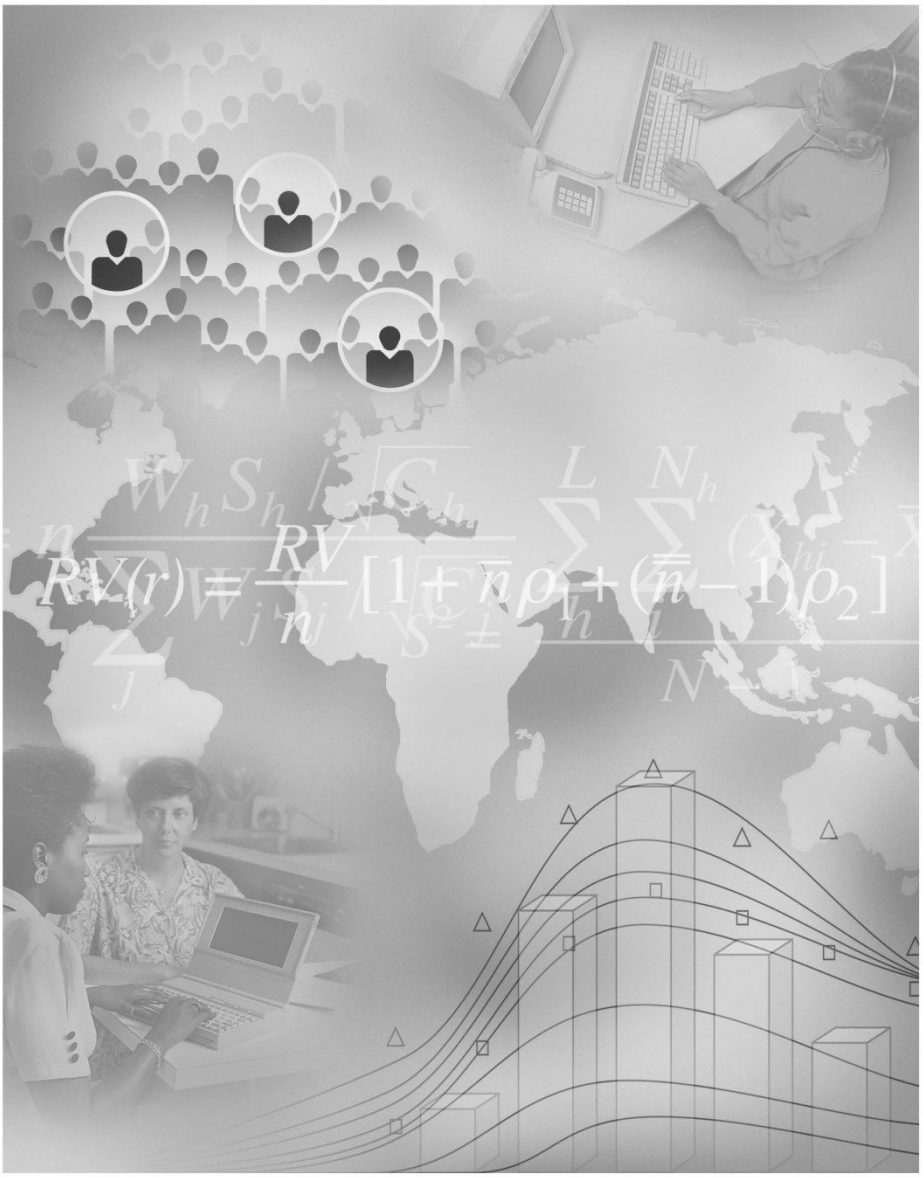
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

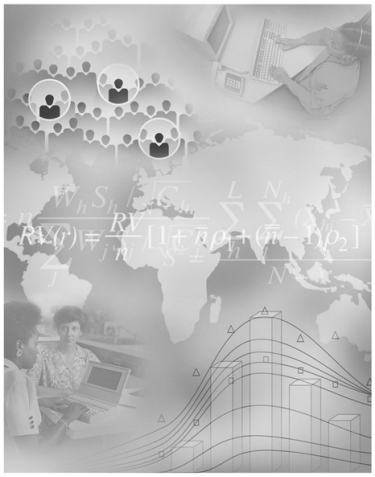
No. 81 **January 2020**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

the Survey Statistician
The Newsletter of the International Association of Survey Statisticians



The Survey Statistician No. 81, January 2020

The Survey Statistician is published twice a year by the International Association of Survey Statisticians and distributed to all its members. The Survey Statistician is also available on the IASS website at
http://isi-iass.org/home/services/the-survey-statistician/

Enquiries for membership in the Association or change of address for current members should be addressed to:

IASS Secretariat Membership Officer
Margaret de Ruiter-Molloy
International Statistical Institute
P.O. Box 24070, 2490 AB the Hague
The Netherlands

Comments on the contents or suggestions for articles in the Survey Statistician should be sent via e-mail to the editors:
Danutė Krapavickaitė (danute.krapavickaite@vgtu.lt) or Eric Rancourt (eric.rancourt@canada.ca).

ISSN 2521-991X

In this Issue

Letter from the Editors

Dear Readers, we would like to wish you health and success in the New Year 2020 and also wish stability in the world: in the behaviour of people and in surrounding environment, because it helps to reach better accuracy in survey statistics, which in turn further helps society.

We would like to express our many thanks to former IASS President Peter Lynn for his support to *The Survey Statistician*. The current President of IASS Denise Britz do Nascimento Silva addresses the readers with her first Letter from the President in this issue. We wish her successful work in this new position.

The IASS Scientific Secretary James Chipperfield introduces himself in the Report from the Scientific Secretary. He brushes up on the conferences that were supported by the IASS in 2019 and those forthcoming in 2020. Preparation to the 63rd World Statistics Congress that will be held on July 11-15, 2021, in the Netherlands is announced. The *News and Announcements* section also contains experiences of statisticians from the 2019 conferences.

Nowadays methodologists in official statistics meet new data sources and algorithms to deal with them. Techniques such as machine learning and artificial intelligence have changed the nature of methodology for data integration and present new research challenges in this field. Two papers in the *New and Emerging Methods* section and one paper in the *Ask the Expert* section address this topic.

The authors of the paper in the *Book and Software Review* section introduce a checklist to evaluate the documentation of public use survey data files, and present an example of such an assessment. The Country Reports section presents a variety of statistical activities across a number of countries.

We would like to acknowledge everyone working hard to put *The Survey Statistician* together; in particular Mārtiņš Liberts from the Central Statistical Bureau of Latvia for preparing the tables of contents in the *In Other Journals* section and making the layout of the newsletter; Margaret A. de Ruiter-Molloy from Statistics Netherlands; Melini Cooper at the Australian Bureau of Statistics; and Olivier Dupriez from the World Bank for their invaluable assistance.

Please let James Chipperfield (james.chipperfield@abs.gov.au) know if you would like to contribute to the *New and Emerging Methods* section in the future. Also, if you have any questions which you would like to see answered by an expert, please send them to Eric Rancourt from Statistics Canada (eric.rancourt@canada.ca). If you are interested in writing a book or software review or suggesting a source to be reviewed, please get in touch with Danutė Krapavickaitė of the Vilnius Gediminas Technical University, Lithuania (danute.krapavickaite@vgtu.lt). The country reports should be sent to Peter Wright of Statistics Canada (peter.wright2@canada.ca).

If you have any information about conferences, events or just ideas you would like to share with other statisticians – please do go ahead and contact any member of the editorial board of the newsletter.

The Survey Statistician is available for downloading from the IASS website at http://isi-iass.org/home/services/the-survey-statistician/.

Danutė Krapavickaitė (danute.krapavickaite@vgtu.lt)

Eric Rancourt (eric.rancourt@canada.ca)

Letter from the President

Dear colleagues

It is a pleasure and an honour to write to you as President of IASS and I am very motivated to contribute to IASS activities and its progress. I welcome our newly elected Executive Committee: Monica Pratesi (president-elect), and vice-presidents, Isabel Molina, James Chipperfield, Lucia Barroso and Nadia Lkhoulf.

I would like to extend my appreciation and sincere thanks to all candidates who agreed to stand for election and to the Nominating Committee: Daniella Cocchi (chair), Christine Bycroft, David Haziza, Jairo Arrow, Paul Smith and Veronica Beritich. Also, I would like to thank Peter Lynn as former IASS President and IASS VPs during 2017-2019, Anders Holmberg, Cynthia Clark, Jean Opsomer and Risto Lehtonen, who kindly agreed to keep their roles as IASS EC until the election process was completed.

In order to fulfil its mission and support survey statisticians, IASS has been actively contributing to conferences and workshops, and has once more awarded the Cochran-Hansen prize to a young statistician. James Chipperfield, our new Scientific Secretary, presents a summary in his report.

We are liaising with ISI and other ISI Associations to enhance collaboration and to participate in workgroups focussed on ISI financial sustainability and on activities related to data science, women in statistics and capacity building. We will keep you posted about these actions.

I would like to take this opportunity to pay my tribute to IASS member Prof. T.M.F. Smith who passed away recently. Fred, as he was known by many of us, made a remarkable contribution to survey sampling and survey methods as well to the statistical society in general. He influenced many lives in a wonderful and inspiring way. Certainly mine. As my PhD supervisor, he changed my life and taught me so much. Not only Statistics, for sure. His knowledge, generosity and friendship made me a better person, researcher and lecturer. I will never forget him.

Let me also highlight that TSS is only possible due to the valuable work of the editors and those responsible for its production and circulation, as well as a team of contributors. Please have a look at page 2 and join me in a big thank you to all. The same happens to our website, maintained by Olivier Dupriez. If you are willing to contribute to TSS or as IASS webmaster, do not hesitate to get in touch.

To secure IASS relevance in the information society environment and to further IASS success, we need to work together to promote IASS goals and to foster new developments in the area. As a community, I am sure we can do it. Each of us can contribute as an IASS ambassador. Please let the us know your views, suggestions and ideas.

As this is the January edition, I wish you all a year filled with peace, health and harmony plus, of course, many productive and interesting activities. Let's reaffirm our wishes for a better world and give our contribution as survey statisticians.

Denise Silva (denisebritz@gmail.com)

Report from the Scientific Secretary

It is my pleasure to take over as Scientific Secretary from Risto Lehtonen, who has been generous with his time during this transition period. By way of introduction, my entire professional career has been spent in the Methodology Division at the Australian Bureau of Statistics. I have worked on many aspects of surveys, from sample design, data collection, confidentiality, analysis and quality assurance of survey data that have been linked with error. Recently I decided to give back to the statistical community that has given me so much. This led me to join the IASS as well as to accept positions such as the Scientific Secretary and Associate Editor for Survey Methodology and the Journal of Official Statistics. I encourage you to share your views on the functions of the Scientific Secretary with me.

IASS Support for Conferences in 2020. Promoting and supporting scientific conferences and workshops has been one of the key activities of the IASS. The IASS published a Call for Requests for Support for Workshops and Conferences to be held during 2020. The IASS Executive Committee received five requests by the November 2019 deadline and decided that all met the criteria as set out in the Request. The IASS-supported events are listed below:

- **Summer School on Survey Statistics**, 24-28 August 2020, Minsk, Belarus. The Organisers include the Baltic-Nordic-Ukrainian Network on Survey Statistics (BNU network) in collaboration with Belarusian State Economic University (BSEU), School of Business of Belarusian State University (BSU), and the National Academy of Sciences of the Republic of Belarus. The key aims include the promotion of survey statistics by sharing and exchange of knowledge and experience of between teachers, students, researchers and practitioners. The funding will cover travel and registration fees of participants from Ukraine and Belarus. Details will be released soon on the conference website: https://wiki.helsinki.fi/display/BNU/Events.

- **11th International Francophone Conference on Surveys**, 13-16 October 2020, Université libre de Bruxelles. Organiser is the: Société Française de Statistique – SFdS. Key aim is to review the state of practice and research in sample survey methodology. Notably, the conference provides scholarships for young students from developing countries to attend and present at the conference. The IASS funding will cover general conference costs. Conference website: http://sondages2020.sciencesconf.org.

- **International Conference on Establishment Surveys (ICES) VI**, 15-20 June 2020, New Orleans, Louisiana, USA. Organiser is the American Statistical Association. Key aims include to "highlight new, improved, and upcoming establishment statistics methodologies and technologies using census data, administrative or other organic data, and sample survey data". The IASS funds provided to ICES were to support travel costs of students attending the conference. Conference website: https://ww2.amstat.org/meetings/ices/2020/.

- **Small Area Estimation Conference – SAE2020 – BigSmall**. 06-08 July 2020, Centro Congressi Federico II, Naples, Italy. Organiser is the Department of Economics & Management of the University of Pisa. Key aim is to assess the current development and usage of small area estimation methods. The IASS funding will cover general conference costs. https://sae2020.org/.

- **Population Data for Informed National Planning and Development**, February 2020, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria. Organisers include the Federal University of Agriculture and the Royal Statistical Society Nigeria Local Group. The key aim is to bring together Census stakeholders (from data collection to data analysts) and to underscore the role of statistics in delivering on the objectives of a Census. The IASS support fully funded the conference.

IASS Support to Conferences in 2019. In line with its intentions outlined in the January 2018 edition of The Survey Statistician, the Executive Committee provided financial support to the following four events in 2019:

- The 6th Italian Conference on Survey Methodology. The key aim was to promote methodological and applied research in survey sampling, in human as well as natural sciences. Conference website: http://meetings3.sis-statistica.org/index.php/ITACOSM2019/ITACOSM2019.
- The 5th Baltic-Nordic Conference on Survey Statistics was a scientific conference presenting development on theory, methodology and applications of survey statistics. Conference website: https://www.oru.se/hh/banocoss2019.
- Sixth biennial European Establishment Statistics Workshop was hosted by EUSTAT with key aim of using modern technology for improving establishment statistics. Conference website: https://statswiki.unece.org/display/ENBES/EESW19.
- Survey Process Design Workshop was organized as a one-day event in February 2019 at the Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.

World Statistical Conference 2019. The IASS Executive Committee organised two Special Invited Presentation Sessions (SIPS) and three Invited Paper Sessions at the World Statistical Conference. The first SIPS, the IASS President's invited lecture, was entitled "A thing of the past? Household surveys in the new global data ecosystem" and presented by Gero Carletto (World Bank). The second SIPS was made up of Dr. Diego Andres Perez Ruiz (University of Manchester) presenting his Cochran-Hansen prize winning paper and Frauke Kreuter (University of Maryland & University of Mannheim) presenting a talk entitled "The international program in survey and data science: an environment for training and cooperation."

The Conference on Current Trends in Survey Statistics held from 13-16th August at the National University of Singapore was a satellite conference to the World Statistical Conference 2019 and was endorsed by the IASS. The satellite conference was well attended by IASS members where Malay Gosh received the 2019 Award for Outstanding Contribution to Small Area Estimation. Previous winners were J. N. K. Rao and Danny Pfeffermann.

Cochran-Hansen prize session/winner. As mentioned in the July 2019 issue of The Survey Statistician (TSS), Dr. Diego Andres Perez Ruiz was awarded the Cochran-Hansen-Prize 2019 for his contribution to the paper "Augmenting Probability-Based Surveys with Nonprobability Survey Information: a Bayesian Approach" that was published in the Journal of Official Statistics in 2019 https://content.sciendo.com/view/journals/jos/35/3/article-p653.xml. It is a paper very relevant to the times, which is awash with so-called Big Data. If you would like to see a picture of the winner with the incoming and outgoing IASS Presidents see the 9th September tweet: https://twitter.com/iass_isi/status/1171062017836429312.

Preparation for 2021 WSC. The next International Statistical Institute's World Statistical Congress in 2021 is just around the corner. Soon the EC will start organising two IASS special invited paper sessions, other IASS-related invited paper sessions, and the short course scheme. You will no doubt hear more on this in the coming months.

Your contribution. *The Survey Statistician* publishes methodological articles on developments in survey statistics in its section called *The New and Emerging Methods*. The format of the article is at most 8-10 pages and should cover the presenting challenge, the methods and their application, and the relevance to the development of survey methods. Please contact me if you are interested in writing such an article or would like to advertise a conference, workshop, or course that would be of interest to IASS members. I look forward to your contributions.

James Chipperfield (james.chipperfield@abs.gov.au)

News and Announcements

My experience attending the 62nd ISI World Statistics Congress 2019

August 18 to 23, 2019. Kuala Lumpur Convention Centre, Kuala Lumpur, Malaysia

The Cochran-Hanse Prize 2019

Diego Andrés Pérez Ruiz

Manchester University, UK, diego.perezruiz@manchester.ac.uk

To attend the 62nd ISI World Statistics Congress 2019, I was awarded with the prestigious Cochran-Hansen Prize in survey methodology. The prize was first stablished in 1999 and is given every two years for the best paper in survey research methods submitted by a young statistician from a developing or transition county. The prize, originally established by the International Statistical Institute (ISI), and now awarded by the International Association of Survey Statistician (IASS), consists of a travel scholarship, books and journal subscriptions, recognising the opportunity to attend international meetings and present the work, putting a special interest in the early career development of young statisticians.

To be considering for this prize, I submitted the paper "Augmenting Probability-Based Surveys with Nonprobability Survey Information: a Bayesian Approach". The paper was a joint collaboration with a team of different experts from different fields and different universities from the United Kingdom and Germany. The paper, published in the *Journal of Official Statistics*, is co-authored with Arkadiusz Wiśniowski, a Senior Lecturer in the School of Social Statistics at the University of Manchester, Prof. Joseph W. Sakshaug, a distinguished researcher, head of the Data Collection and Data Integration Unit, and acting head of the Statistical Methods Research Department (KEM) at the Institute for Employment Research (IAB) and a professor at the University of Mannheim, and Prof. Annelies Blom who is Professor in the School of Social Sciences at the University of Mannheim. Our team worked remotely with occasional visits to the IAB, the University of Mannheim, and the University of Manchester.

The paper proposes a novel idea to combine probability and non-probability surveys by supplementing small probability samples with nonprobability samples using Bayesian inference and is part of an ongoing research agenda. The paper can be accessed using the following link: https://content.sciendo.com/view/journals/jos/35/3/article-p653.xml.

I first read about the Cochran-Hansen prize on the International Association of Survey Statistician web page (in the award section). Since the first moment I read, it caught my attention as it names two of my favourite statisticians with presence in survey methodology: William Gemmel Cochran a well-known statistician with contributions in the area of sample techniques and experimental design, and Morris Howard Hansen with many significant contributions in surveys and censuses.

Getting the prize is not an easy task. The list of previous winners includes researchers from many different countries, including Estonia, India, Brazil, Philippines and Mexico, and all the winners have

different backgrounds and publication records. So I prepared my application and after reading with detail all the necessary documentation I started preparing the submission packet and addressed the submission to the chair of the IASS 2019 Cochran-Hansen Prize Committee, Dr. Anders Holmberg, who sent the paper for review to the other members of the Cochran-Hansen Prize Committee appointed by the IASS. The submission packet included a cover letter and a copy of the paper. I submitted my application and kept my fingers crossed.

Some weeks after my submission, I received an email notifying me that I was awarded the 2019 Cochran-Hansen Prize. I was so happy that I couldn't believe at first. The first thing I did was communicate this news with my family and co-authors who shared my happiness. Once I had made the necessary arrangements to attend the ISI World Statistics Congress, I booked my tickets to Kuala Lumpur and waited anxiously for the trip.

Arriving to Kuala Lumpur was a long journey; it involves several hours of flight and a stop in Dubai. Kuala Lumpur is a city that surprises me. It is a vibrant city with tall buildings and busy streets. I stayed in a hotel near Kuala Lumpur Convention Center (KLCC), where the ISI World Statistics Congress was taking place. Having arrived safely to Kuala Lumpur I was ready to start the congress.

One of the first activities I did was attend a short course before the opening of the ISI World Statistics Congress. I registered for a short course on imputation methods for the treatment of item nonresponse in surveys. The short courses were in a different location, they were in the Sasana Kijang of Bank Negara Malaysia in Kuala Lumpur, these courses are organised by the ISI in cooperation with its associations. The instructor of the course was Prof David Haziza. David is an excellent researcher and instructor. During these two days, we reviewed different techniques to deal with missing data. I really enjoyed the course and the way David introduced some concepts about missing data and techniques to deal with inference in the presence of missing data and resampling methods and calibration.

After the short course ended, I waited for the opening. The opening of the ISI World Statistics Congress was exceptional, it took place on Sunday 18th August in a full Plenary Hall in Kuala Lumpur Convention Center. The opening was officiated by the Prime Minister of Malaysia, Tun Dr. Mahathir bin Mohamad, following the opening speech of the ISI President Helen MacGillivray. It is worth mentioning that this year's congress had a special focus on the women who are at the forefront of statistics and data science. It was a proud moment in the history of the World Statistic Congress to have 5 women presidents from top associations presenting papers on leading vital developments.

The opening was followed by the official presentation of the International Prize of Statistics. On this occasion, the winner was Bradley Efron, professor of statistics and biomedical at Stanford University and author of the bootstrap, a statistical method with high impact across many scientific fields. Prof Efron participated in the ceremony through a video message, while ISI President Helen MacGillivray received the prize certificate on his behalf.

During the congress I had time to listen to different talks on different topics. There were talks in Bayesian statistics, Big Data Analytics, Statistical and Data Literacy, Recent developments in functional data and discrete models. Some other talks were on Financial time series modelling, Official Statistics, and Modelling demographic data. I attended several talks related to my research interests and afterwards I prepared for my talk.

When the day arrived I was nervous at the beginning. The talk was schedule to begin at 10.30am. The organiser of the session was Prof Risto Letonen and the Chair was Prof. Peter Lynn. It also was a shared session with Frauke Kreuter from the University of Maryland, USA and the University of Mannheim, Germany. She talked about the international program in survey and data science: an environment for training and cooperation.

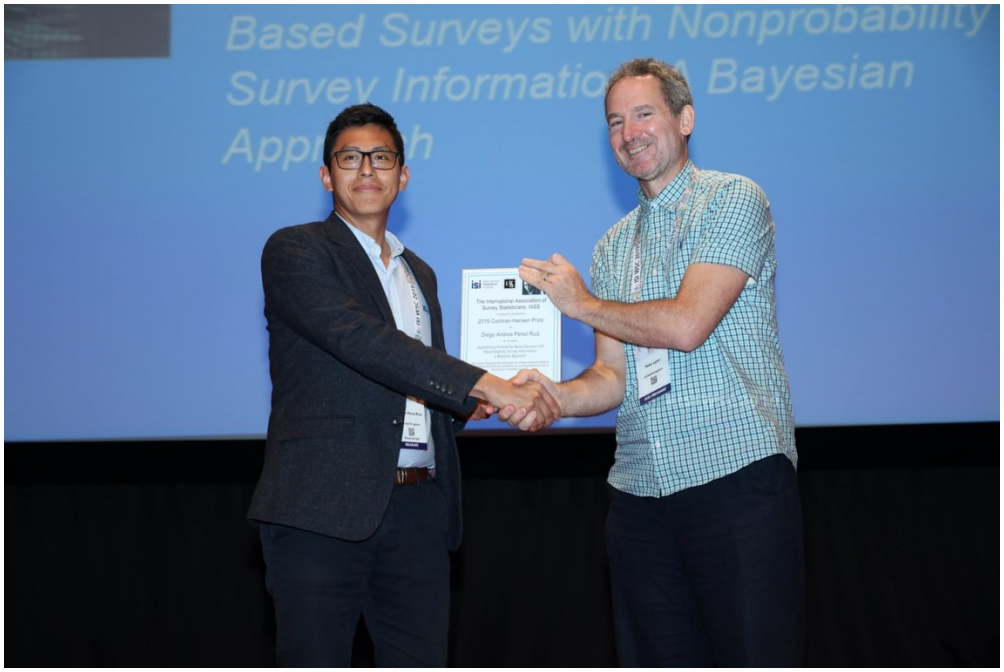


Figure 1: *Dr. Diego Andres Perez Ruiz accepted the Prize from the former president Prof. Peter Lynn in the closing ceremony of the World Statistics Congress in Kuala Lumpur, Malaysia*

During my talk I introduced some of my research aims and I explained and showed the methodology and the results of our application. In addition, I show how our proposed method works with real and simulated data. After my talk I listened to the talk by Frauke Kreuter, where she discussed some results about the international program in survey and data science. At the end of both talks, we listened to the discussion and had time to take some questions from the audience. The discussant of the session was Dr Pedro Luis do Nascimento Silva from the IBGE and former ISI president. Dr Pedro provides a very useful discussion and came with many interesting questions about the paper.

After my talk I had time to relax and visit some of the main attractions in Kuala Lumpur. I went to the top floor of the Petronas tower and enjoyed the view at night. I also visited the Batu caves and ate street food in Jalan Alor, the heart of the city's cultural cuisine. The variety of the food there is amazing with noodles, traditional dishes and delicious desserts. I also did some shopping in the Central Market, where I bought small presents for my family.

The closing ceremony took place on Friday 23rd August, and was very well attended. Dr. Mohd Uzir Mahidin took the opportunity to express his thanks to all the persons from Malaysia and abroad for the successful organisation of the WSC. The audience enjoyed the ISI WSC 2019 Flashback Video. Then it was time for President Helen MacGillivray to say a big thank you to the organisers and to hand over the ISI 'keys' to incoming President John Bailer. He invited all the participants to the next WSC, to take place in 2021 in The Netherlands, by introducing the WSC 2021 welcoming video. ISI Director Ada van Krimpen emphasized the open character of the Dutch society and introduced the Sand Story about WSC 2021 and The Hague city.

This event, on Friday evening, 23 August was the grand finale of the Congress. The congress participants enjoyed an entertaining evening with top artists from Malaysia. We all celebrated this extremely successful and enjoyable World Statistics Congress. A thousand thanks to the Malaysian hosts and all the wonderful people for their hospitality.

I would like to extend my appreciation to the ISI and IASS for giving me this honour, and for making this trip possible. Thanks to the logistic committee, scientific committee and the registration desk. And special thanks to Pedro Luis do Nascimento, Denise Silva and Peter Lyn for looking after me.

Figure 2: *Dr. Diego Andres Perez Ruiz with the incoming president of the IASS Denise Silva at the closing ceremony*

The experience of being the Cochran-Hansen prize winner will become an invaluable and treasured memory and will be very important in my career development. The knowledge, connections and perspectives I found in Kuala Lumpur will continue to inform my thoughts and work and I look forward to the 63rd ISI World Statistical Congress in the Hague, the Netherlands.

Report on the ITACOSM 2019 Conference

Risto Lehtonen

The series of the biannual ITACOSM conferences (Italian Conference on Survey Methodology) provides one of the major international scientific events on survey statistics and methodology. The 6th conference, ITACOSM 2019, took place in 5-7 June 2019 in Florence, Italy. The overall theme of the conference, "Survey and Data Science", is a nice reflection of the fundamental and rapid changes in the current data landscape and of the resulting challenges. By the programme, "ITACOSM 2019 tries to give a response to the increasing demand from researchers and practitioners for the appropriate methods and right tools to face these changes". In my experience, the event fully achieved these goals.

Almost 80 papers were presented in the five plenary sessions, twelve specialized (invited) sessions on specific topics, and six contributed and poster sessions. A wide range of interesting and timely topics were covered. The integration of data from traditional and non-traditional (big data) sources and related estimation problems were addressed by three of the keynote speakers. Several other presentations also dealt with this area, which has become of great interest in official statistics worldwide. Small area estimation was another popular topic, treated sometimes in connection to data integration. Topics covered also included design and analysis of complex surveys, non-sampling errors and missing data problems, environmental surveys and spatial sampling, record linkage, surveys on sensitive issues, the estimation of population size via integration of multiple sources, and quality issues. Just a few highlights are possible in this paper.

In his keynote talk on the future of statistics, *Giorgio Alleva* of Sapienza University of Roma had four keywords: data, capabilities to manage data, methods, and data governance. The following were

identified as prerequisites for a leading role of official statistics and statistical science in the new data ecosystem: full collaboration with substantial disciplines and users, going beyond the traditional probability sampling paradigm, building new skills in university education, recognizing the strategic role of data integration for data from multiple sources, and meeting the challenges of governance of data, their production, processing and communication. Prof. Alleva considered privacy as the most important cross-cutting theme related to new data sources, covering all aspects of the data life cycle. In my opinion, all these factors are crucial for the future survival of statistical science and official statistics.

Li-Chun Zhang of University of Southampton discussed data integration further in his keynote talk, now under a concept "entity ambiguity", referring to the uncertainty of correct identification of units in the combined data when integrating information from multiple sources. Prof. Zhang presented new developments for inference under entity ambiguity, which may arise in particular when integrating information from traditional surveys and non-traditional (big data) sources. A recent book "Analysis of Integrated Data" by Zhang and Chambers (eds.), Chapman & Hall/CRC (2019) provides a timely presentation of data integration and analysis.

Paul Biemer of RTI International also treated data integration and made an excursion beyond the probability sampling paradigm in his keynote entitled "Total error frameworks for integrating probability and nonprobability data". The focus was in integrated datasets that are combinations of survey (i.e. "design") and non-survey (i.e. "found") data. He reviewed several total error frameworks for assessing the quality of integrated datasets, as well as the quality of hybrid estimates obtained from such data. In integrating survey and big data, Prof. Biemer encouraged a careful consideration of all measurable components of the total error, including sample recruitment bias and measurement errors, which are important potential error sources but are often ignored in practice. Related topics are discussed in a forthcoming book "Big Data Meets Survey Science: A Collection of Innovative Methods" (Hill, Biemer, Buskirk, et al. (eds.), Wiley).

As a further example of these areas, a specialized session "Inference from informative and nonprobability survey samples" was arranged with presentations by *Jae-Kwang Kim* on "Combining non-probability and probability survey samples through mass imputation", Changbao Wu on "Sample matching and double robust estimation with nonprobability samples" and *Jean-François Beaumont* on "Data integration of probability and nonprobability samples".

Coming to additional topics, *David Haziza* of University of Montréal concentrated in his keynote talk on imputation techniques for the treatment of item nonresponse in surveys. He identified four key questions of interest in the last two decades: 1) How to obtain asymptotically unbiased and efficient point estimators, 2) How to obtain some protection against the misspecification of the underlying model(s), 3) How to consistently estimate the variance of imputed estimators, and 4) How to preserve relationships between survey variables, and presented approaches for proper dealing with these questions in practice. For example, multiple imputation is often applied for complex surveys involving stratification, clustering and unequal probability sampling. It has been shown that point and variance estimators can be biased if ignoring these complexities in multiple imputation. It is thus advisable for example to include design information as additional covariates in the imputation models, which in addition may involve random effects to account for the clustering. Prof. Haziza discussed also some recent developments in the area including multiply robust procedures.

Environmental statistics constitute an increasingly important area in applied survey statistics. Agricultural productivity is estimated to increase to meet the feeding needs of future generations. The key question is how to manage this growth in a sustainable way and how to monitor development. In her keynote talk, *Elisabetta Carfagna* of University of Bologna introduced methodologies for monitoring agriculture and agri-environment. Different geospatial data sources have become available for monitoring purposes. Prof. Carfagna focused on the main methodological aspects underlying the use of geospatial technology for obtaining reliable and timely agricultural and agri-environmental statistics. The use of remote sensing data was illustrated in both area sampling design and estimation design. Several applications were presented.

Abstracts and most presentation slides are available at the conference website http://meetings3.sis-statistica.org/index.php/ITACOSM2019/ITACOSM2019. Selected papers are going to be published in a special issue of the Statistics & Applications journal.

ITACOSM 2019 was organized by the Survey Sampling Group of the Italian Statistical Society and was supported by the International Association of Survey Statisticians (IASS) and the University of Florence. The Scientific Program Committee was chaired by Alessandra Petrucci and the Local Organizing Committee by Emilia Rocco, both of the University of Florence.

Thanks for an enjoyable conference of a high scientific level and a lively atmosphere. Looking forward to the next ITACOSM!

Report from the sixth European Establishment Statistics Workshop

Mojca Bavdaž, Arnout van Delden, Paul A. Smith

The sixth *European Establishment Statistics Workshop* (EESW19) was held in Bilbao, the Basque Country, Spain, on 24-27 September 2019. It was organised by ENBES, the *European Network for Better Establishment Statistics* and hosted by EUSTAT, the statistical office of the Basque Country. The programme can be found at https://statswiki.unece.org/display/ENBES/EESW19; the abstracts and papers are linked from the programme. Previous workshops were held in Stockholm, Sweden, in 2009; Neuchâtel, Switzerland, in 2011; Nuremberg, Germany, in 2013; Poznań, Poland, in 2015; and Southampton, UK, in 2017.

For the first time, three short courses were offered, that attracted 40 participants: two ECAS (*European Courses in Advanced Statistics*) courses – *Topics in Business Data Collection Methodology* by Ger Snijkers and Gustav Haraldsen, and *Statistical Data Cleaning for Business Statistics with R* by Mark van der Loo, as well as *Machine Learning* by José L. Cervera-Ferri.

The workshop gathered 50 participants from 17 countries largely coming from institutes of official statistics, but also from several academic institutions and some non-profit organizations. Different formats were used with sufficient of time for discussion and networking (regular, panel and poster sessions; small group discussions; social events). The scientific programme covered a variety of topics in establishment statistics.

In data collection, achieving response continues to be a costly endeavour. Several presentations addressed electronic data collection, which is shown to be cheaper and useful for statistical institutions but not necessarily simple or effective (e.g. some business respondents from top management prefer to be called than to participate in a web survey that first requires registration; post typically triggers more response than email). Participants discussed the controversial role of fines that range from a thousand to several tens of thousands of euros in participants' institutes and seem to reduce non-response but may jeopardise the relationship with businesses. Another controversial idea is to replace data collection efforts with statistical modelling based on past data to speed up production and reduce costs, because the modelling itself depends on collected data. Several presentations embraced the idea of tailoring to the business context (e.g. considering the distribution of information in the business when designing and implementing a complex questionnaire, using respondents' language when assigning products to the right classification codes, targeting business needs when providing feedback). Communication with businesses thus seems to be increasingly electronic, diversified and customised but coming out of more centralised internal processes (e.g. centralised data collection).

There were presentations on traditional methodological topics, including an assessment of whether designing samples for best quality estimates of levels or of changes was to be preferred, and a paper on using mixture models to identify unusual observations for data editing. There were papers addressing weighting in indirect surveys, and reweighting to correct frame errors.

A panel held during EESW19 highlighted business investment in intangible assets, an example of a challenging subject to measure. The panellists pointed to the lack of a generally accepted conceptual and operational definition and insufficient theoretical background given different institutional arrangements and developmental levels around the world. Collecting data on intangibles is hindered by the unavailability of adequate data in businesses or difficult access to such data (e.g. spread across many people and departments). Future efforts in the context of establishment statistics should aim at agreeing on a definition and rethinking data collection methodology for intangibles.

The small group discussions covered three topics. A first group discussion was on “how to improve own data collection and influence data collection underlying administrative sources”. Improvements of own data collection include centralisation of data collection (but this needs high level support), keeping the address file up-to-date, and communicating quality better even if official statistics is already trusted. An important shortcoming of administrative data is that their quality is often unknown. Good communication with the data suppliers is essential. The second group discussion was on making questionnaires more intelligent. They discussed that intelligent questionnaires utilise three sources of information: a) information about the respondent, the company and the business context; b) information previously given to or gathered by the NSI and c) process data gathered during the questionnaire completion. With an intelligent questionnaire ideally, there is no need of instructions, error messages or help because it incorporates conversation design, using the best conversation flow and logic. The third group discussion was about network analysis of business statistics. It addressed how new detailed data on transactions can be used to analyse networks of businesses and describe the actors. Some examples of these types of analysis are already appearing, and it would be beneficial to link network analysis and official business statistics more closely. (An ENBES workshop on this topic is in preparation for April 2020.)

Secondary data are increasingly being used in different national statistical institutes for business statistics, highlighted at EESW19. Beside the traditional administrative sources, there are also examples of less conventional data sources such as crowd sourcing (to collect for cannabis prices), web scraping (data for the consumer price index, web site texts for later use with text mining methods), and chemical waste water analysis to estimate drug use. The secondary data were used either as a single source or multi-source. For the multi-source statistics there was a split-population design where administrative data was used for small businesses and survey data for the larger businesses. There was also an example of a split-variable design where each administrative data source covers the full population, but different variables were collected in different administrative sources. The papers gave an overview of four challenges. First, there were coverage errors. If there is a business frame these coverage errors can be estimated, otherwise it is more difficult to estimate them. Second, there were linkage errors because the linkage variables were not unique or because the unit types in the linked data sets differed. It is not so easy to correct for those linkage errors. Third, the variables in the administrative data sets may need to be harmonised, which requires well defined standards for the targeted variables. Fourth, estimation methods were needed to deal with specific problems such as transforming the data to the desired publication period and correcting for late arriving data.

Machine learning is a set of methods that is increasingly used in business statistics. Most of the applications reported at the workshop concerned supervised learning, which means that a set of labelled examples is used to learn the relation between a set of input variables, called features, and an output variable. Often this output variable is a classification variable. The machine learning methods were used in different ways – as a tool to analyse which background variables relate to a certain response behaviour, as a way to impute missing values with a large set of explanatory variables and as a mean to derive information from unstructured sources. Examples of the latter are website texts and satellite images. A main advantage of machine learning methods is their flexibility: all kinds of (non-linear) relationship between features and output can be modelled. Under certain conditions this may lead to more accurate estimated values than with more traditional (regression) models. But there is also a need for a framework to assess the quality of machine learning results.

Elements of such a framework are the generalisability of predictions for unseen cases, the accuracy of estimated aggregates, interpretability of the models, model robustness or for input disturbances and for label errors, stability of the predictions over time, and the validity of the methods used.

ENBES gratefully acknowledges financial support from the IASS for towards holding this workshop.

The Chief Methodologist Network

Anders Holmberg and Sybille McKeown

The Chief Methodologist Network was established in 2013 at the World Statistics Conference in Hong Kong. The network has a focus on strategic issues facing methodology programs in NSOs and provides an opportunity to share experiences and learnings about contemporary methodological leadership. The membership comprises Chief Methodologists of national statistical offices (NSOs) from 13 countries and Eurostat, other senior leaders from methodology programs in NSOs, together with a number of academics connected with official statistics. Network meetings are held twice a year, with the aim of at least one meeting coinciding with a conference or other large gathering with a number of members present.

The most recent meeting was held in Malaysia at the World Statistics Congress in July 2019. The meeting covered approaches to collaboration between NSOs and academics, strategic workforce planning for methodology areas and the role of Chief Methodologists in governance. Previous meetings have considered topics such as the challenges of moving new solutions from innovation to production, the changing nature of methodology for official statistics in the digital world and building new skills in data science. The network has previously also facilitated opportunities for peer review, knowledge sharing and staff exchanges.

The network is currently chaired by the Chief Methodologist of the Australian Bureau of Statistics (ABS), with the ABS also providing the secretariat function. For further information about the network, please contact Anders Holmberg (anders.holmberg@abs.gov.au).

Ask the Experts

Wider applications for dual and multiple system estimation

Peter G. M. van der Heijden[1] **and Maarten Cruyff**[2]

[1]Utrecht University, the Netherlands, and University of Southampton, UK,
p.g.m.vanderheijden@uu.nl
[2]Utrecht University, the Netherlands, m.cruyff@uu.nl

Abstract

Dual and multiple system estimation can be usefully applied in much more general settings than usually considered. We consider settings such as making use of covariates that are not available in all of the lists, and lists that cover different (but overlapping) periods in time, lists that cover different (but overlapping) regions, and lists that cover different (but overlapping) age ranges.

Keywords: Dual system estimation; multiple system estimation; covariates; capture-recapture

1 Introduction

We want to show here that dual and multiple system estimation can be applied in much more general settings than usually considered.

In dual system estimation the aim is to estimate a population size. See Table 1. Two lists of individuals, say lists A and B, are linked. Being in list A (B) is denoted by 1, and not being in list A (B) by 0. There are 259 individuals in A and in B, 539 individuals in A but not in B, and 91 individuals in B but not in A. The individuals not in A and in B are missed by both lists, and the aim is to estimate their number.

Important assumptions are perfect linkage of the lists, a homogeneous probability to be included in at least one of the lists (for the other list the inclusion probabilities may be heterogenous over individuals), and independence between the inclusion probabilities of A and B. Due to the assumption of independence, the estimated missing count for the cell (0,0) is 539 * 91 / 259. The independence model in dual system estimation can also be denoted as a loglinear independence model. This has the advantage that generalizations of dual system estimation to situations where background characteristics of individuals are taken into account, and more than two lists are used, are easily described. We will denote loglinear models by placing the variables that constitute the higher order margins between square brackets. So here the independence model is denoted by $[A][B]$.

Table 1: *Contingency table after linking list A and list B*

	$B=1$	$B=0$
$A=1$	259	539
$A=0$	91	0

Dual system estimation plays a role in the census, where the lists are the census survey (A) and the so-called census coverage survey (B). In a census context the overlap in cell (1,1) is relatively large in comparison to (1,0) and (0,1), and thus the estimate in cell (0,0) is relatively small. Therefore a violation of the independence assumption will have only a minor effect on the population size (compare Gerritsen et al., 2015a, where the resulting bias is quantified). But in other contexts this is not necessarily the case.

As said, dual system estimation assumes independence of the inclusion probability of list A and of list B. This can be easily violated. It may be, for example, that in both lists there is heterogeneity of inclusion probabilities in the sense that some individuals have lower probabilities to be on both lists and other individuals have higher probabilities. Think of young men, that are harder to include in both the census survey as well as in the census coverage survey. Two important ways to deal with such violations are (i) using covariates, and (ii) using more than one list.

For covariates one can think of variables such as age and gender. Then the assumption becomes that inclusion in list A is independent of inclusion in list B for each combination of gender (denoted by, for example, X_1) and age (X_2) separately. In terms of loglinear models this would mean that we would have to fit the loglinear model $[AX_1X_2][BX_1X_2]$. Using this loglinear model would solve the young men problem that we just discussed.

For a third list one may think of using, for example, a police register C with apprehensions. This would mean that we have a $2\times2\times2$ table with one missing cell. Thus there are seven counts, and a loglinear model with seven parameters may be fit: $[AB][AC][BC]$. In this model the independence assumption is replaced by the assumption that there is no three factor interaction. This would mean that the odds ratio between the census survey and the census coverage survey is identical for individuals included in the police register and individuals not included in the police register. This is a much less demanding assumption than independence.

These results are well know and described in detail in, for example, Bishop, Fienberg and Holland (1975), and the International Working Group on Disease Monitoring and Forecasting (1995). We now move to new grounds, namely to settings where the covariates are not available in each list, and to settings where the lists cover populations that only partly overlap.

2 Missing covariates

Consider dual system estimation. When two lists are linked, covariates that are present in only one of the lists will be missing for individuals that are not on that list. Consider Panel 1 in Table 2 (see Van der Heijden et al., 2012, 2018, for details). List A is part of the population register, with individuals born in Iran, Irak or Afghanistan, and B is the corresponding part of the police register. Marital status (denoted as X_1) may be a covariate of interest but it is only collected in the population register A. Hence it is missing for those individuals only in B. Police region where apprehended, covariate X_2, is only collected in the police register B and therefore missing in A. For the individuals both

Table 2: *Covariate X_1 (Marital status) is only observed in population register A and X_2 (Police region where apprehended) is only observed in police register B*

Panel 1: Observed counts of All individuals

		$B=1$		$B=0$
		$X_2=0$	$X_2=1$	X_2 missing
$A=1$	$X_1=0$	259	539	13,898
	$X_1=1$	110	177	12,356
$A=0$	X_1 missing	91	164	-

Panel 2: Fitted values under $[AX_2][X_1X_2][BX_1]$

		$B=1$		$B=0$	
		$X_2=0$	$X_2=1$	$X_2=0$	$X_2=1$
$A=1$	$X_1=0$	259.0	539.0	4,510.8	9,387.2
	$X_1=1$	110.0	177.0	4,735.8	7,620.3
$A=0$	$X_1=0$	63.9	123.5	1,112.4	2,150.2
	$X_1=1$	27.1	40.5	1,167.9	1,745.4

in A and B the relation between marital status and police region where apprehended is know, see the upper left four cells. We would like to estimate the missing values on marital status and police region where apprehended for those individuals that are in only one of the registers. Panel 2 provides these estimates. The estimates are produced using the EM algorithm assuming missing at random assumption. The model used is $[AX_2][X_1X_2][BX_1]$. The model has 8 parameters, and where 8 is also the number of counts in the table in Panel 1. Due to the term X_1X_2 the odds ratio between X_1 and X_2 in Panel 1 is projected to obtain estimates for the missing data. Notice that, for example, 4,510.8 + 9,387.2 = 13,898, so the 13,898 are spread out over the two cells.

A more complicated application concerns Polish individuals in the Dutch population register, see Table 3 taken from Gerritse et al. (2015b). We want to know the number of individuals born in Poland being in the Netherlands. For the census, it is important to split this number up in individuals having usual residence and those who have not. There are three lists, namely the population register, the employment register and the police register, hence we have a triple system estimation problem. For the first two lists the covariate usual residence can be derived, but for the crime suspects register this information is missing. Hence there are two missing cells, for which we know the sum: 1,043. This number is the number of individuals that are only in the crime suspects register. Using a loglinear model in the EM algorithm this number is spread out over the two missing cells, and subsequently for the two (no, no, no) cells an estimate can be found that completes the population size estimation problem.

These two examples illustrate that it is possible to use covariates that are not available in each of the registers. From a substantive point of view this is important as it provides population size estimates broken down over the levels of the covariate, and this provides further insight into the constitution of the population. From a statistical point of view van der Heijden et al. (2018) show that, when the missing at random assumption holds, making use of the covariates lead to estimates with good properties in terms of RMSE.

Table 3: *Polish individuals by the population register, the employment register and the crime suspects register, by usual residence. The counts for the two cells labeled "missing" add up to 1,043.*

			Crime suspects	
Usual residence	Population	Employment	Yes	No
No	Yes	Yes	32	3,523
		No	34	3,225
	No	Yes	149	60,190
		No	missing	0
Yes	Yes	Yes	183	21,309
		No	195	14,052
	No	Yes	81	20,216
		No	missing	0

3 Different but overlapping populations

Dual and multiple system estimation can also be used when lists refer to different but overlapping populations. The key is to consider it as a missing data problem.

As a first example, assume that the lists cover different time periods. In Zwane et al. (2004) we study the size of the population of babies having Spina Bifida. There are six lists and the lists cover different time periods. For the lists not covering the full time period, the missing entries in the contingency table are estimated with EM, assuming missing at random. This latter assumption means that relations between lists found in the full time periode are projected to the time period where some of the lists are missing, an assumption that is plausible.

One may also think of one list covering the north and the middle part of a country and another list the middle and south part. Using EM one can estimate the counts for the south part for the first list, and for the north part for the second list. The relation between both lists in the middle part is used for such projections.

As a last example, one can also think of lists covering different age ranges. For example, one could have a census list and a driving licence list. Here the young will not have a driving licence, but the (non-)overlap between the census and the driving license list can projected to this young age group.

4 Conclusion

We hope to have shown that dual and multiple system estimation can be applied in wider context than is usually considered. We would welcome to see such applications and are happy to advice!

References

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W (1975). *Discrete Multivariate Analysis, Theory and Practice*. McGraw-Hill, New York. doi: 10.1007/ 978-0-387-72806-3.

Gerritse, S. C., van der Heijden, P. G. M., and Bakker, B. F. M. (2015a). Sensitivity of population size estimation for violating parametric assumptions in loglinear models. *Journal of Official Statistics*, **31**(3), 357-379. doi=10.1515/jos-2015-0022.

Gerritse, S. C. , Bakker, B. F. M., and van der Heijden, P. G. M. (2015b). Different methods to complete datasets used for capturerecapure estimation: estimating the number of usual residents in the Netherlands. *Statistical Journal of IAOS*, **31**(4), 613-627. doi: 10.3233/SJI-150938.

International Working Group for Disease Monitoring and Forecasting (1995). Capturerecapture and multiple record systems estimation. Part i. History and theoretical development. *American Journal of Epidemiology*, **142**, 1059-1068. doi: 10.1093/oxfordjournals.aje.a117558.

Van der Heijden, P. G. M., Whittaker, J. Cruyff, M., Bakker, B. F. M., and Van der Vliet, R. (2012). People born in the middle east but residing in the netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics*, **6**(3), 831-852. doi: 10.1214/12-AOAS536.

Van der Heijden, P. G. M., Smith, P. A., Cruyff, M. and Bakker, B. F. M. (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics*, **34**(1), 239-263.

Zwane, E., Van der Pal-de Bruin, K., and Van der Heijden, P. G. M. (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine*, **23**, 2267-2281. doi: 10.1002/sim.1818.

New and Emerging Methods

New Data Sources for Official Statistics – A Game Changer for Survey Statisticians?

Siu-Ming Tam[1] and Anders Holmberg[2]

[1] Australian Bureau of Statistics, University of Wollongong, Australia, stattam@gmail.com
[2] Australian Bureau of Statistics, anders.holmberg@abs.gov.au

Abstract

Faced with declining budgets, rising data collection costs and increasing demand for richer, more detailed and frequent statistics, National Statistical Offices are increasingly looking at using new data sources for the production of official statistics. However, as the inferential value of new data sources is limited by issues such as coverage bias and measurement errors, it is paramount that methods are developed and used to address those issues. In this article, we summarize methods which, given underlying data structures, are advocated in the literature to address under-coverage and measurement error. Finally, the article also proposes 10 "rules" for engaging with new data sources for the production of official statistics.

Keywords: big data, estimation, integrating data, secondary data sources, prediction methods, survey quality.

1 Introduction

Responding to the new Australian Bureau of Statistics (ABS) strategic directions on data leadership, the methodologists in the organisation nominated new data algorithms (also known as machine learning (ML) and artificial intelligence (AI)), new data sources and data integration as the new statistical frontiers, and pledged to prioritise research in these areas.

In his President's Invited paper delivered at the 62nd World Statistics Congress in August 2019 hosted jointly by the Malaysian Statistics Office and the International Statistical Institute, Professor Bradley Efron gave an interesting lecture on Data Science. Our take home messages from the lecture are that, as demonstrated in the examples shown in his talk, ML methods generally perform better than "classical" statistical methods in terms of prediction, but do not generally do well in attribution, i.e. understanding the relationship between response variables and the "features" (also known as auxiliary variables in classical statistics). According to his discussant, Professor Noel Cressie, Professor Efron's talk highlighted the distinction between science (which is to find out the truth) and engineering (which is to make things work). Classical statistics have been developed to deal with the former, whilst ML/AI are developed to deal with the latter.

There is already an increasing trend in National Statistical Offices (NSO) to take up ML/AI in official statistics in, for example, predicting the occupancy status of dwellings for Census operations (Dzhumasheva, 2019), predicting the best time to contact respondents (Wang (2019)), coding

operations (Gweon et al. (2018)), editing and imputation (Richman et al. (2002), Jentoft and Zhang (2019a), Ruiz (2018)), replacing a survey question by predictive modelling using registry data (Burger et al. (2019)) etc.. We further note, in some other NSOs, ML and AI, in combination with new data sources, are used to supplement or replace traditional data sources in the production of official statistics e.g. use of support vector machines and satellite imagery data to predict crop classification (Handbook, 2017).

Data integration, which links two or more data sets together that have overlapping population units, creates new data sets that will have more public value than either of its component data sets. For example, by linking migrants' data from immigration records with census data over time, the analyst can look at the settlement outcomes of different migration cohorts and develop better targeted policies for migrants. The relevance of data integration in this paper is that it is a process that creates new data sources for official statistics. However, as there will undoubtedly be linkage errors in fusing two or more data sets together, many integrated data sets will have characteristics similar to non-probability samples and raise challenges for statistical inference. Lothian et al. (2019) outlined the opportunities and challenges in linking data sets across time, space and sources, and proposed a schema for linking traditional and non-traditional data sets.

In this paper, we will outline some inference challenges and proposed solutions in the literature that come with new data sources. This is done by illustrating four type of data structures that survey statisticians may come across. Naturally, the subject of new data sources is huge and growing by the day, and we will give our, undoubtedly, biased views about the subject. It is hoped that our paper will be a catalyst for other survey or official statisticians to give their perspectives on the subject, to enrich the debate on the future methodological directions of official statistics.

2 New data sources and their inferential value

In this paper we shall use the terms "big data" and new data sources interchangeably. Whilst big data often are characterised by a number of V's, e.g. Volume, Velocity, Variety etc., they are, from the official statistician's perspective, just data sources that, similar to administrative data, censuses or surveys, may be used in the production of official statistics. A diagrammatic illustration of the possible data sources for official statistics is provided in Figure 1.

Challenges in using big data for finite population inference are well documented in the literature – see for example, Couper (2013), Baker et al. (2013), Hand (2018), Tam and Clarke (2015), Japec et al. (2015), Macfeely (2019), Tam and Kim (2018). One often cited misconception of big data is that the size of the data set will compensate for any deficiencies in the data. Using the fundamental theorem of estimation error by Meng (2018), Tam and Kim (2018) showed that when the response variable is binary and if

$$
\begin{aligned}
p &= \Pr(Y_i = 1) \\
b &= \Pr(I_i = 1 \mid Y_i = 1) - \Pr(I_i = 1 \mid Y_i = 0) \\
r &= \frac{\Pr(I_i = 1 \mid Y_i = 1)}{\Pr(I_i = 1 \mid Y_i = 0)},
\end{aligned}
$$

where b and r are different measures of response bias, the effective sample of the big data set is given by $n_{eff} = \dfrac{f^2 N}{b^2 p(1-p) + f}$ where $f = n_B / N$, n_B and N denote the known size of the big data and population respectively. Furthermore, the bias of the sample mean compiled from big data as an estimator (also known as B-sample mean) of the population mean is $\dfrac{-p(1-p)(1-r)}{1-(1-r)p}$. When

$f = 1$ and, using the Bayes' Theorem, it is easily shown that $b = 0,$ and $n_{eff} = n_B$; also it follows that $r = 1$ and the bias of the sample mean is zero, as expected (Tam et al. (2019)).

Figure 1: *Possible data sources for the production of official statistics*

When the response bias is non-negligible, the inferential value of big data is substantially reduced, and the bias of the B-sample mean estimator is non-zero. For example, the inferential value of the big data to estimate the proportion of English speakers at home during the 2016 Australian Census is illustrated in Table 1, and the bias of the B-sample mean is given in Table 2 (Tam and Kim (2018)). Note that in Table 2, as illustrated in the formulae above, the bias, given $r,$ is the same regardless of the size of the big data sample.

Other work to assess and analyse bias in this context has been given by Biemer (2019) (building on Meng (2018)) provided an expression of the estimation error in terms of data encoding error and sample recruitment error, and by Mercer et al. (2017) who developed a framework for a quality assessment of the level of selection bias.

Table 1: *Effective sample size to estimate the proportion of English speakers, with different values of f and b*

Big Data fraction, f	Big Data size	Response bias, b		
		1%	5%	10%
1/10	2,340,189	507	20	5
1/4	5,850,473	3,171	127	32
1/3	7,722,624	5,525	221	55
1/2	11,700,946	12,684	507	127

Table 2: *Statistical bias in estimating the proportion of English speakers at home, with different values of f and r*

Big Data fraction, f	Big Data size	Response bias, r		
		1.1	1.3	1.5
1/10	2,340,189	2%	4%	7%
1/4	5,850,473	2%	4%	7%
1/3	7,722,624	2%	4%	7%
1/2	11,700,946	2%	4%	7%

Note: The proportion of English speakers at home in the 2016 Australian Census was 73%.

3 Validity of descriptive inference from new data sources – Type 1 data structure

By descriptive inference, we mean making inference of the parameters of a finite population, e.g. population means, proportions or totals, which are the "bread and butter" work for official statisticians. For analytic inference with big data, the reader may refer to Kohler et al. (2019). In this section, we assume that presence of an additional data source, A, to assist with inference using the new (big) data source B. We also assume the response variable of interest is only available from B, but the same auxiliary variables are available from both B and A. Data from on-line panels fall in this type of data structure, which is depicted as Type 1 in Table 3 below.

Table 3: *Four types of data structures*

	Source	Response variable, Y	Auxiliary variable, X	Representativeness?
TYPE 1	New data source, B	A	A	No
	Additional source, A	NA	A	No
TYPE 2	New data source, B	A	A	No
	Probability sample survey source, A	NA	A	Yes
TYPE 3	New data source, B	A	A	No
	Probability sample survey source, A	A	A	Yes
TYPE 4	New data source, B	NA	A	No
	Probability sample survey source, A	A	A	Yes

Note: A denotes available, NA denotes not available

Denote by U the population of known size N. Let each population unit be associated with an outcome of interest, denoted by $y_{i,}$ for $i \in U$ and let n_B. denote the sample size of B Here we assume the common assumption that $B \subset U$, and there are initially no duplicated units in B. Let $\delta_i = 1$ for $i \in B$, and 0 otherwise. Assume the response variable, y_i and the auxiliary variable vector,

x_i are observed for $i \in B$, and $X = \sum_{i \in U} x_i$ is initially known. We are interested to estimate $Y = \sum_{i \in U} y_i$, and $\bar{Y} = Y / N$. Let $\bar{Y}_B = \sum_{i \in B} y_i / n_B$.

Zhang (2019a) developed Missing-at-Random (MAR) conditions under a superpopulation (SP) approach, or quasi randomisation (QR) approach for commonly used estimators from B to be unbiased. These are summarised in Table 4.

Table 4: *Conditions for unbiasedness*[1]

Estimator name	Estimator	Unbiasedness condition(s)
$B-sample$ expansion estimator	SP: $\hat{Y} = N\bar{Y}_B$	SP: $E(y_i \mid \delta_i = 1, i \in U) = E(y_i \mid i \in U) = \mu,$ a constant
	QR: $\hat{Y} = \sum_{i \in B} (y_i / \hat{\pi}),$ where $\pi = \Pr(\delta_i = 1)$	QR: $E(\delta_i; y_i, i \in U) = \pi,$ a constant
$B-sample$ calibration estimator	SP: $\hat{Y} = \sum_{i \in B} w_i y_i,$ where $\sum_{i \in B} w_i x_i = X$	SP: $E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U)$
	QR: If x_i is the post-stratum dummy index $\hat{Y} = \sum_{j,k} (y_{jk} / \hat{\pi}_j),$ where $\pi_j = \Pr(\delta_{jk} = 1)$, j denotes stratum index and k denotes sample unit index within stratum, such that $\sum_{j,k} y_{jk} = \sum_{i \in B} y_i$	QR: $E(\delta_{jk}; y_i, jk \in U) = \pi_j,$ where $\delta_{jk} = 1$ if the unit is included in B, and 0 otherwise.
$B-sample$ inverse propensity weighted (IPW) estimator[2]	QR: $\hat{Y}_{IPW} = N \frac{\sum_{i \in B} (y_i / \hat{\pi}_i)}{\sum_{i \in B} (1 / \hat{\pi}_i)}$	QR: $\Pr(\delta_i = 1 \mid x_i; i \in U) = \Pr(\delta_i = 0 \mid x_i; i \in U) = \pi_i$

Notes:

(1) Under the SP approach, the Expectation Operator is with respect to the superpopulation. Under the QR approach, it is with respect to the inclusion probability distribution. Unbiasedness also includes Asymptotic Unbiasedness as defined in Zhang (2019a).

(2) Assuming $\pi_i = \pi(x_i;\eta) > 0$*, a parametric probability of inclusion function, is completely determined by* x_i*, then* $\hat{\pi}_i = \pi(x_i;\hat{\eta})$ *where* $\hat{\eta}$ *is determined by solving (a)* $\sum_{i\in B} x_i - \sum_{i\in U} \pi_i x_i = 0$ *if* x_i *is known for* $i \in U$*, or* $\sum_{i\in B} x_i - \sum_{i\in A} w_i \pi_i x_i = 0$ *otherwise (Chen et al. (2018)). This uses a generalised (pseudo) estimation equation approach and assumes* π_i *is modelled by a logistic regression model; or (b)* $\sum_{i\in B} \pi_i x_i - \sum_{i\in A} w_i x_i = 0$ *(Kott and Chang (2010)), which uses a calibration weighting approach. Note that the IPW estimator proposed by Zhang (2019a) is* $\hat{Y} = \sum_{i\in B} (y_i / \hat{\pi}_i)$.

Bethlehem (2016) used simulation to show that the B-sample calibration estimator may be able to reduce the bias due to under-coverage or self-selection from on-line web panels. However, he concluded that this only works if the proper auxiliary variables are available. His results are also reaffirmed by simulations in Dever et a. (2008) and Schonlau et al. (2009). Noted that the work of Zhang (2019a) showed that a MAR assumption is needed to underpin those simulations.

Lee (2006) examined the performance of B-sample IPW estimator for on-line panel surveys and concluded that it can reduce, but not eliminate bias, at the expense of increasing variance. They also found that the relationship between the covariates and the response variable was important in forming propensity models, as weak relationship not only did not decrease bias, but also increased variance. Similarly, a MAR assumption is required to justify the simulations.

4 Validity of descriptive inference from new data sources – Type 2 data structure

There is a predominant view in recent literature that the best approach to harvest the information of big data is to combine them with probability sample survey data (Elliott and Valliant (2017), Hand (2018), Thompson (2018), Lohr and Raghunathan (2017)). We now consider the case when the additional source, A, comes from a probability sample, with the weight of the sampling units denoted by d_i. The Type 2 data structure is depicted in Table 3. Note that if the new data source, B, also comes from probability sample surveys, there is already a large body of literature covering this subject – see for example, Citro (2014), Kim (2011), Kim and Rao (2012) Kim et al. (2016), Merkouris (2004), Park et al. (2017), Wu (2004) – which will therefore not be covered in this paper.

In what follows, we outline the results known to the authors that describe methods to harness big data for official statistics.

Result 1 (Elliott and Valliant (2017)). Treating the additional source, A, as a reference sample, they proposed to the following to estimate π_i for the B-sample IPW estimator $\hat{Y}_{IPW}$ defined in Table 4:

$$\pi_i \propto \Pr(A_i = 1 \mid x_i, i \in U) \frac{\Pr(\delta_i = 1 \mid x_i, i \in B \cup A)}{\Pr(A_i = 1 \mid x_i, i \in B \cup A)},$$

where $A_i = 1$ if $i \in A$ and 0 if $i \in U \setminus A$. As for the IPW estimator, the conditions for their estimator to be asymptotically unbiased, in the QR sense, are

$$\Pr(\delta_i = 1 \mid x_i; i \in U) = \Pr(\delta_i = 1 \mid x_i; i \in A) = \Pr(\delta_i = 0 \mid x_i; i \in U) = \pi_i$$

(Zhang (2019a)).

Result 2 (Fuller (2009), Theorem 5.1.1). Let $\pi_i = \Pr(\delta_i = 1)$ and suppose that there exists a vector λ such that $\pi_i^{-1} = x_i'\lambda$. Assume that the weights for the units in the probability sample survey are

d_i, then under some regularity conditions, the regression estimator, $\hat{Y}_{\text{Reg}} = X\hat{\beta}$, or $\hat{Y}_{\text{Reg}} = \sum_{i\in A} d_i x_i' \hat{\beta}$ if X is unknown, where $\hat{\beta} = (\sum_{i\in B} x_i x_i')^{-1} \sum_{i\in B} x_i y_i$, is asymptotically design unbiased.

Results 3 (Yang and Kim (2017), Kim (2018)). Suppose $y_i = m(x_i, \beta) + e_i$ for some β with known function $m(.)$, and $E(e_i \mid x_i) = 0$. Under some regularity conditions, the imputation estimator $\hat{Y}_{IM} = \sum_{i\in A} d_i m(x_i, \hat{\beta})$, where $\hat{\beta}$ is a consistent estimator of β, is approximately asymptotically SP unbiased, provided that $E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U) > 0$.

Results 4 (Rivers (2007), Yang and Kim (2018)). If the Horvitz-Thompson estimator from sample A is denoted by $\sum_{i\in A} d_i y_i$ and $\hat{y}_i$ is the "nearest neighbour" (NN) of unit i in A, defined by $\hat{y}_i = y_{k_i}$, where $k_i = \arg\min_{j\in B} \| x_i - x_j \|$, then the nearest neighbour estimator $\hat{Y}_{NN} = \sum_{i\in A} d_i \hat{y}_i$ is, under some regularity conditions, asymptotically design unbiased, provided that $E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U)$. The condition holds if $f(y_i \mid x_i, i \in B) = f(y_i \mid x_i, i \in U)$.

Result 5 (Chen et al. (2018), Kim and Wang (2019). Assume a parametric propensity model $\pi_i = \pi(x_i; \eta) > 0$ and a SP model $E(y_i \mid x_i, i \in U) = x_i'\beta$. The estimator $\hat{Y}_{DR} = \sum_{i\in A} w_i \hat{y}_i + \sum_{i\in B} \hat{\pi}_i^{-1}(y_i - \hat{y}_i)$ is doubly robust (Robins et al. (1994)), where $\hat{\pi}_i$ is determined by one of the three methods in Notes 2 under Table 4 above, and $\hat{y}_i = x_i'\hat{\beta}$ and $\hat{\beta}$ is as defined in Result 2, provided that

$$E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U).$$

Remark 1. Note that all the results in this section require a MAR assumption, with the exception of Result 2 which requires the inverse of the selection probabilities to be of a specific form.

The performance of some of these estimators (NN, IPW and DR) were compared by Yang and Kim (2018). DR work better than NN for the all three investigated scenarios. The IPW is sensitive to non-linearity misspecification but work better than DR in two of the three scenarios.

5 Validity of descriptive inference from new data sources – Type 3 data structure

Big data can be used to substantially increase the efficiency of estimators from a sample survey, if they are used as benchmarks in the estimation process (Kim and Tam (2018), Tam et al. 2019)). The idea is to treat the population as comprising two strata, namely, a big data stratum which is fully observed, and a missing stratum, the information from which will be obtained from a probability sample survey. The data structure Type 3 under this scenario is depicted in Table 3.

Result 6 (Kim and Tam (2018)). The post-stratified estimator, $\hat{Y}_{PS}$, given by

$$\hat{Y}_{PS} = \sum_{i\in U} \delta_i y_i + N_C \frac{\sum_{i\in A} d_i(1-\delta_i) y_i}{\sum_{i\in A} d_i(1-\delta_i)}$$

is approximately design unbiased, where $N_C = N - N_B$. In addition, $Var(\hat{Y}_{PS}) \approx (1-W_B)\frac{N^2}{n}S_C^2$ where $S_C^2 = (N-N_B)^{-1}\sum_1^N (1-\delta_i)(y_i - \bar{Y}_C)^2$, $\bar{Y}_C = \sum_1^N (1-\delta_i)y_i / N_C$, n is the sample size of A, and $W_B = n_B / N$, assuming $n/N \approx 0$.

Remark 2. If $S^2 = N^{-1}\sum_1^N (y_i - \bar{Y})^2$ and $\hat{Y}_A = N\sum_{i\in A} y_i / n$, and assuming simple random sampling for A, then

$$\frac{Var(\hat{Y}_{PS})}{Var(\hat{Y}_A)} = (1-W_B)\frac{S_C^2}{S^2} \ll 1,$$

$\text{if } S_C^2 \approx S^2$. The factor $(1-W_B)$ is the under-coverage rate of the big data. Therefore, we have the expected result that the higher is the coverage rate, the lower will be the sampling variance of the estimator.

Kim and Tam (2018) also show that the Data Integration Estimator under certain circumstances is equal to the post stratified and thereby asymptotically design unbiased.

Remark 3. If there are duplications in the units in B, the definition of δ_i can be modified from zero/one to zero/number-of-times that the unit appears in B. In addition, if auxiliary variables x_i are available for all the units in B and A, the information may be harvested by modifying (1) as follows: $\sum_{i\in A} w_i v_i = (N, N_B, \sum_{i\in U}\delta_i y_i, \sum_{i\in U} x_i)$ where $v_i = (1, 1-\delta_i, \delta_i y_i, x_i)$ (or $\sum_{i\in A} w_i v_i = (N, N_B, \sum_{i\in U}\delta_i y_i, \sum_{i\in U}\delta_i x_i)$, if $X = \sum_{i\in U} x_i$ is unknown, where $v_i = (1, 1-\delta_i, \delta_i y_i, \delta_i x_i)$).

Remark 4. If there are measurements errors in B such that y_i^* is measured instead of y_i, this can be accommodated by modifying (1) as follows: $\sum_{i\in A} w_i v_i = (N, N_B, \sum_{i\in U}\delta_i y_i^*, \sum_{i\in U} x_i)$, where $v_i = (1, 1-\delta_i, \delta_i y_i^*, x_i)$. If the measurement errors occur in the units in A, this can be accommodated by using $\hat{Y}_{RegDI} = \sum_{i\in A} w_i \hat{y}_i$, where $\hat{y}_i$ is estimated from a measurement error model based on the observations $\{(y_i, y_i^*), i \in A \cap B\}$.

Remark 5. If there is non-response in the probability sample survey, A, we can use a QR approach for $\hat{Y}_{RegDI}$ with a not missing-at-random (NMAR) propensity model, as follow. Let r_i be 1 if the ith unit in A is a respondent, and 0 if it is a non-respondent. Even if $r_i = 1$, y_i can be observed from the B sample for units with $\delta_i = 1$. One can therefore assume a more general parametric response model: $P(r_i = 1 \mid x_i, y_i) = P(x_i, y_i; \phi_x, \phi_y)$, where ϕ_x, and ϕ_y are unknown parameters. The parameters can be consistently determined by solving the following estimation equations, provided that $f(r_i, \delta_i) = f(r_i)f(\delta_i)$:

$$\sum_{i\in A} \frac{d_i r_i}{p(x_i, y_i; \phi_x, \phi_y)} \begin{pmatrix} x_i \\ \delta_i y_i \end{pmatrix} = \sum_{i\in A} \begin{pmatrix} x_i \\ \delta_i y_i \end{pmatrix}.$$

Once $\hat{\phi}_x$ and $\hat{\phi}_y$ have been determined, the final weights for $\hat{Y}_{RegDI}$ are determined by minimising $\sum_{i\in A} d_i r_i^{-1}(\frac{w_i}{d_i r_i^{-1}} - 1)^2$ subject to a calibration constraint, e.g. $\sum_{i\in A} r_i w_i v_i = (N, N_B, \sum_{i\in U}\delta_i y_i)$.

Kim and Tam (2018) showed through Monte Carlo simulations that the RegDI estimator outperforms other estimators in a set of scenarios that reflect measurement errors situations and model misspecifications. Result 6 can also be used for Type 2 data structure, by using the following vector, $v_i = (1, 1-\delta_i, x_i)$ in (1).

6 Validity of descriptive inference from new data sources – Type 4 data structure

Many of the big data sets do not have the variable of interest to the official statistician. For example, in agricultural statistics, official statisticians collect information on crops e.g. classifications and yields from agricultural censuses and surveys, but the Satellite imaginary data which may be used to predict crop classifications or yields, only contain information on wavelengths.

The general approach in using this type of big data – refer to Type 4 in Table 3 - is to use a training data set to develop a statistical model (or train an algorithm in data science) for prediction – see for example Handbook (2017). We see an explosion of machine learning and artificial intelligence methods applied to this type of data structure for prediction. It is beyond the scope of this paper to cover this big body of literature. Instead, we give a few relevant references on applications to official statistics for the interested reader (Carfagna and Gallego (2006), Daas and Puts (2014), Daas et al. (2015), Husek (2018) and Richman (2009), Saar-Tsechansky et al (2007), Tam (2015)).

Alternatively, this type of data structure can be handled by the method outlined in Remark 4 above, i.e. treat the satellite imagery data as y_i^*, and use the training data set to build a measurement error model to predict y_i.

7 Other innovative ways of using big data

There are many other innovative methods of using big data for finite population inference. For example, transactions data, also known as scanner data, are being used by a number of national statistical offices for the compilation of price relatives for the consumer price index (CPI). Accompanied with this, new methods, known as multilateral index methods, have been developed which are considered to be one of the most effective ways to exploit the full amount information available in the transactions data – see ABS (2017), Ivancic et al. (2011).

To address huge reporting load from household expenditure surveys, and address reporting errors, Zhang (2019b) proposed to use scanner data to compile the CPI weights, and use the household expenditure survey as an audit sample to assess the accuracy of the scanner data-based CPI weights. He also developed a test for assessing the accuracy, and also a measure for the uncertainty, of these weights.

In another application, Kim et al. (2018) used a two–level structural error model to combine the survey information for small areas, $\hat{Y}_{hi}$, which is subject to non-significant sampling error, with big data sources i.e. x_{hi}, which are subject to coverage and measurement errors. Their objective is to borrow strength from the different small areas to predict Y_i. The probabilistic structure of their model is summarised in Table 5.

Table 5: *Probabilistic structure of the Kim et al. (2018) model*

Model	Data	Parameter	Latent variable
Level-one	$\hat{\mathbf{Y}}_h = (\hat{Y}_{h1}, \ldots, \hat{Y}_{hn_h})$	$\boldsymbol{\theta}_h$	$\mathbf{Y}_h = (Y_{h1}, \ldots, Y_{hn_h})$
Level-two	$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_H)$	$\boldsymbol{\zeta}$	$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$

Level 1 is essentially the two-equation Fay-Harriot model combined using Bayes formulae, i.e. $h(Y_{hi} \mid x_{i,}\hat{Y}_{hi};\theta_h) \propto g(\hat{Y}_{hi} \mid Y_{hi})h(Y_{hi} \mid x_{hi};\theta_h)$. The MLE of $\hat{\theta}_{hi}$ estimated by EM algorithm are then used as "observed" inputs to estimate the MLE of ζ, again using the EM algorithm. The best prediction of Y_{hi} is then given by $\hat{y}_{hi}^{**}$, where $\hat{y}_{hi}^{**} = E\{E(Y_{hi} \mid \hat{Y}_{hi}, x_{hi};\theta_h) \mid \hat{\theta}_h, \hat{\zeta}\}$.

8 Understanding the whole statistical production process

So far, we have only examined one, albeit vital, phase of the statistical process, namely estimation. Lothian et al. (2019) argued that we need to understand the whole statistical production process when dealing with non-traditional data sources. To achieve this understanding, they proposed a strategy for structuring, analyzing problems and answering questions, based on a system of statistical base registers, plus consistent monitoring and maintenance strategies. These statistical registers are to serve as lighthouses for illuminating 'trusted' estimation procedures and provide a benchmark for comparing and investigating representativeness concerns. Their schema includes a framework for:

- structuring the non-probabilistic data;
- making it useful for cause and effect statistical inference;
- incrementally developing, designing and maintaining the database system; and
- inserting total survey error concepts into the schema.

Recognising that non-representativeness is a key issue for new data sources, they recommended the B-sample calibration estimator. We believe that the methods outlined in the earlier sections of this paper will provide a larger tool set for bias reduction to be used in the schema.

9 Concluding remarks

When should big data be used? Tam and Van Halderen (2019) outlined ten rules of big data engagement for the production of official statistics. These are summarised in Table 6 below.

Table 6: *Ten rules of big data engagement in official statistics?*

Non-Negotiable	Essential
1 Use big data as a solution to a well-defined statistical need	8 The use of big data reduces provider load
2 The long-term supply of the big data should be certain	9 The use of big data produces better statistics
3 Social license issues must be addressed	10 The use of big data is a fail safe
4 The big data is impartial	
5 Security and confidentiality issues have been addressed	
6 The big data is a cost effective alternative or supplement to traditional, statistical data sources	
7 Statistics are amenable to valid statistical inferences	

Of these, seven rules are considered as “non-negotiables”, and the remaining three rules are considered essential. Even though in this paper, we have only discussed one of the seven “non-negotiables”, i.e. statistics produced from new data sources are amenable to valid statistical inferences, it should be remembered that there are other important considerations to be made before using them.

From the results presented in this paper, it can be concluded that:

- where the response variable is available from a probability sample, A, and where it is possible to match the units in A with B, the RegDI estimator is the preferred estimator. Where there is no measurement errorin A, the estimator is approximately design-unbiased. If there is partial or unit non-response in A, the non-response can be modelled using NMAR assumption;
- where the response variable is not available in the probability sample, A, but auxiliary variables are available from both A and B, such that MAR can be assumed, the DR estimator is a failsafe estimator and preferred. Alternatively, the RegDI may also be used where matching of the units in A and B is possible;
- where the response variable is not available, and where the new data source does not come from a probability sample, but where MAR can be assumed, the B-sample IPW estimator or the B-sample calibration estimator may be used;
- Regardless, the availability of good auxiliary variables which are correlated with the response variable is vital for bias reduction for these types of estimators (Bethlehem (2016)). In passing, we note the simulation results in Buelens et al. (2018) which, by comparing the B-sample expansion estimator, and calibration estimator, with a number of commonly used machine learning techniques e.g. regression trees, artificial neural networks and support vector machines, showed that the latter techniques perform better in bias reduction; and
- Importantly, it is vital to decide when to engage with new data sources. Where new data sources are used, it is also important to get on top of the whole statistical production process involved in their use and apply the total survey error framework to assessing the quality of the resultant official statistics.

Finally, given the scope of this paper, we have not included any measures of uncertainties with the above estimators. The interested reader should refer to the relevant papers included in the References for the details.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent the views of the Australian Bureau of Statistics nor University o Wollongong.

References

Australian Bureau of Statistics (2016). Information paper: Making Greater Use of Transactions Data to compile the Consumer Price Index, Australia. Catalogue Number 6401.0.60.003. ABS, Canberra.

Baker, R., Brick, J., Bate, N., Battaglia, M., Couper, M., Dever, J., Gile, K., and Tourangeau, R. (2013). Report of the AAPOR Task Force on non-probability surveys. https://www.aapor.org/Education-Resources/Reports/Non-Probability-Sampling.aspx. Accessed 3 November 2019.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review* **78**, 161 -188.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching?, Social Science Computer Review, 34, 59-77.

Biemer, P. (2019) Can a Survey Sample of 6000 Records Produce More Accurate Estimates than an Administrative Data Base of 100 Million? *The Survey Statistician* **80**, 11-15.

Buelens, B., Burger, J. and van den Brakel, J. (2018). Comparing inference methods for non-probability samples. *International Statistical Review* **86**, 322-343.

Burger, J., Buelens, B., de Jong, T. and Gootzen, Y. (2019). Replacing a survey question by predictive modelling using register data. Paper presented to the 62nd World Statistics Congress, Kuala Lumpur.

Carfagna, E. and Gallego, F. (2006). Using remote sensing for agricultural statistics. *International Statistical Review* **73**, 389-404.

Chen, Y., Peng, L. and Wu, C. (2018). Doubly robust inference with non-probability survey samples. https://arxiv.org/abs/1805.06432. Accessed 7 October 2019.

Citro, C. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology **40**, 137-161.

Couper, M. (2013). Is the sky falling? Survey Research Methods 7, p.145-156.

Daas, P. and Puts, M. (2014). Social media sentiment and consumer confidence. *European Central Bank Statistical Paper Series* **5**, 1-29.

Daas, P., Puts, M., Buelens, B. and van den Hurk, P. (2015). Big data as a source for official statistics. *Journal of official statistics* **31**, 249-262.

Dever, J., Rafferty, A. and Valliant, R. (2008). Internet surveys: can statistical adjustments eliminate coverage bias. *Survey Research Methods*, **2**, 47-62.

Dzhumasheva, S. (2019). Improving census occupancy determination – the potential of administrative for the Census. Paper presented to the 2019 ASEARC Workshop, Sydney.

Elliott, M. & Valliant, R. (2017). Inference for non-probability samples, Statistical Science, **32,** 249-264.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2018). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics* **33**, 101-122.

Fuller, W. (2009). *Sampling techniques.* John Wiley and Sons. Hoboken.

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society A 181*, 1–24.

Handbook on earth observations for official statistics (2017). Report prepared by the Satellite Imagery and Geo-spatial Statistics Task Team of the United Nations Global Working Group on Big Data. New York. https://unstats.un.org/bigdata/taskteams/satellite/. Accessed 22 October 2019.

Husek, N. (2018). Telematics data for official statistics: An experience with big data. *Statistical Journal of the International Association for Official Statistics*, 34, 499-504.

Ivancic, I., Diewert, D. and Fox, K. (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics,* 161, 24-35.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015). Report of the AAPOR Task Force on big data.

Jentoft, S., and Zhang L-C., (2018) Two phase and double machine learning for data editing and imputation. UNECE Workshop on statistical data editing. Neuchatel 18-20 September.

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Norway_ZHANG_Paper.pdf.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika* **98**, 119-132.

Kim, J.K. (2018). Unpublished survey data integration lectures delivered to the Australian Bureau of Statistics.

Kim, J.K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation, *Survey Methodology* **42**, 19-40.

Kim, J.K. and Rao, J.N.K (2012). Combining data from two independent surveys: a model-assisted approach. Biometrika, **99**, 85-100.

Kim, J.K. and Tam, S-M. (2018). Data integration by combining big data and survey sample data for finite population inference. Submitted.

Kim, J.K. and Wang, Z. (2018). Sampling techniques for big data analysis. *International Statistical Review* **87**, 177-191.

Kim, J.K, Wang, Z., Zhu, Z. and Cruze, N. (2018). Combining surveys and non-survey big data for improved sub-area prediction using a multi-level model. Journal of Agricultural, Biological and Environmental Statistics **23**, 175-189.

Kohler, U, Kreuter, F. and Stuart, E. (2019). Nonprobability sampling and causal analysis. Annual Review of Statistics and Its Applications, **6**, 149-172.

Kott, P., and Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse, *Journal of the American Statistical Association* **97**, 1265-1275.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for voluntary panel web surveys. *Journal of Official Statistics* **22**, 329-349.

Lothian, J., Holmberg, A. and Seyb, A. (2019). An evolutionary schema for using “it-is-what-it-is” data in official statistics. Journal of Official Statistics **35**, 137-165.

Lohr, S. and Raghunathan, T. (2017). *Combining survey data with other data sources. Statistical Science* **32**, 293-312.

Macfeely, S. (2019). Big data and official statistics. In *Big Data Governance and Perspectives in Knowledge Management*. IGI Global

Meng, X. (2018). Statistical paradises and paradoxes in Big Data (I): Law of large populations, big data paradox, and the 2016 US Presidential Election. The Annuals of Applied Statistics, 12, p 685 -726.

Mercer, A., Kreuter, F, Keeter, S. and Stuart, E. (2017). Theory and practice in non probability surveys – parallels between casual inference and survey inference. Public Opinion Quarterly **81**, 250-279.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, **99,** 1131-1139.

Park, S., Kim, J.K. & Stukel, D. (2017). A measurement error model for survey data integration: combining information from two surveys, Metron **75**, 345-357.

Richman M. B., Trafalis T. B. & Adrianto I. (2009). Missing data imputation through machine learning algorithms. In *Artificial Intelligence Methods in the Environmental Sciences*. Springer.

Rivers, D. (2007). Sampling for web surveys. In *Proceeding of Section on Survey Research Methods*. American Statistical Association.

Ruiz, C., (2018) Improving Data Validation using Machine Learning. UNECE Workshop on statistical data editing. Neuchatel 18-20 September.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Switzerland_RUIZ_Paper.pdf

Saar-Tsechansky M. & Provost F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research* **8**, 1623-1657.

Schonlau, M, van Soest, and Kapteyn, A (2009). Selection bias in web surveys and the use of propensity scores. Sociological Methods and Research, **37**, 291-318.

Tam, S-M. (2015). A statistical framework for analyzing Big Data. The Survey Statistician 72, 36-51.

Tam, S-M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. International Statistical Review, **3**, 436-448.

Tam, S-M. and Van Halderen, G. (2019). The five V's, seven virtues and ten rules of big data engagement for official statistics. Submitted.

Tam, S-M and Kim, J.K. (2018). Big data ethnics and selection bias: an official statistician's perspective. Statistical Journal of the International Association of official statistics, **34**, 577-588.

Tam, S-M., Kim, J.K., Ang, L. and Pham, H. (2019) (in press). Mining the new oil for official statistics in Big Data Meets Survey Practice: A Collection of Innovative Methods, John Wiley and Sons, Hoboken.

Thompson, M. (2018). Combining Data from New and Traditional Sources in Population Surveys, International Statistical Review, **87**, 79-S89.

Wang, S. (2019). Using predictive response propensity scores with the random forests method to direct a responsive intensive follow up strategy. Paper presented to the 62nd World Statistics Congress, Kuala Lumpur.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. Canadian Journal of Statistics **32**, 15-26.

Yang, S. and Kim, J.K. (2017). Predictive mean matching imputation in survey sampling. https://arxiv.org/abs/1703.10256.

Yang, S. and Kim, J.K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation. https://arxiv.org/abs/1807.02817. Accessed 5 November 2019.

Zhang, L. (2019a). On valid descriptive inference from non-probability sample, Statistical Theory and Related Fields, 3:2, 103-113, https://doi.org/10.1080/24754269.2019.1666241. Accessed 3 December 2019.

Zhang, L. (2019b). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big data statistics. https://arxiv.org/abs/1906.11208. Accessed 8 November 2019.

Quality of Multisource Statistics – the KOMUSO Project

Gabriele Ascari[1], Karin Blix[2], Giovanna Brancato[1], Thomas Burg[3], Angela McCourt[4], Arnout van Delden[5], Danutė Krapavickaitė[6], Niels Ploug[2], Sander Scholtus[5], Peter Stoltze[2], Ton de Waal[5], Li-Chun Zhang[7]

[1] Italian National Statistical Institute, [2] Statistics Denmark, [3] Statistics Austria, [4] Statistics Ireland, [5] Statistics Netherlands, [6] Vilnius Gediminas Technical University, Lithuania, [7] Statistics Norway

Abstract

The results of the project on the quality of the multisource statistics launched by the European Statistical System and accomplished by the National Statistical Institutes of eight European countries under the name KOMUSO are described. The work carried out consists of two main documents: the *Quality Guidelines for Multisource Statistics* supplemented with the collection of the quality measures for statistical output and examples to use them; and *Quality Guidelines for Frames in Social Statistics* with the list of quality measures and indicators followed by the examples to use them. An overview of the documents created in the project is presented in this paper.

Keywords: quality guidelines, basic data configuration, quality measure, calculation method, accuracy, coherence, frame, quality indicator.

1 Introduction

Describing the quality of statistics is a core activity at National Statistical Institutes (NSIs). For statistics based on a single source, e.g. a census, a single survey, or a single administrative register, there are well-established frameworks both for calculating various indicators (Daas et. al., 2011; Zhang, 2012) as well as communicating these in quality reports (Eurostat, 2014). However, more and more often official statistics is not based on a single source. Rather they are compiled on the basis of multiple sources, i.e. a combination of two or more of the mentioned sources. The pool of sources may even include other types of data, perhaps some kind of 'big data'. Statistics produced with these characteristics is collectively termed multisource statistics.

Describing the quality of multi-source statistics has proven a daunting task for many NSIs and for good reasons. The literature on the subject has been scattered and highly technical, and it has required a substantial effort from the NSI to implement the proposed methods, both in terms of time and competencies needed.

As a consequence, the European Statistical System (ESS) in 2016 launched a project with the following objectives in relation to multisource statistics:

1) to take stock of the existing knowledge on quality assessment and reporting and to review it critically; to produce recommendations on the most suitable approaches;
2) to develop new measures for the quality of the output based on multiple sources where at least one source is administrative;
3) to produce a methodological framework for reporting on the quality of output.

The project also had to encompass work on the quality of frames, which may be thought of as the backbone of many statistics and thus defining the fundamental quality of the statistical products. Thus, measures relating to the quality of frames themselves and the data whose production is supported by the frames had to be described. More specifically the project had to produce a

methodological framework for assessing the quality of the frames used in social statistics and drafted a proposal for minimum quality requirements for sampling frames for EU social statistics.

The instrument chosen to carry out the project was an ESSnet, which is an applied research project financed by the European Commission. This ESSnet was coordinated by Statistics Denmark and with active participation from the NSIs of Austria, Hungary, Ireland, Italy, Lithuania, The Netherlands and Norway. The project had a duration of 45 months (from January 2016 to October 2019) and was given the title KOMUSO, which is an imperfect acronym for Quality in Multi-Source Statistics (but also the name of specific group of monks within Zen Buddhism).

The project results consist of the *Quality Guidelines for Multisource Statistics* supported by the collection of practical applications to measure quality of the statistical output based on the multiple data sources; and *Quality Guidelines for Frames in Social Statistics* supported by the corresponding collection of measures and indicators with practical examples of their implementation. The entire production of the project is available on the EU CROS portal (European Commission, 2019). This article provides an introduction to the ESSnet KOMUSO in the sense, that the reader will not only be informed about the content of the project but also pointed towards more detailed reports and resources produced within the project.

2 Quality Guidelines for Multisource Statistics (QGMSS)

The Quality Guidelines for Multisource Statistics (QGMSS) manual aims to support the National Institutes of Statistics in the shift from traditional, single-source processes to the often more demanding multisource processes, a passage that has become more frequent in recent years. The guidelines can be used by process managers in charge of multisource statistics in the planning stage or in a self-assessment exercise to verify if all the quality issues concerning the multisource approach have been properly addressed. They can also be of inspiration for seeking support from methodological sectors on advanced issues concerning error estimation. In the manual, the quality of multisource statistics is analyzed from both a theoretical and a practical perspective, corresponding roughly to the two parts that compose the volume.

Part 1 of the manual describes the quality framework that hinges around three main features: output quality (European Statistical System statistics quality dimensions), statistical sources of errors and process quality, the last mapped on the Generic Statistical Business Process Model (GSBPM). Also, some relevant issues concerning quality management in the multisource settings are considered. The errors taken into consideration are listed in the Table 2.1 below, taken from section 1.1 of the manual, also reporting which of the process component (administrative or survey) is affected by each specific type of error of the category.

In the context of multisource statistics, quality management systems must take into account the peculiar challenges that the integration of several data sources implicate. To this aim, special attention should be paid to some specific elements of the common principles shared by the most used quality management approaches. Specifically, the relationship with data providers, the process orientation and the continuous quality approach are the elements on which emphasis has been put and in the first part of the manual a paragraph for each of them has been developed.

Part 2 of the manual, which is the core of the volume, contains recommendations and guidelines. For each Eurostat quality dimension, i.e. relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity (Eurostat, 2003), three general recommendations (one on preventing, one on monitoring and adjusting and one for estimating the error) and the corresponding quality guidelines are considered.

Table 2.1. *Classification of errors in the multisource statistics*

Error Category	Type of error included	Survey	Administrative sources
Validity error	Specification error	X	
	Relevance error		X
Frame and Source error	Under-coverage	X	X
	Over-coverage	X	X
	Duplications	X	X
	Misclassification in the contact variables	X	
	Misclassification in the auxiliary variables	X	X
Selection error	Sampling error	X	
	Unit non-response	X	
	Missing units in the accessed data set		X
Measurement error and Item missingness	Arising from: respondent, questionnaire, interviewer, data collection	X	
	Fallacious or missing information in admin source		X
Processing error	Data entry error	X	
	Coding or mapping error or misclassification	X	X
	Editing and imputation error	X	X
	Identification error		X
	Unit error		X
	Linkage errors	X	X
Model error (examples, non-exhaustive)	Editing and imputation error, record linkage error, …	X	X
	Model based estimation error (Small Area Estimation, Seasonal Adjustment, Structural Equation Modelling, Bayesian approaches, Capture-Recapture or Dual System Estimation, Statistical Matching, …)	X	X

The chapters of the manual part 2 are based on the structure detailed as follows:

1) An introduction detailing the definition of the quality dimension in the context of multisource statistics.
2) A discussion on the main errors affecting the specific quality dimension.
3) The identification of the main GSBPM sub-processes and multisource data configurations where errors may occur (see section 3).
4) The recommendations and the guidelines (i.e. the activities, approaches and methodologies) that can be taken to prevent, monitor and evaluate the potential errors more relevant for the given output quality dimension;
5) Summaries of Quality Measures and Calculation Methods (QMCMs), where present, concerning the corresponding quality dimensions.

Indeed, the manual has links with practical applications on error measurement represented by the QMCMs (see section 3).

It is worth to notice that the guidelines are ordered, generally speaking, from the simplest to most complex or, in other words, from the ones that the NSIs may implement with a relatively small effort to the more demanding ones. For example, in the case of the prevention of processing errors in the chapter on accuracy and reliability, the suggestions range from the use of controlled data entry to

the evaluation of the validity of a rebased or new weighting system. Indeed, some of the actions suggested – especially those concerning the recommendations on the error estimation – are quite complex to implement and may be experimental in some of their features (in this respect, an appropriate warning has been included in the introduction). However, the manual is intended to be a flexible tool to be read and used by researchers and institutions at different levels of maturity and experience. In any case, to aid the reader in the search of the most useful guidelines for their own processes, specific tables have been introduced to categorize them according to the source component to which they can be applied, e.g. the survey component, the administrative component or both of them (including the case of integrated components).

As for the type of errors, while all steps of the process may be at risk of incurring non-sampling errors, in multisource statistics attention should be paid especially to the processing and the model error. The former includes the integration procedures that are especially critical for the quality of a process based on multiple sources; the latter involves all the steps in which a modelling phase occurs, such as imputation, seasonal adjustment, forecasting and so on.

Error estimation methods can be broadly classified according to the types of error, whether affecting the measurement or the representation lines (Groves *et al.*, 2004; Zhang, 2012). The former case is about the errors affecting the variables, typically measurement errors, whereas the latter case concerns errors affecting the units, e.g. coverage. For this reason, where meaningful, some guidelines are marked accordingly.

Finally, where relevant in the guidelines, the computation of Eurostat Quality and Performance indicators (Eurostat, 2014) is suggested. Examples of such indicators are the rate of available statistics for relevance, the rate of over-coverage for accuracy, the time lags for timeliness and punctuality.

Thorough feedback was collected at different stages of the development work and the suggestions received have been included whenever possible. The observations received both during the development work and on the final version of the QGMSS were positive and encouraging, denoting the current needs in the dynamic and still evolving field of multisource statistics. Of course, in such a fast-moving environment additional integrations to the manual will probably be needed, but the flexible structure of the manual itself allows for an easy update process.

3 Quality of statistical output based on multisource statistics

Some of the work carried out in Komuso on measuring the quality of statistical output based on multiple data sources is presented in this section. The work on this topic is subdivided into three steps:

1) Literature reviews and suitability tests are carried out. In the literature reviews quality measures and recipes to compute them are studied and described. In the suitability tests already existing or newly proposed quality measures and recipes to compute them are tested on real or simulated data.
2) 32 Quality Measures and Computation Methods (QMCMs) are produced. A QMCM is a brief description of a quality measure and the corresponding calculation recipe. It also includes a description of the situation(s) in which the quality measure and accompanying recipe can be applied.
3) Hands-on examples to 31 of the QMCMs are provided.

In order to structure the work, collections of the possible data sources used to obtain the statistical results are classified by basic data configurations (BDCs). The following six BDCs are considered:

- BDC 1: Multiple non-overlapping cross-sectional microdata sources that together provide a complete data set without any under-coverage problems.
- BDC 2: Same as BDC 1, but with overlap between different data sources.
- BDC 3: Same as BDC 2, but now with under-coverage of the target population.
- BDC 4: Microdata and aggregated data that need to be reconciled with each other.
- BDC 5: Only aggregated data that need to be reconciled.
- BDC 6: Longitudinal data sources that need to be reconciled over time (benchmarking).

BDC 1 can be subdivided into two cases: the split-variable case where the data sources contain different variables and the split-population case where the data sources contain different units. For more information on BDCs a reader may refer to De Waal, Van Delden and Scholtus (in press). All QMCMs, examples to these QMCMs, literature reviews and suitability tests are available on the EU CROS portal (European Commission, 2019).

In the next sections brief descriptions are given of some situations as well as examples of the quality measures that are examined in Komuso. For descriptions of some other situations and quality measures, please refer to De Waal, Van Delden and Scholtus (2019).

3.1 Basic Data Configuration 1: Complementary variables and no coverage problems

3.1.1 Accuracy of growth rates due to classification errors

QMCM_A_19 examines the estimation of bias and variance of quarterly and yearly growth rates per domain when there are errors in the classification variable that determines the domains. In QMCM_A_19 these domains are industries which are determined by NACE codes, i.e. the business activity classification. The true NACE code of an enterprise is considered to be unknown, and the enterprise activity code in the business register, which is fixed for one year, is considered as possibly erroneous. Let U denote a target population of units (i.e. enterprises). Suppose that two data sets are observed, where the first data set contains a variable y^r for all units of a subpopulation $U^r \subset U$, and the second data set contains a variable y^q for all units of a subpopulation $U^q \subset U$. For units in the intersection $U^{r,q} = U^r \cap U^q$, both variables y^r and y^q are available. It is assumed that the two subpopulations have a large overlap. The situation may arise in a repeated survey of the population.

Suppose that both subpopulations are divided into strata, where the set of possible stratum codes, i.e. NACE codes, is denoted by $\{1, \ldots, M\}$. Let s_i^r be the true stratum of unit $i \in U^r$ in the first data set, and s_i^q the true stratum of unit $i \in U^q$ in the second data set (which may be different from s_i^r, for instance because the two data sets refer to different points in time). Let the indicator $a_{hi}^r = 1$ if $s_i^r = h$ and 0 otherwise, and similarly let $a_{hi}^q = 1$ if $s_i^q = h$ and 0 otherwise. Let Y_h^r be the total of variable y^r in stratum h and Y_h^q the stratum total for variable y^q, with $Y_h^r = \sum_{i \in U^r} a_{hi}^r y_i^r$ and $Y_h^q = \sum_{i \in U^q} a_{hi}^q y_i^q$. The statistic of interest is the ratio $R_h^{q,r} = Y_h^q / Y_h^r$. Unfortunately, the classification of units is prone to errors, and, instead of s_i^r and s_i^q, $\hat{s}_i^r$ and $\hat{s}_i^q$ are observed which may contain errors. Let $\hat{a}_{hi}^r = 1$ if $\hat{s}_i^r = h$ and 0 otherwise and similarly let $\hat{a}_{hi}^q$ be the observed version of a_{hi}^q. The stratum totals and their ratio are estimated by $\hat{Y}_h^r = \sum_{i \in U^r} \hat{a}_{hi}^r y_i^r$, $\hat{Y}_h^q = \sum_{i \in U^q} \hat{a}_{hi}^q y_i^q$, and $\hat{R}_h^{q,r} = \hat{Y}_h^q / \hat{Y}_h^r$, respectively.

In order to derive analytic expressions for the approximate bias and variance of the estimated ratio $\hat{R}_h^{q,r}$, a second-order Taylor expansion is used for this estimator. The case of quarter-on-quarter growth rates within the same year is relatively simple, because the enterprise register is fixed throughout the year. The cases of quarter-on-quarter growth rates for quarters in different years and yearly growth rates are more complicated.

Some assumptions concerning the classification error mechanism are made in QMCM_A_19. The classification errors in $\hat{s}_i^r$ (and in $\hat{s}_i^q$ for units that occur only in U^q) are described by a level matrix $\mathbf{P}_i^L = (p_{ghi}^L)$, with elements $p_{ghi}^L = P(\hat{s}_i^r = h | s_i^r = g)$. If classification errors are assumed to be dependent over time also a Markov-like model is used, where the observed code depends on the true code in the present quarter and the true and observed codes in a preceding quarter, but not on earlier quarters, which leads to probabilities of the following form:

$$p_{gklhi}^C = P\left(\hat{s}_i^q = h \middle| s_i^r = g, s_i^q = k, \hat{s}_i^r = l\right).$$

These probabilities may again be arranged in a matrix, denoted by $P_i^C = \left(p_{gklhi}^C\right)$, where C stands for 'change'. The probabilities in the level matrix and change matrix need to be estimated.

Using these matrices, approximations for bias and variance of the industry growth rates due to the errors in the industry classification code are obtained analytically in QMCM_A_19.

3.1.2 Accuracy of estimated totals when composition and classification of units change

QMCM_A_14 concerns the effect on output quality of changes in business structure, and the consequent measurement errors, that may occur in a sample after the units have been selected. In QMCM_A_14, it is assumed that all statistical units (enterprises) are registered in a business register (BR). This BR is used as a sampling frame for a stratified simple random sample without replacement, where the strata are formed by the NACE code for economic activity and size, e.g. number of employees. The sampled units are used to estimate an overall population total for a target variable y. It is assumed that after the survey data of enterprises from the selected sample have been obtained, information is received from some of the respondents that the sampling units (SUs) have changed their characteristics:

1) the SU has merged with other enterprises, possibly from different strata;
2) the SU have been split into multiple ones with possibly different values for the stratification variables;
3) another value for the classification variables (for example, NACE code or size group) of the SU has been reported.

These changes may occur because of errors in the classification variable that is taken from an administrative data source or because of changes in the population which occurred between the sample selection and data collection. It is assumed that all information about the enterprise changes is known.

The initial population U evolves into the population U' at the time of the observation with target variable y replaced by a variable y' and the population total $t_y = \sum_{k \in U} y_k$ replaced by the population total $t_{y'} = \sum_{k \in U'} y'_k$. The selected sample ω, $\omega \subset U$ is replaced by the observed sample ω', $\omega' \subset U'$. The problems mentioned above are solved in the following ways:

1) first and second order inclusion probabilities are calculated for the sample ω';
2) the available part of the split enterprises is described by a second phase sampling design, or alternatively random imputation is used to fill in the missed part of the split enterprises;
3) a model for the size group or NACE code changes is assumed.

The Horvitz-Thompson estimator is applied for estimating totals and variances of the estimated totals. Relative bias and relative variance are used as accuracy measures in comparison to the case where changes in the population are not taken into account. In the example to QMCM_A_14, a

simulation study is described. The simulation results show that relative bias and relative variance for the estimator of a total increases with increasing size of changes in the observed population.

3.2 Basic Data Configuration 2: Overlapping variables

3.2.1 Quality framework for register-based statistics

QMCM_A_10 presents a framework for a qualitative assessment of output quality when the output is based on several sources with overlapping variables. The framework has been developed for the Austrian register-based population census. This population census was a full enumeration from several administrative data sources.

A central population register is used that is assumed to have no undercoverage. Each source has to deliver data on a micro level. The data sources are overlapping with respect to the units as well as the variables. A procedure for quality evaluation of statistical output starts with the assignment of quality indicators for every variable in every register used for input. These quality indicators are quantitative functions with values in the interval (0,1). A higher value of a quality indicator means higher quality of the variable. Three quality dimensions are evaluated at this first step: documentation, pre-processing and external source. In this first step, expert knowledge may be used to assess the quality of the data. Quality assessment is then expressed in terms of beliefs of correctness of each data source.

Some of the variables are unique for a data source, whereas other variables occur in several data sets. When the initial data sets are merged with the Central Data Base, the quality indicators for the variables occurring in several data sets are combined using the Dempster-Shafer theory, which uses the beliefs of correctness of the various data sources. Some of the variables are derived from other variables, i.e. their values are imputed, and quality indicators for each imputed value are calculated. In a further step, the Central Data Base is compared to an external source to check the quality, and the final quality indicator is derived.

The quality framework shows changes in the quality during data processing of the register-based population census. Despite having been developed for the population census, the method may also be applied to other register-based statistics. Assumptions for successful application of the quality framework include, for instance, independence of the administrative data sources, and the possibility to link them by a unique key-variable on a unit level.

3.2.2 Accuracy of observed data with measurement errors

QMCM_A_13 examines multiple administrative and survey sources that provide the value of the same categorical variable of interest for (part of) the target population, where all measurements may be imperfect. In this case, an approach based on a latent class model can be used to estimate the true values. In this approach, the accuracy of data source g can be evaluated with estimates of the probabilities $P(Y^g = i|X = i)$, where Y^g is the observed value in data source g and X is the true (latent) value. In this approach, quality measures are naturally provided by the conditional distribution of the latent true variable given the available information (e.g., the posterior variance).

3.3 Basic Data Configuration 3: Undercoverage

3.3.1 Sensitivity analysis of population size estimates using capture-recapture models

QMCM_A_9 supposes that we are interested to estimate the size of a population and its accuracy, using two incomplete, linked, registers of the same population. Some units are included in both registers, let us denote their number by m_{11}, some units are included in the first register but not in the second (m_{10}), and some units are included in the second register but not in the first one (m_{01}).

The number of units in the population that are not included in either of the two registers, m_{00}, can then be estimated by $\hat{m}_{00} = m_{10}m_{01}/m_{11}$. Assumptions underlying this estimator are:

1) inclusion of a unit in register I is independent of its inclusion in register II;
2) inclusion probabilities of units are homogeneous for at least one of the two registers;
3) the population is closed;
4) it is possible to link the units of the registers I and II perfectly;
5) neither register contains units that do not belong to the target population (no overcoverage).

The first and the second assumptions are usually violated in human populations. This violation should influence the accuracy of the population size estimates obtained. The approach taken in QMCM_A_9 is to use covariates, the levels of which have heterogeneous inclusion probabilities for both registers. Loglinear models can be fitted to the contingency table of inclusion indicators for registers I and II and the covariates. The first assumption above is then replaced by the weaker assumption of conditional independence of units to be included in registers I and II conditional on the values of the covariates.

Gerritse et al. (2015), on which QMCM_A_9 and its example are largely based, present a study of the impact of violation of the independence assumption on the accuracy of population size estimates. A known level of dependency between the inclusion probabilities for both registers is created, and estimates for the population size under the independence assumption are obtained. These results are compared to the results obtained when there is dependency. In this way, sensitivity of the population size estimates to violation of the independence assumption is studied.

3.4 Basic Data Configuration 4: Micro data and macro data

3.4.1 Variance estimation for the repeated weighting estimator

QMCM_A_6 examines the repeated weighting (RW) estimator and its variances. This estimator ensures numerical consistency among tables estimated from different combinations of surveys and administrative data sets. To apply the RW estimator, the set of target tables to be estimated first has to be specified. Next, all margins of such a target table are added to the set of tables to be estimated. A marginal table is obtained by (i) aggregating over one or more categorical variables of a multi-way table or (ii) using a less detailed classification of a categorical variable. In a second step, each table is estimated by means of the regression estimator from the most appropriate data set. QMCM_A_6 gives variance formulas for the repeated weighting estimator. The quality of the RW estimator is measured by the estimated variance.

The example to QMCM_A_6 discusses an application of the RW estimator to the Structure of Earnings Survey, which is a combination of the Employment and Wages Survey and the Labour Force Survey. The RW estimator has also been used for the Dutch population census in 2001 and 2011.

3.5 Basic Data Configuration 5: Macro data only

3.5.1 Quality measures for accounting equations

Many statistical figures in official statistics should satisfy an accounting equation, where statistical figures have to sum up to a total. Since statistical figures are frequently estimated in different ways or are based on different data sets, such as an accounting equation is often violated. The statistical figures then have to be reconciled in order to satisfy the accounting equation. A quality measure for

such an accounting equation may be based on the variance-covariance matrix of the estimators involved in the accounting constraint. Such a quality measure is proposed in QMCM_C_1.

QMCM_C_1 also proposes an alternative scalar quality measure for accounting equations. Let us briefly discuss this quality measure. Let $Y_1, \dots, Y_p, Z$ be statistical variables which should satisfy a balancing equation $f(Y_1, \dots, Y_p, Z) = 0$ for some aggregation function f, for example $f(Y_1, \dots, Y_p, Z) = Y_1 + \cdots + Y_p - Z$. A practical example of this balancing equation is the situation where Z stands for a total obtained from one data set and $Y_1, \dots, Y_p$ are variables obtained from other data sets that should sum up to this total. Unfortunately, the true values of these variables are unknown. These values are estimated by $\hat{Y}_1, \dots, \hat{Y}_p, \hat{Z}$. We will often find that $f(\hat{Y}_1, \dots, \hat{Y}_p, \hat{Z}) \neq 0$. The estimates $\hat{Y}_1, \dots, \hat{Y}_p, \hat{Z}$ are then adjusted, i.e. replaced by values $\tilde{Y}_1, \dots, \tilde{Y}_p, \tilde{Z}$, in order to satisfy the accounting equation, i.e. $f(\tilde{Y}_1, \dots, \tilde{Y}_p, \tilde{Z}) = 0$. A scalar quality measure proposed in QMCM_C_1 is

$$\Delta A = \mathrm{E}\left(\sum_{k=1}^{p} w_k |\tilde{Y}_k - \mathrm{E}\tilde{Y}_k|^{\alpha} + w_{p+1}|\tilde{Z} - \mathrm{E}\tilde{Z}|^{\alpha}\right),$$

where w_k are some positive weights $(k = 1, \dots, p)$, $\alpha = 1$ or $\alpha = 2$, and $\mathrm{E}\tilde{Y}_k$ and $\mathrm{E}\tilde{Z}$ denote the expectations of $\tilde{Y}_k$ $(k = 1, \dots, p)$ and $\tilde{Z}$ under posited models for $\tilde{Y}_k$ $(k = 1, \dots, p)$ and $\tilde{Z}$. The higher the value of ΔA, the more uncertain is the accounting equation, lower is its quality and the quality of the involved statistical figures.

The quality measures proposed in QMCM_C_1 can be used to compare several adjustment methods that can be applied to the same accounting equation. The proposed quality measures can also be used to measure the coherence between various figures that should satisfy an accounting equation.

3.6 Basic Data Configuration 6: Longitudinal data

3.6.1 Covariance matrix for reconciled low frequency and high frequency data

QMCM_A_21 supposes that low frequency aggregated data of high accuracy is available, for instance annual estimates of some indicator. Also, high frequency aggregated data of lower quality from another source is assumed to be available. Results from the high frequency aggregated data should sum up to results from the low frequency data. For example, sums of quarterly indicators should be equal to values of annual indicators. When the low and high frequency data are based on different data sets, this is often not the case. The problem then is to replace the high frequency data with benchmarked values that sum up to the low frequency data. These benchmarked values should differ as little as possible from the initial high frequency data, which is measured by means of a quadratic distance function. This procedure can be formulated as a quadratic optimization problem with linear restrictions.

QMCM_A_21 proposes to use the variance-covariance matrix (or vector of variances) of the reconciled high frequency data as a quality measure for these data.

4 Quality guidelines for frames in social statistics

The description of social phenomena by statistical figures is one of the main functions of national statistical system. Frames are essential to this function. Therefore, the quality guidelines related to this topic is an important step to complete the quality framework work for the production of social statistics.

The document Quality Guidelines for Frames in Social Statistics (QGFSS) consists of five chapters where the first two ones can be seen as introductory, and the chapters three to five contain the actual guidelines.

Starting with the first introductory chapter setting out the purpose of the document, there are three specific objectives of the document. The first one relates to principle 4 of the Code of Practice for

European Statistics “Commitment to Quality” – and specifically indicator 4.1 – together with the stipulation of indicator 7.3 “The registers and frames used for European Statistics are regularly evaluated and adjusted if necessary in order to ensure high quality”. Accordingly, one of the main objectives of this document is *to deliver a building block for safeguarding compliance with the Code of Practice in terms of the construction, use and assessment of frames in social statistics.*

The second specific objective of the guidelines is to provide producers of social statistics with systematic guidance for all process steps relevant for frames. The procedures of NSIs working with frames in social statistics seem to be more heterogeneous than the procedures for economic statistics, where the development and maintenance of the frame (which in most cases is equivalent to the business register) are in some way generic for NSIs all over the world. If we look again at indicator 7.3 of the Code of Practice, it tells us about registers and the “frames for population surveys”. If we look at various NSIs, they offer different scenarios for constructing, using and maintaining frames. So, the second objective of the document *is to provide basic, generic guidance regarding all relevant processes for frames in social statistics in a systematic way, based on agreed definitions and standards.*

Historically the idea of using frames originates from the investigation of social phenomena through surveys, which showed that some kind of list of the units of interest was needed, mainly dwellings aiming to reach households and/or persons living there, from which you can draw a sample and conduct a survey. As a result, the main interest of NSIs was in sampling frames. In recent years it has become clear that sample surveys – while still of significant importance – are not the only way of gathering statistical information. In this regard, the role of frames as a direct source for delivering statistical information becomes more important. Hence, the third specific objective of this document is *to broaden the perception of frames in social statistics so that they can be used as a possible direct source in a multi-source environment.*

Basic definitions and concepts relevant for the subsequent chapters and for the guidelines are provided in the second introductory chapter. The concept of frame is defined in general subsequently extending the considerations to social statistics. At first glance, everybody believes to have an exact and workable understanding of the term *frame*. But the definitions of a frame vary, according to the needs and the intended use of the frame. A frame is any list, material or device that delimits and identifies the elements of the target (survey) population. Depending on the use case, a frame may allow access to, and/or provide additional characteristics of the element. Some basic characteristics were identified which moved definition of frame in the direction of social statistics arriving at the following list:

- the frame is a list of elements available in an usable IT-format;
- the frame aims to map the population of the country of the NSI as accurately as possible;
- the frame contains persons as basic units and households or dwellings as composite units;
- each person household and dwelling has a unique identifier;
- the frame is enriched by auxiliary information enabling a profound use (i.e. at least with contact variables);
- the variables include linking variables which allow connecting persons, households and dwellings to external registers.

The guideline chapters 3 to 5 can be seen as the core of the document and it makes sense to provide them with three general remarks.

Firstly, chapters 3-5 consist of several sub-chapters which follow the same structure:

1) a general overview and a general description of the topic;

2) challenges with respect to the topic of the sub-chapter: What kind of errors can occur due to problems with the topic of the sub-chapter? Which quality dimensions are affected due to these problems?

3) list of the quality guidelines.

Secondly, the aim of the work was to provide guidelines for frames in social statistics without any advice on the processes themselves. If one looks at the chapter dealing with the use of frames in sampling one does not find any guidelines on sampling, for instance advising to stratify population for a certain survey. This created sometimes a kind of rendering problem because it turned out difficult to separate the one for the other. What now is available is sometimes a compromise solution in this regard.

Thirdly, the question of minimum requirements. The idea behind it is to determine for a guideline some set of basic actions/facts/evidence which are a minimal standard in order to comply with the guideline. Some NSIs might have difficulties to reach them. As an example the creation of a minimum standard regarding the very fundamental question of the units included into a frame (dwellings, dwelling and persons, even households) can cause such difficulties. Furthermore it should be mentioned that not every guideline is suitable to be enhanced by minimum requirements.

The topics covered by the guidelines are manifold and organized into three different chapters which are described in more detail.

4.1 Constructing and Maintaining Frames in Social Statistics

This chapter deals with the aspect of how to arrive at a frame suitable for being used within a statistical institute. Five subtopics are addressed here. Since frame construction can be seen as a process of forming an output on a multisource basis the selection and assessment of adequate sources is one of the key elements for a descent output.

Another important issue is the assignment of relevant responsibilities for the coordination of the construction and/or update of the frame. Who is responsible for what? Which experts have to be involved and when? Who is going to trigger which process strep?

One of the key issues when dealing with multiple sources is record linkage, and methods for construction of frames have a strong focus on this topic.

A fundamental question arises: what kind of output is expected from the frame construction process. Several possible scenarios are possible: master frame, specific frames, single frame, multiple frame, list frame, area frame, direct/indirect frame etc.

The chapter finally addresses the update procedure. Most of the times frames are used periodically and therefore the need for descent procedures of updating is a velar prerequisite for the quality of the resulting outputs. The topic is subdivided into three steps of the update procedure namely receiving the input, processing the update and checking the output.

4.2 Use of frames in social statistics

The QGFSS distinguishes between three different forms of frames usage.

The most common form of frame usage is sampling for surveys. For such use cases the talk is going about a sampling frame. It does not need much prophecy to see that there was need to clarify about some terminological issues here: single frame vs. multiple frame, area frame vs. list frame, etc. The guidelines are mainly focusing on how to obtain decent quality under different scenarios and forms

of sampling. Another subchapter deals with contact variables which are of course of significant importance for the practical work.

The second frame use case included concerns the use of frame data in supporting statistical processing. Two processes are addressed here: weighting & calibration and data cleaning (edit & imputation).

As the last form of frame usage, the possibility to use frame data as direct input to compile statistical outputs as part of statistical products is considered. The idea here was to address directly the quality components as defined in the European Statistical System by regulation 223. So, the guidelines are provided how the aim of optimizing the six criteria (relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability and coherence) can be achieved when frame data are used as direct input source for statistics.

4.3 Assessing and evaluation the quality of frames in social statistics

When talking about quality it has to be distinguished between two questions: “How does the use of frame data impact to the quality of the statistical output?” and on the other hand “How can the quality of the frame as a self-standing internal product be assessed?” While the first aspect to some extent is described at the end of the previous chapter the aim of this chapter was to talk about the second aspect.

Firstly the methods to assess the quality of a frame are described. When looking at the different kinds of non-sampling errors relevant for frames, the repository of measures and indicators for certain error-types was created. It is described in more detail in section 5 of this paper. A further idea developed in the QGFSS is to deliver a combined frame quality indicator by taking a weighted sum over the various quality error indicators as defined in section 5 relevant for the specific error types. Finally, this combined quality indicator may be reduced to a specific application in social statistics.

The second part of the frame quality chapter deals with quality and metadata management, quality improvement and quality reporting. Here a general approach for quality assurance, possible methods for improvement on quality of frames as well as an adequate approach for metadata concerning frames is presented. As most important reference a standardized questionnaire for metadata on frames data which are used by the social and population statistics, which was developed by Eurostat, is introduced. It can be found as an annex to the document

5 Quality indicators for frames in social statistics

The report Quality Measures and Indicators of Frames for Social Statistics provides methodological support to Quality Guidelines for Frames in Social Statistics, with respect to quality assessment and evaluation. It can be useful to methodologists when it comes to the design and implementation of such studies, since the relevant topics appear only scattered in the literature otherwise, and not quite up-to-date regarding the various issues one encounters in the context of multisource statistics, which are beyond the traditional uses of frames for sampling. In particular, it can provide valuable inputs to both managers and methodologists, in connection with the creation, maintenance and improvement of an integrated, rich and accurate frame environment, referred to as the Enhanced Population Dataset (EPD), and the resulting Rich Frames for statistical production. An EPD consists of all the available data about relevant units in Social Statistics, such as person, household, dwelling or address, etc. their classification variables and contact information. It may be further enriched by additional demographic, social and economic variables. Central Population Registers (CPR) that exist in a number of European countries are a special case of EPD. Linking patient register, tax

register and the master address file can yield another EPD, especially if the combined data is updated continuously. Apart from content, the different EPDs may have quite different quality levels.

The report has three main chapters. The first one of them covers the definitions of frame and frame errors. The definition of frame has been given earlier in Section 4. Most important is to notice the extension, from the traditional definition of *sampling frame* (e.g. Lessler and Kalsbeek, 1992; Wright and Tsao, 1983), to situations where multiple sources of relevant data in a statistical system are combined and processed, such that it can be directly used to delineate the target population and to make statistics about it. The formulation increases the relevance of the quality framework to register-based census-like statistics, where e.g. population and housing statistics are produced based on integrating data of different types of units. Five types of frame errors are defined:

- *coverage error* due to missing, erroneous and duplicated frame units;
- *domain classification error* of frame units;
- *alignment error* between different types of frame units;
- *unit error* of composite frame units;
- *contact information error* of frame units.

Alignment and unit errors are introduced as additional types of frame errors to the traditional classification. Despite the relevance and importance of composite units (such as household and dwelling, beside person as the base unit) in frames for social statistics, these errors have received insufficient attention in the literature with few exceptions (e.g. Zhang, 2011). Attention to these errors are necessary in the context of frame as a multisource statistical product, e.g. when the dwelling register is combined with the CPR, and with respect to register-based census-like statistics on the relevant topics.

Table 5.1. *List of frame accuracy measurement items*

CM1	Total under- and over-coverage for the target population
CM2	Total correct domain classification
CM3	Domain-specific population under- and over-coverage
CM4.	Domain misclassification (i.e. cross-domain under- and over-coverage)
PM1 – PM4	Counterparts of CM1 – CM4, due to progressive data in the sources
AM1	Total of correctly aligned base units (i.e. persons typically)
AM2	Domain totals of correctly aligned base units
AM3	Distribution of correctly aligned base units by composite unit types
AM4	Total of correctly aligned composite units (e.g. household, address, etc.)
AM5	Domain totals of correctly aligned composite units
UM1	Total number of population composite units
UM2.	Domain total numbers of population composite units
IM1	Total of frame units of given type with (correct, invalid, missing) contact
IM2	Domain totals of frame units with (correct, invalid, missing) contact

The second main chapter covers the items, approaches and methods for frame quality assessment. First, a list of 17 frame accuracy measurement items are given in Table 5.1. Definitions are provided for the corresponding absolute (or relative) values. Instead of laying down absolute thresholds of ‘minimum quality’, it is proposed to have a *minimum* set of items for frame quality assessment. For each item one can obtain either quality measures (estimates with associated uncertainty) or quality indicators (without quantifying the associated uncertainty) depending on the available data, resource and methods.

Table 5.2. *Summary overview of assessment approaches and methods*

Assessment Approach	Coverage & Domain Classification		Alignment and Unit	Contact
	CM1-CM4	PM1-PM4	AM1-AM5, UM1-UM2	IM1-IM2
Coverage or Quality Survey	Using sample from the population: DSE and TDSE; Using sample from the frame: RRC, Census follow-up	---	Quality survey based on audit sample from the frame	As special case of multi-frame sampling
Modelling (only limited application, experience)	For coverage: log-linear models with 3+ lists, latent class (entity) models, etc. For domain classification: misclassification models, Structural Equation Models, etc.	Few existing examples of models for delays	Allocation error model	As special case of log-linear models
On-going Survey	Existing data collection protocol and quality indicators	---	Lack of standard data collection protocol	
Diagnostics	Net or gross discrepancy checks, Sign-of-Life, Quality Indicator System, etc.			

Next, various approaches to frame quality assessment and the associated methods are reviewed and summarised, as shown in Table 5.2. The most readily applicable methods are described in more details, as it may be e.g. the case that an established quality measure method exists for an item but is costly to implement, such as a population coverage survey for CM1. Three approaches are identified as the most promising for developing regular means of frame quality assessment at a lower cost, which are design-based methods applicable to on-going surveys, modelling and diagnostic methods based on multiple registers. Finally, the last main chapter illustrates the three most readily applicable approaches empirically. Depending on the situation and data available, it is shown practically how they can be combined to yield what may be referred to as the hybrid approach. For each approach, the assessed items and additional background and details of the associated methods are described, and the results are summarised and commented, including a short appraisal of the approach given at the end of each relevant section or subsection.

6 The project has ended, and the work can start

The project has finished after almost four years and the combined effort of many people from a number of NSIs. This means that now is the time for all the countries within the ESS (and possibly beyond) to start using the artefacts created within the project to describe the quality of their statistical products based on multiple sources.

In this paper, we present the main products of the ESSnet. They are manuals facing the quality issues of multisource statistics both from a theoretical and practical perspective. Attention is also given to one of the most important multisource products, i.e. frames for social statistics.

To stimulate the use of the results of the ESSnet, a two-day course on the results relating to Quality Guidelines for Multisource Statistics (QGMSS) and Quality Measures and Computation Methods (QMCMs) was prepared and delivered in September 2019 at the Eurostat premises. This course (in a slightly shortened version) will be offered again in connection with the next European Conference on Quality in Official Statistics (Q2020) in June 2020 in Budapest.

Some actions at NSI and Eurostat level can be further launched to foster the results of this project. Individual NSIs willing to adopt the guidelines can experiment with them and provide feedback for their update. At NSI level, training programs based on the guidelines can be developed and carried

out. Eurostat could sponsor the development of checklists for the guidelines to be used for the assessment at single process level. The quality framework and applications developed for the quality of multisource statistics, has not been designed with big data in mind. However, they can be evaluated in the field of big data in the light of their applicability, extension and alignment with the work that is already being carried out.

In conclusion we can truly state that there is still plenty of work left for the future, both regarding implementation of the results at the national and ESS level, but also further methodological work within the established framework. Overall, the project has ended and now the work can start.

References

Burg, T. and M. Six (2018) *Quality Guidelines for Frames in Social Statistics – The "Making-Of".* Paper presented at the European Conference on Quality in Official Statistics –Q2018 – Cracow.

Daas, P. and S. Ossen (2011). *Report on methods preferred for the quality indicators of administrative data sources*. Blue - ETS Project, Deliverable 4.2. Available at: http://www.pietdaas.nl/beta/pubs/pubs/BLUE-ETS_WP4_Del2.pdf (accessed December 2019).

De Waal, T., A. van Delden and S. Scholtus (2019a), Multi-Source Statistics: Basic Situations and Methods. *International Statistical Review* (in press). (https://doi.org/10.1111/insr.12352).

De Waal, T., A. van Delden and S. Scholtus (2019), Quality Measures for Multi-Source Statistics. *Statistical Journal of the IAOS,* 35(2), pp. 179–192.

Eurostat (2014). *ESS Handbook for quality reports*. Available at: http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf (accessed December 2019).

European Commission (2019). *ESSnet on Quality of Multisource Statistics – Komuso*. https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en.

Eurostat (2003). *Definition of Quality in Statistics*. Luxembourg, October 2-3. Available at: https://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf (accessed October 2019).

Eurostat (2014). *ESS Handbook for quality reports*. Available at: http://ec.europa.eu/eurostat/documents/3859598/6651706/KS-GQ-15-003-EN-N.pdf (accessed October 2019).

Gerritse, S., P. G. M. van der Heijden and B.F.M. Bakker (2015), Sensitivity of Population Size Estimation for Violating Parameter Assumptions in Log-linear Models. *Journal of Official Statistics,* 31(3), pp. 357-379.

Groves R. M., F. J. Jr. Fowler, M. Couper, J. M. Lepkowsky, E. Singer and R. Tourangeau (2004). *Survey Methodology*., 2nd edition, Hoboken, New Jersey, John Wiley & Sons.

Lessler, J. T. and W. D. Kalsbeek (1992). *Nonsampling Error in Surveys*. Wiley.

Wright, T. and H. J. Tsao (1983). A Frame on Frames: An Annotated Bibliography. In Tommy Wright, ed. *Statistical Methods and the Improvement of Data Quality.* Orlando, FL: Academic, pp. 25-72.

Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, 27(3), pp. 415-432.

Zhang L.-C. (2012), Topics of statistical theory for register-based statistics and data integration, *Statistica Neerlandica*, 66(1), pp. 41-63.

Book and Software Review

A Checklist for Assessing the Analysis Documentation for Public-Use Complex Sample Survey Data Sets

Stanislav Kolenikov[1], Brady T. West[2], Peter Lugtig[3]

[1] Abt Associates, USA, skolenik@gmail.com

[2] Survey Research Center, University of Michigan, USA, bwest@umich.edu

[3] Utrecht University, The Netherlands, p.lugtig@uu.nl

Abstract

We document our understanding of, and recommendations for, appropriate best practices in documenting the complex sampling design settings for statistical software that enables design-based analyses of survey data. We discuss features of complex sample survey data such as stratification, clustering, unequal probabilities of selection, and calibration, and outline their impact on estimation procedures. We provide assessment guidelines and a checklist that will aid complex sample survey data providers in aligning their level of documentation with best practices and show how existing surveys and their documentation score based on these guidelines.

Keywords: design-based inference, population surveys, statistical software, complex samples, Total Survey Error.

1 Principal sampling design features

A principal objective in survey research is to develop survey designs that minimize Total Survey Error (TSE; Groves and Lyberg 2010). Sampling and adjustment errors are two of the errors within the larger TSE framework that can be internally quantified in statistical software. When coverage and nonresponse errors can be estimated as well, there are possibilities to adjust errors in order to ensure that the analysis of the survey represents the larger population. If this is done well, the results from the survey analysis are asymptotically unbiased with respect to the sampling design, while uncertainty due to the various errors can be estimated as well. In this paper, we focus on the "big four" features of complex sampling designs: stratification, cluster sampling, unequal probabilities of selection, and weight adjustments. Each design feature is described in more detail below. Although we discuss the possible reasons why one would use a particular survey design, we refer to Groves et al (2011), Lohr (2010), Biemer & Lyberg (2003), Groves (2004) or other textbooks on survey design for a broader context and overview of sampling design decisions.

1.1 Stratification

Stratification divides the population and sampling frames into mutually exclusive groups (strata) before sampling. Common examples of strata include:

- Geographic regions for in-person samples;
- Diagnostic groups for patient list samples; and
- Industry, employment size and/or geographical regions for establishment samples.

Complex sampling designs employ stratification to:

- Oversample subpopulations of interest (e.g. ethnic minorities) if they can be identified on the frame(s);
- Oversample areas of higher concentration of the target rare population;
- Ensure specific accuracy targets in subpopulations of interest;
- Utilize different sampling designs/frames in different strata;
- Avoid outlying samples and spread the sample across the whole population;
- Optimize costs vs. precision via Neyman-Chuprow or more complicated allocations.

Unless stratification is primarily intended to oversample subpopulations of specific interest, it can be expected to lead to reductions in sampling variances. In many human populations, these efficiency gains are modest, but they can be very substantial in establishment populations.

1.2 Cluster Sampling

Cluster, or multistage, sampling design consists of sampling groups of observation units (clusters), rather than the ultimate observation units directly. From a statistical efficiency viewpoint, this is a less desirable feature as clustering of units that have similar characteristics reduces precision of survey estimates. Common examples of randomly sampled clusters include:

- Geographic units (e.g., census tracts, enumeration districts) in face-to-face surveys;
- Entities in natural hierarchies (e.g. health care providers within practices within hospitals, or students within classes within schools).

Why do complex sampling designs employ cluster sampling?

- Complete lists of all units are not available, but survey statisticians can work with lists of administrative units (e.g., states, counties, Census tracts, enumeration districts) for which membership of the next stage sampling units can be clearly established;
- Reduce interviewer travel time/cost in face-to-face surveys;
- Substantive researchers have an analytic interest in multilevel modelling of hierarchical structures.

1.3 Unequal Probabilities of Selection

In practice, sampling designs introduce unequal probabilities of selection for different sampling elements. From a solely statistical perspective, this is a less desirable feature as larger variances in weights across cases reduce the precision of survey estimates.

Complex sampling designs can assign unequal probabilities of selection to different population units to achieve several goals. Commonly, unequal probabilities result from implementation of a primary sample size target. First, when (smaller) subpopulations of interest (e.g., ethnic/racial minorities) that would not have sufficient sample sizes in an equal probability of selection method (epsem) sample are oversampled directly from lists, or indirectly oversampled by selecting geographic areas with a higher concentration of the target rare population, unequal probabilities of selection would result. Second, most samples for face-to-face surveys are designed with probability proportional to size (PPS) sampling at the first stages, with fixed sample size at the final stage to achieve an approximately epsem design. In many cases, however, the sample size at the final of selection (e.g. the size of a household) is unknown in advance, leading to different weights for the units of observation. Third, unequal probabilities are nearly inevitable in multiple frame sampling, where units can be sampled through several possible channels. In phone surveys, dual phone users, i.e., those who have both landline and cell phone service, are more likely to be selected than those who have cell only or landline only service.

1.4 Weight Adjustments

After the data are collected, survey statisticians further adjust the weights to make appropriate corrections (see Valliant and Dever, 2018, for details). These adjustments generally account for:

- (non-)Eligibility;
- Frame noncoverage;
- Frame overlap in multiple frame surveys;
- Statistical efficiency;
- Unit nonresponse.

Weight adjustments are done out of necessity, and typically aim to reduce noncoverage and nonresponse biases. However, these improvements generally come at the expense of an increase in sampling variances. Some exceptions are possible with weight calibration to population totals when outcomes of interest are strongly correlated with the calibration variables.

1.5 Sampling is about doing the best job for the money

All the complex sampling features described above are ultimately employed to collect data in more efficient and cost-effective ways. These efficiencies come with statistical trade-offs, however. While the use of cluster samples would allow survey designers to save on travel costs, precision of the estimates will be worsened due to intracluster correlations. However, if travel costs are reduced by a factor of five, and the reduction in statistical efficiency is by a factor of two, then undoubtedly a cluster sampling design is the more economical one in units of precision per dollar. In most general population samples (except some European countries with excellent population registers), there is no access to the full population listing, forcing survey designers either to use area samples to gradually gain access to individuals, or use an infrastructure created for a different purpose (phone communication or postal service) to contact potential respondents. Obtaining a full population list to sample from would be a prohibitively expensive exercise.

When studying populations that are subsets of the general population (e.g., families with children; religious minorities; military veterans; and many other special populations), survey statisticians may have multiple ways to reach these populations by screening out a larger, general population sample, or through the social systems associated with that population (e.g., daycare centers and schools to reach children). Those different frames may have different costs of identifying eligible units but may have to be used in conjunction to ensure complete coverage of a given population and correct inference. As an example, home-schooled children can only be found in a general population sample that may be more expensive than a school-based sample. In studies of rare populations, the variance in weight factors will inevitably arise as a function of different screening rates, different coverage of the various frames used, and stratification of the frames oversampling areas of higher concentration of the population of interest that would allow to collect data less expensively.

As a result of all the considerations above, population surveys employ complex sampling designs in their fieldwork. Data resulting from such complex surveys cannot be naively analysed as is, and survey weights and possibly other elements of the complex sampling design have to be accounted for. Survey statisticians routinely compute weights for data users. These weights often take the form of a design weight that corrects for eligibility, frame overlap, and unequal selection probabilities in sampling. A separate nonresponse weight corrects for nonresponse, and sometimes for noncoverage errors in the frame used. In some surveys additional weights are provided for the purpose of doing cross-national comparisons (multi-country surveys) or longitudinal analysis (cohort or panel studies). For more information on how modern surveys are efficiently designed, and weights are computed, we refer the reader to Kalton, Flores-Cervantes (2003), Lohr (2010), Bethlehem (2010), Valliant, Dever, Kreuter (2013), Valliant and Dever (2018) or Kolenikov (2016). The weights included in a survey dataset should be accompanied with detailed documentation on how the weights

were computed and how they should be used in practice by applied researchers. We have often found that the documentation of survey weights is inadequate. Sometimes, details on how the weights were designed are missing. More often, the description of the weights is sparse or very technical. This then leads to users not using weights at all or using them incorrectly. West, Sakshaug and Aurelien (2016) have shown for example that analytic errors are prevalent in 145 analyses of the survey 'Scientists and Engineers Statistical Data System' (SESTAT). They reported that "… only 55% of the products incorporated the publicly-available sampling weights into the analyses, only 8% of the products accounted for the complex sampling features when estimating variances, and only 11% of the products presenting design-based analyses performed appropriate subpopulation analyses accounting for the complex sampling". In the medical domain, Khera et al. (2017) reported that "a total of 79 [out of 120] sampled studies (68.3% [95%CI, 59.3%-77.3%]) among the NIS studies screened for eligibility did not account for the effects of sampling error, clustering, and stratification".

Ignoring survey design weights will lead to wrong inferences. Data users therefore will need to know why and how to use weights that are being provided with the public-use files of large survey data. Simultaneously, survey designers and methodologists need to document how these weights are being produced and provide guidance to users on how to use weights in practice. This paper therefore seeks to provide rubrics for how survey weights and sampling design settings should be documented for the ultimate survey data users. We will define a set of assessment guidelines consisting of five main elements and two bonus elements, and then use these guidelines to discuss the survey documentation of several popular surveys originating in the U.S., U.K., and Europe.

1.6 Scale of weights

The purpose of weights in analysis of complex survey data is to provide the foundation for finite population inference, with the Horvitz-Thompson estimators of totals being the basic building blocks. Within the finite population inference paradigm, the sum of weights is the population size (known or estimated). We however have encountered data sets, more likely in social sciences and political polling, that are provided with weights that sum up to the sample size. We believe the use of this convention dates back to early statistical software design and operations, and represents outdated practice. For example, in the early days of SPSS in the 1970s, it introduced frequency weights. In order to get inference approximately right, these frequency weights had to sum up to the sample size to get the standard errors approximately right, as $1/\sqrt{n}$ rather than $1/\sqrt{N}$ where n is the sample size and N is the population size, or the appropriate sum of weights. There is no need to have this as a convention in most of modern software that implements appropriate variance estimation methods such as linearization or replicate variance estimation, and we would advocate that the data providers release weights that sum up to the population size (especially for data users interested in estimating population totals).

2 Selected statistical software for design-based analysis

We briefly review the capabilities of selected software packages for performing design-based analyses of complex sample survey data below. For working examples of the code that one would use in each of these packages to perform common design-based analyses of survey data, please visit the web site https://github.com/skolenik/svyset_manifesto. See West et al. (2018) for additional details regarding other software packages that facilitate design-based analysis.

2.1 R

R (R Core Team 2019) is a free, open-source software environment for statistical computing and graphics. The base R system provides the computational background and a minimal set of statistical computing procedures (e.g., distributions), while most of the functionality exists in third party packages. Implementation of complex sample survey estimation in *library(survey)* (Lumley 2010) separates the steps of declaring the sampling design and running estimation. All typical designs and

variance estimation methods are supported: simple random sample; stratified random sample; unequally weighted designs; two-stage designs; calibrated weights; two phase samples; designs with jackknife, BRR, bootstrap and arbitrary replicate weights. Fundamental statistical methods are supported: descriptive statistics, estimation for domains, generalized linear models, contingency table analysis, survival analysis, quantile and distribution function estimation.

An alternative package is *library(ReGenesees)*. It is not as easily accessible and not as regularly updated as the *survey* package.

2.2 Stata

Stata (StataCorp 2019) is a commercial package that provides most of the functionality through the official release, but also provides ways for the third-party developers to code their commands that are indistinguishable from the native Stata commands, at least by syntax. In Stata, survey settings can be specified once with the *svyset* command, and be used later with the *svy*: estimation prefix. The settings can be saved with the data set, so that the end users do not have to take this step on their end. This is a highly recommended best practice for data providers. All typical designs and variance estimation methods are supported: simple random sample; stratified random sample; unequally weighted designs; two-stage designs; designs with jackknife, BRR and bootstrap replicate weights. Estimation with calibrated weights is supported for versions of Stata 15.1 and above. A broader range of statistical methods is supported: descriptive statistics, estimation for domains, generalized linear models, contingency table analysis, survival analysis, generalized structural equation models; multilevel mixed effects model; finite mixture models; a variety of econometric models such as binary, discrete response, and sample selection models.

2.3 SAS®

SAS software (SAS Institute 2019) is a commercial statistical package. Nearly all statistical functionality is implemented via procedures (*PROC*) developed by SAS Institute. In SAS software, survey settings need to be declared in every *SURVEY* procedure. All typical designs and variance estimation methods are supported: simple random sample; stratified random sample; unequally weighted designs; two stage designs; designs with jackknife and BRR replicate weights. The bootstrap replicate weights can be incorporated by a shortcut (Phillips 2004). A limited range of statistical methods is supported: descriptive statistics, estimation for domains, generalized linear models, contingency table analysis.

3 Documentation on appropriate design-based analysis techniques for complex sample survey data: an assessment checklist

Large scale data collections are nowadays routinely released to the public. They typically include anonymized, public-use survey microdata, along with some variables that include details about the fieldwork itself, and one or several weighting variables that allow any data user to correct for unequal sampling probabilities introduced in the survey design, as well as noncoverage and nonresponse errors. The survey datasets are accompanied with survey documentation that purports to explain the design of the survey and detail the measurements taken. In this section we propose a short checklist to assess the quality of survey documentation concerning survey design features specification in software. At the moment, this list is reactive, i.e. to be used to assess the existing documentation. We hope that in the future, this list can be used proactively, so that organizations producing survey data and its documentation can make sure their data products are sufficiently user-friendly.

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** This would be a person with training on par with or exceeding the level of

the Lohr (2010) or Kish (1965) textbooks, and applied experience on par with or exceeding the Lumley (2010) or Heeringa, West and Berglund (2017) books.

2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** This would be a person who has some background / training in applied statistical analysis, but has only cursory knowledge of survey methodology, based on at most several hours of classroom instruction in their discipline "methods" or “metrics” class, or a short course at a conference.
3. **Is relevant survey information described succinctly in one place, or scattered throughout the document?** It is of course easier on the user when all the relevant information is easily available in a single section. However, some reports put information about weights in one place, e.g. where sampling was described, while information about other complex sampling features (e.g., cluster/strata/variance estimation) only appears some twenty pages further.
4. **Are examples of specific syntax to specify survey settings provided?** Has the data producer provided worked and clearly-annotated examples of analyses of the complex sample survey data produced by a given survey using the syntax for existing procedures in one or more common statistical software packages? And as a bonus, have examples been provided in multiple languages (e.g., SAS, R, and Stata)?
5. **Are there examples given for how to answer substantive research questions?** In all statistical languages, there are specific ways to run commands that are survey-design-aware. In other words, only specifying the design may not be sufficient in ensuring that estimation is done correctly. For instance, are examples provided for both descriptive and analytic (i.e., regression-driven) research questions?
6. (Bonus) **Is an executive summary description of the sampling design available?** Many researchers would appreciate a two-to-three sentence paragraph to summarize the sampling design that they could copy and paste into their papers, e.g.,

{This survey} is a three-stage area sampling design survey with census tracts, households, and individuals as sampling units. The final analysis weights provided by {the organization who collected the data} account for unequal selection probabilities, nonresponse, and study eligibility, and are used in all analyses reported in this paper. Standard errors are estimated using bootstrap variance estimation procedures designed for complex surveys.

7. (Bonus) **What kinds of references are provided?** It is often helpful to the end users if the description of the sampling design features is accompanied by the references to (a) methodological literature describing them in general, and (b) technical publications specific to the study in question, such as the JSM or AAPOR proceedings, technical reports on the provider website, or publications in technical literature describing the study, if appropriate. For example, the description of clustered sampling designs used in the U.S. Census Bureau large scale surveys such as the American Community Survey or Current Population Survey could refer to general descriptions of stratified clustered surveys, to the user Handbooks (Census Bureau 2009), and to the technical papers on variance estimation (Ash 2011).

We now use the seven questions posed above to "score" several existing examples of documentation for public-use survey data files based on these criteria. For example, if the documentation for a public-use data file successfully satisfies / meets the first five guidelines above, the documentation will be scored 5/5. If documentation scores positively on items 6 and/or 7, it will additionally be awarded + for one of these or ++ for both of these items. Thus, for instance, a documentation set that is aimed solely at survey statisticians without any software examples and cites the existing literature extensively will likely get a score of 2+, while a documentation set using simple language with many code examples may get a score of 5.

These scores are designed to be **illustrative**, in terms of rating existing examples of documentation for public-use data files on how effectively they convey complex sampling features and how they should be employed in analysis to users. The scores are designed to motivate data producers to improve the clarity of their documentation for a variety of data users hoping to analyse large (and usually publicly-funded) survey data sets.

3.1 Practical strategies for extracting survey design settings from existing documentation

When asked to analyse an existing data set that features complex survey data, we typically rely on a number of heuristics to figure out what the survey statisticians intend for the ultimate data users to do.

1. Search the documentation for the software footprints as keywords: *svyset* per Stata, *PROC SURVEY* per SAS, *svydesign* per R *library(survey)*.
2. If that fails, search for "sampling weight", "final weight", "analysis weight", "survey weight" or "design weight". You can search for "weight" per se but you should expect that this is likely to produce many false positives (e.g., weight as a physical measurement in kg), especially in health studies.
3. See if there is any description of the sampling strata and clusters near the text where weights are mentioned.
4. Search for "*PSU*" and "*cluster*" and "*strata*" and "*stratification*" to find the variables that needed to be specified in survey settings.
5. Search for "*variance estimation*", the generic technical term to deal with complexities of survey estimation.
6. Search for "*replicate weights*", "*BRR*", "*jackknife*" and "*bootstrap*", the keywords for the popular replicate variance estimation methods.

In reviewing weighting documentation of existing surveys, we have also encountered more obscure language such as “pseudovalues”, “pseudostrata”, “pseudounits”, “variance replicates”, “variance units”, “pseudoreplicates” and some other terms indicating that the variables provided for variance estimation may not be the true sampling design variables. While technically correct, such language does little to help an inexperienced user in identifying the relevant settings to be applied, primarily through a disconnect between the “textbook” terms and the terms used in documentation.

4. Evaluating documentation in practice

In this section, we will evaluate a convenience sample of the documentation for several public use survey data files (PUFs). The goal of this section is not to provide our overall assessment of weighting procedures across all datasets; we merely want to illustrate how several large-scale and much used survey datasets have described what was done in their complex sampling designs and corrections. We will apply the above checklist questions to see how the documentation compares in terms of effectively describing appropriate analysis techniques to data users. Additional examples, including reviews of documentation with lower ratings, are available at the main project webpage, https://github.com/skolenik/svyset_manifesto.

4.1 The National Survey of Family Growth (NSFG), 2013–2015

Rating: ⋆ ⋆ ⋆ ⋆ ⋆

Funding:

- Eunice Kennedy Shriver National Institute of Child Health and Human Development
- Office of Population Affairs
- NCHS, CDC
- Division of HIV/AIDS Prevention, CDC
- Division of Sexually Transmitted Disease Prevention, CDC
- Division of Reproductive Health, CDC
- Division of Birth Defects and Developmental Disabilities, CDC
- Division of Cancer Prevention and Control, CDC
- Children’s Bureau, Administration for Children and Families (ACF)

- Office of Planning, Research and Evaluation, ACF

Data collection: The University of Michigan Survey Research Center (http://src.isr.umich.edu)

Host: The National Center for Health Statistics (http://www.cdc.gov/nchs/)

URL: http://www.cdc.gov/nchs/nsfg

Assessment Checklist:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. Electronic documents like Example 1: Variance Estimates for Percentages linked from the documentation page under *Variance estimation* subtitle make it very easy for survey statisticians and applied researchers alike to correctly declare complex sampling features to survey analysis software for design-based analyses.
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes. See above.
3. **Is relevant survey information that the data user needs to know about the complex sampling contained in one place?** Yes, although very little (if anything) is said about the actual complex sampling design. Instead this information appears in separate electronic files, such as Sample Design Documentation. This is out of necessity, however, given the complexity of the NSFG sampling design, and all of the information that a user needs to compute weighted point estimates and estimate variance accounting for the complex sampling can be found in examples like the one indicated above.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** Yes. Three examples are clearly documented (tabulations for categorical variables; means for continuous variables; analysis with domains/subpopulations) and linked on the main documentation page, and both syntax and output are included in each case. Bonus: syntax and output are provided for both SAS and Stata.
5. **Are examples of analyses given for addressing specific substantive questions provided?** Yes; see previous item.
6. **(Bonus) Is an executive summary of the sampling design provided?** Yes; such an executive summary is given in the first section of the main sample document.
7. **(Bonus) What kinds of references are provided?** There are several references to the most important sampling design literature included in Section 11 of the document linked above.

Score: 5++/5

The NSFG provides an excellent example of the type of documentation that needs to be provided to data users to minimize the risk of analytic error due to a failure to account for complex sampling features. *Accessed on 2018-07-15.*

4.2 The Population Assessment of Tobacco and Health

Rating: ⋆ ⋆ ⋆ ⋆ ⋆

Funding: The Population Assessment of Tobacco and Health (PATH) Study is a collaboration between the National Institute on Drug Abuse (NIDA), National Institutes of Health (NIH), and the Center for Tobacco Products (CTP), Food and Drug Administration (FDA).

Data collection: Westat (http://www.westat.com)

Host: The National Addiction and HIV Data Archive Program

URL: https://www.icpsr.umich.edu/icpsrweb/NAHDAP/series/606

Assessment Checklist:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. Section 5 of the Public-Use Files User Guide provides clear detail

on the calculation and names of the various weight variables that can be used for estimation. This section also discusses variance estimation, and clearly describes the replicate weights that have been prepared for data users enabling variance estimation. Software options are also discussed in this section, and code illustrating the use of multiple programs for the protype example analyses is provided in Appendix A.

2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes. Appendix A of the User Guide is very helpful, given that it provides annotated example code for several different packages. Section 5 is aimed at survey statisticians and will be overwhelming to an audience that is less technically prepared.
3. **Is relevant survey information that the data user needs to know about the complex sampling contained in one place?** Yes; Section 5 provides all the necessary sampling information for analysis purposes, and Appendix A contains all of the necessary code for actual practice.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** Yes. Appendix A of the Public-Use Files User Guide is an excellent example of providing this kind of resource for data users.
5. **Are examples of analyses given for addressing specific substantive questions provided?** Yes. Appendix A illustrates a variety of potential analyses that data users could perform.
6. **(Bonus) Is an executive summary of the sampling design provided?** Chapter 2 of the User Guide provides a detailed summary of the sampling design, which serves as an executive summary.
7. **(Bonus) What kinds of references are provided?** There are several references to the most important sampling design literature included at the end of the User Guide.

Score: 5++/5

The PATH PUF user guide is another excellent, gold-standard example of detailed and useful information designed to make the life of the survey data user easier. *Accessed on 2018-12-17.*

4.3 European Social Survey (Round 8)

The ESS represents an interesting example of a survey on which we had observed tangible improvements in documentation throughout the lifetime of our project. Weighting documentation is provided per round of the ESS, as sampling procedures and nonresponse adjustments differ slightly by round. The nature of the documentation has also changed, however. For rounds 1-7, the documentation that is available is written mainly for survey statisticians. For round 8, there is a different setup, with more elaborate and more accessible information. During our work on this paper, we bumped our rating up for this survey, owing to improvements in the documentation.

Funding: European Commission, Horizon 2020. Rounds 1-8 of ESS have been founded by national science foundations and/or European national governments.

Data collection: coordinated by City University, London, UK. Data collection in separate European Countries coordinated within every country.

Host: European Social Survey, formerly at Norwegian data Archive

URL: http://www.europeansocialsurvey.org/

Rating: ⋆ ⋆ ⋆ ⋆

Weighting documentation (general):
http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf

Round 8 User Guide:
https://www.europeansocialsurvey.org/docs/round8/methods/ESS8_sddf_user_guide_1_1.pdf

Assessment Checklist (Round 8):

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes. The European Social Survey is a repeated cross-sectional study conducted in about 30 different countries in Europe. Sampling is conducted within every country, using either listing methods or registers (of individuals or addresses). Three weights (design, poststratification and population equivalence weights) are included in the main data file. This allows for Horvitz-Thompson estimation, but not the specification of a complex survey design. However, an Integrated Sample data file does include information on stratification or cluster variables, as well as selection probabilities for every respondent. On top of this, a multilevel file adds regional indicators to the main datafile, allowing for multilevel-analysis
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes, three weights are provided: a design weight, a poststratification weight and a population equivalence weight. Guidance is included on how to combine the three weights, and when to use what weight in some examples of analyses. ESS Round 8 documentation discusses the sampling design variables such as strata and clusters.
3. **Is relevant survey information that the data user needs to know about the complex sampling contained in one place?** Documentation is scattered across many different documents and files on the ESS website. One good aspect of the European Social Survey is that the users are explicitly warned that data need to be weighted when data are downloaded from the ESS website. The Round 8 User Guide does compile the description of all the design variables. It is unclear whether users of other rounds will stumble upon it, however.
4. **Are examples of specific syntax for performing correct design-based analyses provided?** Yes. Box 2 in Section 3.2 “Estimating standard errors” of the Round 8 User Guide provides Stata *svyset* syntax.
5. **Are examples of analyses given for addressing specific substantive questions provided?** Yes. Box 3 in Section 3.2 “Estimating standard errors” of the Round 8 User Guide provides Stata syntax to obtain design-adjusted estimates, however the syntax is incorrect as it uses subsetting of the data rather than subpopulation/domain estimation (West, Berglund and Heeringa 2008). The subsequent discussion of the differences between naïve and design-adjusted estimates is very helpful.
6. **(Bonus) Is an executive summary of the sampling design provided?** There is an executive summary that describes the basic sampling methodology. There is no easily accessible executive summary that explains how and why sampling differs over the countries.
7. **(Bonus) What kinds of references are provided?** There are references to standard textbooks on complex survey design, and references to other documents on the ESS website, with more detailed documentation.

Score: 4/5

The ESS provides a mix of legacy documentation written by survey statisticians for survey statisticians, and the more recent documentation aimed at the non-statistical users. The use of multi-country, multi-round data sets remains very complex, however. *Accessed on 2019-09-16.*

4.4 A Portrait of Jewish Americans

Rating: ⋆ ⋆ ⋆ ⋆

Funding: The Pew Research Center’s 2013 survey of U.S. Jews was conducted by the center’s Religion & Public Life Project with generous funding from The Pew Charitable Trusts and the Neubauer Family Foundation.

Data collection: Abt SRBI under contract to Pew Research Center

Host: Pew Research Center http://www.pewresearch.org/

URL: http://www.pewforum.org/dataset/a-portrait-of-jewish-americans/

Assessment Checklist:

1. **Can a survey statistician figure out from the documentation how to set the data up for correct estimation?** Yes; survey documentation explains the differences between the household and the person-level weights, and stresses that the bootstrap weights should be used for variance estimation.
2. **Can an applied researcher figure out from the documentation how to set the data up for correct estimation?** Yes; Stata syntax is provided early in the document, or can be found by search in the PDF file.
3. **Is relevant survey information described succinctly in one place, or scattered throughout the document?** Yes; all the relevant information is contained in the **Key Elements of the Data** section in about 2 pages.
4. **Are examples of specific syntax to specify survey settings provided?** Yes; item 6 of **Key Elements of the Data** section identifies the variables and provides Stata syntax for individual level and household level analyses. (Search for any of *Stata*, *SAS*, *weight*, *svyset* would lead the researcher to this information.) A warning is given that SPSS Statistics Base package cannot correctly compute standard errors.
5. **Are there examples given for how to answer substantive research questions?** No examples are given.
6. (Bonus) **Is an executive summary description of the sampling design available?** Sampling design is described in painstaking detail in about 9 pages. No short summary of the design is available from the technical documentation, although such a summary can be found in the substantive report (Pew Research Center 2013).
7. (Bonus) **What kinds of references are provided?** No additional references are given.

Score: 4+/5

A Portrait of Jewish Americans is a very well described survey that most researchers will be able to analyze correctly by following the instructions of the data provider. Slight limitations of the documentation are that examples of the settings are only given for one package, Stata, and no examples of substantive analyses, e.g. those leading to the headline tables in the substantive report, are provided. *Accessed on 2018-12-11*.

5 Other considerations in data usability

Reviewers of this paper brought up several other dimensions of data usability that applied researchers will face. Among them are documentation of missing data and ease of access to the data itself.

Missing data always presents a threat to research validity. Depending on the research question at hand, researchers may choose to use, or develop new, estimation and testing strategies. For example, a popular econometric model of sample selection, the Heckman model, grew out of a not-missing-at-random missing data problem, and has become a staple in applied social science research. In many typical situations, unit nonresponse (when none of the survey data were collected from a sampled unit) is handled by nonresponse adjustments and weight calibration, performed by the data provider; and item nonresponse (when some but not all variables are observed for a sampled unit) is handled by imputation (which can be performed by the data provider or by the data user). While we agree that documentation of missing data approaches deserves its own discussion, this is not focus of this paper.

The mode of access to the microdata varies vastly across surveys and providers. Some data sets (e.g. the American Community Survey) are available in open access. Some data sets, typically by academic and non-profit providers, require email registration and a minimal data use agreement; the essence of such agreements is captured by the motto on the IPUMS (Integrated Public Using Microdata Series; a collection of data sets by the U.S. federal agencies as well as about 100 censuses from around the world): “Use it for good, never for evil.” Access to the data through secure

facilities such as Research Data Centers in the U.S. or Canada often allows access to the full rather than abridged design variables, e.g., all three or four stages of selection, which usually allows the user to produce smaller standard errors. The process of access to these restricted data is tedious: the researcher needs to write an extensive formal application, which has to pass multiple rounds of confidentiality review; the researcher needs to be present in a special physical facility with secure access; and any statistical output to be taken out of this facility is subject to review by confidentiality officers of the agency that grants access. Again, the ease of access is not the focus of this paper.

6 Online materials

This paper is accompanied by two web sites. The first is a periodically updated repository that contains an earlier version of this text, code examples, and evaluation of the survey design settings in the documentation for a number of surveys. This repository is available at https://github.com/skolenik/svyset_manifesto. The second is an R Shiny app, where applied researchers can paste example code from SAS, Stata and R and generate corresponding code in other software packages to facilitate correct design-based analyses in the future. Please visit https://statstas.shinyapps.io/svysettings/ for details.

References

Ash, S. (2011). Using Successive Difference Replication for Estimating Variances. *Proceedings of the Survey Research Methods Section*, 3534–3548, American Statistical Association, Alexandria, VA.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78 (2), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality* (Vol. 335). John Wiley & Sons.

Bryan, J. (2017) Project-oriented workflow. Technical report https://www.tidyverse.org/articles/2017/12/workflow-vs-script/.

Census Bureau (2009). *What Researchers Need to Know: ACS Handbook*. Technical Report. https://www.census.gov/library/publications/2009/acs/researchers.html.

Groves, R. M. (2004). *Survey errors and survey costs* (Vol. 536). John Wiley & Sons.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology* (Vol. 561). John Wiley & Sons.

Groves, R., and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly, 74 (5), 849–879. https://doi.org/10.1093/poq/nfq065.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis*. Chapman and Hall/CRC.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19 (2), 81–97.

Khera, R., et al (2017). Adherence to Methodological Standards in Research Using the National Inpatient Sample. Journal of the American Medical Association, DOI: 10.1001/jama.2017.17653.

Kish, L. (1965) *Survey sampling*. New York: Wiley.

Kolenikov, S. (2016). Post-stratification or non-response adjustment? *Survey Practice*, 9 (3). https://doi.org/10.29115/SP-2016-0014

Lohr, S. L. (2010). *Sampling: Design and Analysis: Design and Analysis*. Chapman and Hall/CRC.

Lumley T (2010).*Complex Surveys: A Guide to Analysis using R*. John Wiley & Sons, Hoboken, New Jersey.

Pew Research Center (2013). A Portrait of Jewish Americans. Technical report, available at http://www.pewforum.org/2013/10/01/jewish-american-beliefs-attitudes-culture-survey/.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

SAS Institute (2019). Base SAS 9.4; SAS/STAT 14.1 User Guide. Cary, NC.

StataCorp, L. P. (2019). Stata statistical software: Release 16. *College Station TX*.

Valliant, R., & Dever, J. A. (2018). *Survey weights: a step-by-step guide to calculation*. College Station, TX: Stata Press.

Vallian, R., Dever, J., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer, New York.

West, B. T., Berglund, P., & Heeringa, S. G. (2008). A Closer Examination of Subpopulation Analysis of Complex-Sample Survey Data. *The Stata Journal*, 8 (4), 520–531. https://doi.org/10.1177%2F1536867X0800800404.

West, B. T., Sakshaug, J. W., & Aurelien, G. A. S. (2016). How big of a problem is analytic error in secondary analyses of survey data?. *PloS one*, *11*(6), e0158120.

West, B.T., Sakshaug, J.W., and Aurelien, G.A. (2018). Accounting for Complex Sample Design Features in Survey Estimation: A Review of Current Software Tools. *Journal of Official Statistics*, 34(3), 721-752.

ARGENTINA

Reporting: **Verónica Beritich**

Statement by Martine Durand, OECD Chief Statistician and Director of the Statistics and Data Directorate, on the report OECD Assessment of the Statistical System of Argentina and Key Statistics of Argentina

Argentina is living through a period of great and complex challenges. As it confronts these challenges, a strong national statistical system and high-quality official statistics are essential to inform policy and decision makers, academics, the media, and the general public on the state of the country in all its economic, environmental and social dimensions.

To be useful, and trusted, official statistics must respect internationally accepted principles and methods. However, as statistics are constantly evolving to meet new needs and measure new phenomena, ensuring their quality is always a work in progress.

As a contribution towards Argentina's efforts in the statistical field, I am pleased to announce today the publication of a new Assessment of the Statistical System of Argentina and Key Statistics of Argentina, conducted by the OECD Statistics and Data Directorate. This technical report is the result of strong co-operation between the OECD and the Argentinian statistical authorities, in particular INDEC, which started in 2016 with a preliminary review after Argentina requested its adherence to the Recommendation of the OECD Council on Good Statistical Practice: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0417.

The assessment published today covers Argentina's legal and institutional framework for official statistics, important components of its statistical infrastructure such as censuses of population and housing, and Argentina's statistical production in a range of domains including national accounts, price statistics, business statistics, trade statistics, labour statistics, income distribution, and well-being indicators. All these aspects have been carefully evaluated in terms of their coverage and compliance with international standards.

The report provides an overview of the significant progress accomplished by the Argentinian statistical authorities over the last three and a half years and recognises the commitment of INDEC to leading the process of improving the quality of official statistics. While much has been achieved in almost all areas, the report also points to important outstanding challenges in moving Argentinian official statistics closer to international standards. To meet these challenges, a number of recommendations are made to strengthen the national statistical system, establish procedures for efficient coordination within it, implement a quality management system, strengthen the statistical infrastructure, and further align official statistics with international standards.

Special emphasis is placed on the need for a new statistical law that will reinforce the professional independence of INDEC and other producers of official statistics; ensure the impartiality, objectivity and transparency of official statistics; and improve statistical authorities' access to administrative data.

In presenting this report I would like to thank the Argentinian statistical authorities, in particular INDEC, for the strong co-operation they have offered and the professionalism they have presented throughout this process. I hope that this report will encourage Argentina to continue its efforts to strengthen its national statistical system so as to bring it fully into line with international standards, especially those set out in the Recommendation of the OECD Council on Good Statistical Practice. The OECD stands ready to continue its co-operation with the Argentinian authorities and all official producers of statistics in Argentina in the pursuit of this endeavour.

For further information about this statement, please contact stat.contact@oecd.org.

Buenos Aires hosted the 19th Annual Meeting of the Washington Group on Disability Statistics

The National Institute of Statistics and Censuses (INDEC) hosted the 19th Annual Meeting of the Washington Group on disability statistics. It was held on 25-27 September 2019 in Buenos Aires at facilities of the Ministry of National Security. The mission of the Washington Group is to promote and coordinate international cooperation for the measurement of disability in censuses and national surveys. Its main objective is to provide necessary basic information on disability to ensure its global statistical comparability. The president of the Washington Group, Jennifer Madans, highlighted the work of Latin America and the Caribbean, “a region where significant progress has been made identifying current problems and linking them to the socio-economic situation.” In addition, she highlighted INDEC's commitment to "take responsibility for the challenge of measuring disability". In turn, the representative of the Economic Commission for Latin America and the Caribbean (ECLAC), Daniela González, stressed the strengthening of cooperation in the region on this subject and said that "10 years ago it was unthinkable."

On behalf of INDEC, the National Director of Social and Population Statistics and the Director of Population Statistics participated in the opening of the meetings and headed the panel “Measurement of disability in the Census 2020 round in Latin America and the Caribbean.” In their presentation, the directors described the methodology of the survey conducted in 2018 for the National Study on the Profile of Persons with Disabilities.

Some conclusions of the 19th Annual Meeting revolved around the need for updated administrative records in the region because "social and population statistics depend largely on information from censuses." In the case of Argentina, it was pointed out that “administrative records are dispersed and disjointed so far."

The meeting, the second one held in Buenos Aires, was attended by authorities of the Organization of American States, the Department of Foreign Affairs and Commerce of Australia, the International Labor Organization, the Ministry of International Development, and the National Center for Health Statistics of the United States, among others. At the national level, officials of the Vice Presidency of the Nation, the National Disability Agency, the team of the Directorate of the Permanent Household Survey and the technical team of the National Directorate of Social and Population Statistics of INDEC attended. In addition, representatives of statistical institutions from 31 countries were present. The delegates' presentations focused on the census operations to be developed in the next world round.

General information on this survey can be found at www.indec.gob.ar.

For further information, please contact ces@indec.gob.ar.

AUSTRALIA

Reporting: **Summer Wang**

Response propensity modelling with machine learning

The difficulties in achieving and maintaining response rates are leading National Statistical Offices to focus on potential respondents and how to direct data collection efforts most efficiently to maximise the quality of outputs. This area of research is known as adaptive and responsive design and complements efforts to improve the level of response via improving the provider experience. The Australian Bureau of Statistics (ABS) is currently seeking to identify a line of investigation regarding the use of machine learning for response propensity modelling.

For household surveys and before standard workload enumeration, we want to predict the propensities on how a geographic area will respond to different collection protocols. The propensities will inform decisions on protocols to use and the number and location of interviewers needed. Finally, after standard workload enumeration, we want to predict the likelihood of getting responses from those that have not yet responded and use this to assign follow-up protocols.

For business surveys we want to apply response propensity modelling before the Intensive Follow Up stage to identify units that will respond without reminder letters or calls. This is a cost saving initiative, referred to as a "Gold Provider" strategy in previous ABS work.

During the Intensive Follow Up, we want to identify units least likely to respond regardless of how many additional reminders we make. These units would be de-prioritised in the Intensive Follow Up, allowing effort to concentrate on units that are more likely to deliver a response.

It is proposed that the Random Forests Machine Learning method be investigated. This method appears to be well suited to our problem because it deals with small sample sizes relative to the number and type of predictors as well as behaviour that involves factor interactions not known by the researcher at the time the model is estimated. The aggregation across trees is expected to generate more stable estimates compared to those generated from any single tree.

BELARUS

Reporting: **Natalia Bokun and Natalia Bandarenka**

Based on the materials of the National Statistical Committee of the Republic of Belarus

Population Census in 2019

The population census in the Republic of Belarus was conducted in the period from October 4th to October 30th, 2019. The duration of the census was 27 days.

Unlike the two previous censuses (1999, 2009), the population census 2019 was conducted in three ways:

- “face-to-face” interviewing at residences by a specialist, October 21st to 30th;
- “face-to-face” interviewing at stationary sites, October 4th to 30th;
- by the Internet (the principle of self-registration), October 4th to 18th.

The census program included the following thematic clusters:

- contact information (name, identification number, address)
- the census program proper, including:
    - personal demographic characteristics (sex, age and marital status)

    - social and economic characteristics (education level, profession, occupations, sources of livelihood, social status)
    - ethnicity (ethnic origin, native language, language skills)
    - questions concerning the study of population reproduction; and
    - migration
- questions connected with other surveys (employment or unemployment, subsidiary plots)

The innovation of the recent census was the replacement of the traditional paper questionnaire with tablet computers in which census forms were downloaded digitally, as well as maps of sites with addresses and house outlines. Because of the connection to the control system during collection, data quality was checked automatically when entering data into the application on the tablet. Additionally, a set of identifying variables were automatically filled using an automated query of the information system "Register of the Population". Using this approach, 20% of the variables on the questionnaire were filled, including full name, date and place of birth, gender, citizenship, place of residence and place of stay. Using tablets also increased the load per specialist to 750 people (in 2009 it was 300 people).

Coinciding with the population census of 2019 in the Republic of Belarus, the agricultural census took place. According to the recommendations of the FAO, an agricultural census is to be conducted once every 10 years. In the Republic of Belarus, the National Statistical Committee maintains current statistical accounting of the main activities of agricultural organizations. It also selectively monitors the agricultural activity of citizens who live in rural areas and operate personal subsidiary farms. Excluded are citizens with agricultural activities in urban areas or in garden associations, as well as agricultural operations at seasonal houses or summer cottages. Instead, these citizens answer questions about their agricultural activity in the census program. The questionnaire on agricultural activities included a minimum set of indicators to specify the property ownership (possession, use, rent) of the household, its location (urban or rural area, garden associations), the structure of arable land, the number of perennial plantations, the number of livestock, poultry and bee colonies.

For more details, please go to www.belstat.gov.by.

CANADA

Reporting: **Elisabeth Neusy**

Reporting the Quality of Estimates through Confidence Intervals

Coefficients of variation (CVs) are widely used at Statistics Canada to measure the quality of survey estimates in terms of their sampling error. The main advantage of CVs is that they are a relative measure, and therefore allow the quality of estimates of varying size to be compared. They are especially useful for quantitative variables with large positive values. However, CVs are not suitable for measuring the quality of proportions, estimates of change or differences, nor for variables that can take on negative values.

The table C below presents an example that illustrates why CVs are not appropriate for proportions. The table provides CVs and 95% confidence intervals for two proportions under simple random sampling without replacement, ignoring the finite population correction. In this example, the proportion of 90% would be considered to be of better quality than the proportion of 1% under CV-based rules. The opposite conclusion would be drawn based on the length of the confidence intervals and the sample sizes. The example demonstrates that CVs tend to be too conservative for small proportions and too lenient for large proportions.

Table C: *Estimated CVs for proportions*

Estimated Proportion	Sample Size	CV	95% Confidence Interval (Wilson interval)		
			Lower Bound	Upper Bound	Interval Length
1%	500	45%	0.4%	2.3%	1.9%
90%	10	11%	52.3%	98.7%	46.3%

Given these concerns with CVs, Statistics Canada has recently adopted as a best practice the use of 95% confidence intervals to measure and report the quality of estimates in terms of their sampling error. Confidence intervals have many advantages: they are appropriate for all types of estimates and they express sampling error in a form that is clear and easy to interpret.

Several challenges remain for this project. One challenge is to ensure that appropriate methods are used to construct confidence intervals. The most commonly used method of constructing confidence intervals is the Wald interval, which is of the form $\hat{y} \pm z_{1-\alpha/2} \times \sqrt{\hat{\text{var}}(\hat{y})}$ where $\hat{y}$ is the estimate and $\hat{\text{var}}(\hat{y})$ is the estimated variance of $\hat{y}$. Wald intervals are based on the assumption that the sampling distribution of $\hat{y}$ is approximately normal. For proportions, the normality assumption is known to break down for small sample sizes and for proportions near zero or one. Simulation studies have led us to recommend the use of modified Wilson intervals or modified Clopper-Pearson intervals for binomial proportions under complex designs (Kott and Carr, 1997; Korn and Graubard, 1998). Further research is required for other types of estimates.

Another challenge is to develop a standard set of release rules for surveys that report sampling error through confidence intervals. These are rules that specify whether an estimate and its confidence interval should be suppressed for quality reasons, released with a warning, or released with no warning. Some Statistics Canada surveys have begun transitioning to the use of confidence intervals to report quality; these surveys are using preliminary release rules until a standard set of rules is approved for general use.

References:

Korn, E.L., and Graubard, B.I. (1998). Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data. Survey Methodology, 24, 193-201.

Kott, P.S. and Carr, D.A. (1997). Developing an Estimation Strategy for a Pesticide Data Program. Journal of Official Statistics, Vol. 13, No. 4, 367-383.

LATVIA

Reporting: **Ieva Zemeskalna and Mārtiņš Liberts**

Central Statistical Bureau of Latvia celebrates its centenary with the conference "Into the Future – Statistics for Modern Society"

On September 1, Central Statistical Bureau of Latvia (CSB) was celebrating its centenary, and on August 30 organised a conference for data users and data providers titled "Into the Future – Statistics for Modern Society". Two sessions – "Around an Individual" and "Untamed Statistics" – engaged data enthusiasts telling their data stories.

One hundred years ago, a politically independent statistical institution was established in Latvia. It served as a link between data providers and data users by producing reliable statistics. The key goal of the institution has not changed since then.

“Along with the economic development and globalisation of the economy, classical methods based on the direct data acquisition do not allow collecting data without excessive administrative burden, therefore data collection methods have to be changed. Nowadays, statisticians use administrative data, alternative data sources, web scraping, merging of alternative data with survey results, as well as experimental statistics. Moreover, our future plans will engage also mobile operators and satellite data. The CSB has become a dynamic, innovative institution employing professionals willing and able to face these new challenges,” said Ms Aija Žīgure, president of the CSB, in her conference opening speech.

Mr Ojārs Spārītis, president of the Latvian Academy of Sciences, as a guest of honour in his opening speech stressed the overall value of statistics for every society and state. Keynote speaker Mr Andres Vikat, Chief of the Social and Demographic Statistics Section at the UNECE gave a glimpse of development actions and challenges for the international community of official statistics in an era awash of data.

The first session was opened by Ms Rudīte Spakovska, journalist at Latvian Television, sharing the view on data literacy under presentation “Data is available. What's next?” Assoc. prof. Andrejs Ērglis, Head of the Latvian Centre of Cardiology, gave a presentation “Science, medicine and credibility of statistics on heart affairs” to raise awareness of significance of data in development of medical sciences. Ms Zane Driņķe, Dean of the Faculty of Business Administration at the Turība University, presented the statistical portrait of entrepreneurs of Latvia, giving and insight on their different characteristics based on statistics.

During the session “Untamed statistics” invited speakers explored unbelievable ways of using data for decision making, storytelling and beyond. Prof. Signe Bāliņa, Deputy Rector for digital society at the University of Latvia gave a presentation on “Artificial intelligence and role of data”. Mr Aldis Ērglis, Lead of the machine learning laboratory “Emergn Latvia”, presented how statistics are related with machine learning techniques. Mr Kārlis Zālīte, programmer, space technology specialist of Latvian origin from Groningen University, presented an aggregate data drug information system for evidence-based health care policy decision making.

Much of the audience represented state institutions (including line ministries, State Revenue Service, State Language Centre, National Library, etc.), universities, media, and enterprises, and covered quite a large range of data providers and users.

- Video and presentation slides of the conference (the conference language was Latvian, the presentation by Mr Andres Vikat was in English, see it from 24:30):
  https://www.csb.gov.lv/en/basic_page/187
- Latvia. The first 100 (in English):
  https://www.csb.gov.lv/en/statistika/statistikai-100
- Statistics Latvia – feeling the pulse of Latvia for the past 100 years (in English):
  https://www.csb.gov.lv/en/about-us

For more details, please contact media@csb.gov.lv.

Reporting: **Susmita Das**

Innovative methodologies to combine census/survey data and admin data at Stats NZ

Stats NZ has been working on several novel pieces of research that combine data from the field with admin data. In this article, we talk about the innovative methodologies that were developed around the 2018 Census and the NZ Child Poverty workstream.

Administrative Data in the 2018 Census

The 2018 Census (https://www.stats.govt.nz/methods/overview-of-statistical-methods-for-adding-admin-records-to-the-2018-census-dataset) is the first time the New Zealand census dataset has included administrative (admin) records to count people who did not complete a census form, replacing the use of 'substitute' imputed records in previous censuses. The 2018 Census dataset is made up of census forms supplemented by high-quality administrative records. Stats NZ matched 2018 Census forms data with a file of administrative data that provided a good approximation of the New Zealand population. This enabled Stats NZ to add people who were not counted via the census forms into the census dataset. The final 2018 Census dataset consists of 89 percent census responses and 11 percent records sourced from admin data.

Stats NZ also used admin data and 2013 Census data to add information about people's characteristics that are less likely to change over time – for example, Māori descent and birthplace. Careful use of admin data has enabled Stats NZ to add real data about real people to the dataset where we were confident the people should be counted but hadn't completed a census form. We also used data from the 2013 Census and admin sources and statistical imputation methods to fill in some missing characteristics of people and dwellings.

New Zealand's child poverty estimates

Stats NZ has published estimates of New Zealand's child poverty rates for the years 2006/07 to 2017/18 (https://www.stats.govt.nz/information-releases/child-poverty-statistics-year-ended-june-2018). The estimates used different data sources and methods (https://www.stats.govt.nz/methods/child-poverty-statistics-technical-appendix-201718) to produce robust estimates to be used as baseline of government initiatives aimed at reducing child poverty.

A primary source for producing child poverty measures is Stats NZ's Household Economic Survey, or HES (http://datainfoplus.stats.govt.nz/Item/nz.govt.stats/8d483e91-13f9-4a6a-8cea-8c7ce993f103). HES collects information on household income and expenditure, and demographic information on individuals and households. Prior to HES1819, HES was designed to achieve annual sample sizes between 3,500 and 5,000 households. To ensure movements in the child poverty measures can be precisely detected, HES was pooled with another Stats NZ survey, the quarterly Household Labour Force Survey, or HLFS (http://datainfoplus.stats.govt.nz/Item/nz.govt.stats/6a13af44-0057-4a63-835a-c1a0c6f8ef91). HLFS collects responses from around 15,000 households, using the same respondents over eight consecutive quarters and replacing one-eighth (on a rotating basis) by a new set of respondents.

To maximise pooling of HES and HLFS, tax data and benefit income from the Integrated Data Infrastructure, or IDI (https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/) were used to replace income collected directly from HES respondents and to provide income for HLFS respondents. The IDI is a large research database on life events of people and households. IDI links or integrates data from government agencies, Stats NZ surveys, and non-government organisations.

The use of administrative data addresses issues of underreporting of benefit income and some overreporting of salaries and wages in HES.

Although link rates to the IDI, after improving the linking methodology, were relatively high for both HES and HLFS, we ensured that any bias from unlinked records was adjusted for by imputation. Multiple imputation was used to impute investment income for most HLFS respondents. To match the annual HES interview pattern, we used two HLFS rotation groups from each quarter. To produce household income before housing costs, data pooled from HES and HLFS were weighted using HES benchmarks. The lack of good quality administrative data source on housing costs and material hardship required the creation of new benchmarks for weighting HES data to improve estimation.

SWEDEN

Reporting: **Thomas Laitila**

The 5th Baltic-Nordic Conference on Survey Statistics

On June 16-20, the 2019 Baltic-Nordic Conference on Survey Statistics (BaNoCoSS) was conducted at Örebro University, Sweden. The conference was organized by the Baltic-Nordic-Ukrainian (BNU) Network on Survey Statistics and Örebro University in cooperation with Statistics Sweden. This was the fifth BaNoCoSS organized by the BNU network since the first in 2002.

The BaNoCoSS events are scientific conferences presenting developments on theory, methodology and applications of survey statistics in a broad sense. The conference provides a platform for discussion and exchange of ideas for a variety of people. These include, for example, statisticians, researchers and other experts at universities, national statistical institutes, and research institutes and other governmental bodies, and private enterprises. Its topics cover survey research methodology, empirical research and statistics production. University students in statistics and related disciplines provide an important interest group of the conference.

The 2019 conference included three keynote presentations by Roberto Benedetti, University of Chieti-Pescara, and Federica Piersimoni, Italian National Statistical Institute, on the theme “Spatial survey sampling and analysis and GIS”. Seven invited speaker presentations was given by Esa Läärä, University of Oulu on “Uses of sampling methodology in epidemiologic research”, Vera Toepoel, University of Utrecht, “Mobile surveys and sensor data”, Peter van der Heijden, University of Utrecht, “An Overview of Population Size Estimation where Linking Registers Results in Incomplete Covariates, with an Application to Mode of Transport of Serious Road Causalities”, Susie Jentoft, Statistics Norway, “Using integrated geospatial data in official statistics”, Risto Lehtonen, University of Helsinki, “On balanced sampling and calibration estimation in survey sampling”, Danutė Krapavickaitė, Vilnius Gediminas Technical University, “Application of Bayesian Analysis”, and Imbi Traat, University of Tartu, “An Alternative Nonresponse Adjustment Estimator”.

Nine contributed paper sessions, including one poster session, contained 33 presentations of papers. The total number of participants was 60 representing 16 different countries. Of these, 35 were affiliated to universities and 23 to national statistical agencies.

Sponsorship of the conference was received from the IASS, the Nordic Council of Ministers (Nordplus), the Swedish Survey Association, Statistics Sweden, and Örebro University. Endorsement of the conference was received from the ISI.

The next event organized by the BNU network is a summer school in Minsk, Belarus, to take place on 23-27 August, 2020. More information on the BNU network and the 2019 BaNoCoSS are found on the websites wiki.helsinki.fi/display/BNU and www.oru.se/hh/banocoss2019, respectively.

For more information please contact: thomas.laitila@oru.se

UNITED STATES

Reporting: **Mark W. Dumas**

Publication of U.S. state-level labor productivity measures

On June 4th, 2019 the U.S. Bureau of Labor Statistics (BLS) published experimental measures of state-level labor productivity. The output per hours worked series begin in 2007, cover the private nonfarm sector, and are a significant improvement over previous efforts that measured state-level labor productivity as output per worker. The BLS constructs state-level output using published GDP by state and industry data from the U.S. Bureau of Economic Analysis (BEA). The hours worked series is constructed using data from several BLS surveys, including the Current Employment Statistics (CES) Survey, the Current Population Survey (CPS) and the National Compensation Survey (NCS).

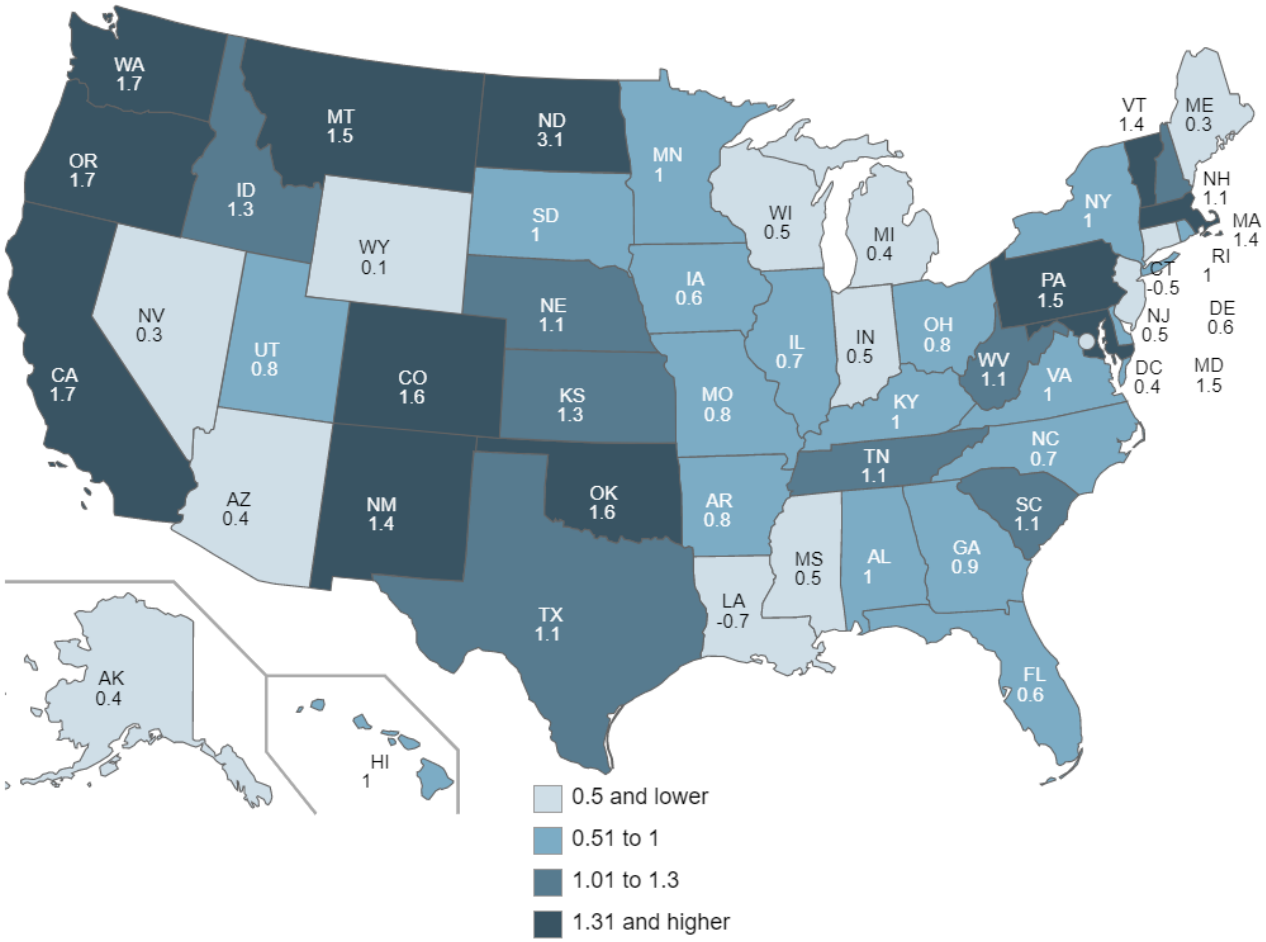


Figure U. *Labor Productivity by State – Average Annual Percent Change (2007-2017)*

The experimental measures demonstrate the variation in labor productivity that exists amongst the U.S. states. Driven by the shale oil boom, North Dakota's average annual productivity growth (3.1 percent) was substantially higher than the other states during the current business cycle. The other fastest growing states were California, Oregon, and Washington, which all grew at a rate of 1.7 percent. The slowest growing states were Louisiana (-0.7 percent), Connecticut (-0.5 percent), and Wyoming (0.1 percent).

The U.S. Bureau of Labor Statistics believes these measures will be of interest to a wide range of data users including policy makers and economic researchers. The analysis of these data over the

long- term may increase the understanding of regional business cycles; the persistence of regional income inequality; contribution of states to national productivity growth; as well as the impact of certain policies, such as regulations and taxes, on economic growth.

Further Information. The complete state-level productivity dataset is accessible on the BLS state productivity web page (https://www.bls.gov/lpc/state-productivity.htm). In addition to labor productivity, the dataset includes measures of output, hours, unit labor costs, and hourly compensation. Furthermore, the BLS published an article "BLS Publishes Experimental State-level Labor Productivity Measures" (https://www.bls.gov/opub/mlr/2019/article/bls-publishes-experimental-state-level-labor-productivity-measures.htm) that describes the data and methodology used to estimate these new labor productivity data.

2020 American Statistical Association Conference on Statistical Practice

The 2020 American Statistical Association Conference on Statistical Practice aims to bring together statistical practitioners and data scientists – including data analysts, researchers, and scientists – who engage in the application of statistics to solve real-world problems daily.

The goal of the conference is to provide participants with opportunities to learn new statistical methodologies and best practices in statistical analysis, design, consulting, and programming. The conference is designed to help applied statisticians improve their ability to consult with and aid customers and organizations solve real-world problems.

- Learn statistical techniques that apply to your job as an applied statistician
- Learn how to better communicate with customers
- Learn how to have a positive impact on your organization

Conference on Statistical Practice 2020 will offer courses (e.g. programming in R and Python), tutorials, a featured speaker, concurrent sessions, poster sessions, a closing session, exhibits, and more.

Organizer: American Statistical Society

When: 20-22 February 2020

Where: Sacramento, USA

E-mail: meetings@amstat.org

Website: https://ww2.amstat.org/meetings/csp/2020/

Workshop on the Concept of Quality for Big Data

This workshop is aimed to start discussion about the concept of quality for organic data and methods to check and improve data quality. Censuses and surveys are results of actively designed data collection exercises and much of the quality criteria address the methods of data collection. As one has no influence on the processes which result in organic data, many of the traditional concepts remain relevant for big data. However, the concept of quality has to move from general quality to suitability for particular analyses, leading to more fine-tuned approaches.

Organizer: Faculty of Social Sciences, Eötvös Loránd University

When: 25-26 February 2020

Where: Eötvös Loránd University, Budapest, Hungary

E-mail: trudas@elte.hu

Website: https://tatk.elte.hu/workshop_concept_of_quality_for_big_data

Population Data for Informed National Planning and Development

Organizers: Federal University of Agriculture and the Royal Statistical Society Nigeria Local Group

When: February 2020

Where: Federal University of Agriculture, Abeokuta, Ogun State, Nigeria

E-mail: olayiwolaom@funaab.edu.ng

The Eleventh Conference on Health Survey Research Methods

The aim of the conference is to discuss new, innovative survey research methods that improve the quality of health survey data. The HSRMC will bring together researchers from various disciplines who are at the forefront of survey methods research, who are responsible for major health surveys, and who use survey data to develop health policy.

Organizers: Westat and US Census Bureau

When: 4-7 March 2020

Where: Williamsburg Lodge, Williamsburg, Virginia, USA

E-mail: leslyger@westat.com

Website: https://hsrmconference.com/call-for-papers

2020 Comparative Survey Design and Implementation (CSDI) Workshop

The main goal of CSDI is improving comparative survey design, implementation, harmonization and dissemination analysis. The workshop provides a forum and platform for researchers involved in research relevant for comparative survey methods.

Organizer: CSDI group

When: 11-13 March 2020

Where: Condorcet Campus, Paris, France

Website: https://csdiworkshop.org/

Conference on the Use of R in Official Statistics

The purpose of the conference is to provide a public forum for researchers from academia and institutes of official statistics organisations to present, exchange ideas and discuss developments in state-of-the-art statistical software commonly used in applied economics and statistics. The most focused debates are expected to be on the use of R in Official Statistics.

Organizer: Statistics Austria

When: 6-8 May 2020

Where: Statistics Austria, Austria

E-mail: contact@r-project.ro

Website: http://r-project.ro/conference2020.html

American Association for Public Opinion Research (AAPOR)

The AAPOR Conference is a collaboration in the scientific community, whose objectives are to provide a training opportunity to attendees; teach the latest methodology and approaches to survey research best practices; make each attendee a better survey researcher, and; maintain and improve professional survey competency.

Organizer: American Association for Public Opinion Research

When: 14-17 May 2020

Where: Hilton, Atlanta, USA

E-mail: mandy@mandysha.com

Website: https://www.aapor.org/Conference-Events/Annual-Meeting.aspx

Better Lives 2030: mobilising the power of data for Africa and the world

The over-arching theme of the conference is “Better Lives 2030: mobilising the power of data for Africa and the world. Bringing together statisticians and all those in government, universities and education who care about the value of statistics to society”.

Organizers: International Association for Official Statistics, International Statistical Institute and Republic of Zambia Central Statistical Office

When: 19-21 May 2020

Where: Livingstone, Zambia

E-mail: zambiastats2020@gmail.com

Website: https://isi-web.org/images/news/Zambia-Call-for-Papers-Announcement.pdf

12th International Conference on Transport Survey Methods

Topics of interest include survey data, passive big data, travel survey and new technology data collection, travel survey tools and methods and emerging or traditional topics.

Organizer: Universidade de Lisboa

When: 31 May – 5 June 2020

Where: Hotel Golf Mar in Porto Novo, near Lisbon, Portugal

E-mail: jabreu@tecnico.ulisboa.pt

Website: https://www.isctsc2020.pt/

Symposium on Data Science & Statistics

Organizer: American Statistical Association

When: 3-6 June 2020

Where: The Westin Convention Center, Pittsburgh, Pennsylvania, USA

E-mail: meetings@amstat.org

Website: https://ww2.amstat.org/meetings/sdss/2020/

International Conference on Establishment Surveys (ICES) VI

The International Conference on Establishment Statistics (ICES) promotes discussion of a broad range of issues related to the statistics of businesses, farms, or institutions. ICES features invited and contributed papers and demonstrations from around the globe that highlight new, improved, and upcoming establishment statistics methodologies and technologies using census data, administrative or other organic data, and sample survey data.

Participants come from academia, government statistical agencies, private businesses, statistical associations, and other sectors with an interest in international best practices in conceptualization, design, data collection, analysis, and dissemination.

The conference will include:

- Short courses at introductory, intermediate, and advanced levels
- Introductory overview lectures about important and timely topics
- Selection of invited and contributed papers
- Software demonstrations

Organizer: American Statistical Association

When: 15-20 June 2020

Where: New Orleans, Louisiana, USA

E-mail: meetings@amstat.org

Website: https://ww2.amstat.org/meetings/ices/2020/index.cfm

SAE2020 – BigSmall

The key aim is to assess the current development and usage of small area estimation methods in a wide range of contexts, including Official Statistics, Big Data Sources, Spatial approaches and satellite imagery, Machine Learning.

The keynote speakers are Graham Kalton (Westat, Rockville, Maryland, USA) and J. Sunil Rao (University of Miami Health System)

Organizer: Department of Economics & Management of the University of Pisa

When: 6-8 July 2020

Where: Centro Congressi Federico II, Naples, Italy

E-mail: info@sae2020.org

Website: https://sae2020.org/

Joint Statistical Meeting (JSM)

The JSM topics range from statistical applications to methodology and theory to the expanding boundaries of statistics, such as analytics and data science. JSM offers the opportunity for statisticians in academia, industry, and government to exchange ideas and explore opportunities for collaboration. Beginning statisticians (including current students) are able to learn from and interact with senior members of the profession.

With a focus on the 2020 theme, Everyone Counts: Data for the Public Good, the JSM program consists of invited, topic-contributed, and contributed technical sessions, poster presentations, roundtable discussions, professional development courses and workshops, award ceremonies, and other meetings and activities.

Organizer: American Statistical Society

When: 1-6th August 2020

Where: Philadelphia, Pennsylvania, USA

E-mail: meetings@amstat.org

Website: https://ww2.amstat.org/meetings/jsm/2020/index.cfm

Summer School on Survey Statistics

The key aims include the promotion of survey statistics by sharing and exchange of knowledge and experience of between teachers, students, researchers and practitioners. Details will be released soon on the conference website.

Organizers: Baltic-Nordic-Ukrainian Network on Survey Statistics (BNU network), Belarusian State Economic University (BSEU), School of Business of Belarusian State University (BSU), and the National Academy of Sciences of the Republic of Belarus

When: 24-28 August 2020

Where: Minsk, Belarus

E-mail: nataliabokun@rambler.ru

Website: https://wiki.helsinki.fi/display/BNU/Events

11th International Francophone Conference on Surveys

Organizer: Société Française de Statistique – SFdS

When: 13-16 October 2020

Where: Université libre de Bruxelles, Belgium

E-mail: catherine.vermandele@ulb.be

Website: http://sondages2020.sciencesconf.org

BigSurv20: Where Big Data Meets Survey Science

BigSurv20 brings together computer and data scientists with an interest in social science and data collection and social scientists, survey methodologists, and statisticians with an interest in computer and data science. During the three-day event, attendees, panelists, and presenters will engage in a variety of discussions on the practical applications for employing Big Data and data science to improve the quality of statistics production. There will be keynote speakers, short courses on cutting edge topics in data science and survey methodology.

There is a focus on how researchers can successfully combine traditional survey data with new data sources, such as registers, social media, apps, and other forms of digital data, the conference will explore different perspectives stemming from computer science and social sciences.

Organizer: Antje Kirchner, University of Nebraska and Peter Lugtig, Utrecht University

When: 4-6 November 2020

Where: Utrecht, Netherlands

E-mail: info@bigsurv20.org

Website: https://www.bigsurv20.org/

In Other Journals

Journal of Survey Statistics and Methodology

Volume 7, Issue 3, September 2019

https://academic.oup.com/jssam/issue/7/3

Survey Statistics

Bootstrap Prediction Intervals for Small Area Means from Unit-Level Nonlinear Models
Andreea L Erciulescu, Wayne A Fuller

Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys
Carolina Franco, Roderick J A Little, Thomas A Louis, Eric V Slud

Survey Methodology

Multidimensional Assessment of Social Desirability Bias: An Application of Multiscale Item Randomized Response Theory to Measure Academic Misconduct
Nils Wlömert, David Pellenwessel, Jean-Paul Fox, Michel Clement

Information Criterion for Nonparametric Model-Assisted Survey Estimators
Addison James, Lan Xue, Virginia Lesser

The Effects of Sampling Frame Designs on Nonresponse and Coverage Error: Evidence from the Netherlands
Ann-Kristin Kölln, Yfke P Ongena, Kees Aarts

Creating Improved Survey Data Products Using Linked Administrative-Survey Data
Michael E Davern, Bruce D Meyer, Nikolas K Mittag

Volume 7, Issue 4, December 2019

https://academic.oup.com/jssam/issue/7/4

Survey Statistics

"Robust-Squared" Imputation Models Using Bart
Yaoyuan V Tan, Carol A C Flannagan, Michael R Elliott

Simultaneous Edit and Imputation For Household Data with Structural Zeros
Olanrewaju Akande, Andrés Barrientos, Jerome P Reiter

Survey Methodology

Who Gets Lost, and What Difference Does It Make? Mixed Modes, Nonresponse Follow-up Surveys and the Estimation of Turnout
Andreas C Goldberg, Pascal Sciarini

Do Sequential Mixed-Mode Surveys Decrease Nonresponse Bias, Measurement Error Bias, and Total Bias? An Experimental Study
Joseph W Sakshaug, Alexandru Cernat, Trivellore E Raghunathan

Panel Effects: Do the Reports of Panel Respondents Get Better or Worse over Time?
Hanyu Sun, Roger Tourangeau, Stanley Presser

Effects of a Government-Academic Partnership: Has the NSF-CENSUS Bureau Research Network Helped Improve the US Statistical System?
Daniel H Weinberg, John M Abowd, Robert F Belli, Noel Cressie, David C Folch, Scott H Holan, Margaret C Levenstein, Kristen M Olson, Jerome P Reiter, Matthew D Shapiro, Jolene D Smyth, Leen-Kiat Soh, Bruce D Spencer, Seth E Spielman, Lars Vilhuber, Christopher K Wikle

---

Survey Methodology
Statistics Canada Statistique Canada
Canada

Volume 45, Number 3 (December 2019)

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2019003-eng.htm

Regular Papers

Estimation of level and change for unemployment using structural time series models
Harm Jan Boonstra and Jan A. van den Brakel

Robust variance estimators for generalized regression estimators in cluster samples
Timothy L. Kennel and Richard Valliant

A note on propensity score weighting method using paradata in survey sampling
Seho Park, Jae Kwang Kim and Kimin Kim

Suggestion of confidence interval methods for the Cronbach alpha in application to complex survey data
Jihnhee Yu, Ziqiang Chen, Kan Wang and Mine Tezal

Cost optimal sampling for the integrated observation of different populations
Piero Demetrio Falorsi, Paolo Righi and Pierre Lavallée

A grouping genetic algorithm for joint stratification and sample allocation designs
Mervyn O'Luing, Steven Prestwich, and S. Armagan Tarim

"Optimal" calibration weights under unit nonresponse in survey sampling
Gösta Andersson

A method to correct for frame membership error in dual frame estimators
Dong Lin, Zhaoce Liu and Lynne Stokes

Short note

On a new estimator for the variance of the ratio estimator with small sample corrections
Paul Knottnerus and Sander Scholtus

Journal of Official Statistics

Volume 35: Issue 3 (Sep 2019)

https://content.sciendo.com/view/journals/jos/35/3/jos.35.issue-3.xml

Probing for Informal Work Activity
Katharine G. Abraham and Ashley Amaya

Correlates of Representation Errors in Internet Data Sources for Real Estate Market
Maciej Beręsewicz

An Integrated Database to Measure Living Standards
Elena Dalla Chiara, Martina Menon and Federico Perali

Connecting Correction Methods for Linkage Error in Capture-Recapture
Peter-Paul de Wolf, Jan van der Laan and Daan Zult

Imprecise Imputation: A Nonparametric Micro Approach Reflecting the Natural Uncertainty of Statistical Matching with Categorical Data
Eva Endres, Paul Fink and Thomas Augustin

A Lexical Approach to Estimating Environmental Goods and Services Output in the Construction Sector via Soft Classification of Enterprise Activity Descriptions Using Latent Dirichlet Allocation
Gerard Keogh

Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach
Joseph W. Sakshaug, Arkadiusz Wiśniowski, Diego Andres Perez Ruiz and Annelies G. Blom

Tests for Price Indices in a Dynamic Item Universe
Li-Chun Zhang, Ingvild Johansen and Ragnhild Nygaard

Volume 35: Issue 4 (Dec 2019): Special Issue on Measuring LGBT Populations

https://content.sciendo.com/view/journals/jos/35/4/jos.35.issue-4.xml

Preface
Nancy Bates, Stephanie Steinmetz and Mirjam Fischer

Are Sexual Minorities Hard-to-Survey? Insights from the 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS) Survey
Nancy Bates, Yazmín A. García Trejo and Monica Vines

Test of a Hybrid Method of Sampling the LGBT Population: Web Respondent Driven Sampling with Seeds from a Probability Sample
Stuart Michaels, Vicki Pineau, Becky Reimer, Nadarajasundaram Ganesh and J. Michael Dennis

Surveying Persons in Same-Sex Relationships in a Probabilistic Way – An Example from the Netherlands
Stephanie Steinmetz and Mirjam Fischer

Comparing Self-Reported and Partnership-Inferred Sexual Orientation in Household Surveys
Simon Kühne, Martin Kroh and David Richter

Asking about Sexual Identity on the National Health Interview Survey: Does Mode Matter?
James M. Dahlhamer, Adena M. Galinsky and Sarah S. Joestl

Measuring Sexual Orientation and Gender Identity in the National Crime Victimization Survey
Jennifer L. Truman, Rachel E. Morgan, Timothy Gilbert and Preeti Vaghela

Intersections between Sexual Identity, Sexual Attraction, and Sexual Behavior among a Nationally Representative Sample of American Men and Women
Emma Mishel

Can They and Will They? Exploring Proxy Response of Sexual Orientation and Gender Identity in the Current Population Survey
Jessica Holzberg, Renee Ellis, Robin Kaplan, Matt Virgile and Jennifer Edgar

---

Survey Practice

Volume 12, Issue 1 (2019)

https://www.surveypractice.org/issue/1155

Differences in Efficiencies Between ABS and RDD Samples by Mode of Data Collection
Carol Pierannunzi, Sonya Gamble, Robynne Locke, Naomi Freedner, Machell Town

Effects of Push-To-Web Mixed Mode Approaches on Survey Response Rates: Evidence from a Randomized Experiment in Emergency Departments
Layla Parast, PhD, Megan Mathews, MA, Marc Elliott, PhD, Anagha Tolpadi, MS, Elizabeth Flow-Delwiche, PhD, William G. Lehrman, PhD, Debra Stark, MBA, Kirsten Becker, MS

Methods for Improving Response Rates in an Emergency Department Setting – A Randomized Feasibility Study
Megan Mathews, MA, Layla Parast, PhD, Anagha Tolpadi, MS, Marc Elliott, PhD, Elizabeth Flow-Delwiche, PhD, Kirsten Becker, MS

"Your Survey is Biased": A Preliminary Investigation into Respondent Perceptions of Survey Bias
Adam Mayer

Vol 13 No 2 (2019)

https://ojs.ub.uni-konstanz.de/srm/issue/view/216

Capturing Multiple Perspectives in a Multi-actor Survey: The Impact of Parental Presence During Child Interviews on Reporting Discrepancies
Bettina Müller

Multivariate Tests for Phase Capacity
Taylor H Lewis

Within-household selection of target-respondents impairs demographic representativeness of probabilistic samples: evidence from seven rounds of the European Social Survey
Piotr Jabkowski, Piotr Cichocki

Does mode of administration impact on quality of data? Comparing a traditional survey versus an online survey via a Voting Advice Application
Vasiliki Triga, Vasilis Manavopoulos

Doing a Time Use Survey on Smartphones Only: What Factors Predict Nonresponse at Different Stages of the Survey Process?
Anne Elevelt, Peter Lugtig, Vera Toepoel

Can Nonprobability Samples be Used for Social Science Research? A cautionary tale
Elizabeth S. Zack, John Kennedy, J. Scott Long

Vol 13 No 3 (2019)

https://ojs.ub.uni-konstanz.de/srm/issue/view/217

Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys
Dominik Johannes Leiner

Linking interview speed and interviewer effects on target variables in face-to-face surveys
Caroline Vandenplas, Koen Beullens, Geert Loosveldt

Qualitative Testing for Official Establishment Survey Questionnaires
Mojca Bavdaz, Deirdre Giesen, Danna L. Moore, Paul A. Smith, Jacqui Jones

Does Benefit Framing Improve Record Linkage Consent Rates? A Survey Experiment
Joseph W. Sakshaug, Jens Stegmaier, Mark Trappmann, Frauke Kreuter

In Search of the Optimal Mode for Mobile Phone Surveys in Developing Countries. A Comparison of IVR, SMS, and CATI in Nigeria
Charles Q Lau, Alexandra Cronberg, Leenisha Marks, Ashley Amaya

Too sophisticated even for highly educated survey respondents? A qualitative assessment of indirect question formats for sensitive questions
Julia Jerke, David Johann, Heiko Rauhut, Kathrin Thomas

- **Statistical Journal of the IAOS**
  - https://content.iospress.com/journals/statistical-journal-of-the-iaos/
- **International Statistical Review**
  - https://onlinelibrary.wiley.com/journal/17515823
- **Transactions on Data Privacy**
  - http://www.tdp.cat/
- **Journal of the Royal Statistical Society, Series A (Statistics in Society)**
  - https://rss.onlinelibrary.wiley.com/journal/1467985x
- **Journal of the American Statistical Association**
  - https://amstat.tandfonline.com/toc/uasa20/current

Welcome New Members!

We are very pleased to welcome the following new IASS members!

Title	First name	Surname	Country
MR.	Andry	Andriantseheno	Ethiopia
DR.	Mary Katherine	Batcher	United States
DR.	Gaia	Bertarelli	Italy
DR.	Chang-Shang	Chen	Taiwan
DR.	James	Chipperfield	Australia
MS.	Monica Richelle	Delos Santos	Philippines
DR.	Pushpal	Mukhopadhyay	United States
MR.	Peter	Ogunyinka	Nigeria
MR.	David	Ojakaa	Kenya
PROF.	C.H.Thomas	Polfeldt	Sweden
MS.	Diane Elizabeth	Ramsay	New Zealand
DR.	Yves	Thibaudeau	United States

IASS Executive Committee Members

Executive officers (2019 – 2021)

President:	Denise Britz do Nascimento Silva (Brazil)	denisebritz@gmail.com
President-elect:	Monica Pratesi (Italy)	monica.pratesi@unipi.it
Vice-Presidents:		
Scientific Secretary:	James Chipperfield (Australia)	james.chipperfield@abs.gov.au
VP Finance:	Lucia Barroso (Brazil)	lpbarroso@gmail.com
Chair of the Cochran-Hansen Prize Committee and IASS representative on the ISI Awards Committee:	Isabel Molina (Spain)	imolina@est-econ.uc3m.es
IASS representatives on the World Statistics Congress Scientific Programme Committee:	Cynthia Clark (USA) in 2017-2019	czfclark@cox.net
	Monica Pratesi (Italy)	monica.pratesi@unipi.it
IASS representative on the World Statistics Congress short course committee:	Nadia Lkhoulf (Morocco)	n.lkhoulf@hcp.ma
Ex Officio Member:	Ada van Krimpen	an.vankrimpen@cbs.nl

IASS Twitter Account @iass_isi (https://twitter.com/iass_isi)

Institutional Members

International organisations:

- Eurostat (European Statistical Office)
- Observatoire économique et statistique d'Afrique subsaharienne (AFRISTAT)

National statistical offices:

- Australian Bureau of Statistics, Australia
- Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistisches Bundesamt (Destatis), Germany
- Israel Central Bureau of Statistics, Israel
- Istituto nazionale di statistica (Istat), Italy
- Statistics Korea, Republic of Korea
- Direcção dos Serviços de Estatística e Censos (DSEC), Macao, SAR China
- Statistics Mauritius, Mauritius
- Instituto Nacional de Estadística y Geografía (INEGI), Mexico
- Statistics New Zealand, New Zealand
- Statistics Norway, Norway
- Instituto Nacional de Estatística (INE), Portugal
- Statistics Sweden, Sweden
- National Agricultural Statistics Service (NASS), United States
- National Center of Health Statistics (NCHS), United States

Private companies:

- Numérika (Asesoría estadística y estudios cuantitativos), Mexico
- RTI International, United States
- Survey Research Center (SRC), United States
- Westat, United States

Save a tree!

Read *the Survey Statistician* online!

http://isi-iass.org/home/services/the-survey-statistician/

Please contact Margaret de Ruiter-Molloy (m.deruitermolloy@cbs.nl)
if you would like to cancel receiving paper copies of this Newsletter.