cessda

CESSDA

Chapter 3. Process

Johana Chylíková <johana.chylikova@soc.cas.cz> |
Jindřich Krejčí <jindrich.krejci@soc.cas.cz>

Data Management Expert Guide
Train the Trainers event, 12-13 April 2018

cessda.eu @CESSDA_Data

Cite as: Chylikova, J., J. Krejci (2020). The CESSDA Data Management Expert Guide. Chapter 3. Process [presentation]. Bergen: CESSDA ERIC.

What is the chapter Process about?

- Data entry and integrity
- Quantitative coding
- Qualitative coding

- Weights of survey data
- File formats and data conversion
- Data authenticity

Subchapter: Data entry

- Data entry and data integrity

- Data integrity =„stored data should correspond to gathered data“

  = assurance of preservationof original information contained in the data.

  - Stored data <=> original information from the research

- „Data authenticity“

- It is important to mention = in the data processing, data integrity is at stake.

Data entry: Quantitative data

Survey data: Data entry procedures have changed over the recent years:

- BEFORE: manual data entry (phases: data collection, data entry, data editing/checking)
- NOW: computer technologies, automated data entry (CAPI)
- Automatic data entry: prevents some types of errors, but produces others. Errors in scripts in CAPI may cause systematic shifts in data; example: gender of a respondent is programmed wrongly – all analyses regarding gender are wrong

Minimizing errors in manual data entry I.

- Reduce burden on those who enter data manually
- Check the completeness of records (cases, variables)
- Conduct data entry twice: approx. 20 percent of questionnaires entered twice by two different persons
  - compare values, if difference in single value, get back to questionnaire and enter the right value.
- Perform in depth checks for selected records: randomly selected records, e.g. 5–10% of all records subject of more detailed, in-depth check.

- **Questionnaire scanning**: (=data not entered manually, but scanned) do scanning twice, compare values

Minimizing errors in manual data entry II.

Perform logical and consistency checks:

- Check minimal and maximal values of variables,
- Check the relations between associated variables (r(education;age));
- Compare your data with historical data (e.g. check the number of household members with the previous wave of a panel survey).

Automated checks - CAPI, data entry software:

- set the range of valid values
- apply filters to manage the data entry

- CAPI software= used by the data collectors; expensive; individuals can't afford it.
- Write your own program or syntax to check for discrepancies.
- **An example of a SPSS syntax to check your data**
- Logical check of income: The household income cannot be SMALLER than individual income

Minimizing errors in manual data entry II.

Variable names:

ide.10 - household income = interval variable, income in Euros, with special values 8 - refused to answer; 9 - don´t know

ide.10a - individual income = interval variable, income in Euros, with special values 8 - refused to answer; 7 - doesn´t have income

CD – respondents identification

Syntax (SPSS):

```
USE ALL.
COMPUTE filter_$=(ide.10a ne 0) and (ide.10 ne 0) and (ide.10a ne 7) and (ide.10a ne 8) and (ide.10 ne
8) and (ide.10 ne 9) and  (ide.10 < ide.10a).
VARIABLE LABELS filter_$ '(ide.10a ne 0) and (ide.10 ne 0) and (ide.10a ne 7) and (ide.10a ne 8) and
(ide.10 ne 8) and (ide.10 ne 9) and (ide.10 < ide.10a) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.  EXECUTE.
FREQUENCIES VARIABLES=CD
/ORDER=ANALYSIS.  FILTER OFF.
USE ALL.
```

What to do with error values? (tricky)

e.g. Suspicious values

- Comparison with respondents’ original answers.
- Inconsistencies can be generated by the respondents themselves (true value not known)

Delete?
Correct?
Keep?

- Single right recipe does not exist Think twice, document changes.

Data entry: Qualitative data

Transcription

- The most common formats of qualitative data = written texts, interview data and focus group discussion data;
- Record -> Transcribe;

Transcription = conversion of audio or video recordings into text.

- If you intend to share your data with other researchers, you should prepare a full transcription of your recordings;
- If the transcription is not in English, prepare summary in English.

Basic rules for transcriptions

- Record in high quality, so you can hear/see everything (equipment, quite place).
- Transcription method:
  - several methods exist
- Speech recognition software:
  - good servant, but must be well set up, expensive

- Alone or in a team: Determine a set of transcription rules/ transcription guidelines:
  - All members of a transcript team should first agree on these rules; you yourself should have rules
- Anonymise – protection
- Choose a file format for the long time preservation
- Anyone has a comment/"transcription" story from their own experience (software, complications, working in a team, anonymisation)?

Subchapter: Quantitative coding

= coding answers to questionnaire items, assigning values to answers

Closed ended questions/questionnaire items

- Codes incoporated in survey questionnaires (individual response categories have values/codes)
- In (CAPI, CATI, etc.) an answer and its code are saved immediately into a computer in the course of data collection

Open ended questions/questionnaire items

- e.g. textual answers in survey questionnaires
- Require an independent coding process with a clearly defined rules

Coding recommendations:

- Make code categories exclusive and coherent throughout the database
- Preserve original information
- Document the coding schemes
- Check verbatim text data (open questions) for data disclosure risk -> privacy
- Distinguish between major and lower level categories - Example ISCO –

  Standardized coding scheme

Example of standardized coding scheme

Example of standardized coding scheme

- Standard Classification of Occupations (ISCO): widespread standard coding scheme; hierarchical category scheme - several dimensions (data collection = one or more open-ended questions)

The current ISCO-2008 uses four-digit codes. In the table below you see some examples.

- **2 Professionals**
- **21 Science and engineering professionals**
- **211 Physical and earth science professionals**
- 2111 Physicists and astronomers
- 2112 Meteorologists
- 2113 Chemists
- 2114 Geologists and geophysicists
- **212 Mathematicians, actuaries and statisticians**
- 2120 Mathematicians, actuaries and statisticians
- **213 Life science professionals**
- 2131 Biologists, botanists, zoologists and related professionals
- 2132 Farming, forestry and fisheries advisers
- 2133 Environmental protection professionals
- **214 Engineering professionals (excluding electrotechnology)**
- 2141 Industrial and production engineers
- 2142 Civil engineers

Exercise:

- Download Part 2: Classification Structure

  http://www.ilo.org/public/english/bureau/stat/isco/isco08/

- Answer to an open question in a questionnaire:
- „I work in a hospital as a nurse and I take care of newborns“

- **Find a code for this profession**
- **Find a code for your own profession**

Coding missing values

Item non-response or when values are missing from other reasons:

- **No answer (NA):** The respondent did not answer a question when he/she should have;
- **Refusal:** The respondent explicitly refused to answer;
- **Don’t Know (DK):** The respondent did not answer a question because he/she had no opinion or did not know the information required for answering.
- **Processing Error:** The respondent provided an answer; interviewer error, illegible record, incorrect coding etc
- **Not Applicable/Inapplicable (NAP/INAP):** A question did not apply to the respondent. For example respondents without a partner did not answer partner-related questions
  - For more missing data situations see ETG

Coding missing values

Numerical codes of missing values:

- Typically negative values or values like 7, 8, 9 or 97, 98, 99 or 997, 998, 999.

- Prevent overlapping codes for valid and missing values.
- Example: „0" may be a real value (income, years of duration of something), do not use „0" as a code for missing value.

Subchapter: Qualitative coding

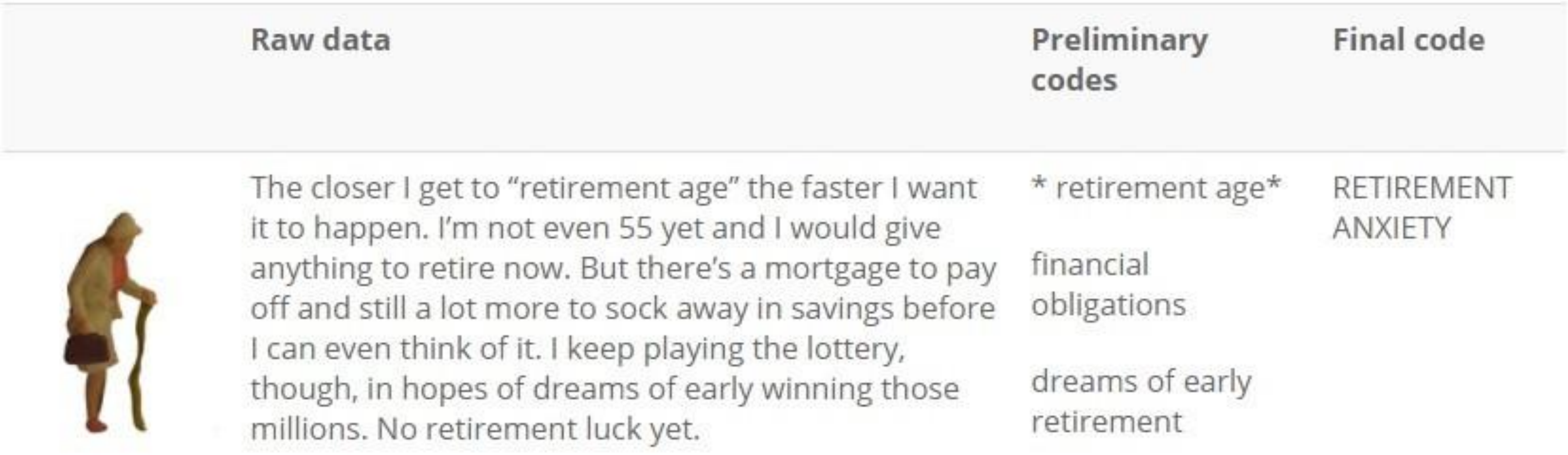
!!! Completely different from quantitative coding!!! What´s in the text? What is it about?

- A process of identifying a passage in the text or other data items (photograph, image).
- Searching and identifying concepts and finding relations between them.
- Codes enable you to organise and analyse data in a structured way, e.g. by examining relationships between codes.
- Qualitative coding dependenst on your expertise!!
  - It is not the aim of this presentation to teach you qualitative coding, only hint some basics.

Concept driven coding vs data driven coding

- Concept driven approach: deductive, you have a developed system of codes representing concepts and look for concepts in the text
- Data driven approach: inductive, you search concepts in the text without a preceding conceptualisation and let the text speak for itself

Example of coding:



	Raw data	Preliminary codes	Final code
	The closer I get to "retirement age" the faster I want it to happen. I'm not even 55 yet and I would give anything to retire now. But there's a mortgage to pay off and still a lot more to sock away in savings before I can even think of it. I keep playing the lottery, though, in hopes of dreams of early winning those millions. No retirement luck yet.	* retirement age* financial obligations dreams of early retirement	RETIREMENT ANXIETY

Documenting codes

- Document the meaning of the codes in a separate file.
- Short descriptions of the meaning of each code.
- Necessary to you and others who will have access to your data/analysis.

- Provide other information to the code:
  - the label or name of the code
  - who coded it (name of the researcher/coder)
  - the date when the coding was done/changed
  - definition of the code; a description of the concept it refers to
  - information about the relationship of the code to other codes you are working with during the analysis.

Prevent coder variance

- The coder has influence on the coding process.
- Establish coding guidelines
- Each coder in the team must be trained according to guidelines

- Several techniques to control coder reliability:
- Checking the transcription
- An independent researcher goes through coded texts and considers a degree to which coders differed from each other.

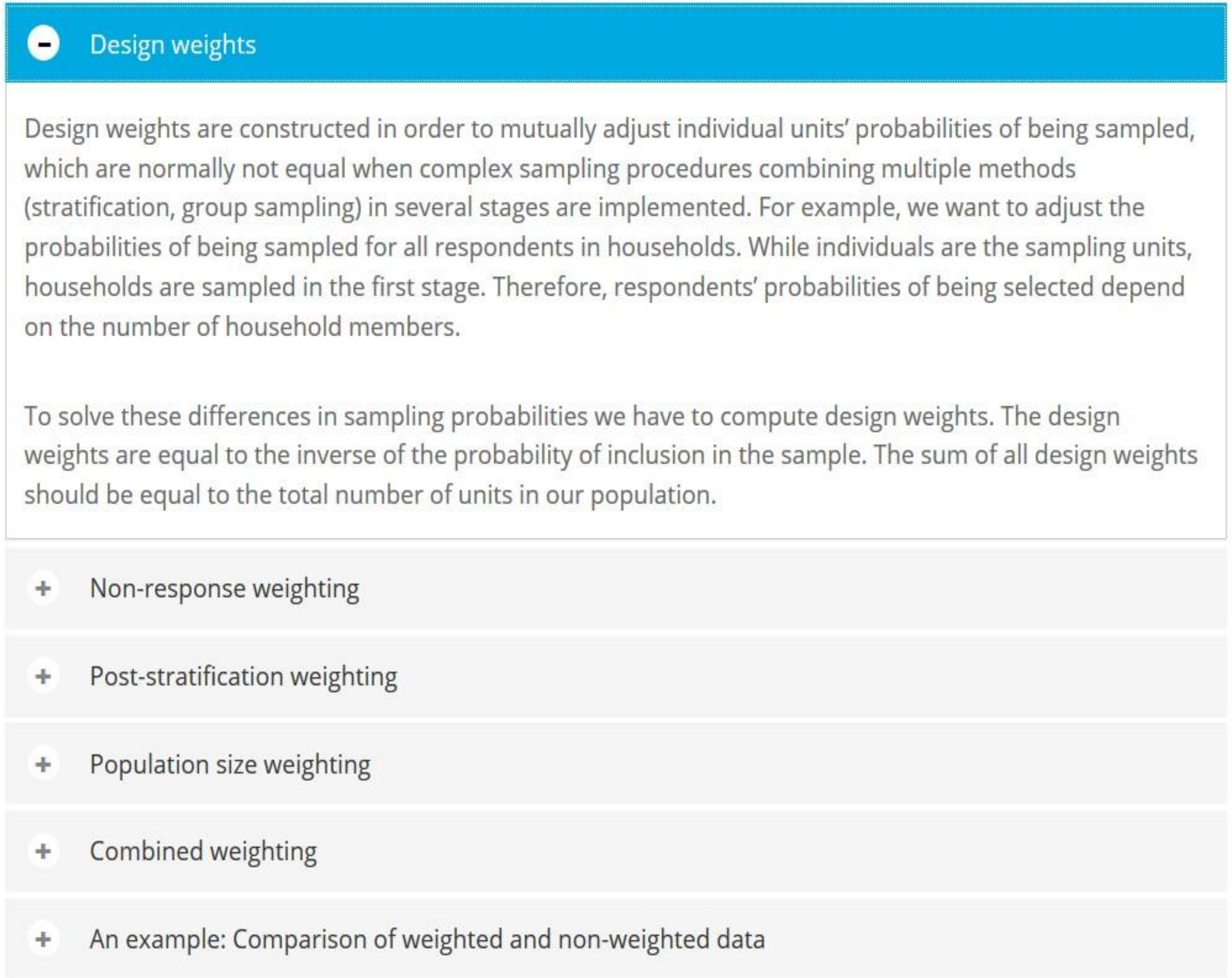
Definitional drift in coding

- Coding large dataset, at the beginnings feelings and perceptions of the coder different from those at the end
- When you finish coding the document, check for the definitional drift
- Have good notes with descriptions of individual codes.

Working in a team:

- If there are more people in the team, individual members can check each other´s coding.

Dive in deep? Weights of survey data



Design weights

Design weights are constructed in order to mutually adjust individual units' probabilities of being sampled, which are normally not equal when complex sampling procedures combining multiple methods (stratification, group sampling) in several stages are implemented. For example, we want to adjust the probabilities of being sampled for all respondents in households. While individuals are the sampling units, households are sampled in the first stage. Therefore, respondents' probabilities of being selected depend on the number of household members.

To solve these differences in sampling probabilities we have to compute design weights. The design weights are equal to the inverse of the probability of inclusion in the sample. The sum of all design weights should be equal to the total number of units in our population.

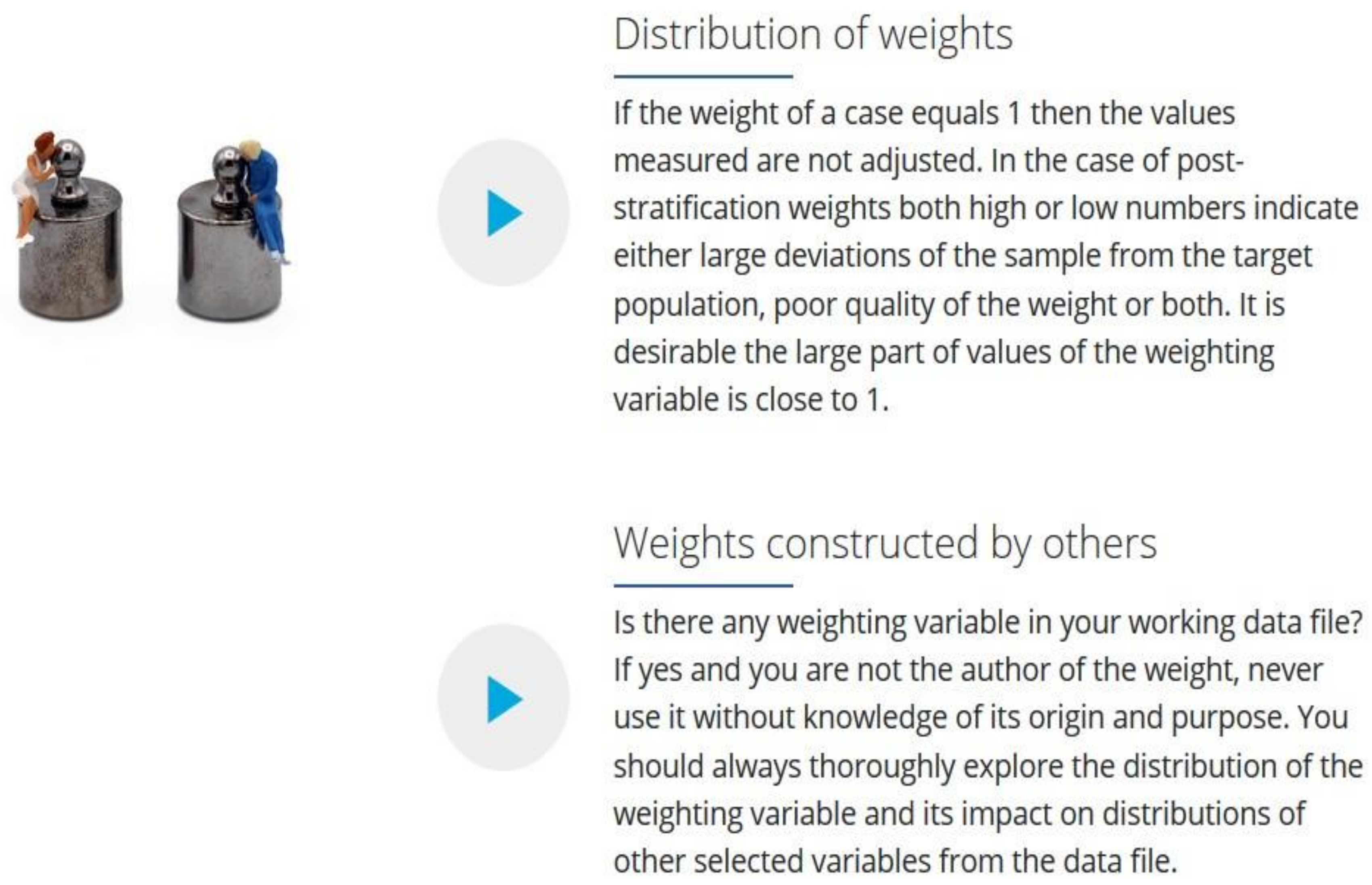
+ Non-response weighting
+ Post-stratification weighting
+ Population size weighting
+ Combined weighting
+ An example: Comparison of weighted and non-weighted data

Adjustment of the sample.

Each individual case in the file is assigned an individual weight which is used to multiply the case in order to attain the desired characteristics of the sample.

- There are different types of weights for different purposes
- Necessary in some sitations
- Issue of quality

Issue of quality



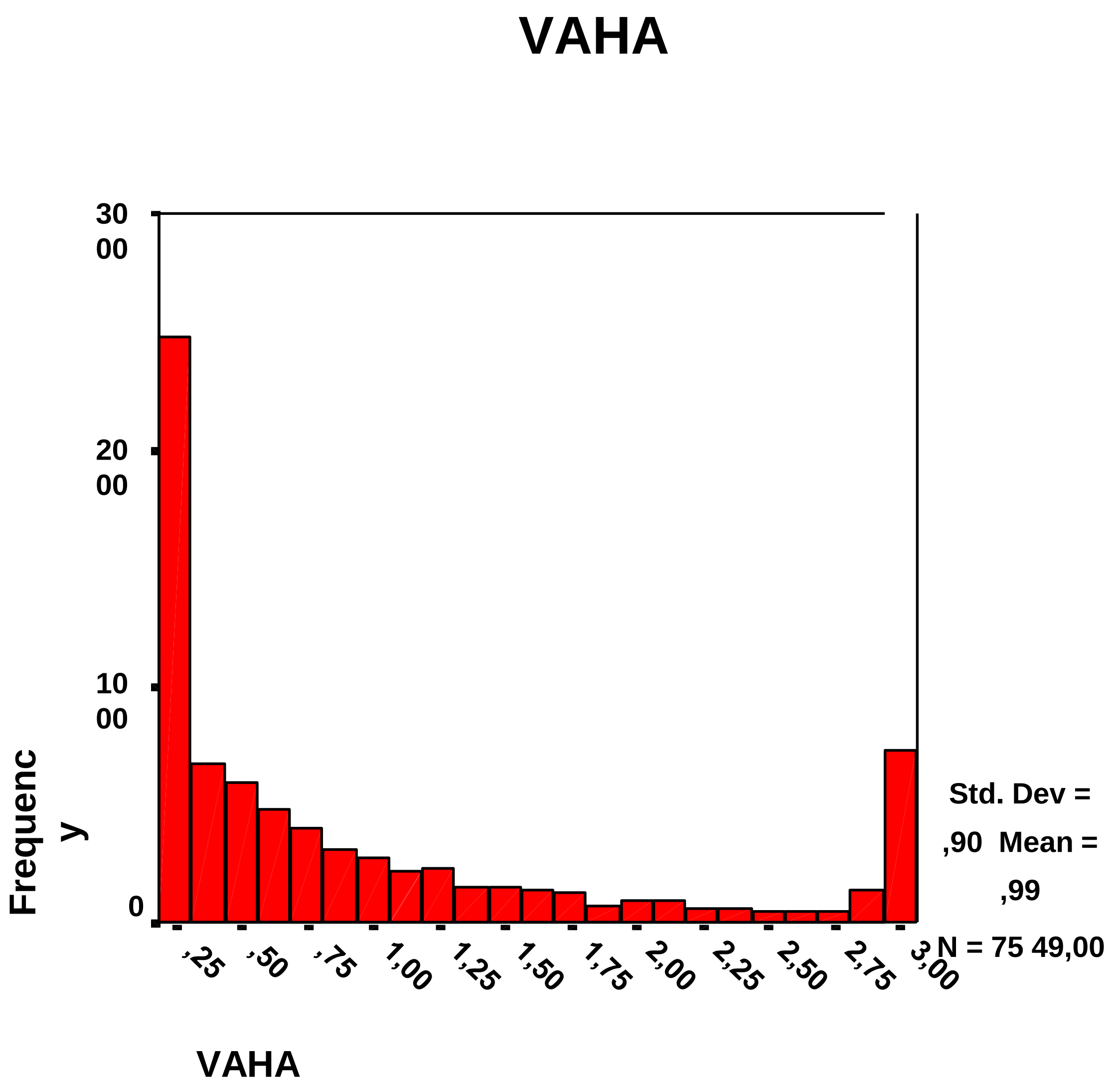
Distribution of weights

If the weight of a case equals 1 then the values measured are not adjusted. In the case of post-stratification weights both high or low numbers indicate either large deviations of the sample from the target population, poor quality of the weight or both. It is desirable the large part of values of the weighting variable is close to 1.

Weights constructed by others

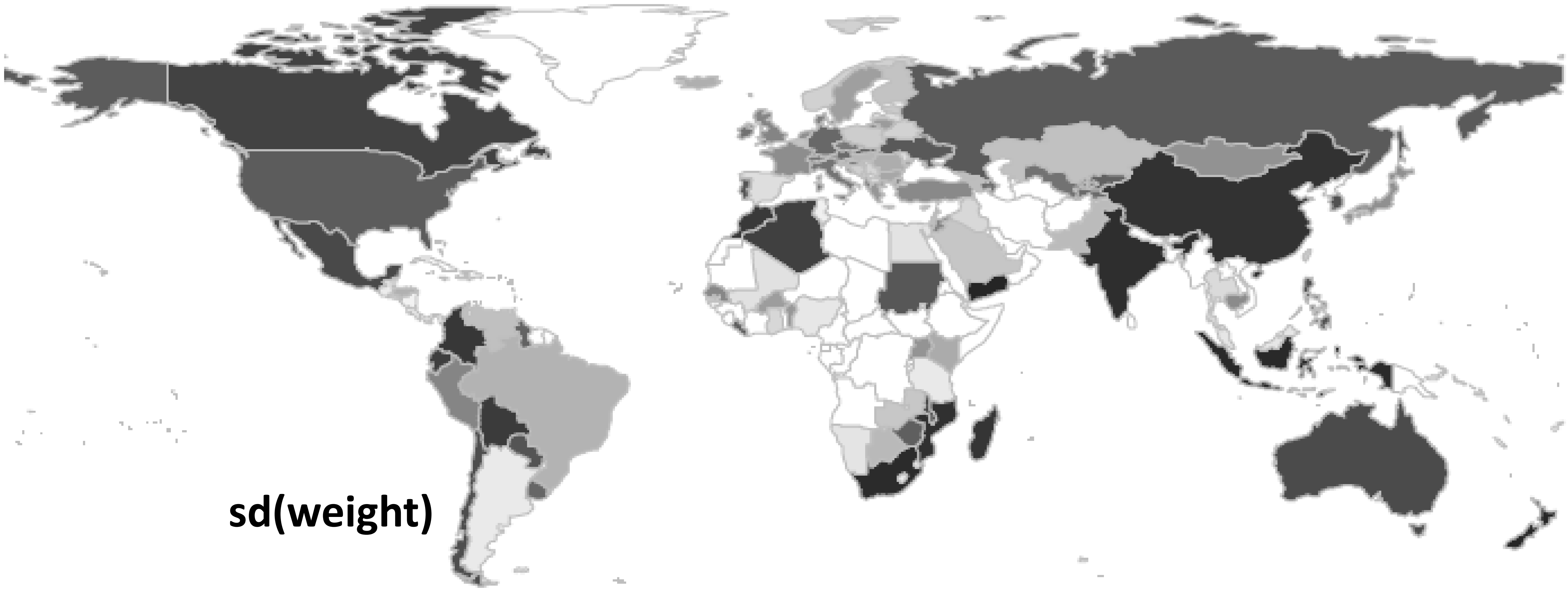
Is there any weighting variable in your working data file? If yes and you are not the author of the weight, never use it without knowledge of its origin and purpose. You should always thoroughly explore the distribution of the weighting variable and its impact on distributions of other selected variables from the data file.

Distribution of weighting variable - how the post-stratification weight should not look



Example - Marcin W. Zieliński: Weighting issue or weighting problem - international survey projects in a comparative perspective; CSDA Workshop on Challenges in the Organisation of International Comparative Social Surveys; http://archiv.soc.cas.cz/sites/default/files/zielinski_weighting_issue.pdf

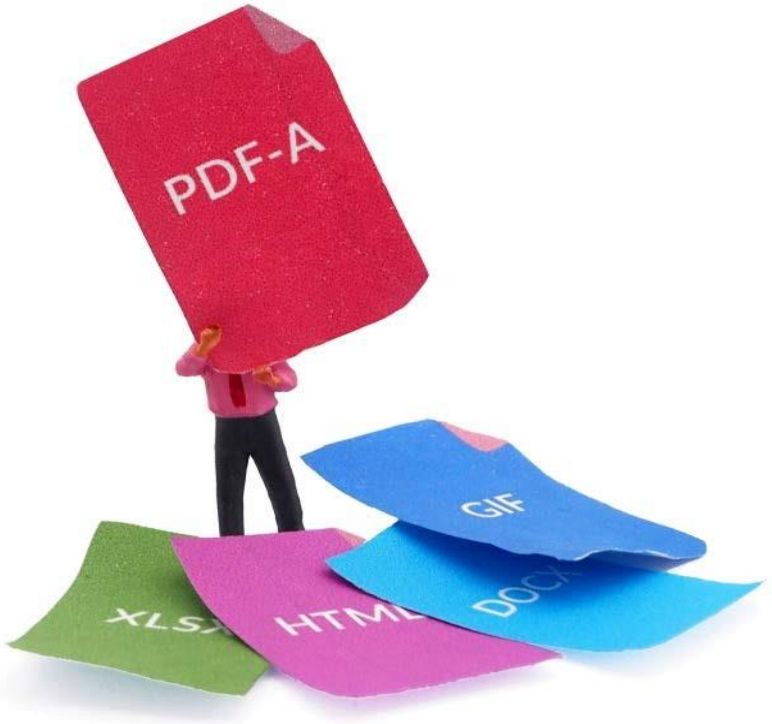
Technically "good weight": mean(wght) = 1; sd(wght) as small as possible; MIN(wght) > 0 and MIN(wght) < 1; MAX(wght) > 1 but small



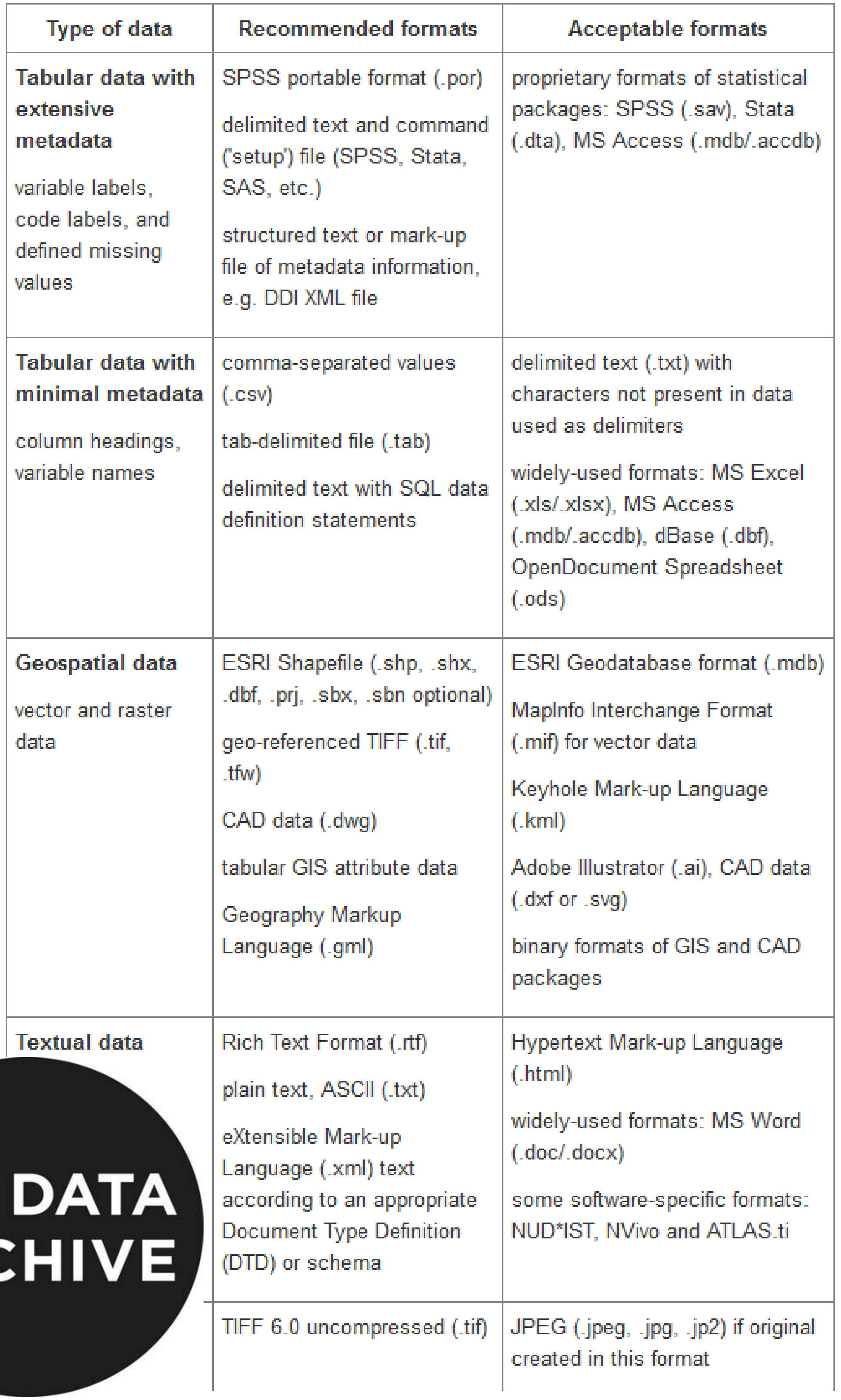
sd(weight)

File formats and data conversion

- Short-term data processing: file formats for operability
  - Proprietary vs. open formats
  - Export / portable formats
- Long-term data preservation
- Link to the table of Recommended file formats



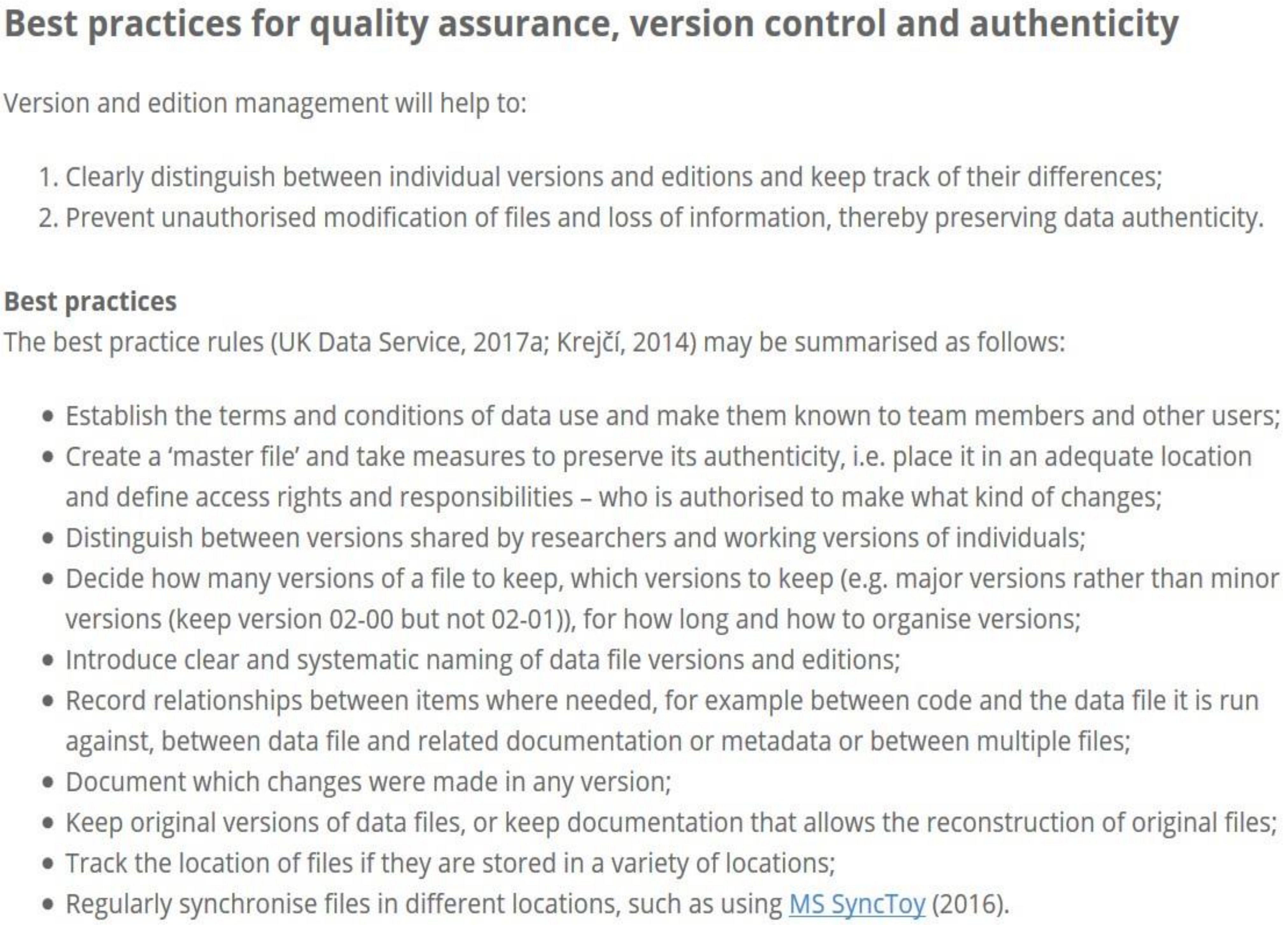
PDF/A, CSV, TIFF, ASCII, Open Document Format (ODF), XML, Office Open XML, JPEG 2000, PNG, SVG, HTML, XHTML, RSS, CSS, etc.



Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata variable labels, code labels, and defined missing values	SPSS portable format (.por) delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) structured text or mark-up file of metadata information, e.g. DDI XML file	proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/.accdb)
Tabular data with minimal metadata column headings, variable names	comma-separated values (.csv) tab-delimited file (.tab) delimited text with SQL data definition statements	delimited text (.txt) with characters not present in data used as delimiters widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Geospatial data vector and raster data	ESRI Shapefile (.shp, .shx, .dbf, .prj, .sbx, .sbn optional) geo-referenced TIFF (.tif, .tfw) CAD data (.dwg) tabular GIS attribute data Geography Markup Language (.gml)	ESRI Geodatabase format (.mdb) MapInfo Interchange Format (.mif) for vector data Keyhole Mark-up Language (.kml) Adobe Illustrator (.ai), CAD data (.dxf or .svg) binary formats of GIS and CAD packages
Textual data	Rich Text Format (.rtf) plain text, ASCII (.txt) eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	Hypertext Mark-up Language (.html) widely-used formats: MS Word (.doc/.docx) some software-specific formats: NUD*IST, NVivo and ATLAS.ti
	TIFF 6.0 uncompressed (.tif)	JPEG (.jpeg, .jpg, .jp2) if original created in this format

UK • DATA ARCHIVE

Data authenticity



Best practices for quality assurance, version control and authenticity

Version and edition management will help to:

1. Clearly distinguish between individual versions and editions and keep track of their differences;
2. Prevent unauthorised modification of files and loss of information, thereby preserving data authenticity.

Best practices

The best practice rules (UK Data Service, 2017a; Krejčí, 2014) may be summarised as follows:

- Establish the terms and conditions of data use and make them known to team members and other users;
- Create a 'master file' and take measures to preserve its authenticity, i.e. place it in an adequate location and define access rights and responsibilities – who is authorised to make what kind of changes;
- Distinguish between versions shared by researchers and working versions of individuals;
- Decide how many versions of a file to keep, which versions to keep (e.g. major versions rather than minor versions (keep version 02-00 but not 02-01)), for how long and how to organise versions;
- Introduce clear and systematic naming of data file versions and editions;
- Record relationships between items where needed, for example between code and the data file it is run against, between data file and related documentation or metadata or between multiple files;
- Document which changes were made in any version;
- Keep original versions of data files, or keep documentation that allows the reconstruction of original files;
- Track the location of files if they are stored in a variety of locations;
- Regularly synchronise files in different locations, such as using MS SyncToy (2016).

Preserving the authenticity

of the original research information contained in the data throughout the whole data lifecycle.

Data cleaning; correcting errors; constructing new variables; adding new information from external sources; changing formats; changing data file structure...

- Different versions of the data file;
- New editions,

Version control

Title				
Description				
Created By				
Date Created				
Maintained By				
Version Number	Modified By	Modifications Made	Date Modified	Status

- Unique identification; naming convention; version numbering
- Record the date
- Don't use ambiguous descriptions for the version (NO: MyThesisFinal.doc, MyThesisLastOne.doc)
- Designing and using a version control table
- Version control facilities; versioning software
- File-sharing services with incorporated version control (but be aware of specific rules at, e.g., Google cloud, Dropbox...)

Versioning new data types

Enable reproducibility and support trustworthiness by allowing all transformations in the data to be traced is more difficult with “new data” as these data are (compared to “traditional data”) more frequently or even continuously updated.

- Collections of Tweets and individual posts may be modified or deleted. As the contents of these data are continuously changing and if archived data are expected to reflect such changes;
- This result is an increasing number of versions.
- Consequently, it is necessary to develop a systematic plan to create and name new versions of constantly changing datasets, or find new solutions for streaming data.

Versioning new data types: emerging issues for data archiving and citing

The most common version control software in software development is Git. Some of the established repositories, such as Zenodo and FigShare or the Open Science Framework, now offer integration with GitHub, so that every version of data sets in those repositories can be recorded through it. A new project called Dolt is developing version control specifically for data which is particularly interesting for dynamic data sets, such as social media data.

To identify the **exact version** of a dataset as it was used in a specific project or publication, the Research Data Alliance (RDA) suggests that every dataset is versioned, timestamped, and assigned a persistent identifier (PID).

In the case of Big Data, however, the RDA warns against excessive versioning: *"In large data scenarios, storing all revisions of each record might not be a valid approach. Therefore in our framework, we define a record to be relevant in terms of reproducibility, if and only if it has been accessed and used in a data set. Thus, high-frequency updates that were not ever read might go - from a data citation perspective - unversioned."*

Wrap up: Data quality

- **Small things matter**: *"The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to [preventing, measuring and] dealing with the many important problems that can arise."* | American Association for Public Opinion Research (2015) (AAPOR)
- *"**In qualitative research**, discussions about quality in research are not so much based on the idea of standardization and control, as this seems incompatible with many qualitative methods. Quality is rather seen as an issue of how to manage it. Sometimes it is linked to rigour in applying a certain method, but more often to soundness of the research as a whole"* | Flick (2007).
- **A complex approach to data quality**: *"The mechanical quality control of survey operations such as coding and keying does not easily lend itself to continuous improvement. Rather, it must be complemented with feedback and learning where the survey workers themselves are part of an improvement process"* | Biemer & Lyberg (2003).

Data Management Expert Guide. Train the Trainers event, 12-13 April 2018

Johana Chylíková <johana.chylikova@soc.cas.cz> |
Jindřich Krejčí <jindrich.krejci@soc.cas.cz>

cessda.eu

@CESSDA_Data