Statistiska centralbyrån
Statistics Sweden

Notes
Apr 19-20 2018

Advisory Scientific Board
Suad Elezović, PMU/MFS
Tiina Orusild, PMU/MIS

Meeting with the Advisory Scientific Board of Statistics Sweden April 19-20, 2018

Board members

Joakim Stymne, Statistics Sweden, chair
Folke Carlsson, Statistics Sweden,
Lilli Japec, Statistics Sweden, co-chair
Tiina Orusild, Statistics Sweden, secretary
Suad Elezović, Statistics Sweden, secretary
Professor Jan Björnstad, Statistics Norway & University of Oslo
Professor Sune Karlsson, Örebro University
Professor Xavier de Luna, Umeå University
Professor Daniel Thorburn, Stockholm University
Professor Thomas Laitila, Statistics Sweden & Örebro University
Professor Natalie Shlomo, University of Manchester
Professor Geert Loosveldt, University of Leuven

Other attendees

Barteld Braaksma, Statistics Netherlands
Mats Bergdahl, Statistics Sweden
Gustaf Strandell, Statistics Sweden
Joakim Malmdin, Statistics Sweden
Eva Elvers, Statistics Sweden

Open session (April 20): Representatives from Swedish public authorities

Day 1

Introduction

The newly appointed Director General to Statistics Sweden, Joakim Stymne opened the meeting by introducing himself. He summarized current issues at Statistics Sweden:

- Vision for Statistics Sweden through the "Strategy 2020" framework based on improving quality, reducing response burden and expanding communications and branding.
- Aiming to increase the coordinating role of Statistics Sweden on the Council of Official Statistics.

- Reduce non-response in surveys, reduce costs of data collection and editing, improve access to data, digitalization, update IT infrastructure, improve production of indicators for Sustainable Development Goals and financing.
- Recent initiatives to move more staff into the Örebro office.

Topic 1: New and alternative data sources in official statistics- Overview and introduction

Speaker: Folke Carlsson
Statistics Sweden meets many challenges at this point in time in the form of higher requirements for new statistics – both nationally and internationally, mostly within the European Union. Another big challenge is the increasing nonresponse within the area of statistics for individuals and households. We experience that it's more and more difficult to make contact with respondents, despite the efforts we make with contact strategies, better telephone numbers, etc. At the same time, we are experiencing rising costs to maintain our IT environment which tends to be more and more complex. Because we don't see a corresponding increase on the income side, our development work requires efficiency gains. Parallel to this, digitalization is increasing in society which provides both opportunities as well as the need for development and adjustment. To manage this, we need to continuously become more efficient and development our operations and activities.
In the strategy that Statistics Sweden has adopted, Strategy 2020, we have set out our task to meet the needs of users for statistics of high quality. We base our statistics on scientific grounds and follow international and Swedish regulations and guidelines for quality. We have said that it should be easy to provide us with the right information. There should be a clear demand among users for our statistics and services.
Against this background, Statistics Sweden has adopted a data collection strategy, which says that we will collect data only when we must. Whenever possible, we will use already collected data or administrative registers. We also need to look into new data sources, regarding existing sources but not yet utilized by Statistics Sweden. This means among other things that we are speaking more of areas and systems of products rather than unique products as a way of circumventing "stove-pipe" solutions.
Statistics Sweden has a long tradition of using administrative data. Already in the 90s, administrative data was a dominant source. Despite this fact, during 2017 we carried out over 200 000 telephone interviews, scanned almost 400 000 paper questionnaires, collected just short of 15 000 files with administrative data and over 800 000 respondents submitted their responses via the web.
At last year's meeting, we discussed the challenges that producers of statistics are facing in general. In this session now, we will present several, more concrete applications where we have worked with what can be called new, or possibly alternative data sources. One can say that last year, we approached the questions more from a general theoretical angle, while we this year have chosen our starting point from concrete applications.

Discussant: Frauke Kreuter

Summary of presentation
Frauke discussed big data issues related to official statistics by presenting some experiences from AAPOR Big Data Task Force, IAB and some other sources. She focused on the role of methodologists and emphasized the need of working in teams with experts from different areas, such as computer (data) scientists, domain experts, SYS ADMIN experts etc. She stressed that methodologists should make use of all available data and learn new non-traditional skills. Privacy and confidentiality will be even more important when facing big data.

Topic 2: Geodata and geospatial applications at Statistics Sweden

Speaker: Jerker Moström

Summary of presentation

Statistics Sweden has a longstanding tradition of geographical applications in the production of official statistics. The first steps towards usage of georeferenced information as a regular component of the statistics portfolio were taken already in 1980s. In late 1980s, real property coordinates together with data from the population register were used to produce the first machine generated population grid of Sweden. Today, GIS and geospatial information, are integral parts of the production chain in many statistical products, especially in the field of land use statistics. The use of geospatial information within Statistics Sweden can be broadly divided into two different categories depending on the purpose of the usage and the properties of the end-use product.

1. Production of geospatial statistics, such as gridded statistics or other small area statistics, where the geospatial statistics itself is released as the end-use product or at least forms an essential part of the result. This category also includes delimitation of localities etc.

2. Production of official statistics where geospatial information and/or geospatial processing is involved at some stage of the production chain but not essentially part of the disseminated result. Stages of production may concern design of surveys, sampling, data collection, processing, analysis and dissemination.

In a broad sense, most statistical products retrieved from administrative records have a geospatial component as many of them rest upon an underlying framework of georeferenced records. However, in terms of production setting, such as data sources used, tools, methods for data processing and analysis, the most "geography intense" field of statistics is land use statistics. As of today, Statistics Sweden is responsible for some twenty official products concerning use of land and water, comprising statistics on land use, land cover, land ownership, protected nature, urban green areas, coastal settlement and development, urban development etc. Practical cases on quality improvement in current production as well as emergence of new statistical application will be described in brief.

A growing usage of geospatial data from a variety of producers also brings about challenges. As the bulk of data from external producers in the production process increases, the harder it is to overview the quality, accuracy and coverage of the data sources involved. Usually metadata typically describing "map products" does not satisfy the needs from a statistical point of view. This is why it is important to invest time to get acquainted with the data producers, to understand the underlying conception of their production as well as the history and original purpose of the information. Recent initiatives within the UN (UN-GGIM) and the EU (ESS), aiming at a better integration between statistics and geospatial data, put an emphasis on the need for a closer cooperation between statistical institutes and mapping agencies. Another challenge is the understanding and recognition of the needs associated with geospatial activities within the statistical offices. Geospatial processing typically demands a technical environment beyond the standards of many statistical offices in terms of software, hardware and storage but also in terms of competence.

Questions to the Board

- What opportunities do you see with using geographic data; for different parts of the statistical processes; new types of statistics, data handling, statistical methods, visualization?
- What are the challenges identified (other than the challenges already mentioned in the paper) and how can we handle them?

- Geospatial and statistical sciences have so far not been very well integrated. Statisticians typically have no or little training in geospatial data processing, and vice versa. How do we make sure to bridge this divide?

Discussant: Frauke Kreuter

Frauke emphasized importance of exploratory spatial data analysis:

- Discover potentially explicable patterns
- Interaction with the data through linked views and dynamic selection
- Examples of classic EDA tools: Boxplots, histograms, conditional plots, parallel coordinate plots
- Extended to ESDA (Spatial) and ESTDA (Space-time)

Challenges other than mentioned:

- Privacy
  - Privacy preserving record linkage
  - Risk with geographic identifiers
- Errors in GIS
  - User error
  - Variability in GPS coordinates
  - Geocoding error

Concerning the question related to the fact that geospatial and statistical sciences are not well integrated, one possible answer is to organize professional training workshops. Frauke referred to the book "Big Data and Social Sciences- A Practical Guie to Methods and Tools".
SCB should invest in combining GIS information with survey data and enhance survey data collection through GIS.

Topic 3: Alternative data sources for the Job Vacancy Survey- On the use of data from job portals for improving the statistics on job openings

Speaker: Ingegerd Jansson and Suad Elezović

Summary of presentation
In 2016, Statistics s Sweden joined the ESSnet Big data, a consortium of 22 countries and organizations for investigating various "Big Data" sources, i e explore new types of data sources that have not yet been used for official statistics production. A part of this work is concerned with the Job Vacancy Survey and the use of related data from external sources. The purpose of the survey is to contribute information about the labour demand but the process is costly and the relevant quantities are difficult to estimate. The response burden could possibly be reduced and quality improved by using alternative data sources as a helping tool in estimating the total number of job openings at a given point in time (month or quarter).

Focus is on online advertising of vacant jobs. There are several possibilities to access such data: to scrape advertisements directly from company web sites, to investigate data from job portals, or to use data collected by third parties.

The legal basis for Statistics Sweden to do web scraping is not clear and we have only carried out tests on a small scale on websites of the public sector. No data has yet been captured using web scraping on private companies. Statistics Sweden thus decided to concentrate on job portals, and in particular data from the portal of the Swedish Employment Agency (SEA), "Platsbanken" (PB). SEA

is a main player on the labour market in Sweden and provided us with data directly from their database. The approach taken is to explore the PB data as a starting point to understand job portal data in general, and to evaluate the quality of the PB data.

Tasks involve quality check and cleaning of data (valid data, data standardization, deduplication, etc.), linking to the Business Register for additional variables, classification, comparisons with the Job Vacancy Survey, and first analysis of data. Some methodological challenges we face are to find suitable indicators of data quality, to determine the coverage of sources, and to solve issues of estimation and inference.

Questions to the board

- Does the Board have any suggestion how to evaluate quality of the data from job portals in general and from PB in particular? Are there other external data sources or statistics that can be used for evaluation?
- In particular, how can the coverage of PB be evaluated?
- How can we treat the fact that the time series from two sources have significant differences with respect to levels? Is it advisable to focus (solely) on the time series properties, such as common growth rates, common trends and common seasonal variation and neglect the fact that the levels differ?
- Is it at all meaningful to use the PB data together with the JV data in a modelling framework, to improve estimates of the number of job openings per quarter? This question is related to the previous questions since we do not know enough about the quality of the PB data but we may conclude that the properties over time look similar for these two data sets.
- Is it advisable to use time series models to make forecasts for the number of job openings per quarter? Here we think about the possibility to thin out the total number of surveys during a year in order to reduce costs and use forecasts instead.
- Would it be meaningful to study the data at a more detailed level, e.g. by the main NACE sections or by some regional division? This issue is addressed here since there are large differences in quality between sections.
- Does the Board have any other suggestion concerning potential use of the data from job portals in general and modelling alternatives in particular?
- Would it be reasonable to perform some kind of outlier correction for those numbers in PB data where we may suspect over-coverage? If yes, does the Board have any suggestion how this may be done in practice?

Discussant: Barteld Braaksma

First of all, it is very important to invest in this topic and these sources! There is much policy demand for information on job openings and the labour market as a whole and the new sources studied here may be able to answer questions that traditional statistics cannot. In addition, if we don't do this as official statisticians, others will increasingly fill that gap, possibly with lower quality standards that may impact policy decisions.

The paper mentions that the legal framework for collecting data from online sources is not entirely clear. To avoid future issues it is important to clarify this legal situation, both at EU and national level.

It is a good idea to focus on existing job portals and use their data, instead of developing own webcrawlers at SCB. Building webcrawlers is a time-consuming activity that requires a lot of maintenance because job portals and other websites

change often. It may still be useful to do limited webcrawling as an NSI, however, to understand opportunities and limitations of these techniques.
Not all job openings are advertised online (an earlier Dutch study found for example an education bias), and the online world is very dynamic and has specific characteristics. Some vacancies may for example be advertised on LinkedIn only, or in novel ways that have yet to be developed. Therefore it is important to keep track of what is happening and determine whether developments reflect changes in recruiting or real developments in the labour market. From that perspective it is also important to establish a good constructive dialogue with data providers to understand the sources they control and possible quality issues. They might even be willing to harmonise and standardise the portals, which would improve their usefulness for statistics.
When reading the paper it was not fully clear what is the goal of the research project: to reduce costs for SCB, reduce burden for survey respondents, improve quality and timeliness of existing statistics or to create new statistics. In the initial research phase the goal can be left open but soon it becomes important to make choices. In making choices it might help to involve users to determine what their needs are. The nature of online information might even require changes in definitions. Levels, trends and structures can usually not be addressed all at the same time with equal quality.
The job portals data may provide many potential uses other than replacing existing statistics. Information on skills or education, for example, seems in heavy demand from users. A focus on specific NACE groups or geographical regions could help both to increase manageability of the data and to satisfy specific user needs. Job portals might give additional insights on the international labour market, possibly in relation with social media like LinkedIn. Reading vacancy texts might give clues on potential uses based on the typical content provided.
This topic is very relevant for international collaboration and is pleased that Statistics Sweden is participating in the ESSnet on Big Data.
Responses to questions to the Board:

1. Quality evaluation of job portal-based statistics and detection of coverage issues will to a large extent rely on common sense. There is no existing theory or specific quality standard that can be applied. It seems best to focus on two approaches: 1) test methods on relatively well-behaved small groups and specific cases, and 2) check plausibility with (external) subject matter specialists and users. The latter may also give insight into 'fitness-for-purpose' from users' perspective: when are results good enough? A related issue is outlier correction. This is always difficult, and often more an art than a science. It depends on good knowledge of the underlying data set. Outliers in a 'big data' source are less influential than in a sample-based survey so it may be wise to be cautious in treating them.
2. The paper identifies serious differences in time series levels between traditional surveys and online portals. These can either be studied in depth or taken for granted: looking at trends makes sense! Recent Dutch studies show that consumer confidence from a traditional monthly survey and a monthly sentiment index constructed from social media follow remarkably similar patterns although there are many differences in populations, methodologies and underlying concepts- in this case it is even unclear what a 'level' would mean. It is advisable to work with time series and small area estimates specialists to make progress.
3. When considering time series-based *forecasts*, the first question is whether the aim is to predict the future (forecasting), or estimate the now or recent

past using projection methods (nowcasting). Obviously, the application of sophisticated methods here is only advisable if the quality of input data is sufficient from a users' perspective, but in principle there is no objection. As mentioned above, the sources considered and the job market are not static, so it is very important to keep looking at the validity of the results.

4. The best way forward in systematic studies of job portals and other related data sources is to develop a modelling framework. All available sources, both classical and big data-type, contribute only part of the statistical puzzle. A suitable framework may integrate all of these pieces and thus deals with the increasing user demand for integrated and coherent statistics. It also helps to reduce dependencies from specific sources, which enables the production of stable outputs from increasingly unstable inputs. When developing a framework it may be helpful to learn from National accountants who have a lot of experience in this area, both at conceptual and operational level. A framework may also be the best way to produce statistics at a more detailed level. Again, the first questions here are what the users want and if the quality can be guaranteed.

Topic 4: Electricity consumption data from smart meters

Speaker: Ingegerd Jansson

Summary of presentation
In 2016, Statistics s Sweden joined the ESSnet Big data, a consortium of 22 countries and organizations for investigating various "Big Data" sources, i e explore new types of data sources that have not yet been used for official statistics production.

Part of this work is concerned with the use of electricity consumption data from smart meters, i e data read from a distance and measuring electricity consumption at a high frequency. The data can possibly be used not only for estimating consumption (and some production), but also for environmental statistics, building and household statistics, etc.

A Swedish data hub for the smart meter data is under development and will be fully developed and in place by the fourth quarter of 2020.. Through the hub, metering data, measuring data, customer data, and contract data, will be managed in one common system. It is in the interest of both Stat Sweden and the grid owners that the hub is able to deliver high quality information for statistical purposes.

Statistics Sweden has recently received a small "hub like" test data set that is being analyzed in the ESSnet Big Data, in order to prepare for future use of data from the hub. With the data set, we are investigating data quality and testing methods for linking data to available register and estimating relevant outputs.

Questions to the board

- How can linking to the registers be refined, in particular, what methods would be useful for linking using coordinates?
- How can models be used to handle the discrepancy between metering points and the object for which we want to produce statistics (households, businesses, local units, buildings, etc.).
- How can quality in general be assessed for smart meter data?

Discussant: Barteld Braaksma

Just as in the case of job portals, it is of strategic importance to build up knowledge in this area. There is much policy demand for energy-related questions, in view of sustainability and the energy transition, emerging market models ('prosumers') and eliciting behavioural changes from energy consumers. There are many players in the energy market and potential users of energy-related statistics like central and local government, energy companies, academics, citizens: talk with them! It is important to include legal considerations when starting to investigate smart meter data. Are there any privacy issues? Who owns the data? Who may access the data?
The paper indicates that SCB has built up constructive relations with providers of smart meter data. It is very important to maintain these relations, use them to better understand peculiarities of the data and, where possible, promote standardisation of the now heterogeneous smart meter data sets.
Responses to questions to the Board:

1. The question how linking between smart meter data and energy users can be refined is difficult to answer generically, because it depends largely on specific knowledge of the data source and local (geographical) situation. It may be useful to consider machine learning techniques or enter into a further dialogue with the data providers, maybe even suggest to them that additional statistically relevant information is built directly into the source where possible. But an important question is whether detailed linking is necessary at all. For traditional survey-based statistics that is the natural way to go, but in this case other approaches might be better. For example, linking at neighbourhood level instead of enterprise level might be feasible, and sufficient for many statistical outputs.
2. It could be useful to consider a modelling approach by developing a coherent framework for energy statistics and beyond since energy affects a lot of different aspects of society (housing, mobility, economy, sustainability, …). Such a framework also serves to assess quality in a systematic way. The framework could include conservation laws on production vs. consumption of energy. Development of a framework could probably best be taken up in an international setting.

Day 2

Welcome

Open Session

Session open to all staff at Statistics Sweden and some other public authorities

1. **Main talk**: Barteld Braaksma discussed.
   Summary of presentation
   The emergence of all kinds of new data sources, often referred to as big data or the data revolution, has a huge impact on the

role of official statistics in society. Statistics Netherlands (CBS) tries to approach the data revolution as an opportunity rather than a threat by looking for new sources, methods and tools on the one hand, and new users, products and services on the other hand. Innovative collaboration models and liaisons with all kinds of partners, both public and private, are crucial to making progress. The seminar discusses the Center for Big Data Statistics and the Urban Data Centers established by CBS, next to instruments such as data camps and the innovation site that shows beta products.

Topic 5: Improving editing at Statistics Sweden

Speaker: Lilli Japec

Summary of presentation

Background

- Statistics Sweden started a project in September 2017 with the objective to reduce costs for editing in products
- Review and improve the data collection and editing process in 53 products
- There have been a number of initiatives at Statistics Sweden over the past 20 years to improve editing and reduce its costs
- From the review 2017: Statistics Sweden spends approximately 76 million SEK on editing per year (on average a product spends about 38 percent of its budget on editing)
- Six percent of the products regularly analyze the effects of editing on statistical estimates.
- 25 percent of the products do not use any weights in production editing
- There have not been any major improvements in the editing process since the previous review in 2004.

Editing in the Occupation Register (OR)

- Occupation is an important variable in the population census
- In order to make a register-based population census we needed to get information about occupation
- It's compulsory for businesses and enterprises to provide information on their employees' occupations.
- The OR consists of 20 different sources with occupation codes on a detailed level (four digit level). Each source has its own method and process to collect and edit information on occupation.
- The data from the OR is available in the Census Hub (Eurostat) and Statistics Sweden's statistical databases
- Data from the OR is also used by researchers to e.g., study the relationship between occupations and diseases and injuries.

Two sources

1. The Statistics Sweden Survey (SSS). Each year Statistics Sweden collects data on occupation (web and paper questionnaires) for a sample of approximately 47 000 small businesses and organizations (1-19 employees).

On a yearly basis this source comprises about 3-4 percent of the persons in the OR (a rotating scheme is applied so that all small businesses and organizations are covered over a four to five-year period).

1. Statistics Sweden also gets information about occupations from the Confederation of Swedish Enterprises (Svenskt Näringsliv). The information comes from businesses and organizations that are not part of the Structural Salary Survey, but which are part of the confederation's collection (we call this the "SLP-rest").

The cost for data collection and editing for 1. and editing for 2. was 5,7 milion SEK in 2017

Questions to the board

Statistics Sweden invites the Scientific Board to provide input to described problems in this paper. In particular Statistics Sweden would like to get input on the following.

A.) Over the years users have been accustomed to get data on a very detailed level and we have relied on their capability to assess the usefulness of the data. We know that users have difficulties doing that and often we do not even provide them with all the information necessary to do so. How should we deal with this problem?

B.) How should we design a recurring evaluation study that could be used to adjust occupation estimates and to provide information to users about the quality of OR? There are a number of methods suggested in the literature which aim at studying the effects of editing. Any advice on preferred methods is appreciated.

A relatively new approach would be to use probabilistic sampling to estimate measurement errors as proposed by Ilves, M. and T. Laitila. 2009. "Probability-Sampling Approach to Editing." *Austrian Journal of Statistics* 38: 171-182. With this method only a small proportion of flagged units from selective editing, selected with known probabilities, should be reviewed. This method has been expanded in Laitila, T., Lindgren, K., Norberg, A. & Tongur, C. (2017) "Quantifying Measurement Errors in Partially Edited Business Survey Data" in the monograph "Total Survey Error in Practice" to model estimation. We see the benefits of this method when there are a limited number of domains of study. Can we use models to estimate bias when we have more domains than the number of reviewed units?

C.) When implementing selective editing we use what we call Relative pseudobias, RPB, defined by Lawrence and McDavitt (1994) as $RPB_{d,j,Q} = T_{d,j,Q} - T_{d,j,Q=100} S(T_{d,j,Q=100})$ where d is domain and j is variable. Q is the proportion of units flagged.

Furthermore, we refer to Särndal, Swensson and Wretman (Model Assisted Survey Sampling, page 165) when we assert that if RPB is less than 30%, a 95% confidence interval has a coverage ratio of 93.96%, which is OK. We therefore reduce the editing, accept nonsampling errors but get a decent level of coverage ratio for the confidence intervals if we make sure that most subgroups have an RPB <30% (we usually require 20% for most and allow 50% for some).
However, when we have total surveys and registers, such as OR and Foreign Trade with goods, we have no variances. In those cases how should we decide the extent of editing? Even more problematic is the fact that some users e.g., Eurostat want us to publish statistics for a huge amount of domains leading relative measurement errors can be unimaginatively large. (In the Foreign Trade with goods each month SCB publish several 10 000 domains. How can we estimate quality for these domains when we flag only 800 units?).
D.) What is the role of editing in a changing survey landscape characterized by multiple data sources including big data, nonprobability sampling, and new data collection methods? Speculations are welcome.

Discussant: Jan Björnstad

Statistics Sweden started an editing project in September 2017 with the aim of reducing costs. The project shall review and improve the editing process and the data collection in 53 products over the next few years. In 2004 these products as well as nine other products were also reviewed. It was found that on average 33 percent of their budgets was spent on editing. In 2017 the amount spent on editing for the 53 products now under scrutiny had increased to 38 percent. The increase could in part be due to the lower cost of data collection, but still is surprising since selective editing has been implemented in at least 11 of these products, using the program SELEKT.

For this meeting at the Scientific Advisory Board, three main topics were raised (A) publication of detailed statistics;(B) editing for registers of categorical variables; (C) using registers to produce detailed statistics for many more domains than can be covered by the editing; and (D) the role of editing in the future with big data and multiple data sources.

Statistics Sweden also asked what to do in the case of selective editing when no variance estimate exists. The suggestion is to use the global points and stop editing for large deviations when the global points flatten out.

Discussion point A. How to deal with publishing detailed statistics

The problem is essentially the same as with all published statistics, namely the level of uncertainty. A measure of uncertainty should be published, such as standard error or the 95% confidence interval. If the uncertainty is regarded as too high do not publish. For example, the coefficient of variation is often used in NSIs as a measure to determine precision and whether an estimate can be published. At the very least, when a detailed statistic is published, always produce interval estimates. This will show if there is any useful information in the detailed statistic.

Discussion point B. Evaluation study of the Swedish Occupation Register (OR)

Statistics Sweden asks if it is possible to first use selective editing in the OR and then edit units in a randomly chosen sample, especially when we have many more domains than the number of reviewed units.

When considering editing in OR it is essential to first consider the uses of OR:

- Provide annual distribution of the working population into different occupational groups for following development of occupations in different sectors and making forecasts for different occupations.
- For research, for example studying the relationship between occupation and diseases and injuries.

In OR there are many small errors and no large errors, so selective editing is not relevant here. One can estimate errors in tables of occupational distributions by editing a sample of businesses or persons. A suggested editing method carried out within the framework of an evaluation study:

1. Micro-editing of a probabilistic sample stratified by main occupational groups;
2. Some form of output- editing for annual distribution of occupational groups at the first digit level
3. Estimate this distribution based on the probabilistic sample with standard errors and prediction intervals.
4. When making micro data available to users, provide them with the OR and the results of the evaluation study.

In general it is better to use editing resources on recurring evaluation studies than the current editing of the Statistics Sweden Survey (SSS).

Discussion point C. Estimating for many domains based on registers

We consider business registers, where we have too many domains compared to the number of edited units. For OR similar considerations apply.

The main recommendation can be summarized as follows:

Step 1: Selective editing in the register to take care of the largest deviations from anticipated predicted values for the whole population. This is important in all cases, and especially for domain estimates.

Step 2: Define the domain statistics as values of random variables and do some form of hierarchical modeling dependent on auxiliary variables such as number of employees and annual sales from the previous year.

Discussion point D. Role of editing in the future with big data and multiple data sources

1. Big data and multiple data sources

Automatic editing would be important in the case of big data, including some sort of selective editing. Otherwise, if relevant and if it is possible, an evaluation study should be carried out based on probabilistic sample. We refer to the article ‘Finding Errors in Big Data’ by M. Putts, P. Daas and T. de Waal in Significance magazine, Volume 12, Issue 3 for some new research in this area coming out of Statistics Netherlands.

2. Non-probability sampling

There is no difference and the same editing processes should be applied as with probability sampling.

Specifically for OR, there are many sources and one can try to evaluate which sources constitute the highest risk for errors and concentrate editing on these sources.

After the general discussion, Joakim Stymne closed the meeting by thanking everyone for participating.

Closed session for the Board members

- Discussion and advice to Statistics Sweden.