TALL SOM FORTELLER

National guidelines on data editing; the foundation for building a solution for the future

ASLAUG HURLEN FOSS OG ANE SEIERSTAD

Statistisk sentralbyrå
Statistics Norway

Technological modernisation – cloud solution

- Storage in Google cloud buckets
- Access management
- Transfer services
- Automatic processing of data from source data to input data.
- Programming languages: Python and R.

Picture: Microsoft

Status of data editing in Statistics Norway

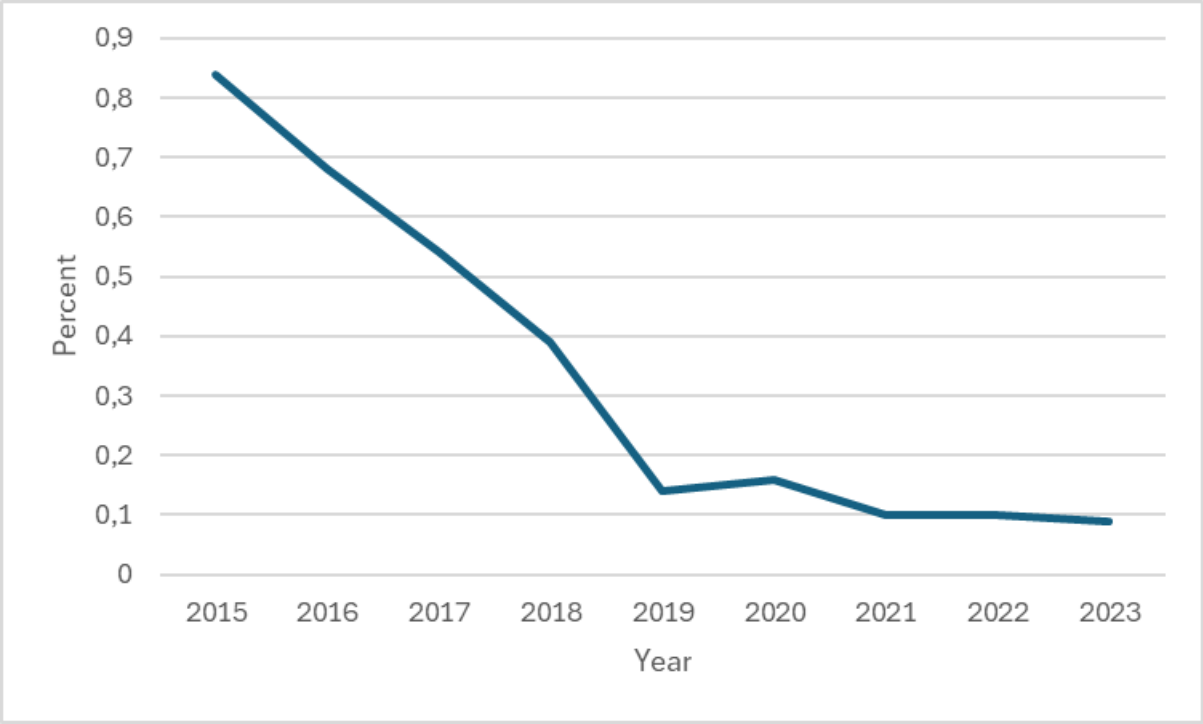
System of data editing:

- Dynarev - built 2005 – 150 surveys
- Driller – built 2011 – 25 surveys

Main functionality:

- Management of controls (edit rules)
- Execution of controls
- Rule-based and manual corrections
- Reports – analyzing

Proportion of manually edited values in percent, Dynarev



Statistisk sentralbyrå
Statistics Norway

Lack of common understanding

- Data editing process
- Terminology
- Human interaction

Picture: Microsoft

Striving for a unified understanding

- We made a draft of the principles and guidelines
- Subject experts: feedback
- Formal consultation to all statistics departments
- Approved by the director's meeting and by the director general

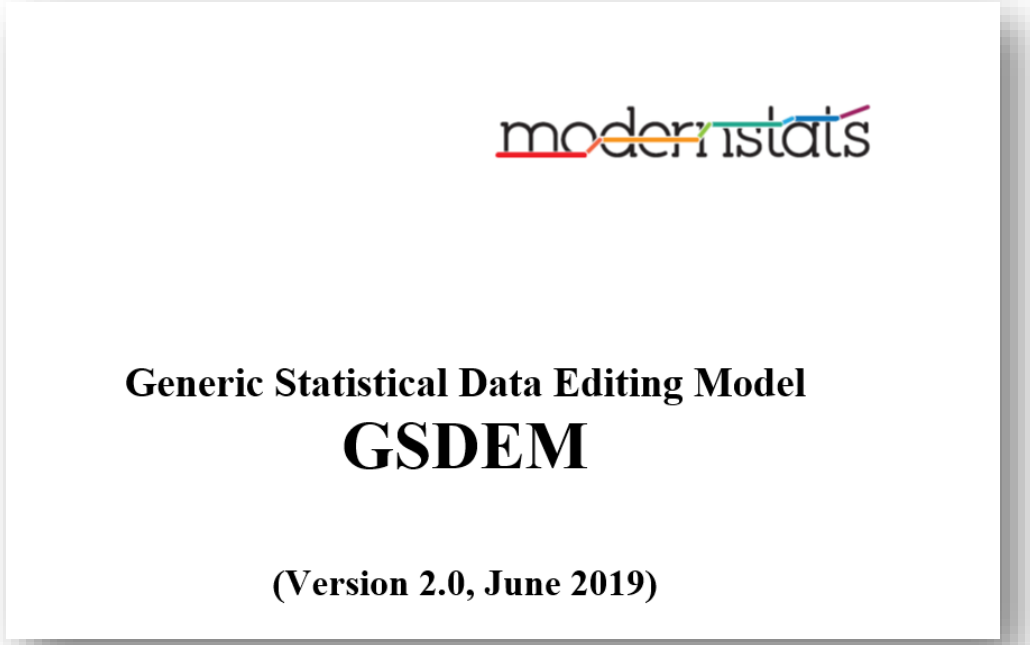
Geir Axelsen Director general of Statistics Norway

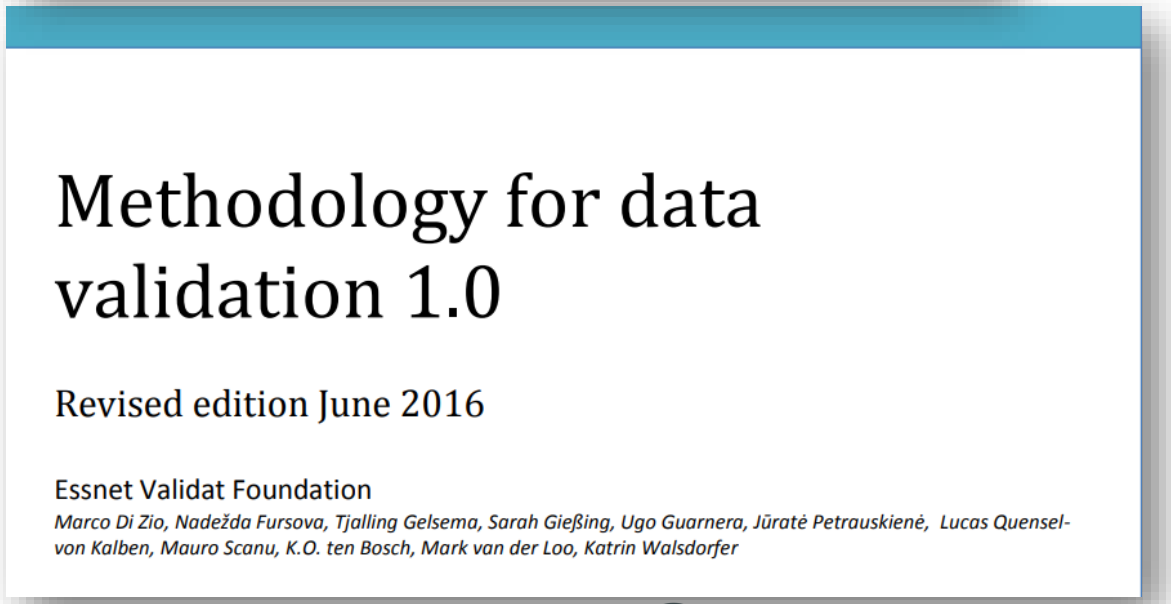
Photo: Karina Rønning

Statistisk sentralbyrå
Statistics Norway

Basis for principles and guidelines

- Apply international methodology
- Adapt to Norwegian condition
- Specification of lower level





Statistisk sentralbyrå
Statistics Norway

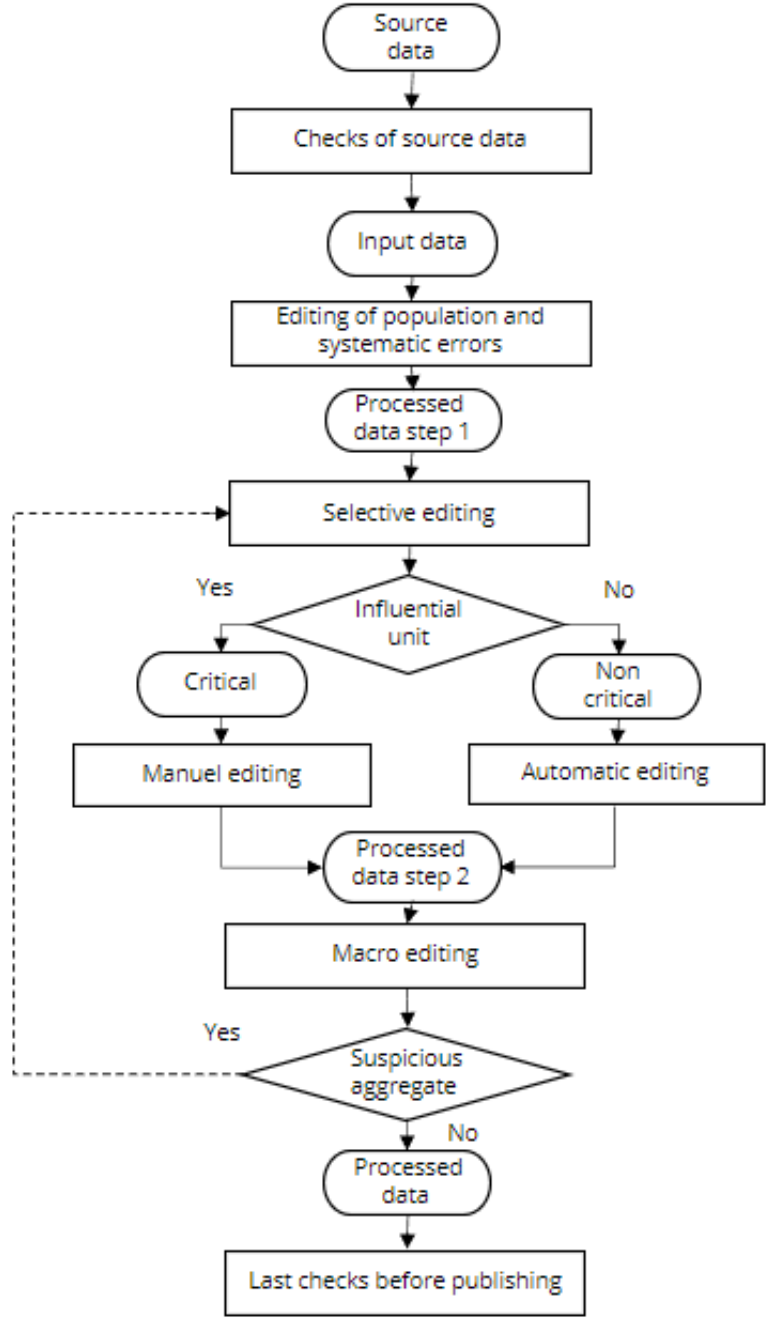
Process model the source for writing guidelines

Added two processes :

- Checks of source data
- Last checks before publishing

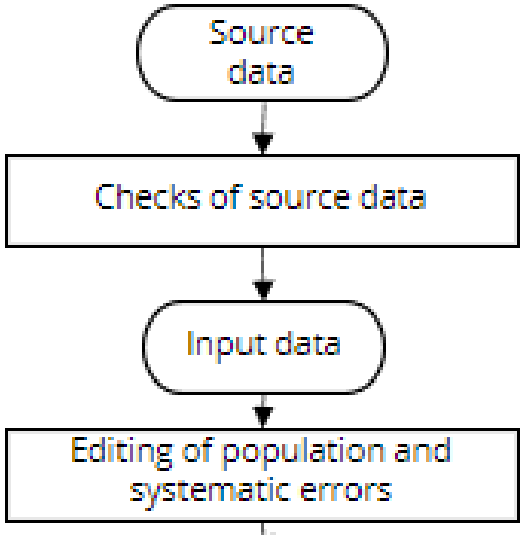
Sometimes separated:

- Social statistics
- Business statistics



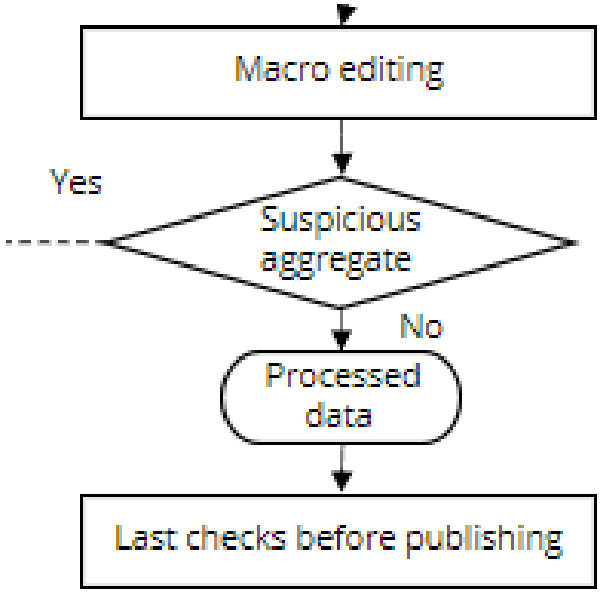
Example: Editing of Personal Identification number

- Control of personal identity numbers is done by cross-referencing against the latest SNR catalogue, containing all personal identity numbers that have ever been issued.
- Expired identity numbers are replaced with the latest identity number to maintain a stable identification.
- Document invalid identity numbers using standardized error codes.



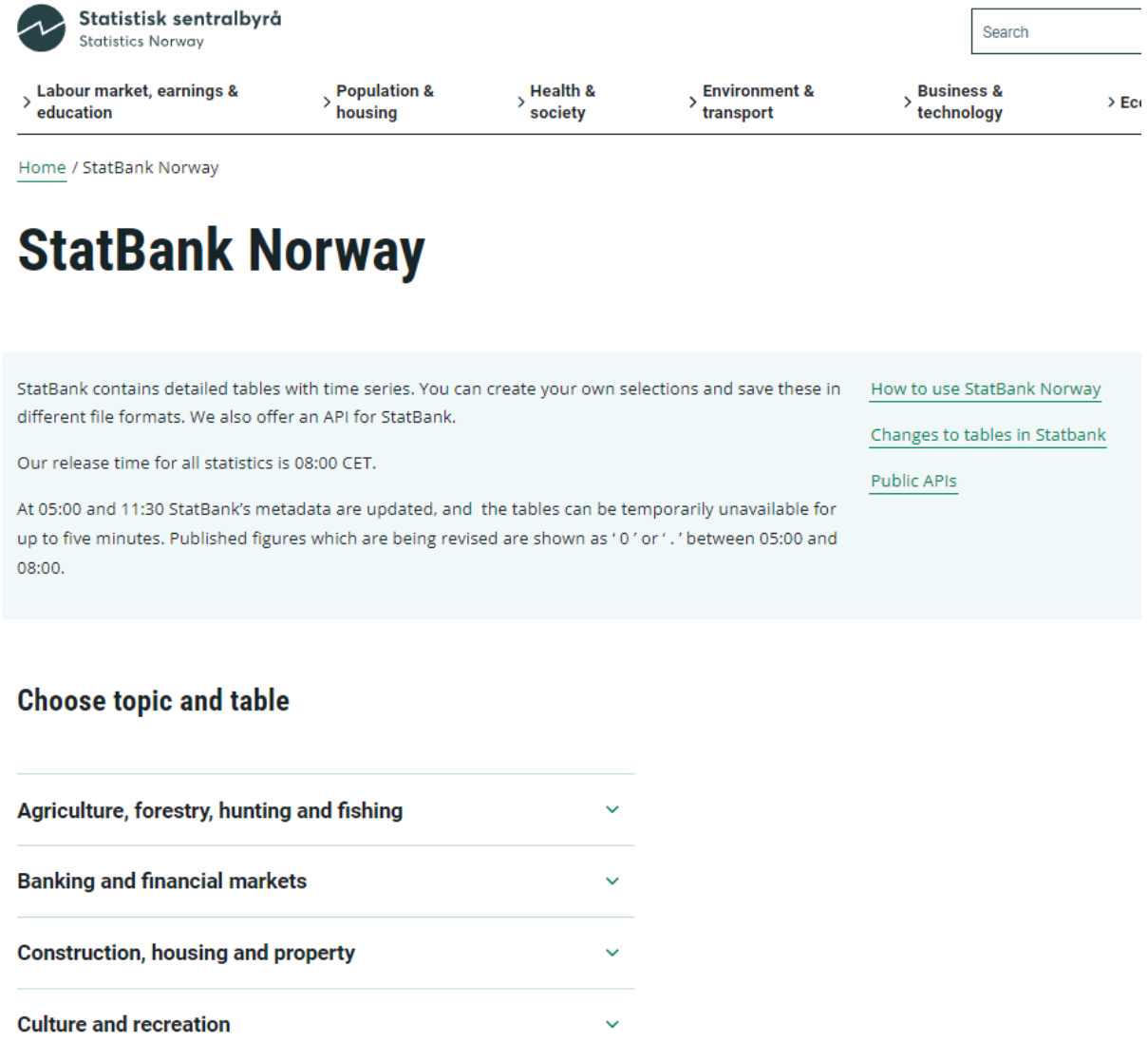
Example: Macro editing

- Assess whether the aggregates are plausible given **historical trends**
- Evaluate the plausibility of the aggregate concerning **derived quantities**, such as ratios.
- Investigate the dataset for **potential errors or explanations** when changes occur beyond what is considered normal.
- Compare aggregates with those of **comparable countries** for assessment.
- Aggregates may undergo **editing** when they are part of a system where certain relationships must be maintained.



Example: Control of Statbank tables

- Statistical producers should perform controls on the tables before submission to Statbank
- Those receiving the tables should control compliance with Statbank rules.



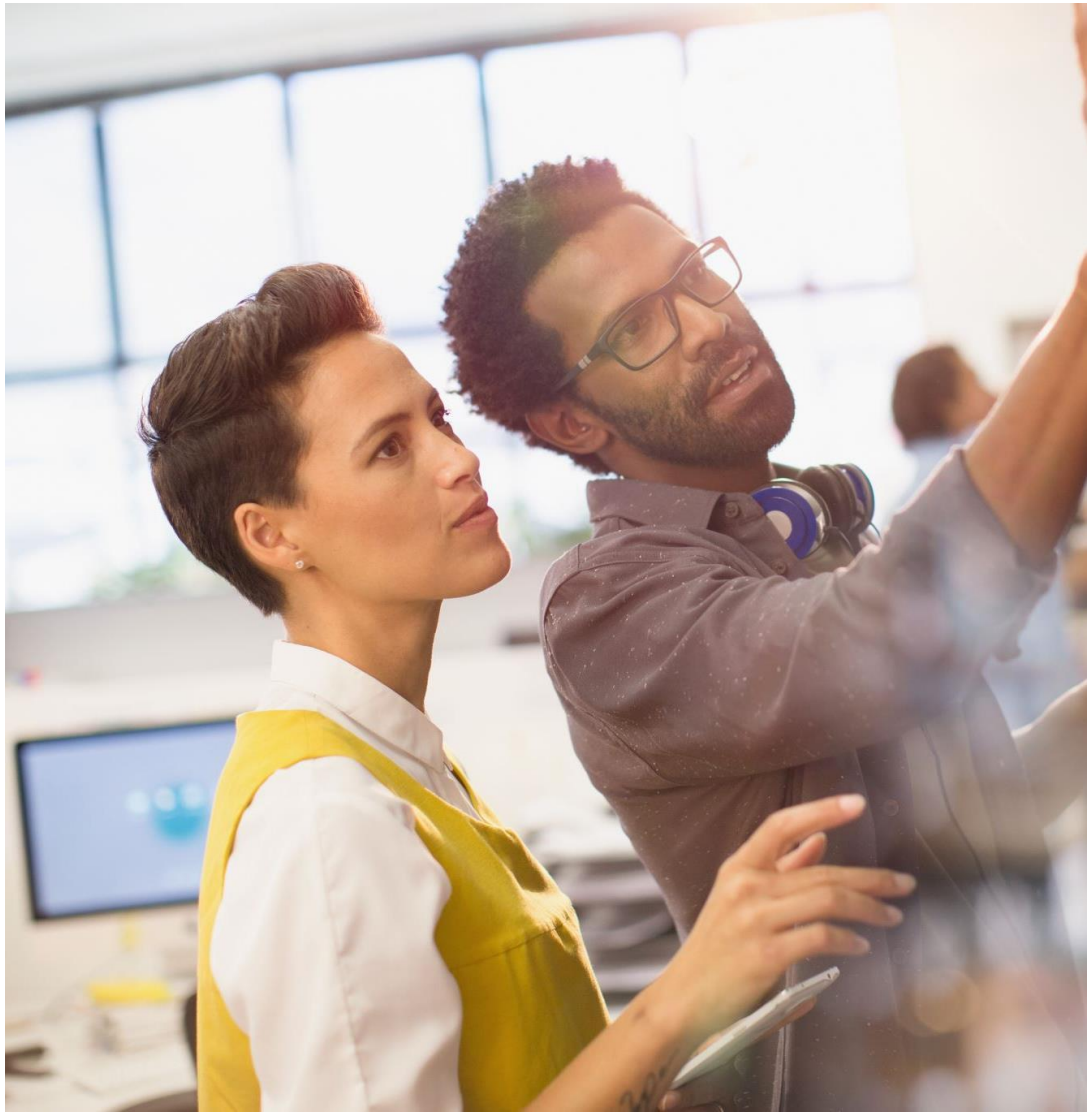
Statistisk sentralbyrå
Statistics Norway

Principles of data editing

- Principle 1: Good knowledge about the subject area of statistics and the background of source data is the basis for a good editing process.
- Principle 2: The objective of data editing should be clearly defined.
- Principle 3: High-quality data input is the best
- Principle 4: Always control the data.
- Principle 5: The earlier, the better.
- Principle 6: The controls, control effects and changes made must be well-documented.
- Principle 7: Automate the editing process as much as possible.
- Principle 8: Streamline the editing work.
- Principle 9: Data editing should be evaluated.

Knowledge

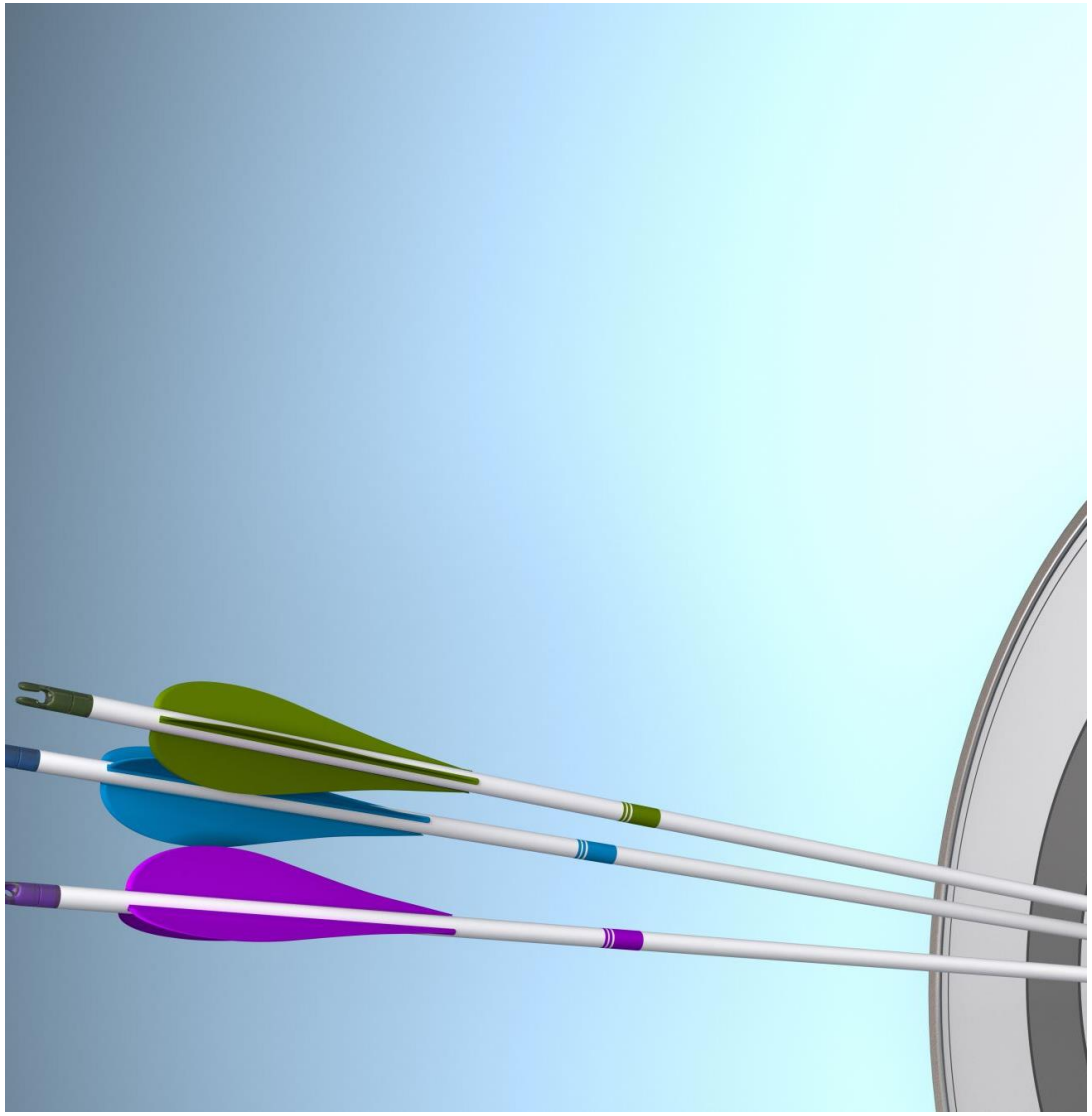
Principle 1: Good knowledge about the subject area of statistics and the background of source data is the basis for a good editing process.



Picture: Microsoft

Objective

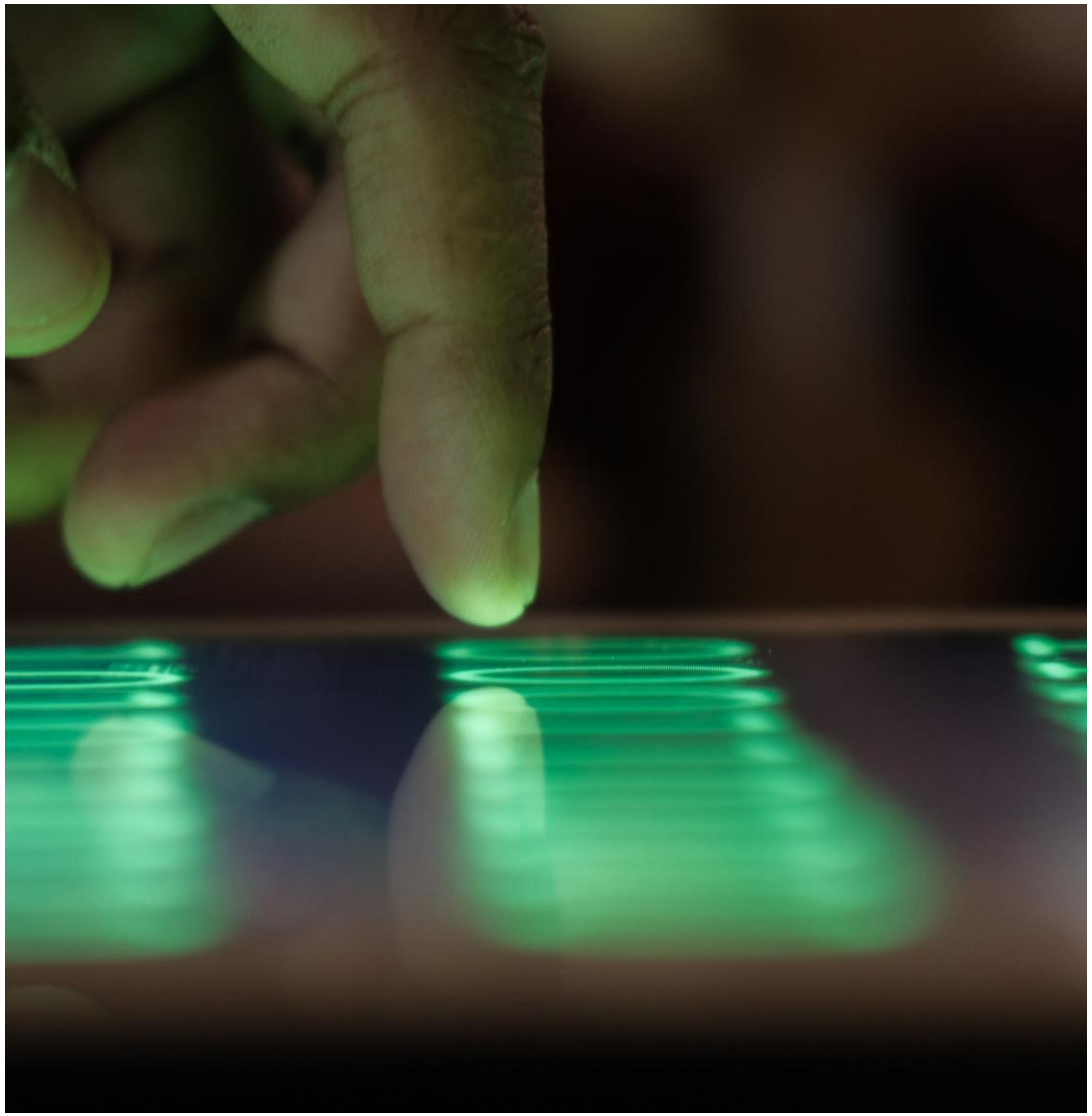
Principle 2: The objective of data editing should be clearly defined.



Picture: Microsoft

High-quality data

- Principle 3: High-quality data input is the best



Picture: Microsoft

Control

Principle 4: Always control the data.

Picture: Microsoft

Statistisk sentralbyrå
Statistics Norway

Place

- Principle 5: The earlier, the better.

Picture: Microsoft

Well-documented

- Principle 6: The controls, control effects and changes made must be well-documented.

Picture: Microsoft

Automate

Principle 7: Automate the editing process as much as possible.

Picture: Microsoft

Streamline

Principle 8: Streamline the editing work.

Improve human interaction:

- Good interfaces
- Visualization
- Selective editing
- Drilling in data between macro and micro

Picture: Microsoft

Evaluate

Principle 9: Data editing should be evaluated.

- Create quality indicators
- Analyze the indicators
- Take action

Picture: Microsoft

Application under construction

- Build in python with some function in R
- Independent of the underlying storage technology
- Possible to manually edit values
- Module based
  - Aggregation and drilling in data
  - Controls – edit rules
  - Influential observation and Hidiroglou-Berthelot function from R
- Process data and quality indicators

TALL
SOM FORTELLER

Thanks to international community for:

Generic Statistical Data Editing Model: UNECE

Metodology for data validation: EUROSTAT

Statistisk sentralbyrå
Statistics Norway