Ingegerd Jansson
Jens Malmros
SCB

Statistics Sweden Scientific Advisory Board

December 13-14, 2022

Attending Board Members

Jan Bjørnstad, University of Oslo
Barteld Braaksma, Statistics Netherlands
Xavier de Luna, Umeå University
Stephanie Eckman, RTI International
Steve Heeringa, University of Michigan
Anders Holmberg, Australian Bureau of Statistics
Per Johansson, Uppsala University
Sune Karlsson, Örebro University
Johanna Laiho-Kauranne, DAIN Studios

Attending Statistics Sweden staff

Mats Bergdahl-Kercoff
Maj Eriksson Gothe
Marie Haldorson
Ingegerd Jansson, secretary
Lilli Japec
Mia Kling
Thomas Laitila
Jens Malmros, secretary
Kristina Strandberg
Gustaf Strandell
Joakim Stymne, chair

Current Issues at Statistics Sweden

Joakim Stymne welcomed the members of the Board. The meeting started with a round of presentations. Mia Kling is the new head of the methodology section at the Executive Office, and Jens Malmros is new as secretary of the Board.

How does an NSI stay relevant in a changing world? Joakim had asked the ChatGPT AI, which had generated a general reply. Real people are still necessary to discuss the question.

Joakim gave an update on the financial situation; Statistics Sweden needs to cut costs the coming years.

Sweden will chair the EU for the first half of 2023, and Statistics Sweden is chairing the statistical agenda for the period. Some questions that will be discussed are population definition (implications for register based statistics), environmental accounts and access to privately held data. A high-level workshop on statistical leadership will be held in Stockholm in April.

Comments from the board

- Will there be a traditional census in Sweden? The new Swedish government has raised the question, but Statistics Sweden has so far not been commissioned to do anything. The last traditional census was in 1990, since then there have been register based censuses. The government has assigned the Tax authority to increase control of where people live, as opposed to where they are registered, to prevent illegal immigration and illicit utilization of benefits and allowances. Illegal immigrants cannot be found by a census, but there are methods for estimation that have been used in Norway. This is also an issue at the European level that has been discussed by the director generals.
- The European statistical act (223/2009) is being changed; one major issue is how to strengthen NSI's access to privately held data. There are a number of legislative acts being put in place supporting the EU Data Strategy which might affect the work on 223/2009.
- Will the financial situation mean that Statistics Sweden will charge more for commissioned services to researchers? Commissioned services are calculated the same way as before, they should cover all costs in relation to the service.

Reply to recommendations

Marie Haldorson and Lilli Japec presented SCB's replies to the recommendations given by the Scientific Advisory Board (SAB) at the May 2022 meeting.

Quality indicators for organic data

Lilli Japec presented the topic and Anders Holmberg introduced the discussion.

Comments from the board

- What do we call these types of data? Several optional names for organic data were brought up during the discussion.
    - Darwinistic data – we don´t know how long they will survive

    - Random data - we don’t know the original purpose of the data
    - Repurposed data – it is designed but for a different purpose
    - Fast data – as opposed to robust and reliable. The main purpose is timeliness.
    - Secondary data – they are for secondary use, data collection is not the main purpose
    - Non-survey data
    - Supplemental data – they only get useful when data are added to other data
    - Non-probability data
    - Unconventional data

    There was also the opinion not to bother to find a name, pick a few examples that will work for statistics and focus on the use-cases.

- There are too many indicators. It is important to be clear on what is being estimated and what can be said about the error in the final estimates. Try to keep it simple, users are likely to prefer only one measure of uncertainty. See reference to Statistic New Zeeland.
- How do we assess if data are of good enough quality? Budget issues need to be added. The cost and the competence required should be considered, so that we get a full picture of the investments required for the purpose. Maybe possible to add a selective criterion, to justify the purpose. Do the data bring value to the statistical system? Clarity, and to understand the original use case is key.
- Quality is not necessarily a question of good or bad, data can be good for one purpose and bad for another. It is also necessary to be clear if the data are intended for a specific use or for multiple use. In addition, there is a difference if you want to make new statistics compared to if you want to improve existing statistics.
- Sustainability, availability over time, continuity, reliability over time, is important. This is connected to who the owner or holder/custodian of the data is. This is a main difference compared to traditional data sources. These questions are discussed at the European level and in some new European legislation.
- A more academic issue is to go even further with trying to unite the different frameworks in the literature.
- Is the work intended to be a framework for a data source or for an estimate? Some confusion regarding this, partly inherent from the literature.

- It is not always clear what the micro data are in the new sources, and usually the NSI wants aggregates. It is important to assess where and when aggregates happen.
- Integrating data is necessary, and the KOMUSO-project is a good source for that. From the user perspective, linking the data to other registers and the microdata created by the linking is very important. It is recommended that Statistics Sweden provides the linking so that microdata become useful for researchers, and then the researchers might be able to help in solving other problems with the quality.
- The users care less about the levels of estimates and more about changes over time. It is important that the data valid over time.
- It is essential to have knowledge about the data generating process. SCB must gather meta data. There might be opportunities to influence the structure, both with public and private data holders.

Imputation of occupation in the Occupational register

Jens Malmros presented the topic and Stephanie Eckman introduced the discussion.

Comments from the board

- Are the measures god enough? It is important to be clear on the purpose of the imputations and of the statistics. For the whole population, the proposed measures are good, but not so good at the individual level. When the statistics are broken down by categories, they need to be evaluated at that level. Consult the subject matter experts. Can existing surveys be used for validation?
- Additional quality control checks
    - Logical checks are very relevant.
    - Make more use of the time aspect
    - Explore how certain the model is in predicting occupation. Something like the R-indicator in sampling? Take advantage of the information on uncertainty.
- Use the hierarchical structure of the occupation variable and start imputations at the highest level, then continue with lower levels. The lowest level is a very small target to start with.
- The complexity of the labour market is increasing. Select variables to reflect that; region, first time access to labour market, time of work experience, still in education, etc could help to clarify. Also, to look into other areas, social insurance for example.

- Why are values missing? The missing mechanism should be modeled. It is important to understanding the reason for missingness. Has missingness been the same over years, or is there a pattern? The missing mechanism can be modeled with machine learning, and the results are used as weights in the imputations (double robust estimator). Another possibility is to introduce variables indicating missingness for the explanatory variables.
- Some additional diagnostics for random forest might be used. Are some results caused by missingness in the prediction variables?
- Have alternative models been tested and compared with machine learning models? Occupation has a different meaning for different age groups. Logit, discriminant, lasso regression, nested logit models are optional models for hierarchical outcomes. An example at CBS was mentioned, where other models turned out to work better.
- Important to realize that there is uncertainty. One category is picked based on probability, but there is also information about other categories. Can the distributional information be used to get at measure of uncertainty? An example at CBS was mentioned.
- Why are imputed data not released?

Need for labour – establishing a new statistical product

Katja Olofsson and Martin Axelson presented the topic and Steven Heeringa introduced the discussion.

Comments from the board

- Several examples of relevant work were given:
    - U. S. Annual Integrated Economic Surveys
    - U.S. Bureau of Labor Statistics: Labor Force Projections and Occupational Outlook Handbook.
    - ABS: merge household expenditure survey and income survey
    - CBS is developing a skills ontology, based on on-line job advertisements.
- A pilot test is strongly recommended
- A modular questionnaire system can be used. Use a core set of questions and modules of questions to cover different needs. Consider a rotating panel sample design.
- It is important but difficult to keep the burden low, need to coordinate as much as possible.
- Combined response burden for the whole statistical system?

- What is most important, long time series or relevant future statistics? Necessary to discuss this with users, including academia.
- Society is changing, and there are new issues, green jobs, migrant workers, etc. Important to look forward. The NSI could assist major users in how the data can be used, data driven decision making, etc.
- A good reference on how to manage time series is Jan van der Brakel on LFS during the pandemic.
- Users always wants better, faster and more detail. Would they be willing to pay for it? Might be a good indicator of the importance of the statistics.
- Difficult to get information from users on what they actually need; reference to an article from JOS special issue on respondent burden Volume 38 (2022): Issue 4 (December 2022) (sciendo.com). The authors looked at how many variables in a survey were used, in downloads, articles etc. 75% of the survey variables were not used.
- Important to have a clear communication with the users on the restrictions, like costs, EU regulations, and if the statistics will be official statistics or not.
- The NSI can take the lead, make a proposal and discus it with the users.
- There will be multiple respondents in large businesses. If management is positive, they will influence lower levels. Ask for information on who will be responding at different levels.
- Prefer local units to legal units, at the local units they will be able to give better and more detailed answers.
- Take a macroeconomic approach to the statistics, calculate the costs of everything and compare it to what will be gained. What can benefit society as a whole? Will reducing costs at some organisations increase the cost of your own or another organisation? Avoid local optimization. A cost-benefit analysis is very important.
- What can be gained with organic (cheap) data? Start with those data and complement with survey data?

ASPIRE

Marie Haldorson presented the results from the 11th round of ASPIRE.

Comments from the board

- Statistical leadership is defined as an internal issue. But the NSI could support other organizations producing statistics and lead the way beyond the NSI. Statistical leadership for society.
- For the methodological strategy, look at EU statistical program for a structure.

- Not only statistical leadership, also data or data science leadership (see the UN statistical division). A larger scope than just statistics. We are better at dealing with data than many others, we can do more.
- Expert users, who are they? “Everyone will be an expert user in the future.” What is a good way to support society when more and more people are able to use data, but don’t know how to do it correctly?
- It is good to focus on change management. Many good decisions are never followed through over time.
- Competence supply? How is that related to the methodology part? Important strategic issue.

Concluding words

Joakim Stymne closed the meeting.