Ingegerd Jansson
Jens Malmros
SCB

Meeting with the Scientific Advisory Board of Statistics Sweden

16-17 May 2024

Attending board members
Jan Bjørnstad, SSB och Oslo universitet
Barteld Braaksma, CBS
Steve Heeringa, University of Michigan (online)
Anders Holmberg, ABS
Per Johansson, Uppsala universitet
Sune Karlsson, Örebro universitet
Johanna Laiho-Kauranne, CSC IT Center for Science
Thomas Laitila, SCB

Attending Statistics Sweden staff
Joakim Stymne, Chair
Lilli Japec
Marie Haldorson
Anna-Maria Kling
Kristina Strandberg
Mats Bergdahl-Kercoff
Jens Malmros
Ingegerd Jansson

Emma Stavås
Jacob Kasche
Gustaf Strandell
Wictoria Widén
Marcus Vingren

Other participants
None

Welcome

Joakim opened the meeting and welcomed all participants.

Current issues at Statistics Sweden

In recent years, Statistics Sweden has seen many changes. A new strategy was developed, a consequence of which was the new organization introduced in 2021. In addition, Statistics Sweden has moved to new premises on both locations. Budget issues also affected the organization.

Externally, Statistics Sweden has improved visibility. For example, the introduction of the new registered based labour market statistics has been very successful. Statistics Sweden has changed the publication time, which has improved visibility. Moreover, the reputation of Statistics Sweden among the public and students seem to improve.

In recent discussions with the state secretary, several forward-looking issues were discussed. Firstly, the strategic goal of Statistics Sweden to reduce response burden by half aligns with the government goal to reduce the administrative burden on companies, which motivates further work in this area. Secondly, the role of Statistics Sweden as data steward is discussed both within and outside of the organisation and is likely to require further investigation. Thirdly, the directive on open data and making high value data sets available will impact Statistics Sweden on several levels. General questions on digitalization and AI were also discussed.

To meet these upcoming challenges, and to revise according to previous results, Statistics Sweden is looking into updating the current strategy.

Reply to recommendations

The two main topics from last meeting was Micro Data Exchange in Foreign Trade Goods Statistics and Imputation of Driving Distances.

For Micro Data Exchange, the Board had several comments and suggestions, mainly concerning mismatch between the current data source and the new data source, and the costs and benefits of the new source. In addition, the Board raised several issues on a detailed level. Most importantly, Statistics Sweden should continue to do the current survey and evaluate the new source in parallel with the current survey. Much work has been invested at the European level, but it is necessary to evaluate if it is worth the effort. A lot of capacity building will be required.

Statistics Sweden agrees that it is important to evaluate any effects before implementing the new source. The international experience of the source is however limited since it has only been implemented in

Denmark. The source shows promise in reducing respondent burden and development will continue. The team would like to return to the SAB in the future.

For driving distances, the board noted that the research was well done and that the proposed methods show improvement compared to the current method. The board pointed out that machine learning relies on a missing at random assumption, which may not be fulfilled. Rather, the board views this as an estimation problem.

Transport Analysis is considering new data sources, for example data from software installed in cars, which could be useful in estimation of driving distances. The recommendation of the SAB to view this as an estimation problem will be followed. However, imputed values (using arithmetic mean imputation) will continue to be used for some statistics, but they will not be used for analysis.

Topic 1: Towards the Vision for Coding: An Accurate and Efficient Coding Process

Emma Stavås introduced the topic.

In 2020, a Vision for Coding was developed at Statistics Sweden. The vision has three parts: Design of data collection, The coding process, and Intended effects. In addition, the vision contains writings on quality assurance.

Recently, the Data department received a mandate from management to streamline the coding process further. The work will consider two case studies: coding of occupation and coding of COICOP. The presented work only considers coding of occupation. The Vision for Coding remains an important input to the goals of this work.

Discussion

Johanna Laiho-Kauranne opened the discussion.

Some comments from the discussion:

- It was noted by the audience that the LFS uses more time for coding than SILC, even though both surveys code according to three standards. This has several reasons, e g the mode used, and that LFS have stricter rules for coding and is a more complex survey.

Vision for Coding

- The vision could be formulated as more than simply coding, it is more generally about quality of relevant information for progress of the economy.
- SCB could continue to rely on the vision for coding to some extent, but it should be revised according to the development during recent years.
- The aims with streamlining the coding process should be clarified. For example, lowering response burden and less use of manual resources.
- The vision could be improved by clarifying the different steps in coding, which could assist in highlighting the difficulties in the process.
- In the vision, the goals on coding are not given with respect to boundaries and quality requirements. The goals would be further enhanced by adding these dimensions. Make it explicit that reduction of burden is a main goal, but with respect to a certain level of quality.
- It should be added to the vision that there should be a common toolbox for coding.

Data

- It would be of interest to consider how the classifications to be coded are used, and consequently, how the coded data are used.
- It can be useful to consider the quality of input data since it has a large effect on the output quality of coded values.
- There is a statistical burden in addition to the response burden. This involves costs and available data. It is useful to map the process and identify prior information that could be of use. Utilise prefilled answers, adaptive design, knowledge of interviewers, information available at other agencies such as the Tax authority, etc.

Methodology

- Continue to develop the understanding of the cognitive processes of respondents in coding.
- The dictionaries used could be extended and possibly jointly developed in the Nordic countries.
- A suggestion is to further explore the use of large language models in coding.
- It may be possible to predict values instead of coding all data. Try to work carefully on one survey and use the result together with registers to impute in other surveys.

International collaboration

- Other agencies also do this; hence it is possible to develop a common solution with other agencies.
- At ABS there is a common government coder project because of the census.
- If information is already present at other agencies, it should be used. This also relates to larger perspectives.

Quality assurance

- A model should be developed to produce the best predictions. However, then the statistics would be produced using data that have been collected from users and coded automatically, and this division is unknown to users. It could be useful to compare the results with and without the automatically coded values. This should also be communicated to users.
- It should be possible to develop a model using survey data that can be used widely.
- Quality control is important. Control coding should be carried out at a few years intervals.

Other

- The organisation of the manual coding work can be considered, for example with respect to centralization.

Topic 2: Quality Assurance of the Re-coding to NACE Rev. 2.1, Combining Model and Manual Coding

Jacob Kasche introduced the topic.

The work concerns the revision to NACE Rev. 2.1 and focusses on quality assurance in the re-coding process utilizing the combined use of different methods. In the previous re-coding process, primarily manual methods were used. In the current process, it is suggested to combine coding using a key, automatic coding using large language models, imputation, and manual coding. A quality assurance process based on these methods is suggested.

Discussion

Barteld Braaksma opened the discussion.

Some comments from the discussion:

Stakeholders

- The description of both internal and external actors is lacking since the process is described in terms of procedures and not in terms of actors. It would be helpful to clarify who the stakeholders are because there might be more stakeholders than expected.
- When the stakeholders are known, their expectations in terms of quality on the revision must be assessed.
- There are a lot of internal stakeholders in this kind of revision. Many products will need update and timeseries will change. It is a good idea to consider this methodological work at an early stage.

Communication

- The quality assurance process has two purposes: to control the re-coding process and to report to stakeholders what has been done and why. The latter part is not clearly present in the current version of the process.
- It is important to communicate with stakeholders and to keep the ministry of finance informed. The revision may affect certain sectors to a large extent, and it may be valuable to communicate changes and how large they might be beforehand.

Related work

- It is possible to use data from websites in the ML part of the coding. Such data are used both at CBS and at the French statistical office INSEE.

- International collaboration may be possible in, e.g., TAILOR[1], AIML4OS[2], ESS Innovation network[3], and HLG-MOS[4]. There may also be other related works, e.g., Austria has developed a hierarchical approach to classification.

Data

- There is little information on the sources of the textual descriptions used. In addition, their quality is not assessed, and one may expect issues such as ambiguity and incompleteness. It would be of interest to investigate if these data could be further controlled and assessed.
- For many NACE groups, there are many similar descriptions. In addition, there are also outliers. Both these characteristics could be used to improve the classification.
- The quality of the data used is unknown; however, it is likely that there are a lot of errors. The quality of the data should be evaluated.
- Webscraping as a source is plausible, and relatively stable once it has been set up. It is a good investment, as the example of the Netherlands show.
- It is important that high value data sets are kept by SCB, but they are expensive to maintain. SCB should seek funding. A motivation for funding could for example be that business register are kept for multiple use by the society.

Models

- Models do not introduce uncertainty; they provide the possibility to estimate uncertainty. There is also uncertainty in manual methods which is not easily estimated.
- It would be useful to consider the hierarchical structure in the model, by utilizing for example a regression tree.
- At ABS, it is likely that both BERT and a SVM approach will be used. It seems like they may complement each other. An issue with LLMs is that their result cannot always be explained.
- The models may degrade if they are used in the future and then trained on model-generated data.
- Be aware of concept drift.
- If you use freely available version of LLMs, your questions are recorded as data to be used in the models. This could reveal

[1] TAILOR - A Network for Trustworthy Artificial Intelligence (tailor-network.eu)

[2] One-Stop-Shop For Artificial Intelligence/Machine Learning for Official StatisticsAIML4OS | Insee

[3] Innovation in statistics | Eurostat CROS (europa.eu)

[4] High-Level Group for the Modernisation of Official Statistics | UNECE

sensitive issues in the organization. It is important to use established software to avoid this.

Methodology

- It is important to consider false negatives and false positives similarly to reduce risk of asymmetry bias.
- There is no justification for the numbers used as threshold for quality values. They should be further explored and motivated.
- It would be useful to include a mechanism to select priority areas.

Process

- A PDCA/Deming cycle approach can be used to improve the process.
- It could be useful to set out from the quality requirements and allow for adaptation during the process.
- It may be possible to evaluate different scenarios in parallel to find out best considerations during the process.

Other

- Consider changing the ambiguous terminology on quality indicators and quality measures. Indicators are auxiliary data and measures are target statistics.

Statistics Sweden's AI policy

Marie Haldorson presented the draft AI policy.

When the current strategy of Statistics Sweden was formed, it included AI and machine learning. However, rapid development in these areas motivate the development of a separate AI policy.

Currently, the machine learning group leads the work on development and implementation of machine learning in statistical production. However, there is an increasing interest in using AI outside of statistical production. There are many ideas in the organization, and in order to make the best use of them it will be necessary to prioritize and avoid unnecessary costs. This is an important motivation for the development of an AI policy.

Discussion

Some comments from the discussion:

Content

- An AI policy is an essential document for an organization, to maintain the trust of the public, the data providers, and the employees. The ultimate purpose should be stated in the policy.
- The portfolio is wide-spread and oriented on the methods used, not on the goals, which are to produce high quality statistics.
- The portfolio lacks activities focussing on user interaction. For example, it would be possible to integrate MS Copilot on the website.
- A clear focus on increasing the internal efficiency would be useful, for example, producing analytical texts and conversion of code. Many of the tools are available and should not be re-invented.
- It would be useful to add references to the guiding policies, for example GDPR and code of practice.
- Other suggestions for extending the policy:
    - Add an introduction where the role of SCB is clarified.
    - Include an approach for how the policy will be revised and communicated.
    - Connect the policy to guidelines for other areas, such as competence.
    - Make explicit that privacy is a primary concern. At NSOs, data protection and privacy are important concerns. Anything that consumes large amounts of data gets hungry for more data. It is important to clarify which data can be accessed for external and internal use, and if other data sources can be combined with data from SCB.

Structure

- It would be useful to distinguish between different purposes in the policy, for example, statistical production and internal efficiency. This would help to understand how things evolve in different areas and what tools can be used.
- A more holistic perspective may be needed to include efficiency gains outside of statistical production, for example in administration or HR.
- It would be possible to go further and have a data strategy.

Data and models

- AI may require us to organise our data better, otherwise it will not be useful.
- The role of data in AI is a specific topic that needs special attention and is often underestimated. The relation between data and models should be better understood. The quality of data has a large impact on the output from the models.
- Understanding the models is important, so that they can be explained to the users.

Terminology

- At Statistics Sweden, AI seems to be mostly machine learning and therefore this terminology should be used. In addition, if the focus is on goals, it will be clearer when machine learning is useful and when it is not.
- The concept of narrow AI is not generally used and may be confusing. The terminology of explainable and non-explainable, as well as responsible AI, may be useful.
- If you would only say that SCB uses machine learning, you will limit your possibilities. For example, the use of LLMs may be hindered.

Topic 3: Household Consumption. Estimation of product composition

Thomas Laitila and Marcus Vingren introduced the topic.

The project on Household Consumption started in 2020 and was previously presented to the SAB in May 2021. Because of low response rates, the household budget survey does not longer meet the standards of the national accounts and CPI. In 2012, the survey was paused and re-designed. However, when a redesigned survey was launched in 2016 response rates were even lower. Therefore, a larger redesign effort was undertaken. In 2022, a new HBS was given. However, the requirements for national accounts and CPI are more detailed and have to be further investigated.

Discussion

Anders Holmberg opened the discussion.

Some comments from the discussion:

General

- This is a multisource-multipurpose project with several methodological questions which are interdependent. There are several purposes. One is to bridge the old and the new design. Another is to provide what is required for NA and CPI. A third thing is the method used for calculation. It would be useful to draw up these purposes and their dependencies.
- The paper is not clear on the requirements of the NA and CPI.
- The goal mentioned to base the statistics on admin sources is not a goal, it is a means to reach a proper goal. The work needs to set out from proper goals and focus on, e.g., modernising statistics production or solving problems.
- The HBS could serve important societal purposes and it is bad that it is not possible to provide information in these matters.

Methodology

- It is possible to put restrictions on the equations, for example, a11=b11.
- The assumption that the errors and covariates are uncorrelated (A1) may not be realistic. The sensitivity of this assumption could be investigated.
- The time dimension is important. For example, transport and accommodation was affected by the pandemic.
- To find the best possible way to solve the problem it is useful to compare to models using, e.g., finite mixture models.

- Not all the weights in the CPI will vary between large intervals. This may be used in the calculations and should be further investigated.
- It is possible to look at how the weights would change if some products or categories are left out.
- It may be possible to model the lower hierarchy parameters to the higher-level parameters.
- There are many products groups that are not covered. It should be investigated what effort is needed to cover these and the cost of not covering them.
- It is a promising and novel approach to look at the SBS, but it should be validated against empirical comparisons.

Related work

- The Necessity and Proportionality framework of Statistics Canada may be useful to consider when deciding which data to collect and use, especially for sensitive data.
- It is important to have a statistical approach, i e how much data do you need and for what. In Norway, there was a pushback because these issues had not been considered.
- There are other projects in other countries, but it is important to clarify the purpose of the current work before looking at these.

Data

- A cash register approach is encouraged. At CBS, there is no data collection on retail trade and only scanner data are used.
- It may be possible to use apps to collect data. There is a European project on using apps to collect transaction data.
- Investigate whether it is possible to include purchases made online from international sites in the statistics.
- Cash may be used more frequently for some small product groups. This should be further investigated.

Concluding words

Joakim thanked the SAB for valuable discussions and comments. The work of the Board is very helpful in moving Statistics Sweden forward in our progress.

Next meeting

The next meeting is November 14-15. The meeting will be held online.