UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
Expert Meeting on Statistical Data Editing
7-9 October 2024, Vienna

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

Aslaug Hurlen Foss and Ane Seierstad (Statistics Norway)
aslaug.hurlen.foss@ssb.no

I. Introduction

1. The goal of writing national guidelines on data editing was to establish a common understanding of the data editing process and to promote best practice. The guidelines are based on the general model of the data editing process described in Generic Statisical Data Editing Model, GSDEM, (UNECE, 2019). In the national guidelines we have added a process before and after. These are: Control of data before it is submitted to Statistics Norway and control of data before transferring them into the Norwegian StatBank[1]. For each process, quality indicators are suggested. In addition to the guidelines, we have drawn up general principles for data editing. These are based on principles of data validation by Eurostat.

2. The guidelines are intended to provide a basis for building a data editing solution for the future. The goal is to develop a generic editing code base that can cover the editing needs of many statistics. Important functionality is the ability to drill down from macro to micro level. It has been established that manual editing of values should be possible, although such interventions should be minimized. Manual editing is also often necessary to train models for automatic correction. All changes to data should be logged. The solution should have access to preapproved functions for selective editing, outlier detection and imputation through the Norwegian Method library (Jentoft, 2023). Other important functionality is the use of visualization and quality indicators. The solution will be built using both Python and R.

II. Background

A. Existing data editing systems

3. In Statistics Norway, there are specific applications designed for data editing in individual statistical productions. Additionally, there are two general applications, Dynarev and Driller, which are utilized across multiple statistical domains. Dynarev is used by approximately 150 surveys in 2024 and was built in 2005. The primary features of this system include the management of controls (edit rules), the execution of these controls, and the configuration and implementation of both rule-based corrections and manual corrections. Additionally, it provides a variety of reports to facilitate efficient oversight of the entire process. All changes are logged, and quality reports are available. The application Dynarev cannot handle big datasets and therefore Driller was built in 2011 and is in 2024 used by approximately 25 surveys. The main difference between Dynarev and Driller is the data structure, view of data and possibility of drilling between macro and micro level.

[1] The statistics bank is the database where all Statistics Norway's figures from the statistics are stored. In the statistics bank you can choose from many different tables and variables.

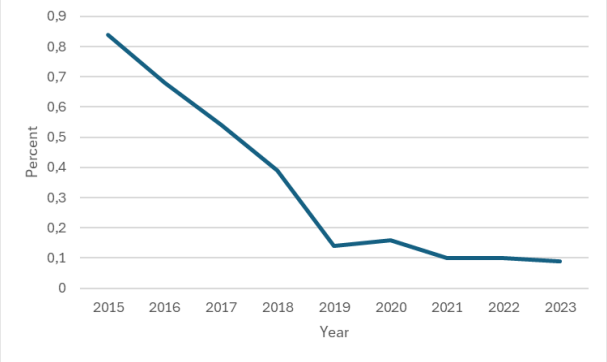
B. Technological modernisation – cloud solution

4. Statistics Norway is now establishing new cloud-based platforms for statistical productions. The main features are data storage in Google cloud buckets, access management, transfer services - for transferring data between buckets and between on-premises and cloud platforms - and automatic processing of data from source data to input data. The programming languages for building statistical productions are Python and R. (Falck, Gløersen and Folkedal, 2019).

C. Change of culture and common understanding of data editing

5. There has been a growing emphasis on reducing the time spent on data editing. Traditionally, the norm has been to correct all erroneous values for each unit, irrespective of their impact. This approach, particularly when executed manually, is labor intensive. Recently, there has been a shift towards minimizing manual work and focusing on units with significant influence. Data from the Dynarev application reveals a substantial decrease in manually edited values, with a drop of 90 percent from 2015 to 2023. In 2015, there were 1.73 million manually edited values, which decreased to 140 thousand in 2023. In 2023, only 0.09 percent of data was manually altered, as shown in Figure 1.

Figure 1. Proportion of manually edited values in percent, Dynarev



6. Data editing has traditionally been associated with the process of manually editing values and not the whole process of data editing as described in the General statistical data editing model (GSDEM). Information regarding this model has been shared in seminars, courses, and internal webpages in Statistics Norway. When discussing the future of data editing when transferring statistics to cloud-based platforms, it became evident that there was little common understanding of this process.

III. Striving for a unified understanding of the data editing process

7. The expert group of data editing of Department of Methodology in Statistics Norway made a first draft of the Norwegian principles and guidelines. It was based on GSDEM (UNECE, 2019), Ess handbook-methodology for data validation (Eurostat, 2016) and previous Norwegian handbook of data editing. Subject experts on different topics were then consulted and asked for feedback. The draft was also published on the Viva Engage channel, inviting all those with a special interest in data editing to provide their feedback. Then a second draft was made with incorporated feedback from subject experts. Draft number two was then sent for a formal consultation to all statistics departments and the IT department. It was organised by the Committee of standards in Statistics Norway. The feedback was included in the third draft that received endorsement by the Committee of Standards. This draft was then approved by the director's meeting and by the director general. The principles and guidelines were then published as a news article on the internal web for Statistics Norway to inform every one of the new guidelines. The document was then published to enable statistical producers outside Statistics Norway to use it. Now it is translated into English (Foss and Seierstad, 2024)

IV. Principles for data editing

8. Th guidelines could not encompass all the aspects of data editing, therefore nine principles were established to complement them. These principles are based on earlier Norwegian principles and Eurostat's principles for validation (Eurostat, 2020). They are general and intended to be applicable across all statistical production.

9. **Principle 1: Good knowledge about the subject area of statistics and the background of source data is the basis for a good editing process.**
With solid expertise in the field, it is possible to assess whether the statistics align with expected trends and to explain any deviations from these. Comprehensive knowledge about the source data helps anticipate potential errors and establish controls that capture these.

10. **Principle 2: The objective of data editing should be clearly defined.**
A clear objective for data editing is crucial for proper prioritization. This can be achieved by defining which tables should be published and what margin of uncertainty these may have, if this is calculable. Maintaining a macro perspective in statistical production and focusing on elements affecting the end results are essential. To ensure impartial processing of data, there should be instructions describing the tasks the producer of statistics needs to perform in the process of data editing.

11. **Principle 3: High-quality data input is the best**
Ensuring high-quality data input into Statistics Norwayensures the best quality in the statistics. The source closest to the data can report it best. For administrative data, the data goes through several steps before reaching Statistics Norway It is most efficient to receive good data initially, as it saves time spent on data correction. Therefore, efforts should prioritize obtaining good data rather than correcting it afterwards. Establishing a good dialogue with data providers and data owners is crucial to obtaining high-quality data. If errors affecting the statistics a lot are discovered, requesting updated figures from the data provider is recommended. For administrative data, feedback on quality indicators is vital for long-term data improvement.

12. **Principle 4: Always control the data**.
Although we trust that data has been controlled before it is received, the data should always be controlled. Successful data exchange is a shared responsibility and cannot be achieved without a reasonable level of trust and understanding of each other's challenges. The responsibility for accuracy lies with the data provider, ensuring that it meets their needs. The responsibility for controlling the data, based on the needs of producing the statistics and providing the data owner with useful feedback to enhance data quality lies with the recipient.

13. **Principle 5: The earlier, the better.**
The data editing process should be designed to detect errors as early as possible. This allows corrections to be made at a stage where knowledge is available. The sooner errors are discovered and corrected in a production

chain, the easier and more efficiently they can be corrected. When errors are identified and corrected early, the remaining processes are less affected by these errors.

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

The controls for a statistic must be clearly and unambiguously defined and well-documented so that they can be communicated to both data owners and other users of the dataset or statistics. This ensures a common understanding among the various stakeholders about what is implemented in the production process and how the statistics are created. The results of the controls, the control effects, must be clearly and unambiguously defined and documented for the dataset under scrutiny. Changes made to the dataset, both manual and automated, must be documented. Describing the actions taken ensures a common understanding of the outcome. Well-documented controls, control effects and changes made form the basis for creating quality indicators for data editing.

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

Automating the editing process involves setting up controls that run automatically, as well as making automatic corrections of data using logical rules or by statistical imputation methods. Additionally, automatic macro-controls, i.e., controls of the tables being published, should be established. Automated processes make production more efficient, but such processes must be monitored.

16. **Principle 8: Streamline the editing work.**

Effective use of selective editing and operating from a macro perspective helps streamline the editing work of the producers of statistics. Additionally, the editing work is streamlined by appropriately managing the tasks performed by the producers of statistics. This includes, for example, managing controls and automatic corrections, assessing data and product quality, and identifying and, if applicable, correcting influential errors. Graphics can provide a quick overview of results, trends, and data structures. Drilling, which is access to deeper levels successively in hierarchical data and figures, assists in explaining data and determining if correction is needed.

17. **Principle 9: Data editing should be evaluated.**

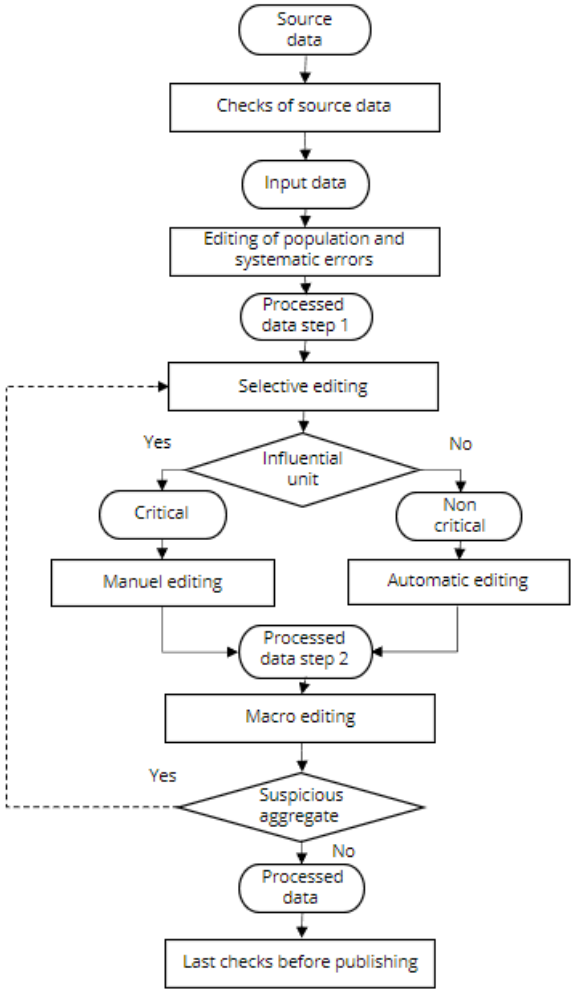
Evaluating data editing using quality indicators is essential to gather knowledge for improvement measures. Evaluation contributes to continuous improvement of the production process and the data. Quality indicators, such as the effect of editing, control effects and imputation rate, can be inputs to such an evaluation. An assessment of costs and use of resources should also be part of the evaluation. If measures are implemented to prevent errors from recurring or to make controls more accurate, the process becomes more efficient, and the quality of the statistics improves.

V. Guidelines for each process including quality indicators.

A. About guidelines

18. The guidelines are a specification of the principles. They are intended to be suitable for as many statistics as possible, hence the differentiation between social and business statistics in some processes. The guidelines are designed to serve as a handbook of good practice in editing work, specifying what should be controlled and how. Additionally, the guidelines include the recommended quality indicators for different parts of the editing process. The guidelines are based on the data editing process model (UNECE, 2019). While the model is a simplification of reality, it effectively illustrates the main processes involved in data editing. Additionally, the Norwegian guidelines include the verification of source data and StatBank tables, which respectively occur before and after the processes in the UNECE model, as reflected in Figure 2. In this chapter we will give some examples of the guidelines, the full guidelines are available (Foss and Seierstad, 2024).

Figure 2 Process model for data editing based on the business statistics model.



B. Example: Checks of source data - identification editing

19. Identification numbers (ID) are crucial in statistical production because most statistics are based on integration of several sources. The goal is to control and document the quality of identification numbers in received data and identify changes in identifiers over time.

(a) General guidelines:

- Verify ID numbers if errors are possible.
- Replace missing or incorrect ID numbers with the correct ones if available.

- Document missing or incorrect ID numbers.

(b) Guidelines for control of personal identity 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.

(c) Guidelines for control of organization numbers:

- Control organization numbers by cross-referencing against the Enterprise and Business Register.

(d) Recommended quality indicators:

- Proportion of invalid identification numbers: A variation of the quality indicator, indicating the percentage of invalid data.
- Proportion of corrected identification numbers: A variation of the imputation rate.

C. Example: Macro editing

20. Macro editing involves analysing aggregates or calculations on data for the entire population with the aim of identifying parts of the dataset that may contain potentially influential errors.

Guidelines:

- Assess whether the aggregates are plausible given historical trends, considering historical trends in other statistics or additional information about the same or related subject areas.
- 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.

D. Example: Last checks before publication

21. Confidentiality Check in Tables

Tables slated for publication must undergo scrutiny to ensure they meet confidentiality requirements. If the tables don't meet these requirements, cells are suppressed using approved functions as specified by the methodology division. These functions primary and secondary suppress cells within the tables.

22. Control of Statbank tables

Once data has been thoroughly controlled and approved, tables destined for Statbank need to adhere to specific rules that are controlled by Statbank. It is most effective for the statistical producer to run similar controls before sending the tables to Statbank, facilitating quicker adjustments if needed.

Guidelines:

- Statistical producers should perform controls on the tables before submission to Statbank
- Those receiving the tables should control compliance with Statbank rules.
- Elements to be controlled include:
  - Number of sub-tables and the number of columns in each sub-table.
  - Correct formatting in time columns.
  - Accurate codes in dotted columns.
  - Use only categorical codes registered in Statbank.

- Unique combinations of time and categorical codes, avoiding duplicates.
- Only numbers (or nothing) for statistical variables.
- All sub-tables should contain the same periods.

VI. Application

23. The first version of the new data editing app is being built to meet the needs of tax data for business statistics, but it is designed in a general way so that it can be used by other statistics. The key concept is to build modules that can be put together in different ways to suit the different statistical productions. Specific requirements were set for the development of the app. Firstly, it must feature a graphical user interface for ease of use. Secondly, it should be constructed using the technologies supported by Dapla, our new data platform. The application must also meet the editing needs of the statistics involved in this initial development phase. Furthermore, the design should be adaptable, allowing for its reuse in other statistical productions. It is essential that the application enables the measurement of editing according to quality indicators. The application should also employ statistical methods to minimize the volume of editing required. In addition, the application should be independent of the underlying storage technology to ensure flexibility. Lastly, the statistical production teams themselves should be able to manage and maintain the application.

Dash Ploty is chosen as the framework for building the graphical interface. The interface in DASH includes a data table that communicates with the Parquet foundation. The foundation is built in Python, utilizing methods from the method library in R. The system supports the use of APIs for queries.

24. Each menu item in the graphical interface triggers code, with each point functioning as a separate module. So far, we have included the following modules: Aggregation module for displaying results and drilling from macro to micro level, Control module to set up and run automatic controls and analyze the results, Modules for the use of more complex functions from approved Method Library (only the Hidiroglou-Berthelot method yet) and a Quality module to look at process data and quality indicators. There is an aggregate/table with a transition to the micro-level. The micro-level has editing capabilities and can be customized. The principle is to correct only what is necessary

25. EimerDB is a Python package developed in Statistics Norway. Unlike traditional database solutions, it employs Parquet as its storage format and utilizes buckets for storage technology. It offers support for partitioning of Parquet and enables SQL queries via DuckDB. EimerDB ensures that all changes are logged in a Parquet context. This feature also facilitates concurrent editing of the same data by multiple users. EimerDB maintains access to unedited data even after changes have been made and provides easy access to all edits made to the data. Every time a value is edited and saved, EimerDB generates a new Parquet file containing only that change, along with information about who made the change. All the statistics which the first version of the app is built for, use data from The Norwegian Tax Administration. It is important that changes are made in one place, so that everyone can benefit from them.

References

Falck, T., Gløersen, R. and Folkedal, J. (2019). Modernizing statistical production at Statistics Norway. Nordic statistical meeting. Access 28.06.2024.
https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fstat.fi%2Fmedia%2Fuploads%2Fajk_en%2FEvents%2Fnsm2019%2Ffalch_-_modernising_statistical_production_at_statistics_norway.docx&wdOrigin=BROWSELINK

Foss, A. H. and Seierstad, A. (2024). Principles and guidelines for data editing. Statistics Norway. Documents 2024/31. Access 29.08.2024
https://www.ssb.no/en/teknologi-og-innovasjon/informasjons-og-kommunikasjonsteknologi-ikt/artikler/principles-and-guidelines-for-data-editing/_/attachment/inline/d403545c-d5d7-471f-9340-9bb01ae1e342:e7399a8716c23f396b7892b6752c8ebce4ae8c32/NOT2024-31.pdf

Eurostat (2016). Methodology for data validation 1.0. Access 29.08.2024
https://ec.europa.eu/eurostat/documents/7755309/7769541/Methodology_for_data_validation_V1.0_REV-2016-06_FINAL.pdf/ed7abb5d-6f1f-4e92-9a72-40d8a7fe0b85

Eurostat (2020). ESS vision 2020 validation, principles for data validation. Access 29.08.2024
https://cros.ec.europa.eu/book-page/principles

UNECE (2019). Generic Statistical Data Editing Model, Version 2.0. Access 29.08.2024
https://unece.org/statistics/documents/2019/06/gsdem-v20

Jentoft, S. (2023). Creating, testing and maintaining a functional library for statistical methods. Abstract at the conference New Techniques and Technologies for Statistics 2023, page 262. Access 30.08.2024
www.iweps.be/wp-content/uploads/2023/03/2023March_NTTS-New-Techniques-and-Technologies-for-Statistics-Eurostat-book_of_abstracts.pdf
Github: https://statisticsnorway.github.io/ssb-metodebiblioteket/

UNECE (2019). Generic Statistical Business Process Model GSBPM, version 5.1. Access 29.08.2024
https://unece.org/sites/default/files/2023-11/GSBPM%20v5_1.pdf