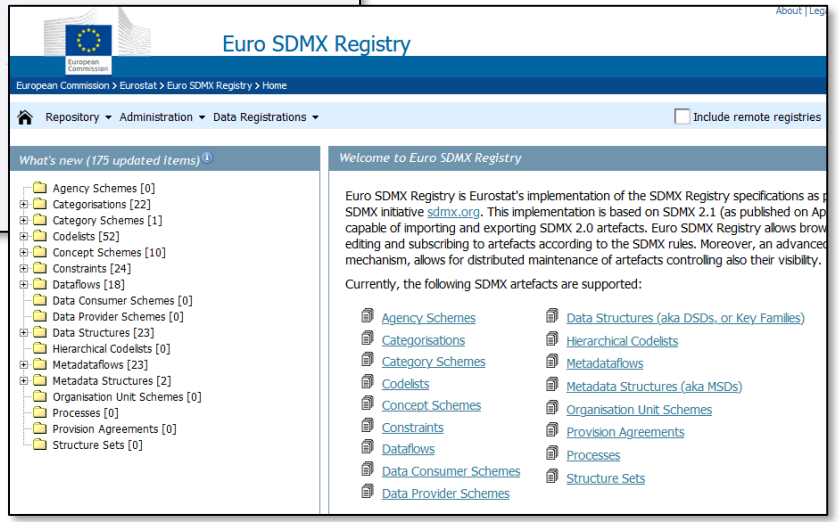
Eurostat, 2018

Implemented in R-package GenericValidationRules:

https://github.com/SNStatComp/GenericValidationRules

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
- Versioning, ownership

SDMX registries (3)

Other & Internal SDMX registries

- Programmatic access via SDMX 2.1 REST API
- Excellent cheat sheet
- Note: changes in SDMX 3.0

https://github.com/sdmx-twg/sdmx-rest

SDMX 2.1 RESTful web services cheat sheet, v1.5.0

Structural metadata queries: `https://ws-entry-point/resource/agencyID/resourceID/version/itemID?queryStringParameters`

		Default
resource	The type of metadata to be returned. Values: datastructure, metadatastructure, categoryscheme, conceptscheme, codelist, hierarchicalcodelist, organisationscheme, agencyscheme, dataproviderscheme, dataconsumerscheme, organisationunitscheme, dataflow, metadataflow, reportingtaxonomy, provisionagreement, structureset, process, categorisation, contentconstraint, attachmentconstraint, actualconstraint, allowedconstraint, structure, transformationscheme, rulesetscheme, userdefinedoperatorscheme, customtypescheme, namepersonalisationscheme, vtlmappingscheme	
agencyID	Agency maintaining the artefact (e.g.: SDMX)	all
resourceID	Artefact ID (e.g.: CL_FREQ)	all
version	Artefact version (e.g.: 1.0)	latest
itemID	ID of the item (for item schemes) or hierarchy (for hierarchical codelists) to be returned.	all
detail	Desired amount of information (allstubs, referencestubs, allcompletestubs, referencecompletestubs, referencepartial, full)	full
references	References to be returned with the artefact (none, parents, parentsandsiblings, children, descendants, all, a resource type)	none

Data queries: `https://ws-entry-point/resource/flowRef/key/providerRef?queryStringParameters`

		Default
resource	data or metadata	
flowRef	Dataflow ref (e.g. ECB, EXR, latest)	
key	Key of the series to be returned (e.g: D.NOK.EUR.SP00.A). Wildcarding (e.g: D..EUR.SP00.A) and OR (e.g: D.NOK+RUB.EUR.SP00.A) supported.	all
providerRef	Data provider (e.g.: ECB)	all
startPeriod	Start period (inclusive). ISO8601 (e.g. 2014-01) or SDMX reporting period (e.g. 2014-Q3).	
endPeriod	End period (inclusive). ISO8601 (e.g. 2014-01-01) or SDMX reporting period (e.g. 2014-W53).	
updatedAfter	Last time the query was performed. Used to retrieve deltas. Must be percent-encoded (e.g.: 2009-05-15T14%3A15%3A00%2B01%3A00)	
firstNObservations	Maximum number of observations starting from the first observation	
lastNObservations	Maximum number of observations counting back from the most recent observation	
dimensionAtObservation	Id fof the dimension attached at the observation level	TIME_PERIOD
detail	Desired amount of information to be returned. Values: full, dataonly, serieskeysonly, nodata	full
includeHistory	Whether to return vintages	false

Legend: Mandatory path parameter / Optional path parameter / Query string parameter (all optional)

For *available data* queries, cf. https://github.com/sdmx-twg/sdmx-rest/wiki/Data-Availability

SDMX registries (4)

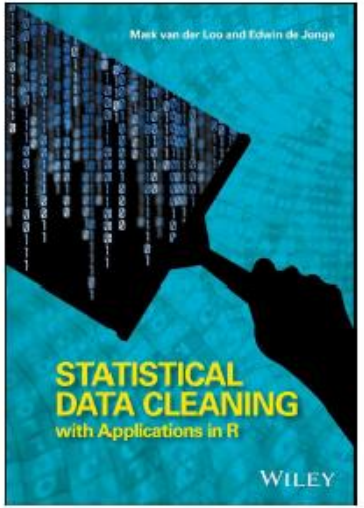
In ValidatFOSS2 we retrieved metadata from SDMX registries using their SDMX API's. Some experiences:

- There is no central list of SDMX endpoints / registries
- Some REST behaviours differ, but we came to one script for all
- For R: rsdmx (cran.r-project.org/package=rsdmx) makes life easy
- Some thoughts about registry content:
  - does a statistician know where to look for which content?
  - is there content overlap among registries?
  - are registry contents consistent?

https://github.com/SNStatComp/validatesdmx (demo notebooks Python, R)

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

Data cleaning with R (3)

R-validate supports the majority of checks needed in today's data validation, such as:

- Variable types, missingness, field length, format, range checks, codelists
- Availability, uniqueness, gaps in time series
- Multivariate checks, balance equalities
- Statistical & groupwise checks

Documented in online ***cookbook***:

data-cleaning.github.io/validate

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

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

- New chapter in cookbook on SDMX:
  data-cleaning.github.io/validate/sect-sdmxrules.html

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

Connecting R-validate to SDMX (2)

- Deriving all rules from a DSD:

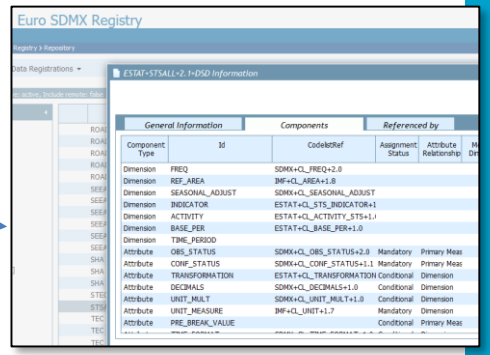
```
# import data
my_data <- read.csv("mydata.csv")

# 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 and checks multiple codelist rules derived from the ESTAT registry
- Easy to integrate in statistical processes



Wrap-up

- The ESS works on improving ***international data validation*** i.a. handbook, principles, main types of rules
- R has a rich ***data-cleaning ecosystem***. ***R-validate*** covers many of today's validation needs in data validation
- SDMX metadata contained in ***registries*** is necessary input for international validation rules
- R-validate now supports ***checks*** against ***any*** SDMX registry in high level functions.
- Some ***experiences*** working with SDMX registries and their APIs have been shared

Questions, ideas, suggestions

Olav ten Bosch o.tenbosch@cbs.nl @olavtenbosch

Mark van der Loo mpj.vanderloo@cbs.nl @markvdloo

and keep an eye on:

awesomeofficialstatistics.org