Ingegerd Jansson and David Sundström
ML/MA/MET

Scientific Advisory Board Meeting May 5th-6th 2022

Attending Board Members

Jan Bjørnstad, Statistics Norway and University of Oslo
Barteld Braaksma, Statistics Netherlands
Stephanie Eckman, RTI International
Anders Holmberg, Australian Bureau of Statistics
Per Johansson, Uppsala University
Annette Jäckle, University of Essex
Sune Karlsson, Örebro University
Johanna Laiho-Kauranne, Statistics Finland
Xavier de Luna, Umeå University

Attending Statistics Sweden staff

Per Appelkvist
Mats Bergdahl-Kercoff
Mattias Björling
Maj Eriksson Gothe
Marie Haldorson
Ingegerd Jansson, secretary
Lilli Japec
Jens Malmros
Andreas Persson
Magnus Sjöström
Kristina Strandberg
Gustaf Strandell
Joakim Stymne, chair
David Sundström, secretary

Current Issues at Statistics Sweden (SCB)

Joakim Stymne welcomed the members of the Board to the first in-person meeting since November 2019. New board members are Stephanie Eckman (RTI international) and Steven Heeringa (University of Michigan).

Joakim Stymne briefly described SCB's new organization. The peer review of SCB was also described along with media impact of (geographically) local statistics.

Reply to recommendations

Lilli Japec presented SCB's replies to the recommendations given by the Scientific Advisory Board (SAB) at the previous meeting.

Statistics Sweden's Strategy and Targets 2022-2026

Marie Haldorson presented SCB's strategy and targets for the coming years. The presentation revolved around the following main points:

- How to gather, guard, grow and give data to society?
- The National Statistical System – backbone or obsolete?
- Statistics vs. Data Science – evolution or revolution?

Business focus, 2022-2024

To put the strategy into action, work in four prioritized areas is pursued, areas which are of major interest to users as well as respondents. The four target areas are the following:

Labour market statistics

SCB provides an actual, deepened, and accurate picture of the Swedish labour market by employing statistics and data in a broad sense – on the national, regional as well as the local level.

Economic Statistics

SCB sheds light on new phenomena within the Swedish economy and meet national as well as international needs of actual and reliable statistics. The opportunities of digitalization are utilized.

Communication

SCB has increased the use of statistics and data by different channels in partnership with data processors and expert users.

Ensured data access

By being a driving partner in the national data ecosystem SCB has secured long-term access to various data sources, to be utilized in the production of official statistics. Simultaneously, the enterprises and organizations' response burden have been halved.

Comments from the board

- It is a very ambitious goal to halve the response burden. Questions were raised regarding how to measure it and if it is even a realistic goal.
- "Sound statistical practice" is not mentioned. This needs to be spelled out in more detail.
- Why the focus on economic and labour market statistics? Environment and economy are related, health and economy are related. A wider picture would be important for the

country. Do the chosen subjects reflect the siloes of statistics?
- What are the costs? Going from surveys to found data will require a lot of work, combining sources, etc. How do we know that it is cheaper to do this transformation?
- What are the new demands new data sources are going to cover? Is something missing in the statistical system? The big things are already there, has there been an analysis of what is missing?
- Regarding use of ML and AI: these techniques are not explainable, and there is a trust-problem for statistics which are not explainable. In general, keeping trust while doing changes is a question of communications. Labour force statistics are particularly sensitive and of high interest to politics.
- It is advisable to create indicators to follow what you do.
- A suggestion is to distinguish better between micro- and macro data and have different strategies depending on the type
- Data science does not care about uncertainty. It is important that point estimates are accompanied by measures of uncertainty. This requires sophisticated modelling. There is a need to teach society about uncertainty.
- Timeliness and relevance are key. The effect of climate change is an example, with uncertainty. Quality statistics are important.

How to develop the cognitive lab's methods to contribute to the evaluation of new data sources

Andreas Persson presented the topic and Annette Jäckle provided a prepared discussion.

Summary of the topic: Statistics Sweden has a cognitive lab that deals with the design and the evaluation of the measurement process in surveys conducted by questionnaires. For this purpose, the methodologists in the lab apply methods such as cognitive interviews, expert reviews, and debriefings. The lab's main purpose, to design and to evaluate the measurement process, is focused on two areas: 1) to describe the data generation process (interactions between respondent, questionnaire, and information systems) and 2) to assess validity.

It was proposed how to extend the cognitive lab's areas of expertise and methods, particularly concerning the data generation process, to include methods that contribute to the evaluation of new data sources. When examining the characteristics of a new data source, descriptions of the data generation process can provide additional insights together

with other methods such as data analysis of test data. The presentation ended with a discussion about future steps and how to put the proposed methods into practice.

Comments from the board

- This is a great idea! In addition, consider the time dimension, what happened in the past? Error properties might change over time. Should the reviews be repeated at certain time intervals? Different variables have different properties. Where is the expert who can answer the detailed questions? Probably not at the highest level in an organisation.
- This work is statistical quality assessment. Pick the most important sources to prioritize. Look at the NZ framework (Reid et al 2017 DOI: https://doi.org/10.1515/jos-2017-0023).
- There could be other organizations doing the same type of work. Why is SCB doing this work, is there a comparative advantage that SCB is doing it?
- Are you paying for data? How will you convince companies to give data? Are there any ongoing negotiations? From new to approved data source – how do you decide?
- A shift from only collecting data for which we know the use to collecting without a clear purpose from the beginning. Can there be quality without knowing the purpose of the data? Is there any way of describing quality without knowing the use?
    - A question of good metadata. It is the only way to describe data without knowing the purpose.
- Check out Datasheets for Datasets for a list of questions you might consider incorporating into your reviews: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4t8QB
- The ideas of concept drift and data drift from machine learning would be valuable here.

A new process for data editing at Statistics Sweden

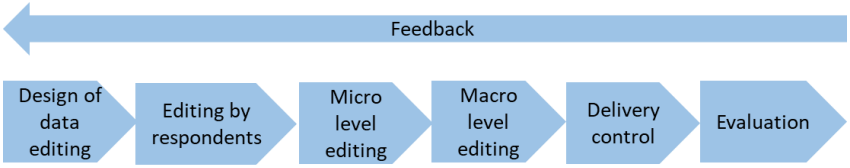
Magnus Sjöström introduced the work with a new editing process and related ambitions at SCB. Gustaf Strandell gave an in-depth presentation and Johanna Laiho-Kauranne provided a prepared discussion on the topic.

Summary of the topic:

As part of the digitalization of SCB's production process and to save resources, SCB has decided to phase out the manual elements of the data editing process. An ambitious plan has been developed according to which a large part of the manual work will be phased out already in 2022 and with continuation in 2023. To support the work, a new data editing process has been developed that place a strong focus on respondent edits, automated edits for micro data and rule-based macro

editing with visualizations. Since the discontinuation of the manual editing process involves some risk-taking, the new process also includes selective editing with the idea that if manual elements are to remain, such as re-contacts with large companies, these should be strictly tailored to needs.

The new editing process is based on the international model *Generic Statistical Data Editing Model* (GSDEM) and is consistent with other such models that are in use at Statistics Sweden.



The first step in the new editing process is called Design of data editing and covers both the implementation of the new editing process for the survey at hand and the continuous improvement of the editing based on the evaluation and feedback to be done after each completed survey round. In the design a holistic perspective on the editing process is strongly recommended to avoid unnecessary work and to keep things manageable throughout.

The second step is called Editing by respondents. Emphasis is on an increased use of interactive respondent controls in our electronic measuring instruments. The basic idea is to give respondents the support they need to avoid errors in the form of well-designed questionnaires, clear instructions, and interactive respondent controls.

The biggest difference from today's editing process, which focuses strongly on traditional manual data editing, is in the Micro-level editing. The focus here is on automatic editing and on selective editing. The latter is used to select the records, if any, to be handled manually.

Macro-level editing is carried out when all the data are collected and most of the statistical production is completed. The focus here is on rule-based macro editing and the use of tools to visualize macro data such as SAS VA or Power BI. In rule-based macro editing, programmed plausibility controls are applied to macro data, like they are currently used in micro editing.

Delivery control is the last check before the statistics are made available to users. The implementation of the new editing process does not currently cover this step as it is of a slightly different nature.

In the final steps, Evaluation together with Feedback, process-data from both editing and data collection are studied. The purpose is to find improvements to be made in the next survey round.

Comments from the board

- This is not a new question. Do not edit for the sake of editing. It depends on what you are producing, e.g., averages might be affected but no effect on robust statistics like the median (robust statistic). It iso not known how editing affects the results in the first place, for example, it most likely generates too narrow confidence intervals.
  How do you make the distinction between editing and robust estimates? Are others using the same data, in other ways? Then a robust estimation-approach could have implications for others' use of the data.
- Be clear on the purpose. It would be useful with indicators if there is a measurement error, and then researchers can use that information and find their own solution.
- The goal is a great one, and a bold one. Concern: A concern is if an imputation model is based on historical data. It would be vulnerable to sudden unexpected changes, such as the pandemic.
- Keep consistency between statistics, and do not use a single source approach. Otherwise you might have conflicting results.
- Surprised to hear about editing without talking about imputation. Previous work at SCB? There is a lot from the old Lotta-project. Some tools were tested 20 years ago. A question of resources and management, how to relocate resources.
- Regarding the use of ML-models: it would be useful to have international cooperation, for example in the ESS. . Some work is also being done at ABS, but the models are much affected by the pandemic.
- Multiple imputation is very difficult to implement, is any statistical agency using it?
- Multiple imputation is problematic if it is deterministic. Too short confidence intervals (a constant is imputed). A more statistically sound line of action would be to draw randomly from predictions using covariates, to measure the uncertainty at the right level. Is anyone using that?
- One fundamental assumption is that observations flagged are like the rest of the data. Some of these might not be erroneously recorded at all. Automated/ML with model need to be careful with that. Some additional checks necessary (for example a small sample that is edited manually).
- Important to prevent errors from the beginning. The ideas for the questionnaire are good, not only a check, and it also generates additional (process) data.
- Collect as much as possible from existing sources, do not adjust questionnaires too much.

The role of methodology and its interplay with future shifts in statistics production (the Australian example).

Anders Holmberg described methodology-related work at the Australian Bureau of Statistics. A discussion regarding training capabilities followed.

Concluding words

Joakim Stymne closed the meeting.