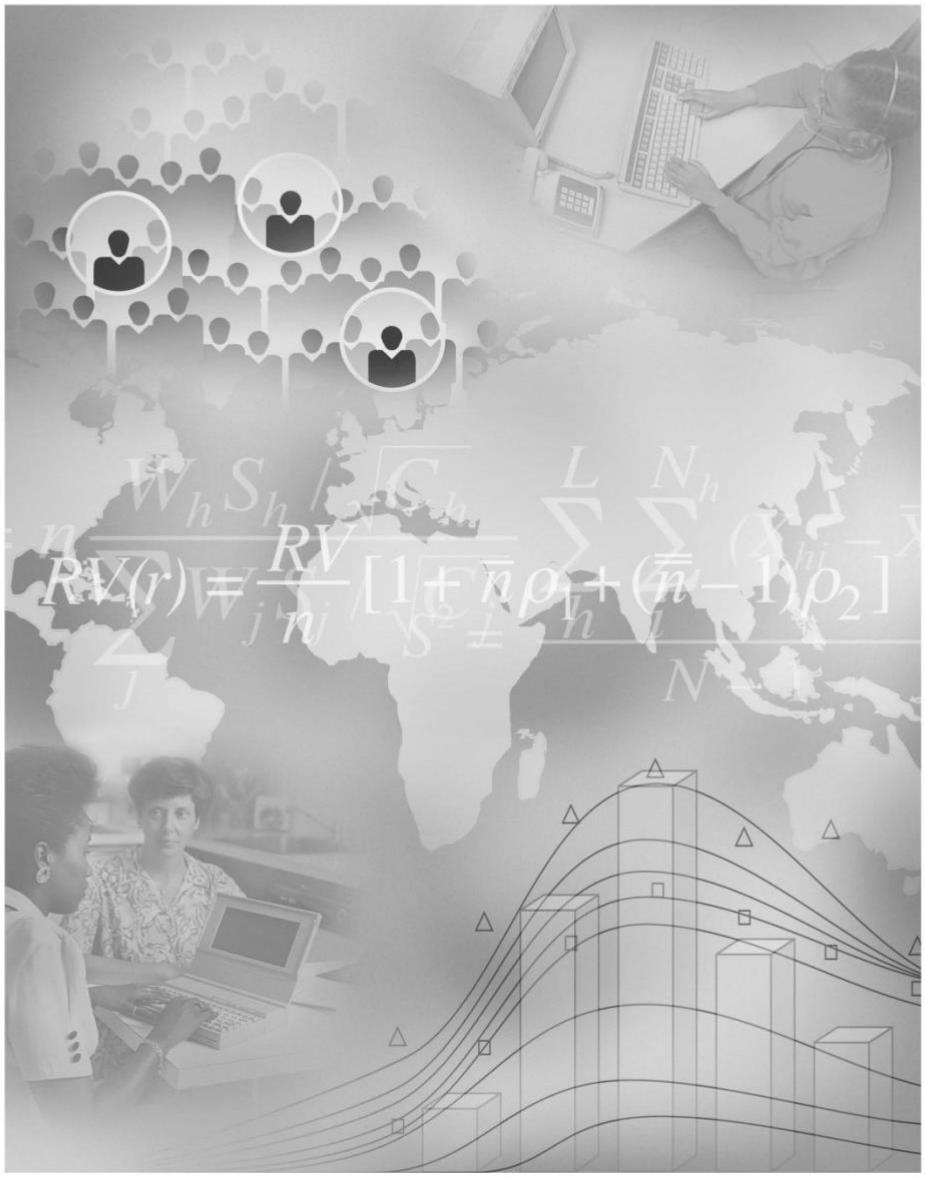
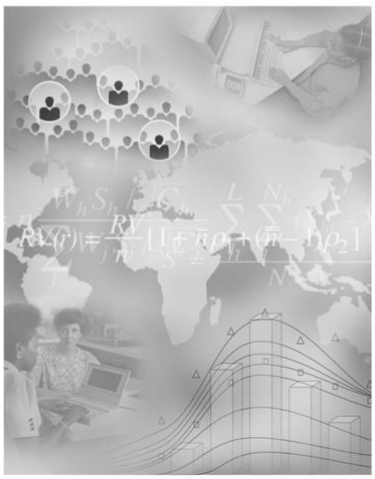
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

No. 88 **July 2023**



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



The Survey Statistician No. 88, July 2023

Editors:
Danutė Krapavickaitė (*Lithuanian Statistical Society*) and Eric Rancourt (*Statistics Canada*)

Section Editors:

Peter Wright	Country Reports
Ton de Waal	Ask the Experts
Maria Giovanna Ranalli	New and Emerging Methods
Alina Matei	Book & Software Review

Production and Circulation:
Maciej Beręsewicz (*Poznań University of Economics and Business*), Natalie Shlomo (*The University of Manchester*)

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 found at the section **Promoting good survey theory and practice around the world** on p.11 of this issue or addressed to: isimembership@cbs.nl

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@gmail.com) or Eric Rancourt (eric.rancourt@statcan.gc.ca).

ISSN 2521-991X

In this Issue

Letter from the Editors

Dear readers,

Congratulations to you on these hot July days (which may be cold in some parts of the world), just before the 64th ISI World Statistics Congress and in the 50th year of existence of the International Association of the Survey Statisticians!

This issue of The Survey Statisticians is especially thick. It contains a lot of old and new information. Each article evoked much of emotions of the editor in preparation of the materials and it is expected that each of you will find something interesting written especially for you.

According to a tradition, the issue starts from the *Letter from the President* by Monica Pratesi. She writes about citizen science and citizen generated data, one of the hot topics nowadays for survey statisticians, and not only for them. As the outgoing president of the executive committee of IASS, she looks forward to monitoring the "changing reality by developing new ways to do surveys". Scientific Secretary Maria Giovanna Ranalli in her last report enumerates main scientific activities of IASS during the second half of the year: IASS sponsored conference in June organized by the Survey Sampling Group of the Italian Statistical Society, webinar series, selection of the Cochran-Hansen prize winners and overview of the IASS participation in the ISI World Statistics Congress in Ottawa. Giovanna stresses the fruitful experience acquired by her participating in the international statistical activities and conferences.

The usual News section is replaced by the special section devoted to the 50-year jubilee of IASS. To this effect, the special section not only contains articles that directly relate to the IASS 50th anniversary, but it also contains articles that celebrate the important and interesting developments that took place in survey statistics in various themes and places. In all, 24 survey statisticians provided inputs by sharing their statistical, historical and some special knowledge with you.

What should the experts be asked? Nowadays with the wide possibilities to spread the knowledge, questions to the experts and new and emerging methods have become closer to each other. This happened because most of the questions are likely to be about new methods, their clarification and possible application. This is confirmed by the article of Sixia Chen and David Haziza on the doubly robust procedure for missing data and the article by Changbao Wu on calibration techniques for doubly robust estimation.

University teachers and practitioners should find useful the article by Thomas Lumley on the enhancement of the R *survey* package over the last 15 years and the review by Camelia Goga of Sharon Lohr's third edition of the book *Sampling: Design and Analysis*, an author who is also produced SAS and R software companion books for sampling design and analysis. The books include many exercises and examples with real data, illustrating the methodological elements.

The issue continues with country reports written by survey statisticians from 14 different countries.

We would like to thank the section editors for their creative work: Giovanna Ranalli, Ton de Waal, Alina Matei, Peter Wright and technical editor Maciej Beręsewicz.

Please inform the editors of the TSS of the news in the world of survey statistics, ask the questions which should be answered, propose topics for the articles or submit your own papers.

Danutė Krapavickaitė (danute.krapavickaite@gmail.com)

Eric Rancourt (eric.rancourt@statcan.gc.ca)

The Survey Statistician, 2023, Vol. 88, 5-6.

Letter from the President

Dear IASS Members,

Our first 50 years have led to an IASS aware of the potential of surveys, scattered within a worldwide network of representatives and with more and more members.

We are ready to look at the future, at innovation in times where technology and sustainability issues path the way to a world of measures characterised by a reality which is often 'oversized', due to climate shocks, pandemics, conflicts.

The demand for timely, more detailed, less burdensome and costly statistics, with wider coverage, is increasing. In our digital era, data are everywhere: new sources, such as mobile phones, social media interactions, electronic commercial transactions, sensor networks, smart meters, GPS tracking devices, or satellite images, produce new information at an incredible speed. Digital technologies offer new opportunities for data collection, processing, storage.

Big data, smart statistics, digital administrative data and citizens are inseparable: from smartphones, meters, fridges and cars to internet platforms, the data of most digital technologies is citizen data, that is the data of the citizens and on citizens.

Survey participation must be conquered, as contact patterns and response rates change (Calogero et al, 2022). Citizen generated data, web scraping and Citizen Science are opportunities to be taken onboard in survey design and call for the development of new data quality profiles and new properties for estimators.

The term "citizen science" means a form of open collaboration in which individuals or organizations participate voluntarily in the scientific process in various ways, including also collecting and analyzing data; and interpreting the results of data. Also known as “public participation in scientific research” the overarching motivation for Citizen Science is to bring together public participation and knowledge production. It includes the generation of scientific data, engages volunteers over a large area and addresses a politically relevant issue (Ruppert et al 2018).

Beside citizen science also the reuse of data generated by citizens on themselves (in processes like using services, making telephone calls or when they populate administrative archives during their lives as students, workers, retired persons etc.) is a challenge. It can reduce the respondent burden to produce statistics and Official Statistics. We refer to this approach as “Citizens as users and co-producers”, both in terms of big data/digital traces and administrative data, especially focusing on the latter for Official statistics production purposes (Pratesi, 2022).

Citizen generated data, web scraping and Citizen Science are opportunities to be taken onboard in survey design and call for the development of new data quality profiles and new properties for estimators.

For survey statisticians this echoes a wider issue: exploiting non-traditional data sources to produce new statistical information. This is especially relevant for SDGs, given the global perspective of the Agenda 2030 (Fraisl et al 2020).

How do survey methodologists react and have reacted? By integrating sources and surveys, revolutionizing censuses, using new administrative sources and Big data to obtain increasingly timely, granular and detailed information. Examples are the statistical production process from the Italian Permanent Population and Households Census, and the integration process of utility data

files (e.g., electricity, gas) in the future design of different activities, like for instance the Household Energy Consumption Survey (Bernardini et al, 2022).

What remains to be done? We need to continue along this path, monitoring this changing reality by developing new ways to do surveys. We need experiments with the use of new sources and the application of innovative methods in producing and analysing survey data. Experiments conducted with a multidisciplinary approach as advanced data collection technology nowadays has often made inferences from diverse data sources easily accessible.

Intensified use of Big Data by survey methodologists is a challenge also in NSIs: this requires a new approach to statistical production process, assigning a crucial role to the integration of traditional and new sources, with a special focus on digital technologies. All this requires a paradigm shift in methodology, from traditional sample survey-based estimates to a system applicable to a multi-source environment using design-based, model-based and model-assisted approaches to inference.

IASS is now on social media and participates in the debate on a variety of old and new issues with its IASS Webinars on the frontiers of survey methodology.

My birthday wishes are to celebrate the history of the association and look to its future, full of opportunities as outlined by the excellent next President and the new Executive Committee.

Don't worry: surveys are here to stay!!!

With my best wishes,

Monica Pratesi

IASS President

References

Bernardini et al (2022). Evolution of the Italian Permanent Population Census: lessons learned from the first cycle and the design of the Permanent Census beyond 2021. In: *UNECE-Conference of European Statisticians, Group of experts on Population and Housing Censuses,* Geneva, 21-23 September 2022

Carletto, Calogero et al. (2022). Positioning Household Surveys for the Next Decade. *Statistical Journal of the IAOS*, **38**, 923-946.

Fraisl, Dilek et al. (2022). Mapping citizen science contributions to the UN sustainable

development goals. *Sustainability Science*. **15,** 1735–1751

Pratesi, M. (2022). Citizen Data and Citizen Science: A Challenge for Official Statistics. In: *Studies in Theoretical and Applied Statistics. SIS 2021. Springer Proceedings in Mathematics & Statistics*, (eds: Salvati, N., Perna, C., Marchetti, S., Chambers, R.), Springer, Cham, 167-173.

Ruppert, E., Grommé, F., Ustek-Spilda, F., Cakici, B. (2018). Citizen data and trust in official statistics. *Economie et Statistique/Economics and Statistics*, **505–506**, 179–193

The Survey Statistician, 2023, Vol. 88, 7-8.

Report from the Scientific Secretary

I am writing these lines on the way back from the IASS sponsored conference ITACOSM2023, organized by the Survey Sampling Group of the Italian Statistical Society and held at the University of Calabria in Italy between June 7 and 9. Almost 100 registered participants from many countries have gathered and discussed over 60 presentations on the theme "New Challenges for sample surveys: innovation through tradition". Many IASS members were among the delegates and our president, Monica Pratesi, addressed the audience at the opening ceremony. I think that the paper of the **New and emerging methods** section of this issue of The Survey Statistician is an example of "innovation through tradition": I am very grateful to Prof. Changbao Wu (University of Waterloo, Canada) for having accepted my invitation to write a paper on **Calibration Techniques for Model-based Prediction and Doubly Robust Estimation** in which he presents an overview of calibration techniques for model-assisted estimation for probability survey samples and their extensions for model-based prediction and doubly robust estimation to missing data problems, causal inference, and analysis of non-probability samples.

When you will read these lines, the **ISI World Statistics Congress** in **Ottawa** will be approaching: we have kept working in the past six months to organize the several appointments important for IASS. First, our **General Assembly**, scheduled for Wednesday, 19th July 2023 at 12:10-13:50 pm (EDT) in hybrid format. Other than a long list of IASS sponsored invited sessions (the link to the whole program can be found here https://www.isi2023.org/conferences/15/programme/), there are three early morning Special Invited Paper Sessions (SIPS) sponsored by the IASS of particular interest: the SIPS **Waksberg Award** with a presentation by the 2023 Waksberg Award Winner, Prof. **Ray Chambers**, on Wednesday 19th July 2023 08:30-09:40, the SIPS **IASS President's Invited Speaker**, Prof. **Fulvia Mecatti**, Tuesday 18th July 2023 08:30-09:40, and the SIPS **Cochran-Hansen Prize** with presentations by our two Cochran-Hansen prize winners: **Alejandra Arias-Salazar (Costa Rica) and Ziqing Dong (China)**, Thursday 20th July 2023 08:30-09:40. There were many good candidates for the prize, and the committee chaired by the IASS VP Nikos Tzavidis had the honor and the burden to decide the winners in February. Alejandra Arias-Salazar has won with a paper on "Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach", and to Ziqing Dong with a paper on "Linearization and Variance Estimation of the Bonferroni Inequality Index".

The organization of the monthly **Webinar series** has continued, and we are particularly thankful to Andrea Diniz da Silva for her engagement. We have now reached Webinar number 30 and we are happy to have made it a monthly appointment that has attracted an audience of up to **five hundred registered participants**. Please, visit the webinar section of our website http://isi-iass.org/home/webinars/ for slides, that of ISI https://www.isi-web.org/courses-webinars-workshops for upcoming and recorded webinars. Those held in the first six months of 2023 have covered the use of state-space models for unemployment estimates, data integration for the production of official agricultural statistics, new sources of data, such as big data and sensor data in surveys and official statistics, classification trees for nonresponse in complex surveys.

During these past months, our social media activity has increased thanks to Annamaria Bianchi, our social media manager, who has continued to post about webinars, conferences, books, articles, prizes, the newsletter release and its contents, and the recruitment drive. The number of followers to the pages are increasing and in particular, we now have more than **one thousand LinkedIn followers**.

We have also worked for the future of IASS and worked for the **election of the Executive Committee 2023-2025**. We are grateful to the **Nominating Committee** chaired by our past president, Denise Britz do Nascimento Silva and including Sanjay Chaudhuri (Asia), Carolina Franco (North America), Enrico Fabrizi (Europe), James Chipperfield (Australia), and Ralf Munnich (Ex officio, Chair of Nominating Committee 2021-2023). The outcome of the elections will be communicated at the IASS General Assembly. As my duty as Scientific Secretary of IASS is coming to an end, please let me take this opportunity to express my gratitude for the fruitful experience working for IASS in the past two years. I am glad that my duties end exactly on this special edition for the IASS 50th birthday! The tradition of sampling statistics and estimation in Italy is well grounded and I have had the luck of conducting my research in a large team (almost a family) of Italian survey statisticians ever since I was a PhD student. Nonetheless, I have to say that being a member of IASS, participating to IASS sponsored sessions at World Statistics Congresses and to IASS supported conferences over the years has brought my experience and my web of scientific relations to an international level and this has greatly enhanced it. I am therefore very grateful for all the years spent together and looking forward for those to come!

Maria Giovanna Ranalli

maria.ranalli@unipg.it

IASS Scientific Secretary

The 50th Anniversary of IASS

Foreword

Dear Readers,

Congratulations to everybody who are studying, teaching, working in survey statistics, the members of IASS, one of the Associations of the International Statistical Institute, which celebrates its 50th anniversary!

This section of the newsletter consists of a collection of the congratulatory utterances and a collection of short articles depicting parts of the history of survey sampling. The section aims to overview various aspects of IASS activities from a historical perspective from different parts of the world.

The short articles tell the IASS story from its creation until current days, going through the development of survey sampling methods, achievements of young statisticians, publication of the scientific journals of the Association; history of survey statistics in some countries around the world coming to the urgent problems of today and taking a look at the future.

Some materials devoted to the jubilee of the IASS are postponed to the January 2024 issue of TSS. Historical topics of survey statistics will be continued. It may demonstrate why traditions of survey statistics are so different among the countries. They depend not only on the people and their work, but also on the economic and political situation in the country for science cannot develop during wars, repressions and famine. I invite you to share your stories on the history of the survey statistics. Please submit them before December 1, 2023.

IASS is an important international association that is very active on many fronts in this fast-paced changing world. Scientific production in survey statistics is huge. It is present in articles of scientific journals and books, implemented in software packages, presented at workshops and conferences, and applied in various fields of our life. Covid pandemic generated one more form of communication: virtual, through online meetings, webinars and conferences. This form successfully continues to exist in parallel to the traditional arrangements. It is impossible to overview everything that has been created and accomplished over 50 years by survey statisticians. The authors who contributed to this issue share their knowledge in a variety of fields covering multiple activities of the IASS. It is expected that the texts included here will evoke some thoughts, including willingness to attract young people to survey statistics.

We would like to thank the authors for their contributions and articles, for their kind attitude to the invitation of the TSS editors and for their sincere and careful work in preparation of the material. Thanks also go to the reviewers for their helpful inputs. Thank you to Jūratė Karasevičienė from Vilnius Gediminas Technical University for a jubilee emblem, to co-editor Eric Rancourt for cooperation and to Maciej Beręsewicz for putting the materials together. It was a pleasure to communicate with all of you during the preparation of this section, and I am thankful for this.

Have a pleasant reading.

Danutė Krapavickaitė

Editor of *The Survey Statistician*

The Survey Statistician, 2023, Vol. 88, 10-11.

Present at the Creation?

Ivan P. Fellegi

Statistics Canada, IASS President 1985-1987

(IASS editors: Paper re-published from the IASS web site with permission from the author)

The ISI has, for decades now (probably since its beginnings), been searching to find ways to remain relevant. During its 37th Session held in London in 1969 a fundamentally important reappraisal of the mission and modalities of ISI was tabled; the committee which authored it had been chaired by M. G. Kendall ("Report of the Reappraisal Committee"). It contained wide-ranging recommendations which are well worth rereading even today. A segment of the report dealt with the possible creation of new sections within the Institute. It suggested that "if a strong feeling arises that it would be an advantage to create an international association in a new field, we suggest the Institute ought to take the lead and set up such a society as a section of the Institute".

I don't know whether the Reappraisal Committee was initiating a new discussion about sections of ISI or whether it was reacting to ideas already circulating then. As it happens, several of us had, in fact, been agitating for a new section, one devoted to survey methodology, a field that was, in most countries in the late 1960s, still quite an undeveloped field. A very influential advocate was P. C. Mahalanobis. As a "young Turk", indeed a very young one, I was also actively agitating in favour of such a development at every available opportunity, whispering in whatever influential ear was polite enough to let me do so.

It was during the ISI session in London in 1969 that the Bureau of the ISI (chaired at that time by W. G. Cochran) asked me to join them for a discussion about the possibility of forming a section of the ISI devoted to this new and aggressively evolving discipline. They listened to me making a brief pitch, but it was my sense that they already had their minds made up to give it a try – so I cannot claim that silver-tongued oratory, even less my impeccably argued case, convinced them. As is usual, when you talk too much, you end up being asked to do what you have been agitating for. So I was asked identify to chair a small committee whose first task would be to draft terms statutes for the putative new association.

I suggested to the Bureau that the committee consist of J. P M. R. Desabie, Leslie Kish, M. N. Murthy, M. R. Sampford and S. Zarkovich and my recommendations were accepted. The drafting committee was in business. Our task was helped by an early draft of the statutes prepared by P. C. Mahalanobis although we did, in fact, draft the new statutes according to what we thought was needed. We worked by correspondence over a period of many months, and I must underline what a particularly valuable member Leslie Kish was: he never failed to respond to correspondence and he was full of good ideas – anyone surprised?

We submitted our work to the Bureau – not for approval of the statutes themselves, since we thought (and the Bureau agreed) that only the new association could adopt its own statutes – but to secure their blessing for the formation of a new Section of the ISI. They did, indeed, put forward our recommendations to the General Assembly of ISI during its 38th Session held in Washington in 1971 and this was unanimously accepted (Bulletin of the International Statistical Institute, 1971). We were asked to reconstitute ourselves as the new leadership of the formally yet to be created IASS: its Bureau, its Program Committee, its Nominating Committee and whatever else might be needed.

Had we been legalistic, we would have been facing an unsolvable problem. First of all, since there was no IASS, we lacked any legitimacy (we were unelected). Second, the Statutes clearly had to be approved by the members of the new Association, but since there was no Association until the Statutes were accepted (and until people enrolled in the new organisation), who would approve the draft statutes? We were aware of these legal niceties but not unduly worried about them. We received a number of time slots within the ISI program for sessions of the new IASS and we set about creating a program. Anders Christianson is quite right in remarking that the program in Vienna encompassed quite a broad range of topics. This was by conscious design: we wanted to establish a precedent for the broad scope of activities of the new organisation.

We also developed a list of people we wanted to nominate as the new (elected) leadership of the yet to be established IASS and we proceeded to sound out informally some leading lights of our profession.

During the first of "our" scientific meetings we set aside part of the available time for the first General Assembly of what became the IASS. Those present were asked to regard themselves as the founding members and they subsequently approved right then and there both the statutes and our proposed slate for the leadership of the new Association, under the presidency of Morris Hansen. The rest, as they say, is history; except that in this case the entire narrative is that: the early history of IASS.

Looking Back

Graham Kalton

IASS President 1991-1993

The 1973 ISI Congress in Vienna was the first Congress that I attended. It was a very enjoyable meeting, with many good sessions on survey sampling and survey methodology. I attended the first IASS General Assembly that took place at the Congress and that formally created the IASS as an organization affiliated with the ISI. The agenda for the General Assembly had been well prepared and the meeting went very smoothly. France's statistical office INSEE offered to run the IASS's secretariat and look after its finances, and that offer was gratefully accepted.

Although a number path-breaking texts on survey sampling had appeared in the 1950's and 1960's (e.g., by Hansen, Hurwitz, and Madow; Sukhatme; Murthy; Cochran; Yates; Deming; and Kish), the distinct discipline of survey statistics was not widely recognized and acknowledged by the broader statistical community 50 years ago. Looking back, I see the creation of the IASS as part of the more general emergence of the identity of survey statistics as a distinct discipline within the field of statistics. For example, around that time, the American Statistical Association's Social Statistics Section created a subsection of survey research methods in 1973 that became a separate Section on Survey Research Methods in 1977 and, in the UK, the Royal Statistical Society established a Social Statistics Section in 1976, encompassing survey methodology.

The IASS played a distinctive role in this emergence of survey statistics by establishing the field as an internationally recognized discipline. It provided—and still provides—a means for survey statisticians in different countries to share their expertise and experiences. More than that, particularly in the early days, the IASS sought to provide assistance to survey statisticians in countries that had very limited experience of sample surveys, running workshops and short courses offered at ISI Congresses with funds for statisticians from developing countries obtained from various sources. The IASS has contributed greatly to the enormous enhancements in the methods of survey research that have taken place around the world over the past fifty years, and its contributions will undoubtedly continue into the future.

Reference

Bulletin of the International Statistical Institute (1971). *Proceedings of the 38th session.* Washington.

The Survey Statistician, 2023, Vol. 88, 12-13.

Congratulations

My **very best wishes** to the members of the IASS on the occasion of its 50th anniversary. May the institution continue to grow, and be effective in promoting good survey methods all over the world.

With warm regards,

Nanjamma Chinnappa. India

IASS President 1997-1999

I have been a **member of IASS since its early days**. When I worked at the Government Social Survey in London in the early 70s, my boss Percy Gray – a brilliant survey statistician – attended the 1973 ISI meeting in Vienna where the first sessions organised by IASS were held.

Following that ISI meeting, there were several occasions when Percy told me how much he had enjoyed his discussions there with Jack Harewood, then the director of the Institute of Social and Economic Research (ISER) in Trinidad.

This highlights the advantages that a body like IASS can bring. Quite apart from any important papers presented at IASS sessions, these events provide a wonderful opportunity for networking with like-minded people with similar professional interests.

Peter Wingfield-Digby, UK

Former consultant of statistical surveys in Africa, Asia, the Pacific and the Caribbean

---

Dear IASS member,

As a **long-standing member**, I wish you can enjoy being a member of the Association as much as I have over the years. Membership to the IASS and the ISI produced tremendous impact on both my professional and personal life. I could learn and get support from the best in the field, thanks to opportunities and connections established via the Association. May this be true also for you on this 50th anniversary of the IASS and beyond. And do not forget: if you enjoy being a member, tell your best friends about it and get them enrolled! It will make it even more enjoyable having your best friends also in the fold.

Pedro Silva, IASS President 2007-2009

Society for the Development of Scientific Research (SCIENCE), Brazil

---

On the date to celebrate the **50th anniversary of the establishment of the IASS**, I offer my warmest congratulation to its leadership team, past and present, for their wonderful services to our profession, not only in providing a platform for networking amongst us, but also the opportunities offered to us to continuously improve our professional knowledge and capabilities as well.

My best wishes to IASS for its future endeavours.

Siu-Ming Tam, Ex-Chief Methodology

Australian Bureau of Statistics

Honorary Professorial Fellow, University of Wollongong, Australia

IASS is to be congratulated for its successful work during the 50 years since its creation. I am proud of having been a member for many years now.

I hope that that IASS will continue to play an important role also in the next 50 years to ensure the quality of survey statistics both in national governmental organisations and in the commercial field. I also hope that it will continue to be an association capable of handling the new challenges that is put forward by the advent of big data, machine learning and artificial intelligence. There is an increasing need for an organisation like IASS that works for correct and objective information also in the future and counteracts the increasing amount of fake news and information.

I also hope that IASS will continue to encourage and enthusiasm young students and statisticians for the field of high-quality surveys and statistics. There will always be a need for good and motivated statisticians also in the future.

Daniel Thorburn

Professor emeritus of official statistics

Stockholm University, Sweden

Dear IASS colleague members!

It is a **pleasure to celebrate with you the IASS 50th Jubilee**. Since its beginning, IASS has been promoting the development of survey statistics, providing opportunities for capacity building, bringing together people from different places, and fostering a welcoming network for survey statisticians. All of that because we have been working and learning together, enjoying a sense of belonging, as IASS matters for us. It is also time to pay tribute for those members who, before us, paved the way. Let us keep this collaborative environment for the long-lasting progress of IASS.

Congratulations to all and commemorative hugs.

Denise Silva

National School of Statistical Sciences (ENCE)

of the Brazilian Institute for Geography and Statistics (IBGE);

Society for the Development of Scientific Research (SCIENCE), Brazil

IASS President 2019-2021

I would like to convey my sincere **congratulations to the International Association of Survey Statisticians** on the 50th anniversary. I also express my gratitude to the founders of the Association, – distinguished survey statisticians I. Fellegi, late T. Dalenius, P. C. Mahalanobis, M. H. Hansen and L. Kish, – who all these years ago started extensive and fruitful discussions on survey statistics.

Over the years, the Association has had an essential role in promoting the development of the theory and practice of sample surveys. I am extremely grateful for all the publications written by the members of the Association in the Survey Statistician newsletter and beyond that help me and my fellow young statisticians to better understand the sampling methodology. The latter works inspire me as a PhD student, and guide toward further development of the sampling theory and its applications in the rapidly changing environment.

On this celebratory occasion, I wish the Association continued success for many years ahead.

Ieva Burakauskaitė

PhD Student of Mathematics at Vilnius University, Lithuania

The Survey Statistician, 2023, Vol. 88, 14-15.

Some Memorable Recollections of IASS First Meeting

J. N. K. Rao

Carleton University, Canada, jrao34@rogers.com

Abstract

International Association of Survey Statisticians (IASS) evolved during the biannual meeting of International Statistical Institute held in Vienna, Austria, 1973. Invited and contributed paper sessions were organized prior to the start of the ISI meeting, devoted to survey sampling theory and methods. I give a brief account of my participation as organizer of an invited paper session and as a speaker in another invited paper session. The first IASS meeting attracted several distinguished survey statisticians.

Keywords: International Association of Survey Statisticians, International Statistical Institute, survey sampling.

1 Introduction

IASS evolved during the biannual meeting of ISI held in Vienna, Austria, 1973. Several invited paper sessions and contributed paper sessions were organized and presented prior to the start of the ISI meetings. A variety of topics were covered during the meetings followed by stimulating discussions.

2 Two invited paper sessions

I organized an invited paper session entitled “Analytic uses of and inferences from sample surveys”. Four leading survey statisticians presented papers in this session. The first IASS Newsletter (now called Survey Statistician) lists the following presentations: (1) J. Sedransk (USA): “Design and analysis of analytical sample surveys. (2) G. Nathan (Israel): “Tests of independence in contingency tables from complex surveys. (3) W. Fuller (USA): Regression analysis for sample surveys. (4) K. R. W. Brewer and R. W. Mellor (Australia): “The effect of sample structure on analytical surveys. H. O. Hartley (USA) and T. M. F. Smith (UK) acted as invited discussants. Fuller’s paper appeared in Sankhya C (Fuller 1975), and it received a lot of attention as judged from subsequent citations.

I participated as invited speaker in another session entitled “Foundations of survey sampling” organized by C. E. Särndal (Canada). The IASS Newsletter lists the following presentations: (1) W. A. Ericson (USA): “A Bayesian approach to two-stage sampling”. (2) V. P. Godambe and M. E. Thompson (Canada): “Philosophy of sample survey practice”. (3) J. N. K. Rao (Canada): “On the foundations of survey sampling”. M. R. Sampford (UK), G. A. Barnard (UK) and R. M. Royall (USA) acted as invited discussants

I might also mention another session entitled “Sampling from imperfect and multiple frames”, organized by A. Sunter (Canada). One of the speakers in that session was H. O. Hartley (USA). W. A. Fuller (USA) acted as invited discussant. Hartley’s paper on unified theory of multiple frame

surveys appeared in Sankhya C (Hartley 1974) and this paper is also highly cited. Sankhya C started in 1974 and I was invited to serve as co-editor. I invited both Fuller and Hartley to submit their papers and both graciously agreed to my invitation. Sankhya C was devoted to survey sampling theory and methods. Unfortunately, it was discontinued after 1978.

All in all, the first IASS meeting was highly successful. The second IASS meeting was held in Warsaw, Poland in 1975 prior to the start of ISI meeting. Subsequently, IASS became a part of ISI and IASS sessions were organized as part of ISI program.

References

Fuller, W. A. (1975) Regression analysis for sample surveys. *Sankhya C,* **37,** 117-132.

Hartley, H. O. (1974) Multiple frame methodology and selected applications. *Sankhya C,* **36**, 99-118.

The Survey Statistician, 2023, Vol. 88, 16-22.

Survey Sampling During the Last 50 Years

Ton de Waal[1] and Sander Scholtus[2]

[1] Statistics Netherlands & Tilburg University, t.dewaal@cbs.nl
[2] Statistics Netherlands, s.scholtus@cbs.nl

Abstract

In this short paper we sketch how survey sampling changed during the last 50 years. We describe the development and use of model-assisted survey sampling and model-assisted estimators, such as the generalized regression estimator. We also discuss the development of complex survey designs, in particular mixed-mode survey designs and adaptive survey designs. These latter two kinds of survey designs were mainly developed to increase response rates and decrease survey costs. A third topic that we discuss is the estimation of sampling variance. The increased computing power of computers has made it possible to estimate sampling variance of an estimator by means of replication methods, such as the bootstrap. Finally, we briefly discuss current and future developments in survey sampling, such as the increased interest in using nonprobability samples.

Keywords: model-assisted sampling, mixed-mode survey designs, adaptive survey designs, variance estimation, nonprobability samples.

1 Introduction

When the editor of The Survey Statistician asked us to write this short paper on survey sampling during the last 50 years we were both honoured and intimidated. We are users of sampling theory rather than developers of new sampling theory, and many others could far better describe the ins and outs of sampling theory. We accepted the invitation anyway when we realized that most survey statisticians are actually like us: users, rather than developers, of sampling theory. Another reason for us to accept the invitation to write this short paper is that we work in official statistics. Official statistics has always been and still is a driving force behind the application of survey sampling theory in practice and the development of innovative survey sampling methods.

Sampling theory focuses on how to select a set of units, such as persons, enterprises, households, or dwellings, from a larger (finite) population of interest, and, after data collection, on how to conduct research, analyse the observed data and infer unknown properties of the population of interest.

Although we will focus here on the last 50 years, of course the history of survey sampling goes back a lot further. The seminal paper by Neyman (1934) is generally considered as the starting point of modern sampling theory. In that paper Neyman showed the benefits of using stratified simple random sampling (SRS) compared to the then popular representative approach, which essentially consisted of constructing a sample that was a miniature version of the population. Another seminal paper was Horvitz and Thompson (1952) in which they derived their well-known estimator for population totals that can be used when units are drawn with different inclusion probabilities. With hindsight, their insight may seem surprisingly simple: give each unit a weight inversely proportional to its inclusion

probability, but the apparent simplicity is probably due to the fact that the Horvitz-Thompson (HT) estimator is so often used nowadays, for instance as an essential element of a more complicated estimation process. Historically, the importance of this result – as well as the analogous result by Hansen and Hurwitz (1943) for with-replacement samples – is that it showed that unbiased estimation is possible when units are included in a sample with different probabilities, as long as these inclusion probabilities are known (and non-zero). This supported the development of other probability sampling methods than stratified SRS.

Nowadays, many different sampling methods are used, such as SRS, stratified sampling, cluster sampling, and probability proportional to size (PPS) sampling in order to obtain valid and accurate population parameter estimates in an efficient way. Sampling theory plays an important role in many different fields, such as official statistics, marketing research, epidemiology, environmental studies, and political and social sciences.

Section 2 of this paper discusses how sampling theory changed during the last 50 years. Section 3 ends with the present and some concluding remarks.

2 How sampling theory changed during the last 50 years

As we all know the computing power has increased immensely over the last 50 years. What was impossible to do 50 years ago is often quite easy – and quick – to do nowadays. These advancements in computer technology facilitated the implementation of more complex sampling designs in common practice, and improved the accuracy of estimates as well as the measurement of the accuracy. They have also inspired survey statisticians to come up with more evolved and much more complex sampling approaches than would have been possible 50 years ago.

2.1 Model-assisted survey sampling

Model-assisted survey sampling aims to combine the best of both worlds: the design-based world and the model-based world. The term 'model-assisted' is used for estimation methods that employ a model for the target variable but yield consistent estimators from a design-based point of view, even when an incorrect model is assumed (Särndal, Swensson and Wretman, 1992). The models used in model-assisted survey sampling generally rely on the availability of additional information on auxiliary variables that are related to the target variable to be measured. Such additional information often consists of population totals or population means that are known from other data sources than the survey at hand. These population totals or means can then be used improve estimates for the target variable. Regression models are often used in this context. 50 Years ago, computing power was just reaching a point where it became practical to estimate parameters of regression models during regular statistical production (Rao and Fuller, 2017) and a lot of work on model-assisted estimation was done over the next two decades.

A very important and nowadays widely used estimator is the generalized regression estimator (GREG). This is a model-assisted estimator designed to improve the accuracy of estimates when auxiliary information is available at unit level. It utilizes the relationship between the target variable and the auxiliary variables, while calibrating the sampling weights to known totals of the auxiliary variables. The GREG estimator (Cassel, Särndal, and Wretman, 1976, Särndal, Swensson and Wretman, 1992, Lohr, 1999) can be expressed as a sum of the HT estimator and a weighted difference between known totals and their HT estimators. The ratio estimator is a special case of GREG assisted by a particular model with only one covariate (Deville and Särndal, 1992). Also non-linear GREG estimators have been developed (see, e.g., Lehtonen and Veijanen, 1998). In an influential paper, Deville and Särndal (1992) introduced the family of calibration estimators, which contains many existing estimators such as GREG and procedures based on raking as special cases.

Originally, the main motivation of the theoretical work on model-assisted estimation was variance reduction. Over the past decades, GREG and other calibration estimators have been adopted widely in practice: sometimes to reduce variance, but probably more often to try to mitigate possible bias

due to selective non-response or undercoverage; see, e.g., Bethlehem (1988). Here, a slight increase in variance due to calibration is actually often anticipated in practice (Kish, 1992). In the presence of non-response, calibration estimators should be considered as model-based rather than model-assisted, since the choice of model can be crucial for bias reduction.

2.2 Complex survey designs, especially mixed-mode and adaptive survey designs

In the early years of survey sampling, a sampling design (i.e., the procedure used to select the sample) was typically used in a relatively simple survey design (i.e. the more general procedure of how to collect data). In most cases, surveys were collected by one mode only, for instance by personal interviewing, paper questionnaires, or by telephone interviewing, and only one sample had to be drawn. Nowadays mixed-mode survey designs and adaptive survey designs are often used.

Response rates have been steadily declining during the last 50 years, whereas survey costs have been steadily increasing. This has triggered the development of mixed-mode survey designs and adaptive survey designs.

Mixed-mode surveys combine different modes of data collection, such as in-person interviewing, telephone interviewing, paper questionnaires, and web questionnaires. Mixed-mode surveys aim to increase response rates, improve the representativeness of the sample, and reduce survey costs. For these reasons, mixed-mode surveys have become more common in practice in recent years. A drawback of mixed-mode surveys is that each data collection mode can introduce its own mode effect, for instance due to the fact that different groups of persons respond differently to different modes. When using mixed-mode designs, it can be hard to disentangle real changes in the population from mode effects (Schouten et al., 2021).

Adaptive survey designs are closely related to mixed-mode surveys and their aims are the same as those of mixed-mode surveys, but they take the idea a step further. Instead of deciding beforehand which data collection mode will be used for each unit selected into the survey sample, the data collection mode may be adjusted during data collection based on the data already observed. For instance, when elderly people are underrepresented in the data observed so far, one may switch to more in-person interviewing and more paper questionnaires and fewer web questionnaires than were originally planned, since elderly people are generally more likely to respond to in-person interviewing and paper questionnaires than to web questionnaires (Schouten et al., 2021).

In both mixed-mode surveys and adaptive survey designs, several sampling designs have to be used (at least one for each mode). The various sampling designs have to be aligned with each other in order to obtain accurate estimates, preferably at low costs. This obviously complicates the construction of these sampling designs.

2.3 Variance estimation

The area in survey sampling theory that probably changed the most during the last 50 years is the estimation of sampling variance. When the computing power of computers was low, the only feasible approach in practice was deriving analytical expressions for the sampling variance (or at least a good approximation thereof) for a certain sampling design and a certain estimator, and estimating these expressions. Deriving such analytical expressions actually still is the preferred approach, whenever this is possible. The problems with this approach are that this has to be repeated for each specific sampling design and estimator, and that this is often too complicated, especially for more complex sampling designs and estimators.

The increased computing power of computers has made it possible to estimate sampling variance of an estimator by means of replication. Balanced half-samples have been used by the U.S. Bureau of the Census since the late 1950s (Wolter, 2007, Rao, 2012).

The jackknife is another replication method. Although some earlier theoretical work has been done on the jackknife, Durbin (1959) seems to be the first who used the jackknife in finite population estimation.

Probably the best known and most often used replication method is the bootstrap proposed and developed by Efron (1979) (see also Efron and Tibshirani, 1994). The use of the bootstrap approach for without-replacement samples from finite populations is not straightforward and quite some work has been done to make it possible to apply the bootstrap approach in this setting. In their excellent overview paper, Mashreghi, Haziza and Léger (2016) classify the bootstrap methods for survey data of finite populations in three groups: pseudo-population bootstrap methods, direct bootstrap methods and bootstrap weights methods. In pseudo-population methods one or more pseudo-populations are constructed by copying the units of the observed sample. Next, bootstrap samples are drawn from the constructed pseudo-population(s) by mimicking the original sample design (see, e.g., Booth, Butler and Hall, 1994). Direct bootstrap methods – as their name suggests – rely on selecting bootstrap samples from the observed sample or a rescaled version thereof (see, e.g., Rao and Wu 1988, Sitter, 1992). Finally, bootstrap weights methods modify the original survey weights to obtain a new set of weights that are then used for estimation purposes (see, e.g., Rao, Wu and Yue, 1992, Beaumont and Patak, 2012).

Traditionally, sample survey theory has considered inference for target parameters of a given finite population. An area that has received increasing attention over the past 50 years is the use of survey data for analytical purposes, i.e., where the finite population itself is not of particular interest. In practice, variance estimation and inference for analysis on complex survey data often was – and occasionally still is – done using simple ad hoc solutions. Nowadays, well-founded approaches are available in the literature (see, e.g., Chambers and Skinner, 2003) and also in statistical software, such as the R package `survey` (Lumley, 2010). A concept that is necessary in this context is that of a superpopulation model. We suppose that a finite target population of size N is drawn from this model. A survey sample of size n is then drawn, possibly by some complex design, from this finite population. Often, the same design-based estimator can be used to estimate either a parameter of the finite population (e.g., "the number of serious traffic accidents that occurred last year") or a parameter of the superpopulation model (e.g., "the expected number of serious traffic accidents to occur within one year"), but the associated sampling variance is different. This distinction becomes relevant for inference when the sampling fraction n/N is not negligible or, more generally, when some units in the population have large inclusion probabilities. The latter situation is quite common for business surveys. Standard design-based bootstrap methods do not capture the overall variability (due to the model and sampling design) when the sampling fraction is large. Beaumont and Charest (2012) developed a bootstrap variance estimation method for model parameters that can be used for large (or small) sampling fractions.

3 The present and concluding remarks

There is one important recent development that we have not discussed so far: the use of nonprobability samples, alone or in combination with probability samples. Probability samples, which are drawn according to a well-designed sampling design, enable statisticians to draw valid conclusions about population parameters of interest by using well-known estimators such as the HT or the GREG estimator. Unfortunately, the collection of probability samples is time-consuming, expensive and affected by non-response. Nowadays, many nonprobability samples, which do not come from a known sampling design, are available at low cost and within a short time. Examples are Big Data, register data and opt-in online surveys. Since the "sampling design" (if any exists) of such a nonprobability sample is unknown to the statistician, it is a major challenge to produce valid and accurate estimates for population quantities of interest.

Nonprobability samples have been used for many decades already, for instance in marketing research where quota sampling and snowball sampling are often used. However, nowadays many

more nonprobability samples, and many other applications besides those in marketing research, such as applications in official statistics, are considered.

The main problems of nonprobability samples are that they are likely to be selective regarding the population and that the selection probability of units is usually unknown (Elliott and Valliant, 2017). This means that estimators for population quantities of interest are likely to suffer from selection bias. To solve the issue of selection bias, some approaches focus on predicting the target variables or parameters at the population level, whereas other approaches focus on estimating the inclusion probabilities of the units in the nonprobability sample. The two approaches can also be combined to achieve doubly robust estimation (Chen, Li and Wu, 2020). For reviews of existing methods, we refer to Elliott and Valliant (2017), Cornesse et al. (2020), Valliant (2020), Rao (2021) and Wu (2022). Research on the use of nonprobability samples is very much alive and seems a promising way to improve quality of survey estimates and at the same time reduce costs.

Nonprobability samples also generate a lot of related research. For instance, since some nonprobability samples are quite large, 'sampling' variance becomes less important, whereas selection bias, coverage bias and measurement bias become more important (see, e.g., Rao, 2021). Another rather new field of research is combining a nonprobability sample with a traditional survey sample when the target variable is available in both samples (see, e.g., Wiśniowski et al., 2020).

Given the limited space, we hardly discussed non-response in this paper (see, e.g., Little and Rubin, 2002, Raghunathan, 2016). We point out that non-response is obviously closely related to survey sampling. In fact, a sample survey can be seen as missingness by design, since the units not included in the sample are 'non-respondents' by design. We did not discuss small area estimation at all, even though this has become an important topic ever since the seminal paper by Fay and Herriot (1979) and small area methods are nowadays widely used at national statistical institutes (see Rao and Molina, 2005).

In this paper, we have given a brief overview of survey sampling during the last 50 years. Due to space restrictions, we had to limit ourselves to describing only some of the most important papers on this topic. We realize that this does not do justice to the work done by many excellent survey statisticians. For more extended reviews of survey sampling, we refer to Rao (2005), Rao and Fuller (2017) and to the first sections in Rao (2021).

Acknowledgement

We thank Jean-François Beaumont for his very useful and valuable comments on our paper.

References

Beaumont, J.-F., Charest, A.-S. (2012) Bootstrap Variance Estimation with Survey Data when Estimating Model Parameters, *Computational Statistics and Data Analysis*, **56**, 4450–4461.

Beaumont, J.-F., Patak, Z. (2012) On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, **80**, 127–148.

Bethlehem, J.G. (1988) Reduction of Nonresponse Bias through Regression Estimation. *Journal of Official Statistics*, **4**, 251–260.

Booth, J.G., Butler, R.W., Hall P. (1994) Bootstrap Methods for Finite Populations. *Journal of the American Statistical Association*, **89**, 1282–1289.

Cassel, C.M., Särndal, C.-E., Wretman, J.H. (1976) Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika*, **63**, 615–620.

Chambers, R.L., Skinner, C.J. (eds.) (2003) *Analysis of Survey Data*. John Wiley & Sons, Chichester.

Chen, Y., Li, P., Wu, C. (2020) Doubly Robust Inference with Nonprobability Survey Samples. *Journal of the American Statistical Association*, **115**, 2011–2021.

Cornesse, C., Blom, A.G., Dutwin, D., Krosnick, J.A., De Leeuw, E.D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J.W., Struminskaya, B., Wenz, A. (2020) A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, **8**, 4–36.

Deville, J.C., Särndal, C.-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**, 367–382.

Durbin, J. (1959) A Note on the Application of Quenouille's Method of Bias Reduction to the Estimation of Ratios. *Biometrika*, **46**, 477-480.

Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, **7**, 1–26.

Efron, B., Tibshirani R.J. (1994) *An Introduction to the Bootstrap*. Chapman and Hall/CRC, New York.

Elliott, M.R., Valliant, R. (2017) Inference for Nonprobability Samples. *Statistical Science*, **32**, 249–264.

Fay, R.E., Herriot, R.A. (1979) Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **85**, 398-409.

Hansen, M.H., Hurwitz, W.N. (1943) On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, **14**, 333–362.

Horvitz, D.G., Thompson, D.J. (1952) A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663-685.

Kish, L. (1992) Weighting for Unequal P_i. *Journal of Official Statistics*, **8**, 183–200.

Lehtonen, R., Veijanen, A. (1998) Logistic Generalized Regression Estimators. *Survey Methodology*, **24**, 51–55.

Little, R.J.A., Rubin, D.B. (2002) Statistical Analysis with Missing Data (second edition). *John Wiley & Sons*, New York.

Lohr, S.L. (1999) *Sampling: Design and Analysis*. Duxbury Press. Pacific Grove.

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

Mashreghi, Z., Haziza, D., Christian Léger, C. (2016) A Survey of Bootstrap Methods in Finite Population Sampling. *Statistics Surveys*, **10**, 1–52.

Neyman, J. (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97**, 558-625.

Rao, J.N.K. (2005), Interplay between Sample Survey Theory and Practice: An Appraisal, *Survey Methodology*, **31**, 117–138.

Rao, J.N.K. (2021) On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*, **83**, 242–272.

Rao, J.N.K., Fuller, W.A. (2017) Sample Survey Theory and Methods: Past, Present, and Future Directions. *Survey Methodology*, **43**, 145-160.

Rao, J.N.K., Molina, I. (2015) Small Area Estimation (second edition), John Wiley & Sons, New York.

Rao, J.N.K., Wu, C.F.J. (1985) Inference from Stratified Samples: Second-Order Analysis of Three Methods for Nonlinear Statistics. *Journal of the American Statistical Association*, **80**, 620–630.

Rao, J.N.K., Wu, C.F.J., Yue, K. (1992) Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, **18**, 209–217.

Raghunathan, T. (2016) *Missing Data Analysis in Practice*. CRC Press, Boca Raton.

Särndal, C. E., Swensson, B., Wretman, J.H. (1992) *Model-Assisted Survey Sampling*. Springer-Verlag, New York.

Schouten, B., Van den Brakel, J. Buelens, B., Giesen, D., Luiten, A., Meertens, V. (2021) *Mixed-Mode Official Surveys: Design and Analysis*. Chapman and Hall/CRC, New York.

Sitter, R.R. (1992) A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association*, **87**, 755–765.

Valliant, R. (2020) Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, **8**, 231–263.

Wiśniowski, A., Sakshaug, J.W., Perez Ruiz, D.A., Blom, A.G. (2020), Integrating Probability and Nonprobability Samples for Survey Inference, *Journal of Survey Statistics and Methodology*, **8**, 120-147.

Wolter, K.M. (2007) *Introduction to Variance Estimation (second edition)*. Springer Science+Business Media, New York.

Wu, C. (2022) Statistical Inference with Non-Probability Survey Samples. *Survey Methodology*, **48**, 283–311.

The Survey Statistician, 2023, Vol. 88, 23–35.

Historical Overview of Small Area Estimation in the 50th Birthday of the IASS

Isabel Molina[1] **and J. N. K. Rao**[2]

[1]Institute of Interdisciplinary Mathematics, Department of Statistics and Operations Research, Faculty of Mathematics, Complutense University of Madrid, Spain, isabelmolina@ucm.es

[2]School of Mathematics and Statistics, Carleton University, Ottawa, Canada, jrao@math.carleton.ca

Abstract

To celebrate the 50th birthday of the IASS, this paper presents a historical overview of SAE methods, focusing on the main ideas and theories that have had a significant impact in the SAE methodology. Starting from estimators obtained under design-based theory, we describe simple indirect methods, including synthetic and composite estimation procedures. Then we go through model-based SAE methods, starting with area-level models, then going through unit-level models and finally describing the more up-to-date procedures for the estimation of complex non-linear indicators, such as poverty and inequality indicators. Due to the applied nature of SAE, we enhance applications of the methods, describing important government programs that regularly produce SAE estimates.

Keywords: Area effects; Mixed models; Model-based inference; Poverty mapping; Small domain.

1 Introduction

Launched at the 39th ISI conference held in Vienna in August, 1973, the IASS was founded as a section of the ISI by Tore Dalenius, Ivan Fellegi, Morris Hansen, Leslie Kish and P.C. Mahalanobis, so this paper is written to celebrate its 50th birthday.

As Anders Christianson notes in “Aims and history” of the IASS (http://isi-iass.org/home/aims/), apart from being devoted to promote survey sampling, “the most important reason for the creation of the IASS was to address major limitations of sampling theory”. The field of small area estimation (SAE) was actually born to address a major limitation of traditional design-based sampling theory, to meet the (public and private) demands of estimates at more disaggregated levels than those for which surveys were originally planned. “Quick and cheap” disaggregated yet reliable statistical information was needed worldwide in policy making, for the formulation of assistance and development programs, or directly for the allocation of government funds in an efficient way.

SAE came out of the shelter of sampling theory in the hands of other disciplines and theories, such as demography and model-based inference, and started growing exponentially by the second half of 20th century, partly thanks to the development and expansion of linear and generalized linear mixed regression models. By the 21st century, this growth has been also stimulated by the pressure that international organizations, like The United Nations, have put on countries to meet the Millennium Development Goals from 2000 to 2015, and the Sustainable Development Goals after 2015. Appropriate monitoring of the progress of these goals relies on timely, accurate disaggregated statistical information.

The IASS has also promoted the expansion of SAE by co-sponsoring several SAE conferences and organizing SAE sessions and short courses on SAE within the IASS meetings. The importance of SAE within the IASS is also witnessed by the many SAE researchers that have been or are currently involved in the IASS committees.

This increasing need of detailed statistical information has lead to the development of a variety of SAE methods that are specific for the type of estimates that need to be produced and the possibilities offered by the information that is available for that. In the US, many of these SAE methods have long been used in official programs to produce regular estimates. For example, the Small Area Income and Poverty Estimates (SAIPE) Programme of the U.S. Census Bureau (https://www.census.gov/programs-surveys/saipe.html), which started back in 1993, produces estimates of school-age children in poverty, regularly for the counties and school districts. The Local Area Unemployment Statistics (LAUS) Program of the Bureau of Labour Statistics produces local monthly and annual employment, unemployment, and labor force statistics. The County Estimates Program of the National Agricultural Statistics Service (NASS) produces county crop yield estimates. The Substance Abuse and Mental Health Services Administration produces estimates of substance abuse in states and metropolitan areas. The US Department of Health and Human Services produces health status, health care access and family income estimates. The latter estimates are used to formulate an energy assistance program for low-income families. For an excellent account of the use of indirect estimators in US Federal Programs, see Schaible (1996).

In Canada, reliable monthly unemployment rates for small areas are used to determine the rules used to administer the employment insurance (EI) program. In Latin America, many countries are currently producing small area estimates of poverty. For example, in Mexico, there is a mandate to produce poverty estimates by municipality every 5 years, and the Mexican National Survey of Household Income and Expenditure (ENIGH in Spanish) alone cannot provide estimates for the municipalities with adequate quality. In Europe, the SAE methodology expanded greatly thanks to projects funded by the European Commission, like EURAREA (https://cros-legacy.ec.europa.eu/content/eurarea_en), SAMPLE (http://www.sample-project.eu/) and AMELI (https://cros-legacy.ec.europa.eu/content/ameli_en). The Italian National Statistical Office uses since 2006 unit-level SAE methods to obtain employment and unemployment indicators for Labour Market Areas(https://www.istat.it/en/archivio/276035). Central and eastern European countries, which moved away from a centralized decision making, have also played a prominent role in the expansion of the SAE methodology, participating in European projects and organizing several conferences related with SAE; for example, the first two international conferences on SAE were held in Warsaw in 1992 (Poland) and in Riga (Latvia) in 1999. Worldwide, the World Bank and the United Nations, specially the Economic Commission for Latin America and the Caribbean (ECLAC), the Economic and Social Commission for Western Asia (ESCWA), UN Statistics Division (UNSD) and UN Population Fund (UNPF), among possibly other, have sponsored multiple activities aimed at building capacities for countries to produce accurate disaggregated socio-economic statistical information, see e. g. the UN Toolkit on SAE (https://unstats.un.org/wiki/display/SAE4SDG).

Here we make a very limited historical overview of the literature on small area estimation, starting from direct methods based on area-specific survey data, going through simple indirect methods that include synthetic and composite estimators, more advanced indirect methods based on models at the area and unit levels, with the many different variants depending on the target indicators and the available data, and finishing with procedures designed for the estimation of general, possibly non-linear, area parameters. We place emphasis on the ideas and theories that represented breakthroughs in SAE, focusing specially on mainstream model-based SAE methods and mentioning practical applications of many of the methods.

Books on SAE include Mukhopadhyay (1998), Rao (2003), Longford (2005), Chaudhuri (2012), Rao and Molina (2015), and the recent book by Morales et al. (2021). Good accounts of SAE theory are also given in the books by Fuller (2009), Chambers and Clark (2012), Pratesi (2016), Jiang (2017) and Sugasawa and Kubokawa (2022).

Important reviews on SAE are given in Ghosh and Rao (1994), Pfeffermann (2002, 2013), Jiang and Lahiri (2006), Datta (2009), Lehtonen and Veijanen (2009) and Ghosh (2020). Reviews focused on SAE for welfare and poverty are given by Guadarrama, Molina and Rao (2014), Pratesi and Salvati (2016), Rao and Molina (2016), Molina (2019), Molina, Rao and Guadarrama (2019) and, more recently, Molina, Corral and Nguyen (2022).

2 From direct estimation to early indirect methods

The first estimates based on sample surveys that were intended for subpopulations were "direct", in the sense that they used only the survey data from the subpopulation of interest without "borrowing strength". These estimates are developed under the umbrella of sampling theory, which has long history. For nice accounts of this theory, see the books by Cochran (1977), Särndal, Swensson and Wretman (1992), Thompson (1997), Lohr (1999) and Wu and Thompson (2020). Direct estimators have several advantages, when applied to areas with large sample sizes. The usual direct estimators have good design properties (at least design consistency as the area sample size n_d increases) and avoid making distributional assumptions for the study variable. Another important advantage of direct estimators is that they use "all-purpose" expansion weights, in the sense that the same expansion weights are used for the estimation of totals or means of whatever variable of interest, making the production of large amounts of statistical information automatic.

Generalized Regression (GREG) estimators and more general calibration estimators (Deville and Särndal, 1992; Lehtonen, Särndal and Veijanen, 2003) applied to domains were designed to improve the efficiency of direct domain estimators, owing to the knowledge of the domain totals of some auxiliary variables. These procedures adjust the sampling weights, and the adjusted weights can be used similarly to estimate totals or means of other variables of interest. Nowadays, expansion weights are typically calibrated using the known totals of certain auxiliary variables and are also adjusted for non-response. However, the resulting calibration estimators are still inefficient for areas with small sample size n_d. Even if a more efficient allocation of the total survey sample size n among the different areas at the design stage of the survey (which is recommendable if estimates need to be produced for those areas) might ameliorate the SAE problem, "the client will always require more than is specified at the design stage" (Fuller 1999; p. 344).

The way of addressing the scarcity of data within some of the areas is to obtain indirect estimates, which "borrow strength" across areas, by making some homogeneity assumptions that link the areas through common parameters. These common parameters are estimated with a larger sample size, which leads to more efficient small area estimators. The idea of sharing information within a larger area appeared already in the first demographic methods dating back to 1950, such as the Vital Rates (VR) method due to Bogue (1950). This method assumed that the ratios between the birth/death

rates in two time periods in the small area of interest were constant within a larger area covering that small area. These first indirect methods used only census data and demographic information from administrative records, and were absent of sampling. Detailed accounts of the traditional demographic methods are given by Purcell and Kish (1979), National Research Council (1980), Rives, Serow, Lee and Goldsmith (1989), Statistics Canada (1987), Zidek (1982) and Rao (2003).

The VR method is "synthetic", because the change in the birth/death rate between two time periods is assumed to be the same for all the small areas contained in the larger area, without allowing for specific area behaviour. According to Gonzalez (1973), "An estimator is called a synthetic estimator if a reliable direct estimator for a large area, covering several small areas, is used to derive an indirect estimator for a small area under the assumption that the small areas have the same characteristics as the large area". Post-stratified synthetic estimators, which assume that the means of the study variable do not vary within large post-strata and only vary between post-strata, are perhaps the simplest synthetic estimators based on survey data. The US National Center for Health Statistics (1968) pioneered the use of synthetic estimation for developing state estimates of disability and other health characteristics from the National Health Interview Survey (NHIS), because NHIS sample sizes in many states were too small to provide reliable direct state estimates. Synthetic estimators can have very small design variances, but their design bias can be substantial because the assumptions behind synthetic estimators are typically strong and unrealistic. Since their design bias is not negligible, design MSE estimates that account for both bias and variance should be used to accompany the synthetic point estimates. Apart from the potentially large bias, a problem is that obtaining efficient and area-specific design MSE estimates is still a challenge for these estimators.

Composite estimators, defined as a weighted average of a synthetic estimator and a direct estimator for the same area, were proposed as a compromise between the small design variance but potentially large bias of synthetic estimators and the small design bias but inefficiency of direct estimators. Curiously, averaging different predictors is nowadays one of the main ideas behind modern machine learning procedures.

In the composite SAE estimators, optimal weights are sought from a design-based standpoint. However, the optimal weight depends on the true design MSE estimates of the two estimators involved, encountering again the problem of estimation of the design MSE for synthetic estimators. Griffiths (1996) studied composite estimators and applied them to the estimation of labor force characteristics for US congressional districts.

Purcell and Kish (1979) considered a common weight for all the areas and obtained the optimal weight that minimized the total design MSE for all the D small areas. The resulting composite estimators have good overall efficiency for the D areas, but not necessarily for each small area. In SAE, it is desirable to reduce the largest MSEs, which typically correspond to the areas with the smaller sample sizes, and this is not ensured by these composite estimators.

Composite estimators shrink direct estimators toward the synthetic ones. The idea of shrinking appears already in the James-Stein (JS) method proposed by James and Stein (1961), see also Efron and Morris (1972) and the famous application by Efron (1975) to the estimation of batting averages of major league baseball players in US during 1970 season. In the JS method, direct estimators are shrunk toward a fixed guess of the true quantity for area d, which can be taken as the average across areas of the direct estimators in the absence of auxiliary information, or to the regression-synthetic estimator when auxiliary information is available. This method applies again a constant weight to the two estimators involved, but in SAE, it is much more appealing to consider area-specific weights, with weight attached to the synthetic estimator that grows for the areas with small area sample sizes and decreases for the areas with large sample sizes (giving then more weight to the direct estimator). Following this idea, Drew, Singh and Choudhry (1982) proposed the sample-size dependent

(SSD) estimators, which are composite estimators defined with simple weights that depend on the area sample size. They applied these estimators to produce estimates for Census Divisions from the Canadian Labor Force Survey. In practice, as it happened in the application by Drew, Singh and Choudhry (1982), SSD estimators borrow little or no strength, because the weights attached to the direct estimators often turn out to be either equal or close to one.

The advent of computers produced an explosion in the number and complexity of SAE procedures, most of them based on regression models. The first SAE models actually lead also to composite estimators, but with optimality properties under model assumptions for the study variable. These estimators dominate the above composite estimators by borrowing substantial strength from the other areas. They can achieve large efficiency gains, provided that the model assumptions hold. An important drawback of model-based estimation procedures is that all the modelling and estimation process, including model validation, is specific to each variable of interest, not allowing for automatic production. This might be one of the reasons why there is a delay in the introduction of SAE procedures in the production processes of National Statistical Offices.

3 From the first explicit model to modern area level models

Perhaps the first application of a model for SAE is due to Hansen, Hurwitz and Madow (1953), p. 483, based on the 1945 Radio Listening Survey. The target was to estimate the median number of radio stations heard during the day in the family houses from 500 U.S. counties. They had estimates x_d, $d = 1, \ldots, D = 500$, obtained from a mail survey conducted in the 500 counties, which were biased due to only 20% response rates and incomplete coverage. Unbiased estimates y_d were obtained from an intensive survey conducted in 85 of the counties. A linear regression model for y_d with x_d as auxiliary variable was used, by regarding the y_d as true values for the 85 sample counties. The fitted regression parameters were then applied to predict the number of radio stations heard during the day in the remaining 415 counties, where the mail survey estimates x_d were available. The resulting predicted values do not account for the fact that y_d are subject to sampling error.

The use of linear mixed models (Searle, 1971; Searle, Casella and McCulloch, 1997; Jiang, 2007) that account for unexplained area heterogeneity really represented a breakthrough in the SAE methodology. The best linear unbiased predictor (BLUP) of a mixed effect (a linear combination of fixed and random effects) under a linear mixed model was obtained by Henderson (1950), in a different context from SAE, related with the prediction of the milk yield of dairy cows. On a completely different context, dealing with estimation of mean per capita income in US areas with less than 1,000 inhabitants, Fay and Herriot (1979) also considered a linear regression of the true area means μ_d in terms of certain area-specific covariates x_d (linking model). However, to account for the (important) sampling errors of the direct estimators y_d of μ_d, they considered an additional sampling model for y_d in terms of μ_d, which, together with the linear regression for μ_d, yields a linear mixed model, known popularly as the Fay-Herriot (FH) model. Based on this model, Empirical BLUPs (EBLUPs) of the true area means μ_d were obtained.

The FH model is still very popular nowadays, because it requires only aggregated data at the area level, accounts for the survey design, through the direct estimators, and accounts for potential unexplained between-area heterogeneity. As a consequence, the resulting EBLUP is a weighted average of the direct and the regression-synthetic estimator, with area-specific weights. Actually, the weight attached to the regression-synthetic estimator is larger for areas where the direct estimator is inefficient (large sampling errors) and smaller for areas where the direct estimator is efficient. The property of approaching the direct estimator as the area sample size grows is appealing, because it ensures design consistency as the area sample size n_d grows. Moreover, if the model parameters were known, EBLUPs based on FH model cannot be less efficient than the direct estimators in

terms of MSE. FH model parameters are estimated by fitting the model to the direct estimators for all the areas (hence borrowing strength). As a consequence, the efficiency of the estimated parameters increases as the number of areas grows. Perhaps the main issue with FH model is that the sampling variances of direct estimators need to be given and are typically deemed as fixed values (without sampling error). Generalized variance function (Vaillant, 1987) is typically applied to smooth these sampling variances, and the smoothed variances are then treated as the true ones. However, when comparing the resulting EBLUPs with direct estimators in applications, it is unclear whether the comparison should be done using the estimated sampling variances or the smoothed versions.

The FH model is regularly used in the US Census Bureau, within the SAIPE project, see Bell (1997). It was also used by Ericksen and Kadane (1985) and Cressie (1989) to estimate the decennial census undercounts in each US state, and Dick (1995) employed the model to estimate Canadian census undercounts. To mention just a few applications of the FH model to estimate welfare indicators, Molina and Morales (2009) estimated poverty rates and gaps in Spanish provinces by gender, Jedrzejczak and Kubacki (2013) estimated income inequality and poverty rates by regions and family type in Poland, and Casas-Cordero Valencia, Encina and Lahiri (2015) estimated poverty rates in Chilean comunas based on the FH model with arcsin transformation of the direct estimators.

Other ways of "borrowing strength" were explored in multiple extensions of the FH model, like the multivariate versions, and models including temporal and/or spatial correlation. Recently, the FH model was extended to include area level covariates obtained from "big data" typically based on non-probability sampling. Marchetti et al. (2015) used big data based on mobility comprised of different car journeys in Italy automatically tracked with a GPS device.

The introduction of Generalized Linear Models (GLMs) by Nelder and Wedderburn (1972) (see also McCullagh and Nelder, 1989), represented a huge step that expanded the use of statistical models in general. After that, two-level GLMs were then applied to estimate mortality or disease rates and obtain corresponding mortality/disease maps. The first proposal, based on a Poisson-Gamma model, was perhaps due to Clayton and Kaldor (1987), who also introduced a model with Conditionally Autoregressive (CAR) area effects. Generalized linear mixed models, or the more general two-level GLMs, have then long been used in many disease mapping and small area applications, with many variants developed, e.g. multivariate versions, or including temporal and/or spatial correlation, etc.

4 Unit level models

The concept of a superpopulation model for two-stage sampling introduced by Scott and Smith (1969) led to important advances in SAE, specially when estimating non-linear area parameters based on unit-level data. The first unit-level model for SAE was proposed by Battese, Harter and Fuller (1988), which was a linear regression model with random area effects, popularly known as the nested error model. They used this model to obtain EBLUPs of county means of crop areas under corn and soybeans, using farm-interview data and auxiliary information obtained from LANDSAT satellite images. Although EBLUPs under a linear mixed model were derived by Henderson under the "infinite" population setup, Royall (1970, 1976) developed EBLUP theory under the finite-population setup without focusing on small areas, see Vaillant, Dorfman and Royall (2001). Current mainstream SAE procedures apply this theory to small areas, by assuming a superpopulation model that links all the areas through common parameters. These common parameters are estimated with the overall survey data from all the areas, which yields substantial increases in the efficiency of model-based estimators compared to direct estimators.

When the area sampling fractions are negligible, the EBLUP of an area mean $\bar{Y}_d$ obtained under the finite population setup with superpopulation model defined by the nested error model, approximates the EBLUP of a mixed effect from the same model under Henderson's infinite population setup, but

this is not the case for non-linear area parameters. To mention just a few other applications of the nested error model, it has been used by Militino et al. (2006) to estimate the area occupied by olive trees in non-irrigated areas at the central region of Navarra in Spain and by Mauro et al. (2015) to estimate means of forest variables of interest by forest regions, based on remote sensing auxiliary data.

Until the first decade of the current century, model-based SAE procedures had focused mainly on means or totals of the variable that is used as model response, since EBLUPs were designed to estimate only linear functions of the model response variables. However, many poverty and inequality indicators cannot be expressed as linear functions of the response variable. Even if the interest was to estimate simple area means of a given variable of interest, once a non-linear transformation (such as log) is taken as response in the model (often done for monetary variables to achieve approximate normality), EBLUPs might not be useful anymore. Note that taking the inverse transformation of EBLUP predictions might lead to severe bias, see Molina and Martín (2018).

Probably the first SAE procedure that was designed for the estimation of general parameters is that of Elbers, Lanjouw and Lanjouw (2003), known as ELL method. This method was based on the nested error model of Battese, Harter and Fuller (1988), but where the random effects in the model were associated to the sampling clusters (or 1st stage units), and including heteroscedasticity. ELL method was used until 2020 as the default method for mapping poverty or inequality at the World Bank and perhaps was the most extensively used method across the globe for that purpose. This is partly because of the simple point and click software `PovMap` software (Zhao, 2006), which was also extremely computationally fast and efficient in terms of memory.

Banerjee et al. (2006) reviewed the research conducted at the World Bank and did already raise concerns about the ELL method, suggesting that it was not accounting for potential area effects. Actually, as Molina and Rao (2010) showed, even if taking the clusters as the small areas of interest in the ELL method, the ELL estimators of the welfare means under a nested error model for the welfare without any transformation, are synthetic. Banerjee et al. (2006) also raised concerns about the ELL estimated standard errors, which were not accounting for the correlation between the observations in different clusters within the same area. These two problems were solved by the Empirical Best (EB) method and the bootstrap MSE estimation procedure proposed in Molina and Rao (2010), work that was developed under the support of the SAMPLE project.

Similar to the ELL method, EB combines survey data with census (or administrative records) auxiliary data, uses a unit-level model for the welfare variable (or a one-to-one transformation of it) and it is able to estimate very general (and several) indicators that depend on the welfare, based on the same model. Nevertheless, apart from being approximately unbiased, EB estimators are nearly optimal, in the sense of minimum mean squared error under the model. Consequently, EB provides estimators with better efficiency than ELL estimators when the nested error model assumptions hold, and in certain cases the gains in efficiency with respect to ELL may be quite large, as illustrated by Molina and Rao (2010) and later in Corral, Molina and Nguyen (2021). The EB method was implemented within the `sae` R package (Molina and Marhuenda, 2015) in the homoscedastic case, as well as in Stata (Nguyen et al., 2018, https://github.com/pcorralrodas/SAE-Stata-Package). Many SAE methods have been implemented in multiple R packages, as well as in other software packages, but a software review is out of the scope of this paper.

The EB method has been applied to estimate poverty indicators in Spanish provinces by gender (Molina and Rao, 2010), mean income in Mexican municipalities (Molina and Martín, 2018), mean income and (non-extreme) poverty rates for census tracks by gender in Montevideo, Uruguay, and poverty rates and gaps in Palestinian localities by gender (Molina Peralta and García Portugués, 2020).

Corral, Molina and Nguyen (2021) extended the model-based simulation experiment of Molina and Rao (2010) to more realistic scenarios with a much better explanatory power of the model and including also contextual variables, with much larger area population sizes and much smaller sampling fractions, generating errors from a Student's t_5 instead of a normal distribution, and also decreasing the overall sample size and the area sample sizes. Additionally, Corral et al. (2021) performed a design-based validation study, using the Mexican Intracensal Survey as a fixed census, and then drawing from it 500 samples using a realistic sampling method. The superiority, in terms of MSE, of the EB over the traditional ELL in all these experiments lead to a revision of the World Bank methodology for poverty mapping and the corresponding software (https://github.com/pcorralrodas/SAE-Stata-Package). This revision incorporates several variants of the EB estimators of Molina and Rao (2010) and the parametric bootstrap procedure for MSE estimation of González-Manteiga et al. (2008).

The nested error linear regression model has been extended to models with non-parametric mean functions. Opsomer et al. (2008) proposed penalized spline regression models. Recently, Krennmair and Schmid (2022) have used machine learning methods; in particular, mixed-effects random forests, for SAE.

5 Concluding remarks

We have made an overview of SAE methods, going from the basic direct and indirect methods to the modern model-based procedures for SAE, including methods developed for the estimation of non-linear area indicators and variants of the basic methods. Really important topics in SAE like model fitting methods and their properties, methods for MSE estimation or calculation of prediction intervals, have not been covered owing to space-time restrictions, details of those topics can be found in Rao and Molina (2015). Moreover, we have mainly focused on frequentist or empirical Bayes procedures. Descriptions of Hierarchical Bayes (HB) SAE methods can be found in Ghosh and Meeden (1997), Malec et al. (1997), Ghosh et al. (1998) and also Rao and Molina (2015).

Even if the usual SAE models that include area effects are more flexible than the corresponding regression models without the area effects (which lead to synthetic estimators), we cannot forget that properties of all model-based estimators depend on the model assumptions. Hence, the assumed model needs to be carefully checked with the available data, e.g. by using customary residual plots, see Rao and Molina (2015) for model diagnostics in the basic SAE models, although more research is probably needed on this important issue.

In the case of clear model departures, the model should be changed to accommodate to data features or the final estimates should be taken with a lot of caution. This is related to another important issue, which is the estimation of area parameters in non-sampled areas. Note that the model assumptions cannot be checked for non-sampled areas and, unless additional information is available, we cannot be sure that these areas satisfy the assumed model. Moreover, as already discussed, synthetic estimators used for those areas are inefficient if area effects are significant. Hence, unless legally bound, a general recommendation is not producing estimates for non-sampled areas.

Once the sample is drawn from the population, the model for the sample part $\mathbf{y}_s$ of the population vector $\mathbf{y} = (\mathbf{y}_s', \mathbf{y}_c')'$ (for which a superpopulation model is assumed) is simply obtained by marginalization; that is, integrating out with respect to the sample complement part $\mathbf{y}_c$. The sample model for $\mathbf{y}_s$ then has the same shape as the superpopulation model when sampling is ignorable, but this does not hold for non-ignorable (informative) sampling. Similarly, the model for the respondents might be different from the model for the sample units under non-ignorable non-response. Methods for SAE accounting for the sampling design have been discussed already. However, concerning model checking, skeptical survey samplers might raise the concern that the superpopulation model cannot be checked under informative selection and/or non-ignorable non-response, because population data

are not available. In this regard, it is important to point out that only the sample/respondents model needs to be checked with the available sample/respondents data.

Another important point is that, when estimating non-linear area parameters based on unit-level models, the values of the auxiliary variables are required for each population unit. This microdata is typically obtained from the most recent census or from administrative records, which are usually protected for privacy reasons, and this protection limits the practical applicability of these methods. Another important issue is that outdated information in the census file for inter-censal years might yield severely biased small area estimators. Corral et al. (2021) analyzed the empirical properties of the common approaches for that case, but further research is probably needed on this important issue.

Finally, conventional MSE estimates of model-based estimators are obtained assuming that the corresponding model assumptions hold, even if we know that "All models are wrong, but some are useful". Hence, these MSE estimators might be understating the real uncertainty. Molina and Strzalkowska-Kominiak (2020) and others proposed to use the same idea of "borrowing strength" behind SAE, for the estimation of the design MSE of small area means, which accounts for model uncertainty. Design MSE estimation for general non-linear indicators is an interesting topic that also deserves further research.

References

Banerjee, A. V., Deaton, A., Lustig, N., Rogoff, K., and Hsu, E. (2006) An evaluation of World Bank research, 1998-2005. Available at SSRN 2950327.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988) An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Bell, W. (1997) Models for county and state poverty estimates. Preprint, Statistical Research Division, U.S. Census Bureau.

Bogue, D. J. (1950) A technique for making extensive postcensal estimates. *Journal of the American Statistical Association*, **45**, 149–163.

Casas-Cordero Valencia, C., Encina, J., and Lahiri, P. (2016) Poverty mapping for the Chilean comunas. In: *Analysis of Poverty Data by Small Area Estimation*, (ed. M. Pratesi), Wiley, New York.

Chambers, R., and Clark, R. (2012) *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford

Chaudhuri, A. (2012) *Developing Small Domain Statistics: Modelling in Survey Sampling*. LAP LAMBERT Academic Publishing GMbH & Co. KG, Saarbrücken.

Clayton, D., and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd ed. Wiley, New York.

Corral, P., Himelein, K., McGee, K., and Molina, I. (2021) A map of the poor or a poor map? *Mathematics*, **9**(21), 2780; https://doi.org/10.3390/math9212780

Corral, P., Molina, I., and Nguyen, M. (2021) Pull your small area estimates up by the bootstraps. *Journal of Statistical Computation and Simulation*, https://doi.org/10.1080/00949655.2021.1926460

Corral, P., Molina, I., Cojonaru, A., and Segovia, S. (2022) Guidelines to small area estimation for poverty mapping. The World Bank.

Cressie, N. (1989) Empirical Bayes estimation of undercount in the decennial census, *Journal of the American Statistical Association*, **84**, 1033–1044.

Datta, G. S. (2009) Model-based approach to small area estimation. In: *Sample Surveys: Inference and Analysis*, (eds. D. Pfeffermann and C. R. Rao). *Handbook of Statistics*, Volume 29B, North-Holland, Amsterdam, 251–288.

Deville, J. C., and Särndal, C. E. (1992) Calibration estimation in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Dick, P. (1995) Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, **21**, 45–54. Drew, D., Singh, M. P., and Choudhry, G. H. (1982) Evaluation of small area estimation techniques for the Canadian Labour Force Survey. *Survey Methodology*, **8**, 17–47.

Efron, B. (1975) Biased versus unbiased estimation, *Advances in Mathematics*, **16**, 259–277.

Efron, B., and Morris, C. E. (1972) Limiting the risk of Bayes and empirical Bayes estimators, Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, **67**, 130–139.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003) Micro-level estimation of poverty and inequality. *Econometrica*, **71**, 355–364.

Ericksen, E. P., and Kadane, J. B. (1985) Estimating the population in census year: 1980 and beyond (with discussion). *Journal of the American Statistical Association*, **80**, 98–131.

Fay, R. E., and Herriot, R. A. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **85**, 398–409.

Fuller, W. A. (1999) Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, **4**, 331–345.

Fuller, W. A. (2009) *Sampling Statistics*. Wiley, New York.

Ghosh, M. (2020) Small area estimation: its evolution in five decades (with discussion). *Statistics in Transition*, **21**(4), 40–44.

Ghosh, M., and Meeden, G. (1997) *Bayesian Methods for Finite Population Sampling*. Springer, New York.

Ghosh, M., and Rao, J. N. K. (1994) Small area estimation: an appraisal (with discussion). *Statistical Science*, **9**, 55–93.

Gonzalez, M. E. (1973) Use and evaluation of synthetic estimates. In: *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., and Santamaría, L. (2008) Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, **78**(5), 443–462.

Griffiths, R. (1996) Current population survey small area estimations for congressional districts. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 314–319.

Guadarrama, M., Molina, I., and Rao, J. N. K. (2014) A comparison of small area estimation methods for poverty mapping. *Statistics in Transition*, **1**(17), 41–66.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953) *Sample Survey Methods and Theory I*, Wiley, New York.

Henderson, C. R. (1950) Estimation of genetic parameters (Abstract). *Annals of Mathematical Statistics*, **21**, 309–310.

James, W., and Stein, C. (1961) Estimation with quadratic loss. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 361–379.

Jedrzejczak, A., and Kubacki, J. (2013) Estimation of income inequality and the poverty rate in Poland, by region and family Type. *Statistics in Transition-New Series*, **14**(3), 359–378.

Jiang, J. (2007) *Linear and generalized linear mixed models and their applications*. Springer-Verlag, New York.

Jiang, J. (2017) *Asymptotic Analysis of Mixed Effect Models: Theory, Applications and Open Problems*. CRC Press, Boca Raton, FL.

Jiang, J., and Lahiri, P. (2006) Mixed model prediction and small area estimation. *Test*, **15**, 1–96.

Krennmair, P., and Schmid, T. (2022) Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society, Series C*, **71**(5), 1865–1894.

Lehtonen, R., Särndal, C.-E., and Veijanen, A. (2003) The effect of model choice in estimation for domains including small domains. *Survey Methodology*, **29**, 33–44.

Lehtonen, R., and Veijanen, A. (2009) Design-based methods of estimation for domains and small areas. In: *Sample Surveys: Inference and Analysis*, (eds. D. Pfeffermann and C. R. Rao). *Handbook of Statistics*, Volume 29B, North-Holland, Amsterdam, 219–249.

Longford, N. T. (2005) *Missing Data and Small-Area Estimation*. Springer, New York.

Lohr, S. L. (2010) *Sampling: Design and Analysis*. Duxbury, Pacific Grove, CA.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., and Gabrielli, L. (2015) Small area model-based estimators using big data sources. *Journal of Official Statistics*, **31**, 263–281.

Mauro, F., Molina, I., García-Abril, A., Valbuena, R., and Ayuga-Télleza, E. (2015) Remote sensing estimates and measures of uncertainty for forest variables at different aggregation levels. *Environmetrics*, **27**, 225–238.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. Chapman & Hall, Cambridge.

Militino, A. F., Ugarte, M. D., Goicoa, T., and González-Aud'icana, M. (2006) Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 450–461.

Molina, I. (2019) Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas. Series of the Economic Commission for Latin America and the Caribbean (ECLAC) from United Nations, Estudios Estadísticos LC/TS.2018/ 82/Rev.1, CEPAL.

Molina, I. (2020) Discussion on "Small area estimation: its evolution in five decades", by M. Ghosh. *Statistics in Transition*, **21**(4), 40–44.

Molina, I., Corral, P., and Nguyen, M. (2022) Poverty mapping methods: a review. Test, DOI: 10.1007/s11749-022-00822-1

Molina, I., and Marhuenda, Y. (2015) sae: An R package for small area estimation. *The R Journal*, **7**, 81–98.

Molina, I., and Morales, D. (2009) Small area estimation of poverty indicators. *Boletín de Estadística e Investigación Operativa*, **25**, 218–225.

Molina, I., and Rao, J. N. K. (2010) Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **38**, 369–385.

Molina, I., Rao, J. N. K., and Guadarrama, M. (2019) Small area estimation methods for poverty mapping: a selective review. *Statistics and Applications*, **17**, 11–22.

Molina, I., and Strzalkowska-Kominiak, E. (2020) Estimation of proportions in small areas: application to the labour force using the Swiss Census Structural Survey. *Journal of the Royal Statistical Society, Series A*, **183**, 281—310.

Molina Peralta, I., and García Portugués, E. (2020) Short guide for small-area estimation using household survey data: illustration to poverty mapping in Palestine with expenditure survey and census data. UN Economic and Social Commission for Western Asia (ESCWA), E/ESCWA/SD/2019/TP.4

Morales, D., Esteban, M. D., Perez, A., and Hobza, T. (2021) *A Course on Small Area Estimation and Mixed Models: Methods, Theory and Applications in R*. Springer, Cham, Switzerland.

Mukhopadhyay, P. (1998) *Small Area Estimation in Survey Sampling*. Narosa Publishing House, New Delhi.

National Center for Health Statistics (1968), *Synthetic State Estimates of disability*, P.H.S. Publications 1759, Government Printing Office, Washington DC, U.S.

National Research Council (1980) *Panel on Small-Area Estimates of Population and Income. Estimating Income and Population of Small Areas*. National Academy Press, Washington DC, U.S.

Nelder, J. A., and Wedderburn, R. W. M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Nguyen, M. C., Corral, P., Azevedo, J. P., and Zhao, Q. (2018) Sae: A stata package for unit level small area estimation. World Bank Policy Research Working Paper 8630.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008) Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, **70**, 265–286.

Pfeffermann, D. (2002) Small area estimation-new developments and directions. *International Statistical Review*, **70**(1), 125–143.

Pfeffermann, D. (2013) New important developments in small area estimation. *Statistical Science*, **28**(1), 40–68.

Pratesi, M. (2016) *Analysis of Poverty Data by Small Area Estimation*, Wiley, New York.

Pratesi, M., and Salvati, N. (2016). Introduction on measuring poverty at local level using small area estimation methods. In: *Analysis of Poverty Data by Small Area Estimation* (ed. M. Pratesi), Wiley, New York, 1–18.

Purcell, N. J., and Kish, L. (1979) Estimates for small domain. *Biometrics*, **35**, 365–384.

Rao, J. N. K. (2003) *Small Area Estimation*, 1st. Ed. Wiley, Hoboken, NJ.

Rao, J. N. K., and Molina, I. (2015) *Small Area Estimation*, 2nd. Ed. Wiley, Hoboken, NJ.

Rao, J. N. K., and Molina, I. (2016) Empirical Bayes and hierarchical Bayes estimation of poverty measures for small areas. In: *Analysis of Poverty Data by Small Area Estimation* (ed. M. Pratesi), Wiley, New York, 315–324.

Rives, N. W., Serow, W. J., Lee, A. S., and Goldsmith, H. F. (Eds.) (1989) *Small Area Analysis: Estimating Total Population*, National Institute of Mental Health, Rockville, MD.

Royall, R. M. (1970) On finite population sampling theory under certain linear regression. *Biometrika*, **57**, 377–387.

Royall, R. M. (1976) The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, **71**, 657–664.

Särdnal, C.-E., Swensson, B., and Wretman, J. H. (1989) The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, **76**, 527–537.

Scott, A., and Smith, T. M. F. (1969) Estimation in multi-stage surveys. *Journal of the American Statistical Association*, **64**, 830–840.

Searle, S. R. (1971) *Linear Models*. Wiley, New York.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992) *Variance Components*. Wiley, New York.

Schaible, W. A. (1978) Choosing weights for composite estimators for small area statistics. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 741–746.

Statistics Canada (1987) *Population Estimation Methods in Canada*, Catalogue 91–528E, Statistics Canada, Ottawa.

Sugasawa, S., and Kubokawa, T. (2023) *Mixed-Effects Models and Small Area Estimation*. Springer, Singapore.

Thompson, M. E. (1997) *Theory of Sample Surveys*. Chapman & Hall, London.

Valliant, R. L. (1987) Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, **82**(398), 499–508.

Valliant, R., Dorfman, A. H., and Royall, R. M. (2001) *Finite Population Sampling and Inference: a Prediction Approach*. Wiley, New York.

Wu, C. and Thompson, M. E. (2020) *Sampling Theory and Practice*. Springer Nature, Switzerland.

Zhao, Q. (2006) User manual for povmap. World Bank.
http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf.

Zidek, J. V. (1982) *A Review of Methods for Estimating Population of Local Areas*. Technical Report 82–4, University of British Columbia, Vancouver, Canada.

The Survey Statistician, 2023, Vol. 88, 36–37.

Is it Time for Young Survey Statisticians to Shine in the Society?

Mahmoud Torabi[1]

[1]University of Manitoba, Canada, Mahmoud.Torabi@umanitoba.ca

Abstract

In survey sampling, policy decisions regarding allocation of resources to subgroups in a population, called small areas, are based on reliable predictors of their underlying parameters. However, in some subgroups, due to small sample sizes relative to the population, the information for reliable estimation is typically not available. Consequently, we need to predict the characteristics of small areas based on the coarser scale data. Mixed models (including cross-sectional, spatial data, and so on) are the primary tools in small area estimation (SAE) and also borrow information from alternative sources (e.g., previous surveys, administrative, and census). In this short paper, I will review my scientific background in this subject with also providing some comments and suggestions for young researchers.

Keywords: small area estimation, spatial statistics

1 Scientific background. Fist of all, this is my honor to write this short paper regarding my Hukum Chandra prize. I was born and raised in Tehran which is the capital city of Iran. In my time, there were three main streams in high school to choose. I chose Mathematics-Physics while other options were Experimental Sciences and Human Sciences. After high school graduation, I participated in the national entrance exam (AKA Konkoor) for a university program. We had 100 options to choose a program and a university after writing the national entrance exam. I was accepted to Statistics program at the National University of Iran (AKA Shahid Beheshti University). Although I chose Statistics, however, I had limited information regarding the program; I should say that I was not accepted to other popular programs in those days such as Engineering (electronic, communication, civil, mechanic). I successfully graduated with BSc and MSc from the National University of Iran before pursing my PhD in Statistics at Carleton University in Canada under supervision of Dr. Jon Rao. In my PhD program, I worked on some interesting problems in small area estimation which resulted in 6 publications in statistics journals. After PhD graduation, I accepted a post-doctoral fellowship (PDF) from University of Alberta, Canada, to investigate the impact of various health research topics in the province of Alberta, Canada. I then joined the Department of Community Health Sciences at the University of Manitoba, Canada, in 2010 as an Assistant Professor of Biostatistics. I am currently Professor of Biostatistics while I hold this position since 2020.

At the University of Manitoba, I established an excellent research team including my collaborators (methodologists and health clinicians) and also my high quality personnel (MSc, PhD, and post-doctoral). It is evident that my transition from applied statistics to Biostatistics started when I accepted a PDF position at the University of Alberta. As a biostatistician, we have developed various bio/statistics models/methods to answer various questions from the public. The team I lead is developing new models and techniques in the context of population health research to effectively integrate the knowledge we generate into health care practice which is fundamental to the health and well-being of the population.

2 Research background. I have developed an original and innovative research program in small area estimation and spatial statistics. In small area estimation (SAE), policy decisions regarding the allocation of resources to subgroups of a population depend on reliable predictors of their underlying parameters. However, in some subgroups, called small areas due to small sample sizes relative to the population, the information needed for reliable prediction is typically not available. Consequently, survey (or administrative) data on a coarser scale is used to predict the characteristics of small areas. Mixed models, which are the primary tools in SAE, are used to borrow information from alternative sources (including survey, administrative, and census) to provide reliable prediction. Such predictions have many applications, e.g. in disease mapping the main objective is to find reliable rates of disease such as cancer in small areas. It also has other applications in agriculture, economics, policymaking, and allocation of funds. The team members I lead have developed novel statistical methods in the context of SAE and applied our innovative approaches to population-health data such as asthma and cancer. In spatial statistics, my program of research is on the development of new and original biostatistics methods for big data over space and time. In population and public health, the identification and measurement of patterns of disease are important goals. These patterns facilitate the understanding of disease and better understanding may lead to the formulation of etiological hypotheses. We may be able to explore the causes of different diseases by identifying the characteristics that increase disease risk (e.g., pollution) and improve disease control. My team members under my direction have developed novel biostatistics methods to better understand big and complex spatial and temporal data. Our innovations have allowed us to better predict spatial and temporal trends of disease, identify corresponding risk factors, and plan for interventions/preventions.

3 Interaction with late Hukum. As explained above, my main research focus has been in SAE and spatial statistics. In particular, my research areas were aligned with late Hukum who was unfortunately died during the covid. I met Hukum in different occasions and in particular in SAE conferences. His personality was unique; he was very kind and a humble person. We discussed few projects from time to time for possible collaborations, but we got busy and could not pursue those ideas. He was a good researcher and made valuable contributions in the context of SAE and spatial statistics. He was attentive in scientific sessions with smile and also open for research discussion. Truly, our SAE community missed him as his character was unique. May his soul rest in peace.

4 Conclusion. As a researcher who has been in academia for more than 10 years, I can attest that the SAE community is growing rapidly as the subject is applicable to many professional organizations and sectors. Many young, energized, and strong researchers are currently working in this important subject, and I can anticipate even more researchers will be involved in this subject area. As it is also evident from the SAE community, senior researchers are mentoring junior researchers, and the future of community is very bright in this direction. Shortly, we will see a transition that young researchers take a full responsibility of the community with support of senior researchers.

The Survey Statistician, 2023, Vol. 88, 38-39.

Cochran-Hansen Prize – Memories from the Beginning

Maiki Ilves[1], Kristiina Rajaleid[2], and Imbi Traat[3]

[1]Statistics Estonia, Maiki.Ilves@stat.ee
[2]Stockholm University, Kristiina.Rajaleid@su.se
[3]University of Tartu, imbi.traat@ut.ee

The Cochran-Hansen Prize of the IASS is awarded every two years for the best paper on survey research methods submitted by a young statistician from a developing or transition country. The history of the prize goes back to 1999, and 14 persons have received the prize so far (http://isi-iass.org/home/cochran-hansen-prize).

Estonia became re-independent in 1991. Before that, during Soviet power, no survey statistics nor official statistics, neither in theory nor in practice, was dealt with. Gradually, this area started to develop. **Imbi Traat** in Tartu University started courses in survey sampling in 1993. Baltic-Nordic network in survey statistics (https://wiki.helsinki.fi/display/BNU/Home), initiated by **Gunnar Kulldorff** from Umeå University, helped a lot in a rapid development of the new area. Students had a great interest in a newly launched subject that had practical applications, requiring specialised statisticians. They wrote good Bachelor and Master theses, and as soon as the information about Cochran-Hansen prize arrived to Tartu University, we were ready to apply for it. Estonia was classified as transition country that time.

In fact, the information about the prize came to us from the Tallinn University, where **Enel Pungas** was the one of the two first receivers of the Cochran-Hansen prize in 1999. She was a master student in demography that time. She submitted her study on the data collection aspects and effects in the Estonian Family and Fertility Survey 1997. Enel received the award – participation in the IASS Summer Courses in Jyväskylä, and possibility to buy scientific literature. Now she works in the Ministry of the Interior as head of the Population Facts Department. She is thankful for those possibilities and recognition in her young days.

The two students of Imbi Traat who received the prize tell their memories below.

Kristiina: I was awarded the Cochran-Hansen prize in 2001 for the paper “On the order sampling design” which was based on my Bachelor thesis. I presented the study in the meeting of the International Statistical Institute in Seoul, South Korea. In conjunction with the meeting, I had the possibility to attend two courses on topics in survey sampling which I appreciated a lot. Also, this was my first journey to such a faraway country with a culture so different from my own. I still remember a breath-taking concert with Korean dance and music, the food (first time to eat with chopsticks!), the city (an amazing mix of modernity and tradition). And my first ever jetlag… After 2001, I continued my studies in mathematical statistics at MSc level. In 2010 I earned a PhD in Medicine (public health science / epidemiology) from Karolinska Institutet in Stockholm, Sweden.

Currently I am working as researcher at Stockholm University, among other projects I am involved in a large national longitudinal survey on work environment and well-being.

Maiki: I received Cochran-Hansen prize in 2005 for my research paper “Variance and its estimator for a practical self-weighting two-phase design”. The research for this paper I carried out at Statistics Estonia as a part of my bachelor studies, based on a real-life problem in the Estonian Labour Force survey. My supervisor for this work was Imbi Traat, at the time an associate professor at Tartu University. She was the one who brought my attention to the IASS competition for young survey statisticians. I am very grateful for her believing in me and encouraging me to apply.

My prize included a plane ticket to Sydney to attend the 55th WSC (then called Session) of the ISI in April 2005. In addition, I was given the possibility to attend two short courses given adjacent to the congress, and was as well awarded a check to buy books of my choice. As a result, I became the owner of a copy of the famous “yellow book” by Särndal, Swensson and Wretman.

Even so many years later, still working in Statistics Estonia, now working more with people as Head of the Development Department, I remember very well that journey, the 33 hours one-way trips and the event itself. It was my first plane trip, my first visit to Australia, my first ISI experience, and my first presentation to so large an audience. What I remember of Sydney was the permanently cloudless sky, the very friendly local people, and the beautiful nature. For a young person without international conference experience, this ISI congress was overwhelming: so many people and so many sessions to choose from, and not to mention presenting my contribution in front of a large, highly knowledgeable audience. I was very grateful to the attendees who had supportive and encouraging comments on my work and presentation. I am also glad that I made some memorable contacts at the congress and stayed in contact with some of them even after the event. All in all, it was a very inspiring experience, and I am very grateful to IASS for this opportunity.

The Survey Statistician, 2023, Vol. 88, 40-45.

The History and Impact of the *Survey Methodology* Journal

Jean-François Beaumont[1]

[1] Statistics Canada, jean-francois.beaumont@statcan.gc.ca

Abstract

Survey Methodology is a peer-reviewed statistical journal that was founded in 1975 at Statistics Canada to provide a venue for discussing practical issues arising from the implementation of sample surveys. After a brief introduction, I will describe the historical context in which the journal was established, its evolution over the years and then its impact on survey practice in Canada and around the world. I will conclude by announcing some special discussion papers and special issues that are currently being planned for publication in forthcoming issues.

Keywords: History, *Survey Methodology*.

1 Introduction

Survey Methodology is a biannual peer-reviewed statistical journal founded in 1975 at Statistics Canada. As currently stated on its website (www.statcan.gc.ca/surveymethodology), the journal aims to publish innovative theoretical or applied research papers, and sometimes review papers, that provide new insights on statistical methods relevant to National Statistical Offices and other statistical organizations.

Survey Methodology was initially established to provide a venue for discussing practical issues arising from the implementation of sample surveys. Indeed, the editorial policy of the very first issue of the Journal states:

"*The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and inter-relationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys.*"

The scope of the journal has significantly expanded over the years. It now covers a wide range of topics of interest to survey methodologists and statisticians around the world, including more modern topics such as the use of multiple data sources, statistical data integration, as well as research, development and application of machine learning methods for the production of official statistics. A current list of topics of interest is provided on the *Survey Methodology* website. The journal would not have flourished without the contribution of dedicated Associate Editors who provide invaluable

recommendations for determining the suitability of papers submitted to *Survey Methodology*. Like the scope, the Editorial Board has significantly expanded over time, starting with three internal members in 1975 and now including over 30 internationally renowned Associate Editors whose fields of expertise cover the diversity of topics included in the journal's scope.

The first issue of *Survey Methodology* appeared in June 1975. It has been published every year, in June and December, ever since. From June 1975 to December 2022 inclusively, a total of 909 papers were published, including 8 discussion papers, 6 special issues and 23 special sections of an issue. In recent years, between 50 to 70 papers are submitted annually, mostly from authors outside Statistics Canada, with an acceptance rate around 30%. Authors are welcome to submit their paper in either French or English. Starting with the December 1981 issue, all accepted papers have been translated and published in both languages.

2 Historical context

Research and development in survey methods was booming at Statistics Canada at the end of the 1960s and early 1970s (Platek, 1999; Platek, 2009) and also throughout the world (Kalton, 2000). Some of the research efforts undertaken by Canadian methodologists at the time led to publications in well-known mainstream peer-reviewed statistical journals (e.g., Beynon, Ostry and Platek, 1970; Fellegi, 1972; Fellegi, 1973; Fellegi, 1974; Fellegi and Holt, 1976; Fellegi and Sunter, 1969; Gray and Platek, 1968; and Ostry and Sunter, 1970) such as the *Journal of the American Statistical Association*. However, there was no international journal dedicated to methodological issues arising when conducting sample surveys (Kalton, 2000), except for the brief existence of *Sankhya* C from 1974 to 1978, which covered survey sampling theory and methods (Rao, 2023). There was a growing need for such a venue, which would allow survey methodologists, not only those from Statistics Canada but also in other statistical organizations, to disseminate their theoretical and empirical research findings in this field. The time was ripe for the launch of a new journal; *Survey Methodology* was thus established in 1975 by Richard Platek. Its first editor was Mangala Prasad Singh, known as M.P. Singh, who remained in this position for 30 years until his death in 2005. The June 2006 issue of *Survey Methodology* contains a special article, with testimonials by a few colleagues and friends, to honour the memory of M.P. Singh and recognize his numerous accomplishments during his career at Statistics Canada, in particular those related to *Survey Methodology*.

The 1970s also saw the emergence of several professional associations for survey methodologists (Kalton, 2000), such as the International Association for Survey Statisticians (IASS), founded in 1973 and celebrating its 50th anniversary this year, and the Survey Research Methods Section of the American Statistical Association (ASA), established in 1978. Subsequently, other journals focussed on sample surveys were launched for the greatest benefit of the community of survey methodologists and statisticians around the world (among others, *The Survey Statistician* in 1978, the newsletter of the IASS, the *Journal of Official Statistics* in 1985, published by Statistics Sweden, and the *Journal of Survey Statistics and Methodology* in 2013, sponsored by the ASA and the American Association for Public Opinion Research). Rancourt (2023) provides a portrait of the history of *The Survey Statistician* and points out the close connections between the IASS and *Survey Methodology* at the end of the 1970s; *Survey Methodology* was distributed at a preferential rate to IASS members and used as the prime vehicle for the publication of papers presented at the International Statistical Institute conference.

3 Milestones through the years

Papers published in the first issues of *Survey Methodology* were mainly written by authors from Statistics Canada, but the journal flourished rapidly under the leadership of M.P. Singh. Within 15 years or so, authors from all over the world, including famous statisticians such as Wayne Fuller (e.g., Fuller, 1990), Graham Kalton (e.g., Kalton, 1986), Leslie Kish (e.g., Kish, 1988), Danny Pfeffermann (e.g., Pfeffermann and Burck, 1990), J.N.K. Rao (e.g., Rao, 1985; and Rao, Wu and Yue, 1992), Don Rubin (e.g., Rubin, 1986) and Carl-Erik Särndal (e.g., Särndal, 1992), were

submitting their papers for consideration in *Survey Methodology*. This allowed the journal to acquire an international stature and become a key source of information for survey methodologists at Statistics Canada and around the world.

M.P. Singh implemented many initiatives to raise the profile of the journal and make it more interesting to readers. For instance, he would frequently arrange for the publication of discussion papers or special issues/sections on important topics such as the special section of the June 2001 issue on composite estimation for the Canadian Labour Force Survey. He also initiated the short notes section, which allowed authors to submit shorter papers without the full development of a regular paper and with a streamlined review process. The first short notes section following his initiative was published in the June 2005 issue; it contained three short articles. To be more accurate, a short communications section had already appeared once, in the December 1987 issue, but this idea was never repeated until it was reinstated permanently in 2005.

A major initiative taken by M.P. Singh, in collaboration with the American Statistical Association and Westat, was the introduction of the Waksberg Award in 2001 in honour of Joseph Waksberg, who made outstanding contributions to survey statistics and methodology. Since 2001, this prestigious Award is given annually to a prominent survey statistician chosen by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. The recipient of the Award writes a review paper for *Survey Methodology* and usually presents it at Statistics Canada's Symposium.

The journal continued to thrive under subsequent editors, namely, John Kovar (2006-2009), Mike Hidiroglou (2010-2015), Wesley Yung (2016-2020) and myself since 2021, with the publication of other discussion papers and special issues/sections, among others, the special section of the December 2011 issue on alternative survey sampling designs organized in collaboration with the U.S. Census Bureau. More recently, a special discussion paper on statistical inference with non-probability survey samples (Wu, 2022), a topic that has increasingly been drawing attention of survey statisticians in the past 20 years, was published in the December 2022 issue. The paper was accompanied with five discussions by international experts in the field. It is also worth mentioning the joint special issue with the *International Statistical Review*, published in May 2019, in honour of Prof. J.N.K. Rao's contributions. Incidentally, Prof. Rao is by far the most prolific author for *Survey Methodology*, as he has written or co-written an impressive number of 31 papers during the period covering the first issue in June 1975 to the December 2022 issue. This includes a recent invited review paper on the major developments in sample survey theory and methods covering the past 100 years (Rao and Fuller, 2017), which was discussed by four eminent survey statisticians.

In 2006, the journal became available online and completely free of charge. The printed version continued to be produced and made available via a paid subscription until December 2012. Historical issues have then been gradually added to the free online catalogue. In 2019, *Survey Methodology* adopted the Scholar One system for a more efficient management of papers submitted to the journal.

4 Impact

The journal has had a significant impact on Statistics Canada's programs over the years. For instance, the stratification method of Lavallée and Hidiroglou (1988) is still implemented in many business surveys. It allows for efficient stratification that improves the quality of survey estimates for skewed variables typically encountered in economic surveys. Another example is the method of Rao, Wu and Yue (1992) for computing bootstrap weights. It is used in a large number of social surveys and allows for estimating the variability of survey estimates for stratified multistage sampling designs. A third example is the method of Särndal (1992) for estimating the precision of survey estimates in the presence of imputation. It is the methodological foundation of the System for Estimation of Variance due to Nonresponse and Imputation (SEVANI), which was developed between 2005 and 2010 (Beaumont and Bissonnette, 2011). There are many other examples where papers published in *Survey Methodology* had a direct influence on methods implemented in statistical programs of Statistics Canada and certainly other statistical organizations.

The journal has also been an important learning source for methodologists at Statistics Canada. A number of papers have long been known to be key readings for young methodologists, either learning on the job or preparing for competitive processes. The papers by Brackstone (1987) on the use of administrative data and the discussion paper by Singh, Gambino and Mantel (1994) on small area estimation have probably been the most circulated among them, especially during the 10 to 15 years following their publication.

The international impact of *Survey Methodology* is perhaps more difficult to assess without further investigation. Platek (1999) states: "In a number of countries, the journal, almost from the beginning, provided a base for teaching and training new statisticians.". It is also my perception, strengthened by a few conversations with survey methodologists or statisticians from different countries, working in National Statistical Offices, universities or other statistical organizations, that *Survey Methodology* has been known to be an essential tool for showcasing and sharing innovative ideas and experiments related to sample surveys. This is confirmed by the yearly number of views (about 50,000 per year) and downloads (about 20,000 per year) of *Survey Methodology* papers (excluding views and downloads from Statistics Canada's network) and by noting that authors come from different countries in the world. Another useful indicator is the number of citations for papers that had a significant impact in the survey practice. For example, according to Google Scholar, Kalton (1986) and Rao, Wu and Yue (1992), two of the most influential papers published in *Survey Methodology*, had both been cited 699 times as of May 1, 2023.

5 Conclusion

Survey Methodology is recognized as a high-quality journal in the international community of survey statisticians. This is not taken for granted and efforts are continuous to keep the journal relevant, attractive and increase its readership. For instance, the review process has recently been revised and streamlined to remain competitive and attractive for authors considering *Survey Methodology* to showcase their research findings. I am sincerely grateful for the cooperation and commitment of all the Editorial Board members and referees who have made tremendous efforts to keep the review process as efficient as possible.

Over the next few years, we plan to increase the frequency of special discussion papers and special issues, as well as continue the publication of the yearly Waksberg Award paper. For instance, in the June 2023 issue, a special paper by Natalie Shlomo on statistical disclosure control and privacy will be published to honour the memory of Chris Skinner, a giant in survey statistics. Chris was the winner of the 2019 Waksberg Award, but could not write his paper and present it before he passed away in 2020. Shlomo's paper will be accompanied with testimonials from Danny Pfeffermann, J.N.K. Rao and Jae-Kwang Kim. The paper and testimonials were presented at Statistics Canada's 2021 International Methodology Symposium.

Another special paper, by Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé, is being planned for the December 2023 issue to honour the memory of another giant in survey statistics, Jean-Claude Deville, who passed away in 2021. The paper will review the most important of his contributions to the field, which include among others, calibration and cube sampling. It will be followed by discussions/testimonials from colleagues and friends. The December 2023 issue will also feature a special section with a few selected papers presented at the 2021 Colloque francophone sur les sondages. The Guest Editor for this special section is Alina Matei.

In 2024, a special issue is planned for three papers that were presented at the 2022 Morris Hansen Lecture event on the use of non-probability samples. All three papers will be discussed by international experts in the field. An introduction by Partha Lahiri, the Guest Editor for this special issue, will precede the papers. A special discussion paper by Carl-Erik Särndal, entitled "Progress in survey science: yesterday – today – tomorrow", is also currently being planned for publication in a future issue in 2024 or 2025 along with discussions from eminent survey statisticians. Finally, the June 2025 issue will be dedicated to celebrate the 50th anniversary of *Survey Methodology*. Stay tuned!

At last, I would like to express my sincere gratitude for the kind invitation to write this article from Danuté Krapavickaité and Eric Rancourt, the Editors of *The Survey Statistician*. Let me conclude by taking this opportunity to thank all the readers as well as all the authors who considered *Survey Methodology* for the publication of their research papers. It goes without saying that the journal would not have been the same without their contributions and continuous interest in its content.

I look forward to reading your future submissions to *Survey Methodology*!

Acknowledgements

I would like to thank an anonymous referee, Edward Chen, Jack Gambino, Eric Rancourt and Wesley Yung for their relevant and useful comments that improve the readability and accuracy of this article.

References

Beaumont, J. F., and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology*, 37, 171-179.

Beynon, T. G., Ostry, S., and Platek, R. (1970). Some Methodological Aspects of the 1971 Census in Canada. *The Canadian Journal of Economics/Revue canadienne d'Economique*, 3, 95-110.

Brackstone, G. J. (1987). Issues in the Use of Administrative Records for Statistical Purposes. *Survey Methodology*, 13, 29-43.

Fellegi, I. P. (1972). On the Question of Statistical Confidentiality. *Journal of the American Statistical Association*, 67, 7-18.

Fellegi, I. P. (1973). The Evaluation of the Accuracy of Survey Results: Some Canadian Experiences. *International Statistical Review/Revue Internationale de Statistique*, 41, 1-14.

Fellegi, I. P. (1974). An Improved Method of Estimating the Correlated Response Variance. *Journal of the American Statistical Association*, 69, 496-501.

Fellegi, I. P., and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical association*, 71, 17-35.

Fellegi, I. P., and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Fuller, W. A. (1990). Analysis of Repeated Surveys. *Survey Methodology*, 16, 167-180.

Gray, G. B., and Platek, R. (1968). Several Methods of Redesigning Area Samples Utilizing Probabilities Proportional to Size When the Sizes Change Significantly. *Journal of the American Statistical Association*, 63, 1280-1297.

Kalton, G. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1-16.

Kalton, G. (2000). Developments in survey research in the past 25 years. *Survey Methodology*, 26, 3-10.

Kish, L. (1988). Multipurpose Sample Designs. *Survey Methodology*, 14, 19-32.

Lavallée, P., and Hidiroglou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33-43.

Ostry, S., and Sunter, A. (1970). Definitional and Design Aspects of the Canadian Job Vacancy Survey. *Journal of the American Statistical Association*, 65, 1059-1070.

Platek, R. (1999). Survey Methodology: The first 25 years. *Survey Methodology*, 25, 109-111.

Platek, R. (2009). The evolution of survey methodology in Statistics Canada up to 1986. Methodology Branch Working Paper METH-2009-004E, Statistics Canada.

Pfeffermann, D., and Burck, L. (1990). Robust Small Area Estimation Combining Time Series and Cross-Sectional Data. *Survey Methodology*, 16, 217-237.

Rancourt, E. (2023). 50 years of keeping survey statisticians of the world informed: *The Survey Statistician*". *The Survey Statistician*, 88, 46-50.

Rao, J. N. K. (1985). Conditional Inference in Survey Sampling. *Survey Methodology*, 11, 15-31.

Rao, J. N. K., and Fuller, W. A. (2017). Sample survey theory and methods: Past, present, and future directions. *Survey Methodology*, 43, 145-160 (with discussion).

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209–217.

Rao, J.N.K. (2023). Some memorable recollections of IASS first meeting. *The Survey Statistician*, 88, 14-15.

Rubin, D.B. (1986). Basic ideas of Multiple Imputation for Nonresponse. *Survey Methodology*, 12, 37-47.

Särndal, C. E. (1992). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. *Survey Methodology*, 18, 241-252.

Singh, M. P., Gambino, J., and Mantel, H. J. (1994). Issues and Strategies for Small Area Data. *Survey Methodology*, 20, 3-22 (with discussion).

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 283-311 (with discussion).

The Survey Statistician, 2023, Vol. 88, 46-50.

50 Years of Keeping Survey Statisticians of the World Informed: *The Survey Statistician*

Eric Rancourt[1]

[1] Statistics Canada, eric.rancourt@statcan.gc.ca

Abstract

Since its inception, the International Association of Survey Statistician (IASS) has regularly produced a biannual newsletter called *The Survey Statistician (TSS)* to inform its members. *TSS* is now conceptualized as the IASS sponsored journal. This article provides a summary of the history of *TSS*, from its format and content to the list of editors and important sections such as Ask the Experts, Country Reports, New and Emerging Methods and Book/Software Reviews. The paper provides some counts and facts related to the development of *TSS* without pretending to be an exhaustive account of all that surrounds its history and extensive devotion of its numerous contributors.

Keywords: Ask the Experts, Book/Software Reviews, Country reports, Editors, New and Emerging Methods.

1 Introduction

The Survey Statistician has been one of the primary tools to maintain IASS members' connections and inform them about the varied activities of the association. Since its modest debut in 1974, *TSS* has constantly evolved, enriching its content with news, articles on new and emerging methods, book and software reviews, survey activities in countries and questions from members being answered by experts. Through the years, in addition to format and style changes, *TSS* changed content to address the preference and demands of members thanks to dynamic editors.

The role of *TSS* was modified since the first version which was more of an account of the meetings that took place. Nowadays, it is a solid newsletter that provides information on recent survey statistics developments, activities with news on some members as well as forthcoming conferences, events and activities of the IASS. Linking with other developments, it also includes tables of content of leading journals relevant to survey statistics. Since the first issue of *TSS*, 96 issues (including the current one - #88) of *TSS* have been produced during its 50-year existence. This corresponds to producing two issues per year, except for one occasion in each of the 70s, 80s, 90s and 2000s.

2 The Early Days

When the IASS started, its newsletter was simply called Newsletter and was produced eight times from 1974 to 1978. Information on this can be found in the 25 years history book of the IASS (IASS, 1999). These provided a summary of the invited papers at the IASS meetings as well as a few other information items. In the first issue, IASS President Morris H. Hansen, made a strong call to members

to provide information for the bulletin and be active in the association. The issue also provided the list of executives and the list of sessions of the first meeting in Vienna. Some details can be found in Rao (2023). Issue #1 of the Newsletter contained an announcement about the new *Sankhya C* which would publish articles on survey sampling. Finally, it included the 1975 program for the Warsaw ISI meeting and a questionnaire about IASS members.

As the Newsletter improved, it was decided that it would be revamped into a more comprehensive and informative product. For this purpose, issue #8 contained a first version (called issue #0) of what became *The Survey Statistician*, with issue #1 in 1979. From that point on, the sections of the newsletter became almost permanently set to include News, information on the Annual General Meeting, Country information, Question/Answers, paper summaries and some articles.

3 Editors and Structure

3.1 Editors

Many people contributed to the success of *TSS*. We present below a table of all the editors to date. Other members have contributed as specific Section Editors. Some will be presented later, but it is not possible to be exhaustive in a short article such as this one.

Table 1. Editors of *The Survey Statistician*

1979-1985	Jacques Desabie
1985-1990	Gildas Roy
1990-1992	Anne-Marie Vespa-Leyder
1992-1995	Denise Lievesley
1995-1999	Mike Brick
1999-2000	Mike Brick and Leyla Mohadjer
2000-2003	Leyla Mohadjer and Jairo Arrow
2003-2008	Steve Heeringa
2008-2010	Dan Hedlin
2010-2014	Natalie Shlomo and Frank Yu
2015-2018	Natalie Shlomo and Eric Rancourt
2018-2023	Danutė Krapavickaitė and Eric Rancourt

Today, the *TSS* editorial board consists of two editors, three section editors and a technical editor. Also, the IASS scientific secretary serves as an editor of the section *New and Emerging Methods*.

3.2 Structure

TSS has been produced twice a year since 1974 except for a few years (only 4) where there was only 1 produced. The main reason was that originally, the issues normally came out in June and December, but a few late issues created some drifting. Eventually, the change of reference period to January and July helped stabilize production and became permanent.

In terms of language, the bulletins have been produced in English with a French version added after a few issues. This continued with *TSS*, and Spanish was also added at some point. Spanish versions were produced until late 1990s and French versions until 2012. INSEE and Statistics Canada took care of the French versions and the *Instituto Nacional de Estadistica* in Spain the Spanish version.

In terms of distribution, thanks go to INSEE, the US Bureau of the Census, the Bureau of Labour Statistics, Statistics Canada, the Australian Bureau of Statistics, the Central Statistical Bureau of Latvia and Poznań University of Economics and Business who have in turn been involved and responsible for production and circulation.

As far as the appearance of the newsletter is concerned, the cover displayed the Table of contents from the first issue until issue #15. That is also when country reports started, and sections became more regular. Then issue #32 in July 1995 was the first one with the IASS logo and 5 years later, issue #45 introduced the design of the world map overlaid by people, equations, and a chart. This cover has stood the test of time as it still constitutes today's cover of *TSS*.

4 Major Sections

Over the years, TSS has presented several interesting sections. From the President's letter and the Letter from the Editors (published in the early days, then stopped for a long time and reinstated in 2010) to the News and Conference listings and many other sections, these have provided much needed and interesting inputs to members. The core of TSS has been made of the following sections: Ask the Experts; Articles (New and Emerging Methods since July 2010); Book & Software Review; Country Reports; and In other Journals. This structure has remained unchanged since issue #62, July 2010.

4.1 Ask the Experts

The Question-and-Answer section later becoming the Ask-the-Experts section was created and coordinated by Leslie Kish from 1978 to 1994. During that time, 42 questions were published. Eventually they were republished as a stand-alone booklet (Kish, 1995). Kish took his questions from people he met during conferences and also included a fair number himself, thanks to his acute awareness and understanding of member needs. In 1995, Vijay Verma took it over until 2001 at which point there was a pause. Then in 2004, Anders Christianson re-birthed the series and with Steven Heeringa as *TSS* editor and I looking after the Web site, we worked as a team to help each other with our three responsibilities. During that time, other special guests (e.g., Bill Winkler and Phil Kott) were invited by Anders to contribute answers on specific topics. Following another pause, Robert Clark revived the section from 2010 to 2014. From 2015 to 2018, Kennon Copeland handled it, and I continued it in 2019 and 2020. Since 2021, Ton de Waal has been coordinating the section.

Over the years, a wide array of topics covering all aspects of the work of survey statisticians was covered. These dealt with questionnaires, coverage, designs, editing and imputation, estimation, confidentiality and more recently on-line surveys, non-sampling error, bias, and the use of non-probability samples just to name a few. In total, about 95 questions were posed and answered, an average of about one per issue overall but closer to two per issue when taking into account the years without them. The Ask the Experts Section also has a bespoke tab on the IASS website given its importance on communicating survey statistics: http://isi-iass.org/home/ask-the-experts/.

4.2 Articles – New and Emerging Methods

The presence of articles in *TSS* varied greatly depending on objectives of the executive and editors but also on the active offers of scientific journals as options to publish articles. In the first Newsletters from 1974 to 1978, IASS session structures with a summary of papers presented were included, and authors were encouraged to use the then-existing *Sankhya C* for their complete paper. It was later decided not to include articles in *TSS* and rather use *Survey Methodology* journal from Statistics Canada as the prime vehicle for papers. At that time, it was made available at a preferential rate to IASS members (see Beaumont, 2023 for more details on the history of *Survey Methodology*). Then from issue #30 to #53, short papers came back to *TSS,* and this was stopped again from #54 to #61. Since issue #62, *TSS* has featured at least one article per issue. Of note is the fact that in recent years (since issue #81 in 2020), people have been encouraged by the editors to submit articles that undergo a review. The goal is not to create a refereed journal but rather to elevate the rigor in presentations of new and emerging topics. Only a very small number of articles which are not in line with the direction are rejected. At the same time, *TSS* constitutes a medium for announcements, short communications, questions, experience sharing. The *News and announcements* section is devoted to this.

Overall, 133 articles have been published in *TSS* on current topics of interest to IASS members. This does not include the 36 book and software reviews (see next section) that were published. So, in all, *TSS* has produced close to 170 articles, an average of well over 3 per year during the existence of the IASS. This is very good given that a much larger number of articles have been published by members in numerous journals.

Since issue #81 in January 2020, open access reviewed articles started to be published. This has allowed access to and possibility to download single articles at a time (rather than the complete *TSS*). Also concerning articles, a new *TSS* invited session was organized at the World Statistics Congress in 2021 on *Issues with big amounts of data for survey statistics.*

4.3 Book & Software Review

The Book and Software Review section started (as Software Review) in issue #35 in 1996 with a review paper by Jim Lepkowski and Judy Bowles on Sampling Error Software for Personal Computers. For seven years this section covered the most well-known survey software and tools such as *SAS*, *SPSS*, *STATA*, *SUDAAN*, and *WesVar* to name a few. Then, until 2011, the section was left out except once to present *R*.

With issue #64, the section was renamed and enhanced to include book reviews as well as approaches, guidelines, checklists that are useful to survey statisticians. Since then, almost all issues presented either a book, a software or a survey tool. Books covered a wide variety of modern and classic parts of survey sampling theory and methods such as disclosure, designs, treatment of nonresponse, registers, record linkage, Small Area Estimation, and web surveys. Software presented included *R*, *RShiny*, *Python*'s *samplics* and the *R* package *survey*. In total, 16 software, 15 books and 5 other tools have been reviewed.

4.4 Country Reports

Informing members on survey activities that are taking place in other countries has been a very popular section of *TSS*. In the 50 years of the IASS, over 600 country reports have been produced by more than 70 countries. Moreover, some of the reports often covered more than one survey and have highlighted interesting implementations of survey methods and approaches. Early editions of *TSS* had countries providing articles which sometimes were in the form of reports and sometimes more in the form of an article. This could happen thanks to a network of country representatives established in 1976. Country reports as we know them today really started in 1989 with issue #21. By 1993, the tasks had become in need of a coordinator and Gordon Brackstone (Statistics Canada) managed the section from 1993 to 2001. Ever since, Canada looked after this, the responsibility being passed on to John Kovar (2002-2009), Pierre Lavallée (2010-2016) and Peter Wright (2016-2023).

Contributions by countries were many. Through the years, seven countries (Australia, Canada, New Zealand, The Philippines, Poland, Spain, and US) provided over 25 reports with Australia and New Zealand each providing over 40 and Canada more than 65. Twelve countries provided between 10 and 25 reports (Argentina, Brazil, France, Germany, Hungary, India, Israel, Italy, Japan, Latvia, Malaysia, and UK), while 31 countries did between 2 and 9 reports and 23 other countries provided one.

4.5 In other Journals

Immediately with *TSS* issue #1, people started listing articles related to survey methodology. At that time, it was simply in the form of a bibliography. Then, starting with issue #15, abstracts were produced and listed. It is with issue #34 in 1996 that the section *In other journals* was created to present the table of contents of some of the main journals publishing articles in survey methodology. The main journals in this section have been *Survey Methodology*, the *Journal of Official Statistics* and the *Journal of Survey Statistics and Methodology*. Today the section continues to inform on recently published articles of interest.

5 Conclusions

The Survey Statistician has been a strong presence uniting the members of the IASS. From its modest origin as a few-page leaflet to a solid almost magazine-style paper newsletter, *TSS* has been available on-line since 2000 and exclusively in this format since 2021 when the last paper issue was published in July of that year. It has also been officially registered with an International Standard Serial Number (ISSN 2521-991X) for the world periodicals since issue #67 in 2014. This has made *TSS* an official periodical, a higher status than a manuscript.

Under the leadership of many editors, *TSS* has thrived and hopefully will continue to keep many survey statisticians informed and interested in the IASS. Further, many other members of the associations have contributed to make it what it is today. With continued participation by many members and the new generation of statisticians writing, asking, submitting, congratulating, recognizing, sorrowing and reading, it will continue to succeed.

Acknowledgements

I would like to thank Jean-François Beaumont, Mike Brick, Danutė Krapavikaitė and Natalie Shlomo for their helpful comments. I would also want to thank Gordon Brackstone for helping me complete my access to all issues of *TSS* through his collection as well as Tara Gagnon, the Statistics Canada Library and the Library of Congress for access to issue #1 of the bulletins that pre-dated *TSS*. Finally, an immense thank you to everyone who contributed to any of the 96 issues.

References

Beaumont, J.-F. (2023) The History and the Impact of the Survey Methodology Journal. *The Survey Statistician*, 88, 40-45.

IASS (1974) *Newsletter*, 1, International Association of Survey Statisticians.

IASS (1979-2023) *The Survey Statistician*, 1-87, International Association of Survey Statisticians.

IASS (1999) 25 *Years of the History of the IASS*, International Association of Survey Statisticians.

Kish, L. (1995) *Questions / Answers from the Survey Statistician, International Association of Survey Statisticians.*

Rao, J.N.K. (2023) Some Memorable Recollections of IASS first Meeting. *The Survey Statistician*, 88, 14-15.

The Survey Statistician, 2023, Vol. 88, 51–57.

Survey Sampling History at Iowa State University

Jae Kwang Kim

Iowa State University, USA, jkim@iastate.edu

Abstract

Iowa State University (ISU) has played an important role in research and education in survey sampling. In this article, we give a brief history of the Statistical Laboratory Survey Section (now known as the Center for Survey Statistics and Methodology) and its impact on the survey sampling community. Some reflections on my journey in survey sampling are also presented.

Keywords: Center for Survey Statistics and Methodology, National Resources Inventory.

1 Introduction

Iowa State University (ISU) has played an important role in research, practice, and education in survey sampling. The Statistical Laboratory (Stat Lab) at Iowa State College was established in 1934 to promote statistical research and provide consulting to other university units, and led by George Snedecor. The Survey Section of the Statistical Laboratory, which later became the Center for Survey Statistics and Methodology (CSSM), was established in 1938 as a result of a cooperative agreement between the Statistical Laboratory and the U.S. Department of Agriculture (USDA).

In this article, I give a brief history of Stat Lab and its impact on the survey sampling community. Some reflections on my journey in survey sampling are also presented.

2 History

2.1 Early Years, 1938-1948

The Department of Agriculture was one of the early organizations in the U. S. to initiate research and development work on probability sampling, and they established a cooperative research program with the Statistical Laboratory at Iowa State University in 1938. Initial work under the cooperative agreement with the USDA led to the development of the Master Sample of Agriculture (King and Jessen, 1945), a national area sample of land that was subsequently used in numerous economic surveys of American agriculture, as witnessed by Fuller (1984).

Jessen (1942) investigated the problem of approximating the optimum sizes of sampling units for agricultural studies. The paper guided the development of the Master Sample of Agriculture and stimulated the later development of designs and theory for rotating samples for surveys taken on successive occasions for time series estimation.

Iowa State is also one of the few universities that has featured training in sample survey theory and methods as an important part of statistical training, as noted by Hansen (1987). Under George Snedecor, Iowa State was also a center for statistical treatment of experimental work. The emphasis in applied statistics at Iowa was then on sample surveys and experimental design. William G. Cochran then joined the Rothamsted Experimental Station and lectured on sample surveys and experimental design. *Sampling Techniques* by Cochran (1953) was developed from the lectures at Iowa State University. In 1946, Cochran left Iowa to organize and head the graduate program in experimental statistics at North Carolina State College at Raleigh. Theodore Bancroft took over from George Snedecor as the Head of the Department of Statistics (established in 1947) and the Director of the Statistical Laboratory during 1950-1972.

2.2 1950s-1960s

Survey research and consulting continued to grow through the efforts of the groups in many disciplines and areas. The early 1950s saw survey statistics research highlighted in Ph. D. dissertations in other disciplines on campus, a practice that has grown significantly since. Other efforts in the 1950s and 1960s included survey research, methodology, and practice in areas such as employment trends in the state of Iowa, home economics studies across the nation, farm practice surveys, and surveys for researching social welfare in the state of Iowa, just to name a few. Interest from international venues also increased after the 1950s, as research fellows from across the globe began visiting the department to take courses in survey methodology and work with faculty and students on applied research projects. For example, the seminal paper of Horvitz and Thompson (1952) was published when the authors were graduate students at ISU and they were influenced by the lectures from Midzuno who visited ISU from Japan. P. V. Sukhatme also visited from India and wrote his famous book (Sukhatme, 1953) at ISU, which was published by Iowa State College Press. His brother, B. V. Sukhatme, joined as a faculty member of ISU later and revised the book (Sukhatme and Sukhatme, 1970) and advised graduate students.

In 1956, the Survey Section began cooperating with the US Soil Conservation Service (now the USDA Natural Resources Conservation Service) to develop survey methods and provide operational support for the National Resources Inventory (NRI), a longitudinal survey of agricultural and other natural resources on nonfederal lands. CSSM continues to work on the NRI project today, and the survey has been the inspiration for many methodologies related to sampling and estimation (Nusser and Goebel, 1997). Results from these surveys are used extensively in the construction of Farm Bills in the US. A second major, ongoing collaboration began in this period with the US Census Bureau on improving survey sampling and methodology in census practices.

H. O. Hartley joined ISU in 1953 and became deeply involved in research and teaching. Hartley made several major contributions, including the famous paper on unbiased ratio estimation published in Nature (Hartley and Ross, 1954). This paper motivated Mickey to develop a whole class of unbiased regression estimators (Mickey, 1959), which was done at ISU. Hartley also wrote seminal papers on domain estimation (Hartley, 1959) and dual frame surveys (Hartley, 1962) while he was at Iowa State. Hartley's contribution to survey sampling is well summarized by Rao (1983). Hartley is the Ph. D. advisor of J. N. K. Rao. Rao stayed at ISU for 5 years (1958-1963), three years as a student and two years as Assistant Professor (AP). Rao, Hartley, and Cochran (1962) published a paper on a very simple procedure of unequal probability sampling scheme without replacement that allows them to estimate the variance of the resulting estimator of total. During his AP tenure, Rao shared an office with Wayne Fuller, who just joined the Statistics Department as an AP, and they remained good friends to each other throughout their professional careers. Hartley also advised Edward Bryant on two-way stratification (Bryant et al, 1960), which received a lot of attention at that time. Edward Bryant later founded Westat.

Figure 1: Wayne A. Fuller

2.3 1970s-1980s

With the arrival of more powerful mainframe computers, namely the IBM 360, as well as advances in statistical computing, the Survey Section began to develop and implement more sophisticated survey and data analysis. SUPER CARP, the first mainframe computer program developed by Mike Hidiroglou in the Survey Section, allowed the implementation of many estimation methods used in survey sampling in an automated manner. The program used the software developed to compute regression estimation in the context of survey sampling. The software was also expanded to allow for estimation and estimated standard errors for totals, ratios, means, and proportions for subdivisions of a sampled population. It also contained several procedures appropriate for data observed subject to measurement errors. It was operational by 1976: See Hidiroglou, Fuller, and Hickman (1978). In 1985, SUPER CARP underwent major revisions and updates of algorithms that allowed its deployment on IBM-PCs. Later, EV (errors in variables) CARP was released as companion software to Wayne Fuller's book, *Measurement Error Models* (1987). Many graduate students, including Cary Isaki, Mike Hidiroglou, Kirk Wolter, Elizabeth Huang, Yasuo Amemiya, Sastry Pantula, David Dickey, and John Eltinge, worked on this project and its related topics under the supervision of Wayne Fuller. Kirk Wolter served as the president of the International Association of Survey Statisticians (IASS) during 1999-2001.

Wayne Fuller has since made many important contributions to survey sampling. Wayne Fuller can be credited for introducing the regression idea to adjust the design weights to construct calibration weights. Huang and Fuller (1978) developed an iterative method for constructing range-restricted weights that meet the benchmarking constraints and the design consistency. Isaki and Fuller (1982) laid the foundation for establishing the optimality of the regression estimator. Battese, Harter, and Fuller (1988) developed a framework for a unit-level model approach to small area estimation.

2.4 1990s-2000s

New faculty members, including Sarah Nusser, F. Jay Breidt, and Jean Opsomer, joined the Survey Section in the 1990s. During this period, survey statistics education and research continued to flourish at CSSM. The wide diversity of projects undertaken included survey consulting on projects such as local efforts to analyze and improve ISU campus services; consulting projects with departments and bureaus of the state of Iowa to develop statistical pictures and gather information about Iowa residents' behaviors and preferences to farm production surveys and Iowa business and economic survey research; ongoing long-term research projects with national agencies such as the USDA / NCRS and the Census Bureau, as well as new projects with the National Cancer Institute, the Centers for Disease Control, the National Institutes of Health, the Bureau of Land Management, and the

National Science Foundation, amongst many others. In November 2002, the Survey Section of the Stat Lab officially became the Center for Survey Statistics and Methodology. Wayne Fuller officially retired in 2001, but he continued working as a part-time consultant at the CSSM.

On the research side, a measurement error model was applied to estimate the usual daily intake distribution (Nusser Carriquiry, Dodd, and Fuller, 1996). This work served as the basis of ongoing work by Alicia Carriquiry and Fuller that still influences approaches to dietary assessment in the United States and many other countries. The NRI's longitudinal 2-stage stratified cluster sample, which is observed every 5 years, was redesigned to an annual survey with supplemented panels. Design and estimation of the supplemented panel survey became very important for NRI application (Nusser, Breidt and Fuller, 1998; Fuller, 2003). Breidt and Opsomer (2000) developed nonparametric regression estimation methods and collaborated on many research problems in survey sampling. Regression weighting methods were further developed for the U. S. Census (e. g., Isaki, Tsay, and Fuller, 2004). Fractional hot deck imputation was developed by Kim and Fuller (2004). Emily Berg wrote a Ph. D. thesis on small area estimation under the supervision of Wayne Fuller (Berg and Fuller, 2014). An advanced-level textbook on survey sampling written by Fuller (2009) was finally published.

2.5 2010s-present

Cindy Yu, Jae Kwang Kim, Zhengyuan Zhu, and Emily Berg joined ISU and became a new generation of CSSM faculty. Each has influenced continuous methodological developments in different ways - Kim via missing data analysis, Zhu via spatial data models, and Berg via her expertise in small area estimation. CSSM provided statistical consulting to other agencies, including the National Agricultural Statistical Service, the Bureau of Land Management, the Bureau of Justice Statistics, and the Food and Agriculture Organization (FAO) of the United Nations.

Kim expanded the departmental curriculum with a graduate-level course on handling missing data, and the lecture notes matured into a textbook (Kim and Shao, 2021). Kim used his expertise in missing data and developed a series of methods for data integration (Kim, 2022). Under the directorship of Zhu, the CSSM continues to expand, and the funding size is now about 5 million USD per year.

3 Reflections

I consider myself to belong to the third generation in the survey sampling community. The first generation at ISU includes William Cochran and H. O. Hartley. The second generation at ISU includes J. N. K. Rao and Wayne Fuller. Standing on the shoulders of giants, I learned the essence of survey sampling theory and methods. When I was a graduate student at ISU, the Statistics Department offered a very specialized curriculum in survey sampling, with separate MS and Ph. D. courses in sampling, and Jay Breidt's lectures were very clear and excellent. I wrote my dissertation under the supervision of Wayne Fuller and have benefited a lot from Fuller's excellent insights and rich research experience.

The second generation flourished in the "Golden Age of Survey Research" (Singer, 2016) when the response rates were high and other data sources were unavailable. As a third-generation member, as Kalton (2019) pointed out, I faced two main challenges in survey sampling. One is the declining trend in response rates and the related increases in the costs of surveys based on probability samples. The other challenge comes from the emergence of an alternative source of information, including large administrative data and low-cost web panel samples. Thus, naturally, I became interested in the research topics addressing these new challenges: handling missing data and adjusting selection bias in the voluntary samples through data integration or weighting.

Imputation for handling item nonresponse is a topic of my Ph. D. thesis. I worked on a consulting

project for the U. S. Bureau of Census on estimating the variance of the census long-form survey estimates after nearest neighbor imputation (Kim, Fuller, and Bell, 2011), which is based on ignorable missingness assumption. Before joining ISU in 2008, I had the opportunity to work on a project related to election exit polls in Korea. This sparked my interest in nonignorable missing research and led to several papers over the years at Iowa (Kim and Yu, 2012; Morikawa and Kim, 2021). An invitation from J. N. K. Rao to visit Ottawa in 2007 was also an eye-opening experience for me. Combining information from two independent surveys (Kim and Rao, 2012) started with the visit to Rao. Consulting projects from Statistics Korea and the USDA National Agricultural Statistics Service, data integration methods were developed using the measurement error model (Kim, Park, and Kim, 2015) and multilevel models (Kim, Wang, Zhu, and Cruze, 2018), respectively. A visit to the Australian Bureau of Statistics in 2016 sparked my interest in data integration research incorporating big data (Kim and Tam, 2021).

As a sampling statistician in academia, I now see another challenge approaching us: how to teach survey sampling and educate the next generation so that they can understand the value of survey sampling. In the era of machine learning and AI, students are more interested in learning modern techniques than classical subjects. Thus, in addition to the decline in survey participation, we are facing a decline in interest in survey sampling research among the next generation. How do we improve our teaching, modernize our textbook and find interesting research problems to attract young "smart" students into survey sampling? I think these questions should be seriously addressed by the survey sampling community in academia.

Acknowledgements

I would like to thank Danutė Krapavickaitė for the invitation to write an article on the history of survey sampling at ISU. I also thank Jon Rao, Sarah Nusser, Mike Hidiroglou, and Wayne Fuller for their very helpful comments on the earlier version of the article. Some contents are taken from the website of the CSSM. The research was partially supported by the cooperative agreement between NRCS-USDA and the CSSM at ISU and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28–36.

Berg, E. J. and Fuller, W. A. (2014). Small Area Prediction of Proportions with Applications to the Canadian Labour Force Survey. *Journal of Survey Statistics and Methodology* **2**, 227–256.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics* **28**, 1026–1053.

Bryant, E. C., Hartley, H. O., and Jessen, R. J. (1960). Design and Estimation in two-way stratification. *Journal of the American Statistical Association* **55**, 105–124.

Cochran, W. G. (1953). *Sampling Techniques*. John Wiley & Sons, New York.

Fuller, W. A. (1984). The Master Sample of Agriculture. In David, H. A. and David, H. T., Eds. *Statistics: An Appraisal*, Iowa State University Press, Ames, Iowa, 583–602.

Fuller, W. A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data* (edited by R.L. Chambers and C.J. Skinner), Wiley, Chichester, UK, 307–322.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

Fuller, W. A. (2009). *Sampling Statistics*. John Wiley & Sons, Hoboken, New Jersey.

Hansen, M. H. (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science* **2**, 180–190.

Hartley, H. O. (1959). Analytical Studies of Survey Data, *Rome: Istituto di Statistica. Volume in Honor of Corrado Gini*. pp. 1–32.

Hartley, H. O. (1962). Multiple Frame Surveys, *Proceedings of Social Statistics Section*. pp.203–206. American Statistical Association.

Hartley, H. 0., and Ross, A. (1954). Unbiased ratio estimators. *Nature* **174**,270-271.

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1978). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Huang, E. T. and Fuller, W. A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1897–1904.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association* **77**, 89–96.

Isaki, C. T., Tsay, J. H., and Fuller, W. A. (2004). Weighting sample data subject to independent controls. *Survey Methodology* **30**, 35–44.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin* 304.

Kalton, G. (2019). Developments in Survey Research over the Past 60 years: A personal perspective. *International Statistical Review* **87**, S10-S30.

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *The Survey Statistician* **85**, 19–29.

Kim, J. K. and Fuller, W. A. (2004) Fractional hot deck imputation. *Biometrika* **91**, 559–578.

Kim, J. K., Fuller, W. A., and Bell, W. R. (2011) Variance Estimation for Nearest Neighbor Imputation for U.S. Census Long Form Data, *Annals of Applied Statistics* **5**, 824–842.

Kim, J. K., Park, S. and Kim, S. (2015). Small area estimation combining information from several sources, *Survey Methodology* **41**, 21–36.

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 Shao, J. (2021). *Statistical Methods for Handling Incomplete Data* (2nd Edn). Chapman & Hall / CRC, Boca Raton, FL.

Kim, J. K. and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference, *International Statistical Review* **89**, 382–401.

Kim, J. K. and Yu, C. L. (2011). A semi-parametric estimation of mean functionals with non-ignorable missing data, *Journal of the American Statistical Association* **106**, 157–165.

Kim, J. K., Wang, Z,, Zhu, Z., and Cruze, N. (2018). Combining survey 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.

King, A. J. and Jessen, R. J. (1945). The master sample of agriculture. *Journal of the American Statistical Association* **40**, 38–56.

Mickey, M. R. (1959). Some finite population unbiased ratio and regression estimator. *Journal of the American Statistical Association* **54**, 594–612.

Morikawa, K. and Kim, J. K. (2021). Semiparametric Optimal Estimation With Nonignorable Nonresponse Data, *Annals of Statistics* **49**, 2991-3014

Nusser, S. M., Breidt, F. J., and Fuller, W. A. (1998). Design and estimation for investigating the dynamics of natural resources. *Ecological Applications* **8**, 234–245.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual daily intake distribution. *Journal of the American Statistical Association* **91**, 1440–1449.

Nusser, S. M. and Goebel, J. J. (1997). The National Resources Inventory: a long-term multi-resource monitoring programme. *Environmental and Ecological Statistics* **4**, 181–204.

Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society: Series B*, **24**, 482–491.

Rao, J. N. K. (1983). H.O. Hartley's Contributions to Sample Survey Theory and Methods. *The American Statistician*, **37**, 344-350.

Singer, E. (2016). Reflections on surveys' past and future. *Journal of Survey Statistics and Methodology*, **4**, 463–475.

Sukhatme, P. V. (1953). *Sampling Theory of Surveys with Applications*. The Indian Society of Agricultural Statistics, New Delhi. Iowa State College Press, Ames, Iowa.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling theory of surveys and its applications*. Iowa State College Press, Ames, Iowa.

The Survey Statistician, 2023, Vol. 88, 58-61.

Sample Surveys in Post-Apartheid South Africa

Jairo Arrow

Retired Deputy Director-General, Statistics South Africa, Vice President IASS
jairo.arrow@gmail.com

Abstract

South Africa conducted its first non-racial housing and population census in 1996. Statistics South Africa (StatsSA) currently conducts two major household sample surveys: the General Household Survey (GHS) and the Quarterly Labour Force Survey (QLFS). International support has played a key role in the development of statistical production in South Africa. StatsSA still faces challenges in terms of building and maintaining a robust statistical infrastructure.

Keywords: sample surveys, post-apartheid, Statistics South Africa, sampling frames, international support, CPI crisis

1 Introduction

Over the past fifty years, survey statistics in Africa have evolved significantly, driven by changing socio-economic and political dynamics on the continent. The earliest surveys in Africa were conducted by colonial administrations. The statistics so collected were not intended to inform policy outcomes for the betterment of the local population but to strengthen the colonial grip on agricultural production, trade, and the local population. The minority Apartheid state in South Africa was not any different from the British, French, or Portuguese colonial powers in the collection and compilation of statistics. My contribution covers the South African post-apartheid period. That said, I would nevertheless touch briefly on the history of sample surveys on post-colonial African during the last fifty years.

2 A brief overview of sample surveys in Africa

The first Sub-Saharan African country to obtain independence from Britain was Ghana in 1957. But it took another thirty years before the country ran its first sample household survey, the Ghana Living Standards Survey (GLSS), in 1987. Although Ethiopia was never colonized the country conducted its first sample household survey, the Ethiopian Rural Household Survey (ERHS), in 1980, which came two decades after the Ethiopian Housing and Population Census of 1960. In North Africa the earliest sample household surveys were conducted in Egypt and Morocco in 1958/59 and 1960, respectively. The first sample household survey conducted in the Southern African region was the Living Standards Measurement Survey (LSMS), which was carried out in Botswana in 1984. South Africa conducted its first sample household survey, the October Household Survey, in 1993.

The first business sample survey conducted on the African continent was the so-called Ghanaian Manufacturing Census in 1962. Since then, many African countries have conducted business sample surveys to collect data on the structure and performance of their economies.

When compared to other continents, in terms of the conduct of household sample surveys, independent African countries do not fare badly. The USA, for example, conducted its first household sample survey in 1940. In Latin America, the first household sample survey was undertaken by Mexico in 1951. India, on the Asian Continent, conducted its first household sample survey in 1950/51, which was even earlier than the UK Family Expenditure Survey (FES) which took place in 1957.

3 Statistics South Africa – Constructing a modern statistical system in post-apartheid South Africa

On April 27, 1994, South Africans lined up in meandering queues patiently waiting to cast their vote for a parliament in "*which the colour of a man's skin*" was immaterial. But what was the turnout? Nobody could tell because there was no comprehensive population register for the whole country. Neither was there a dwelling nor a business sampling frame. Post-apartheid South Africa was still facing the lack of these fundamental building blocks of a modern statistical system when I joined the Central Statistical Services (CSS) in 1997. The CSS, which later morphed into Statistics South Africa (Stats SA), had hitherto served only the white population. The CSS counterparts in the Bantustans, or homelands for blacks, were statistical agencies only in name.

On 9 October 1996, South Africa conducted its first non-racial housing and population census. It was estimated that 10.7% of the people had been missed in the count. This was the beginning of regular population censuses intended to be conducted every five years. Accordingly, the second housing census took place on 9 October 2001. The count was characterised by a relatively high undercount of the population, estimated at 16.7%. The quality of the census, judged by the undercount had deteriorated. This poor performance was blamed on the lack of resources at StatsSA, both human and financial. The next round of the population count was postponed to 2011, and to fill the gap, Stats SA conducted a large household sample survey, the Community Survey (CS 2007), in October 2007. The third census took place in October 2011. The sampling frame used was a collection of geographical units, called Enumeration Areas (EAS). Enumeration areas were created by dividing the country into small geographic areas. In the runoff to the count, South Africans were keen to have better results, but they were again disappointed when the undercount was in the two-digit range of 14.6% The second large sample household survey (CS 2016) was conducted in October 2016. Due to the COVID-19 pandemic the population census planned for 2020 was delayed and took place in February 2022. The results will be published this year, 2023. StatsSA currently conducts two major household sample surveys: the General Household Survey (GHS) and the Quarterly Labour Force Survey (QLFS).

4 The Business Registration Reform Project

When I started at Stats SA in 1997, as a director in the Economic Statistics department, the sampling frame for Economic Statistics was the Business Address Register (BAR). Through a careful analysis of business and economic surveys, we concluded that the BAR was inadequate and fell short of the main characteristics of a sampling frame: regular updating, comprehensive coverage, and proper classification of units by size and industry. These concerns led to important legislative changes. In 1999, Parliament amended and passed the Income Tax Act, giving Stats SA access to all tax categories (Income Tax, VAT, PAYE, and Customs) for statistical purposes only. The Department of Trade and Industry (DTI) was responsible for business registration through its agency, Companies and Intellectual Property Registration Office (CIPRO), now called the Companies and Intellectual Property Commission (CIPC).

There was no single and authoritative view of the business landscape. An interdepartmental project team was set up and President Zuma in his state of the nation address to Parliament in 2007 announced the establishment of a Business Registration Reform Project as one of the goals of his administration in the coming years. This project was to be executed by five government departments:

SARS, the National Treasury, the Department of Trade and Industry, the Department of Labour alongside Stats SA. The objectives of the project were:

- The establishment of a single registration authority for primary business registration.
- Review of the current legal definition of 'primary business registration' to include tax registration.
- Expansion of the type of business entities included in the legal definition of primary business registration.
- Compliance with all primary business registration requirements through a single transaction.
- The introduction of a mandatory unique business identifier for all legal and commercial transactions.
- Expanding the number and type of registration channels available for the purpose of primary business registration.
- Increased data and information sharing across government entities.

A legislative framework was proposed by an inter-department project team, which envisaged a law governing the registration of businesses in South Africa under a single government agency. But the project was abandoned in 2011 when departmental rivalries stood in the way.

5 International support in establishment of sampling frames

Important changes that occurred in the framework of statistical production perhaps would not have occurred had it not been for the material support from the international statistical community. The Australian Bureau of Statistics (ABS) was the first international statistical agency on the scene. Senior ABS staff provided advisory support for extended periods to Stats SA statisticians and senior management. Statistics Sweden also played a key role in the transformation of statistical production environment by dispatching advisors to Stats SA. Together with Statistics New Zealand, Statistics Sweden seconded senior staff to help set up the current Business Sampling Frame which is now the omnibus sampling frame for the business surveys at Stats SA.

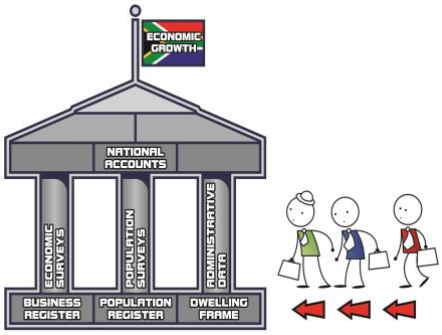
6 The CPI crisis

As StatsSA was feeling more and more confident in its role as the 'top statistical agency' in Africa, as the newly appointed head of the agency, the Statistician-General, Mr Pali Lehohla, claimed. Around 2002-03 a less well-known economic researcher at a small investment bank, Investec, was consistently casting doubts on the quality of the consumer price index (CPI). StatsSA fiercely denied the allegations of misstating the price index as baseless and mischievous. A small team of Stats SA methodologists carried out a detailed analysis of the CPI and came to the same conclusion that the CPI was indeed misstated. This was a huge setback in the trustworthiness of the national statistical collection.

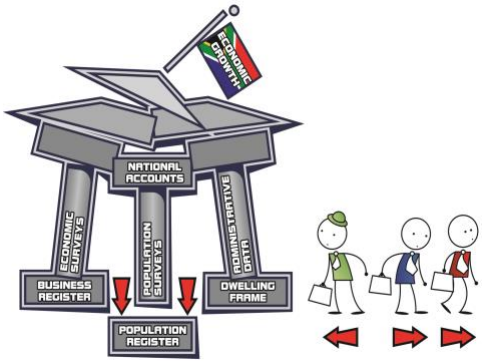
The CPI crisis led to a general distrust in the official statistics produced by the statistical office. Thereafter employment statistics came under strong criticism. A small firm of labour brokers even launched its own series called the Adcorp employment index, which for several years served as a parallel source for employment statistics in the private sector.

7 The role of the sampling frames as a foundation for production of quality statistics

The creation of a credible resilient statistical system is like building a house. The statistical office mandated to compile national statistical indicators becomes synonymous with the factfinder of the society. The statistical office is then the authoritative source of statistical indicators. It becomes the primary destination users seeking reliable and trustworthy statistics would turn to. Many countries do not have a population register but nevertheless successfully conduct household surveys based on robust statistical standards and classifications.



On the other hand, a weak statistical system is like a house with faulty foundations (sampling frames) and leaking roof (National Accounts). The resultant statistical indicators are not trusted by users, who turn to myriad other sources of information, some which might not stand rigorous scrutiny.



In the 1970s, most African countries focused on conducting large-scale national surveys to collect data on key economic and social indicators, such as income, employment, education, and health. However, data collection was often hampered by logistical challenges, including poor infrastructure, limited resources, and political instability.

8 The World Statistics Congress 2009

The 57th Session of the International Statistical Institute, as the World Statistics Congress was then known, was held in Durban, South Africa. It was unmistakable proof of a South African statistical system that had come of age. It was an historical event in other respects: it was the first in Africa and under the first female ISI President, Denise Lievesley. I was fortunate enough to serve as the Executive Secretary of the Organizing Committee. This event was not just an important moment for South Africa but also for the entire African statistical system.

9 Some challenges facing the statistical system.

The past fifty years years have indeed witnessed significant progress in the development of survey statistics in Africa. Many African countries have invested in building stronger statistical systems, including the development of national statistical plans, the establishment of statistical agencies, and the adoption of innovative technologies and methodologies for data collection and analysis.

However, there are several challenges still facing statistical agencies in Africa in terms of building and maintaining intellectual human capital. StatsSA has attracted skilled staff from across the continent, e.g., Ethiopia, Nigeria, Kenya. Additionally, it runs an internship programme which has attracted young graduates from South African universities, who undergo a one-year training programme at Stats SA. This programme has helped StatsSA grow its own crop to address skills shortages in the statistical system.

The Survey Statistician, 2023, Vol. 88, 62-67.

Development of Sample Surveys in Australia and New Zealand over the Last 50 Years

Dennis Trewin

Australian Statistician (2000-2007) and IASS President (1995-1997)
dennistrewin7@gmail.com

Abstract

This note outlines the development of sample surveys in official statistics in Australia and New Zealand since the inception of the IASS. It highlights the increasing need to not just rely on sample surveys although they remain important as a component of mixed mode methods. Accordingly, sample survey methodologists need to broaden their skills to maintain their relevance.

Keywords: sample surveys, mixed mode, non-sampling errors.

1 Introduction

Both countries have centralised statistical systems and the majority of sample surveys are conducted by the national statistical offices. Hence, this paper concentrates on developments in these two offices. Given the context of this issue of *The Survey Statistician*, the paper is also more about past history than recent history. A good reference for Australia is ABS (2005).

As an inaugural member of the IASS in 1973, I have seen many changes in survey methods over the last 50 years. Nevertheless, both Australia and New Zealand (NZ) were relatively mature in their adaption of probability-based sample surveys by 1973 under the leadership of Ken Foreman at the Australian Bureau of Statistics (ABS) and Steve Kuzmicich at Statistics New Zealand (SNZ). The importance of their work was recognised because, for much of their careers, they held quite senior positions and were regarded as part of the executive team as well as being chief methodologists.

By 1973, a national household survey was in place in Australia and had been held since 1960. It was used to conduct the quarterly labour force survey (LFS) and a series of supplementary surveys. In New Zealand, the LFS came later whilst they relied on registered unemployment data. In both countries, a number of special household surveys such as household income and expenditure surveys (HIES) had already been conducted. Multistage area frameworks were used with a mesh block as the primary sampling unit (PSU) in NZ whilst Australia used the larger Census Collectors Districts (CDs) and used field work to create blocks within selected CDs.

Probability sampling methods were also used to conduct surveys of businesses such as the monthly retail surveys. In Australia, the first business surveys using probability sampling were conducted in 1947 and in 1956 in NZ. Sampling methods were also being used in the Australian Census to conduct post-enumeration surveys and to support quality control of Census processing.

2 Developments in Household Survey Methods

Household surveys enabled a massive increase in social statistics in both countries. Following the 1971 Australian Population Census, a significant effort was made into developing a multi-stage area sampling frame that could be used for multiple surveys as well as providing a rotating sample for the main survey – the quarterly labour force survey which had started in the early 1960s and became monthly in 1978. A stratified self-weighting design was used. Households remained in the sample for eight consecutive surveys before being rotated out. Independent population age x sex benchmarks were used to adjust for under-enumeration which was particularly high for young male adults. The new design provided for a 40% reduction in variance without any increase in sample size.

During this time a program of special household surveys was developed in both countries on topics such as household expenditure and income, health, time use, disability and education.

The enumeration methods have also evolved over time. Fifty years ago, face to face interviewing was exclusively used. Telephone interviewing was introduced in Australia in 1994 but a face to face interview was still used for the first time households were in the survey in order to obtain approval to conduct future interviews by telephone and obtain contact details. Before implementation, considerable effort was put into measuring any accuracy impacts from the change in collection mode. The existence of an identified impact delayed the introduction of telephone interviewing, but it turned out to be largely a first survey effect. Computer-Assisted Paper Interviewing (CAPI) was introduced in 2004 also enhancing the availability of para data to assist with survey design. Telephone interviewing was introduced in New Zealand a little later but face to face interviewing remained the dominant interviewing method. The accuracy of telephone lists was never high enough to be seriously considered as a sampling frame. The possibility of respondents completing the survey over the internet was introduced in 2014 in Australia following analysis showing measurement effects were relatively small.

An address frame of residential dwellings was developed for the 2016 Australian Population Census. It proved to be quite accurate and has been updated since then largely using external lists such as Australia Post and machine learning methods with satellite imagery as the data source. Since 2018, this frame has been used for the household survey program enabling more sophisticated means of controlling sample rotation and overlap between surveys. SNZ retains its 2 stage area sampling method for household surveys.

Methods for making estimates of aggregates and their sample errors have changed considerably over time as methods developed and more benchmark data became available. Model assisted GREG estimators are now mostly used.

3 Developments in Business Survey Methods

In Australia, probability surveys were first used for the business surveys (e.g. capital expenditure) in the 1940s and a little later in New Zealand. In the early years, the major cause of inaccuracies were missing units on the sampling frames. There was no great source for updating the frame. In Australia, a major effort in the late 60s and the early 70s was the development of an Integrated Business Register, providing for the hierarchy of enterprise groups, enterprises and establishments complete with industry codes, that could be used as a frame for all economic censuses and surveys. Tax data became available to support updates but still required considerable follow-up effort before new businesses could be added to the frame with confidence. Similar developments happened in New Zealand but it was not until the mid-1980s that the required tax data access was available. Prior to then, field checks by CPI staff were commonly used.

The existence of a Business Register enabled the application of a method, referred to as collocated sampling in Australia, to provide for the rotation of sampled businesses after they had been in the sample for a designated period of time (e.g., 3 years) (Brewer et al., 1972). Furthermore, it enabled

the control of overlap across surveys. Independently, Statistics Sweden had developed a very similar method. A simplified version was used in SNZ.

All the traditional economic censuses (e.g., manufacturing census) were converted to sample surveys in both countries across the following decade. More recently, most economic surveys tend to be economy-wide rather than industry specific.

Mail was the dominant data collection method with telephone interviewing mostly used to follow up non-response. The exception was the monthly retail survey where most of the data collection for small businesses was by telephone. Consequently, it was the first application of Computer-Assisted Telephone Interviewing (CATI) in the ABS, introduced after an extensive study of possible mode effects. About the same time, in both countries, the very largest enterprise groups were profiled so that their enterprise group-enterprise-establishment structure matched their financial accounting arrangements to the extent possible and their data collection was managed by so called large business units.

Despite the considerable efforts to integrate the data collections that provide source data for the national accounts, the statistical discrepancies within the accounts were still larger than desired. In the late 1980s, there was an extensive investigation into the reasons in Australia. It was found that a major reason was the inconsistent treatment of the missing units on the frame (mostly new businesses), businesses that were no longer operating, and non-response. Among other things, this led to a delay in detecting a turning point in the important survey of capital expenditure impacting government fiscal policy at the time. Standard procedures were developed by Methodology, including the estimation of new business provisions by industry, which were provided to all the surveys by an organisational unit especially set up for this purpose. Further data confrontation methods across collections, using the national accounting framework, were also put in place prior to the finalisation of the national accounts. This was a form of output editing but looking at multiple collections simultaneously. It resulted in considerably more accurate national accounts. An important outcome of the investigations was Ministerial support for greater access to tax data. Similar issues in SNZ led to proposals for a more systematic approach to economic collections.

4 Population Census

Statistical quality control methods for processing were first put in place for the 1961 Australian Population Census. Acceptance sampling methods were used which involved the acceptance or rejection of coded forms based on error counts determined by expert coders. Rejected lots were reprocessed. Studies showed that expert coders were not necessarily more accurate. At the 1976 Census, this was changed so the sample checks were used to provide information on the level and type of errors only. This information was used to identify ways in which the quality of processing could be improved e.g., retraining or improved coding instructions.

A Post-Enumeration Survey (PES) has also been conducted by the ABS since the 1966 Census. Its importance increased substantially following the 1976 Census. Following detailed demographic analysis, it was determined that Census counts were far more accurate estimates of the population if they were adjusted for undercount (at an age group x sex x State level) using PES data. Subsequently, the official population estimates for Australia have been adjusted using PES data after every Census. It was the first country to do so. The analysis also found, contrary to conventional wisdom at the time, the Censuses in the latter half of the twentieth Century were far more accurate than the first half. SNZ conducted their first PES in 1991. In SNZ, the PES is used for Census evaluation purposes only.

SNZ conducted surveys on Disability and Use of Maori Language using frames determined by responses to questions placed in the 2016 Census.

SNZ had field enumeration problems in the 2018 Census resulting in large non-response. They used administrative data to ‘rescue’ their 2018 Census supported by major methodological work. Based on this positive experience with the use of administrative data, the 2023 census is being designed

to make more use of this data and thereby reducing the reporting load on the public and possibly improving the accuracy of some aspects.

5 Increasing Interest in Non-sampling Errors

There has been long standing interest in both countries in measuring, understanding and controlling non-sampling errors. As part of the effort to control non-sampling errors, pilot testing was a standard procedure for new surveys or when introducing new methods. Until relatively recently, there was always a strong emphasis on maintaining high response rates to minimise non-response bias. They remain high by international standards although now survey designs focus more on ensuring samples are sufficiently representative of the population by using techniques such as adaptive sampling.

A research study into editing for the Retail Census showed that it introduced more errors than it discovered as the editing clerks learnt how to game the editing system so that each record passed the computer edits. Studies of other collections were consistent with this finding. The resources devoted to editing across all collections were considerable and did not contribute much to overall accuracy. These studies led to the introduction of more cost-effective macro-editing approaches that focussed on the most significant errors reducing costs as well as improving accuracy.

The increased effort into understanding and quantifying measurement errors led to consideration of the optimisation of Total Survey Error (TSE) rather than sampling error. This was inspired by Dalenius' work (1967) and preceded the more recent interest in TSE. In the ABS, it was applied to the design of the new Construction Industry Survey (see Linacre and Trewin, 1993) where one important decision from the TSE work was to use a more expensive field enumeration method for smaller businesses but with a smaller sample size. Subsequently, it has been recognised in design work that there are far better returns from methodological investments in frame maintenance and other non-sampling errors than clever work on sample designs.

This research also showed the importance of good management to reducing non-sampling errors. It was not just about design (see Trewin, 2001).

6 Use of Administrative and Big Data

In recent decades, there has been many innovative uses of administrative data. It has always been used to compile statistics in subject fields such as Foreign Trade and Births Deaths and Marriages. It has also been used to provide proxy indicators for compilations like the national accounts. Furthermore, it has been used to develop and maintain sample frames and benchmark data to help improve the efficiency of sample surveys. In more recent years, innovative uses include:

1. Data substitution (tax data). Considerable effort has been put into maintaining a good and trustworthy relationship with the Tax Offices. Access to tax data has increased considerably over time especially with the introduction of a Goods and Services Tax in both countries which provided monthly and quarterly data. One important use was data substitution. Studies showed the tax data was reliable (perhaps more reliable than data collected by the ABS and SNZ) especially if edited for the more significant anomalies such as coding errors.
2. Linked Data Sets. The links may be between two administrative data sets or between administrative data and Census/survey data sets. This has resulted in the creation of new richer data sets for the production of official statistics and supporting research;
3. Longitudinal Data Sets. A specific application has been the creation of longitudinal data bases; and
4. Big Data. There have been no real applications to date but its use in small area estimation is being actively investigated.

7 Longitudinal and Linked Data Sets

Neither the ABS nor SNZ have conducted many longitudinal surveys, but they have provided support to other agencies. However, in recent years administrative data has been used to create longitudinal data sets, sometimes using their own data sets, using linkages at the individual level across data sets. For example, Business Longitudinal Data Bases have been created in both countries. SNZ has created an Integrated Data Infrastructure combining administrative data for individuals.

As an example of longitudinal data sets involving survey data, longitudinal data files have been created from the monthly labour force survey taking advantage of the fact that 7/8th of the sample is common from one month to the next. A longitudinal data base of Census records has also been developed in Australia using statistical matching techniques. Starting with the 2006 Population Census, a 5% sample of Census records was retained without name and address identifiers but with sufficient information to allow statistical matching across the individual data sets. It also enabled linking across Censuses thereby establishing a longitudinal data set. Linkages with Death Records has enabled much more detailed morbidity analysis including for Indigenous persons in both countries.

Even when linking variables are available, they are subject to error or linking data not being specified consistently. Therefore, the development of algorithms to maximise the accuracy of linking has become a very important job for methodologists.

8 Researcher Access

Among the major changes to ABS and SNZ legislation in the 1980s were legal provisions to enable them to release unidentifiable microdata. This provided constraints on access which some researchers found too limiting. Recognising inadequate use of microdata has high costs, in the early 2000s data laboratories were introduced where researchers could work in a safe setting with supervision and checks to ensure the confidentiality requirements were met. This was later extended to use of Remote Access Data Laboratories so that it was not necessary for researchers to visit the Statistical Offices.

It is also important that researchers have good quantitative knowledge of measurement and other errors so they can be taken into account in the analysis (see Biemer and Trewin, 1997).

A more recent development has been ABS and SNZ becoming custodians for linked data bases data sets (including links with some of its own data sets). Data laboratories are often the only way to access these valuable data sets.

9 Non-ABS and Non-SNZ Surveys

Increasingly, surveys are being conducted by other government agencies. Following the introduction of the 1975 Statistics Act in New Zealand, a survey control function was introduced. Every proposed survey by other agencies had to be submitted to the Minister of Statistics for approval of the sample and survey design. SNZ did the analysis necessary to make a recommendation to the Minister. At the request of Government, the ABS introduced a similar function in 1997. The emphasis was very much on ensuring these surveys were fit for purpose rather than a design that was up to the standard of the official statistical agency.

10 The Future

The role of the survey methodologist has changed massively over the 50 year period. It is no longer sufficient to be an expert in sample design. Mixed mode data collection techniques will become very prevalent generating new methodological challenges. Sample surveys are only one source of data for official statistics and are often used in combination with other data sources. For example, further use of administrative data and big data (e.g. scanner data and satellite images) can be expected, sometimes involving machine learning applications. This would include linked data sets.

Maintaining the quality of surveys to be fit for purpose will be a big challenge. Non-response is already a big issue and will become even more of a challenge as will the maintenance of good quality frameworks. Sample designs and methods that adjust for these types of deficiencies will grow in importance. The demand for data from researchers will increase requiring the development of methods to improve access with confidentiality protection arrangements that meet public scrutiny.

11 Acknowledgements

I would like to acknowledge the very significant contributions made by Siu-Ming Tam (formerly ABS) and Vince Galvin (SNZ) and Len Cook and Paul Maxwell formerly of SNZ.

References

ABS (2005), *Informing a Nation: The Evolution of the Australian Bureau of Statistics*, ABS, Canberra

Biemer, P and Trewin, D. (1997), A review of measurement error effects on the analysis of survey data, in *Survey Measurement and Process Quality* by L. Lyberg et al, pp 603-632, John Wiley & sons, New York.

Brewer, K., Early, L.J., and Joyce, S. (1972). Selecting several samples from the same population. *Aust. J. of Statist*., 14(3), 231-239

Dalenius, T. (1967). Nonsampling errors in census and sample surveys. *Report no 5 in the research project Errors in Surveys.* Stockholm University.

Linacre S., and Trewin D. (1993), Total Survey Design - Application to a Collection of the Construction Industry, *Journal of Official Statistics*, 9(3), 611-621

Trewin, D. (2001), The Importance of a Quality Culture, *Statistics Canada International Symposium Series: Proceedings* 2001.

The Survey Statistician, 2023, Vol. 88, 68-71.

The Contributions of Italian Statisticians to the Development of Survey Statistics

Luigi Biggeri

Emeritus Professor of Economic Statistics (University of Florence, Italy)
Past President of SIS, IASS (2003-2005) and ISTAT, luigi.biggeri@unifi.it

Abstract

In this short article I present, through flashes, a historical synthesis of the contribution of Italians to the development of survey statistics, from the starting stage up to the most recent years. The synthesis refers to both the practical and theoretical developments of the survey statistics, without claiming to be neither detailed nor complete.

Keywords: survey statistics, official and academic survey statisticians, Italy.

1 I start from the Middle Ages and the Grand Duchy of Tuscany to 1920

The first documented censuses show that in 1552 Cosimo I de' Medici organized the first population census of the then Florentine duchy. During the Grand Duchy of Tuscany various scientific societies of a general nature were also established, which included statistical studies. However, they were relatively short-lived due to their prohibition by the sub-sequent governments on the ground that the results of their research were subversive.

From the beginning of the 1800s and during the Italian Risorgimento, statistical knowledge and statistical activity developed a lot. Many scientists (philosophers, sociologists, economists, statisticians, demographers, and so on) devoted themselves to the establishment and management of statistical offices and subsequently participated very actively in the international congresses of statistics and demography.

In 1807 the Kingdom of Italy was one of the first European states to create a Statistical Office under the direction of the great statistician and philosopher Melchiorre Gioia. Then in 1826 a statistical society called the "Tuscan Society of Statistical Geography of Natural History" was established in the Grand Duchy of Tuscany. Both the Office and the Society were short-lived, for the reasons already mentioned above. Subsequently, starting from 1832, within a few years Statistical Offices (or similar) were created in the various states of Italy (Sicily, Subalpine Kingdom, Sardinia, Tuscany, Naples and the Papal State).

Finally, in 1861, when the Kingdom of Italy was formed as a unit, the Division of General Statistics was born (of which the first director was Pietro Maestri) who, assisted by a Superior Council of

Statistics (first president Cesare Correnti), which lasted until 1926 when the Central Institute of Statistics (Istat) was established.

It should be noted that in this period the statistical function had a close link with the government which is summarized in the formula "knowing to administer" or perhaps more precisely "knowing to govern".

Two annotations concerning this period seem important to me: the birth of the Civil State and the behavior and vision of the statisticians of the Risorgimento.

Civil Status office, as we still understand it currently today, was introduced in Tuscany during the period of French domination starting from 1808. Its birth was one of the great changes that marked the transition from the ancient regime to the contemporary age: with this institution indeed, civil institutions occupied land that had traditionally belonged to the Church.

For the statisticians of the Risorgimento, statistics were a fundamental tool of "civilization" which also served to evoke the "body" of that nation which they wanted to be the basis of a future independent State. Statistical knowledge, the collection and dissemination of statistics in the public sphere also constituted an indispensable tool for the transparent functioning of a power legitimized by popular consensus: the "discipline for democracy", in fact. In expressing these ideas some of them, in particular Maestri and Correnti, shared an almost utopian vision of statistics.

With regard to the participation and contribution of Italian statisticians to international congresses and international statistical bodies, let me remind you that the international congresses of statisticians began in 1853 in Brussels with the participation of around 150 scholars, 11 of whom were Italian. It was not until Florence was the capital of the kingdom of Italy that a congress of European statisticians (the sixth in the series) could be held there, attended by 632 Italian and 85 foreign statisticians. The contribution of the Italians has always been relevant, it is enough to refer to the performances of Maestri and Correnti and to the methodological contributions of Angelo Messedaglia.

As is well known, in 1885, during the celebration of the jubilee of the Statistical Society of London, the proposal was made to found the International Institute of Statistics (ISI). Luigi Bodio, who was head of the Italian statistical services which he had led to being among the best in Europe, assumed a leading role in the establishment of the ISI and in its development.

Bodio supported the proposal, but past experience led him to ask that the association should be free and independent of government decisions and that it should include the most eminent European and American statisticians, regardless of their nationality. Due to his well-known qualities as a scholar and his organizational skills, Bodio was elected general secretary of the ISI and remained in office for 20 years. In 1909he was elected President of the Institute by acclamation and was re-elected in the two successive elections, remaining in office until 1920, when he died. It is proof of the prestige Bodio enjoyed, but also of the level at which Italian statistics was assumed. In 1885, 13 Italians out of 106 nominations were nominated as members of the ISI. And in the elections of 1886, 23 Italians were elected out of the 154 elected. Furthermore, most of the first volumes of the ISI Bulletins were written in Italian.

2 The period from 1920 to the end of World War II

It was also characterized in Italy by important events in the field of the survey statistics, both from a methodological and institutional point of view. On the occasion of the ISI meetings, but not only, a dispute began between the conduct of only censuses versus the use of sampling to which Italian statisticians also contributed. As Leslie Kish wrote in the paper presented in 1995 at a meeting arranged by the Italian Statistical Society (SIS), "Neyman's 1934 paper marks a turning point for survey sampling...It was based on a 1929 paper of Gini and Galvani. In which the authors compared the results of the 1921 population census with the results of the same surveys carried out using a representative statistical sample of those surveyed. The results did not coincide and, perhaps, also

for this reason Istat, of which Gini was president, shelved the use of the sampling technique in the public statistics surveys.

At the beginning of the period, the Statistical Service was considered no more and no less than any bureaucratic body. The Fascist government had an interest in reorganizing the service conceived as a government service for the government and not a public service for the citizens, with a centralized arrangement. Therefore, in 1926 the government created the Central Institute of Statistics of the Kingdom of Italy, appointing Corrado Gini as president. In this way, and with a subsequent law of 1929, the main problems of the public production of statistics were resolved for the time, making available much statistical information necessary to carry out research in many fields of application (mainly demographic, economic and social) allowing many researchers whose results were also presented at international conferences, which confirmed the thesis of the originality and autonomy of Italian statistics.

3 Representative statistical sample

In 1944 Istat undertook, at the request and guidance of the Allied Commission, studies to carry out surveys using a representative statistical sample, to the satisfaction of its most representative statisticians including Benedetto Barberi, Lanfranco Maroi and Francesco Brambilla. From 1947 Istat intensified the study of the sampling technique and in 1948 established the "working group for sample surveys" within the Center for Research and Econometric Applications. Sample technique that was extended to many investigations in multiple fields of application. Many eminent statisticians collaborated with Istat, confirming the typical fruitful interaction in Italy between official statisticians and Academia. Among these we deem it appropriate to mention Marcello Boldrini who was also president of the ISI from 1959 to 1963.

In the following decades, the use of representative sample surveys developed more and more, and, in particular between the 1980s and 1990s, Istat made important advances in the field of surveys on families and individuals, launching "multipurpose" surveys. Progress in survey statistics has been continuous including the integrated system of registers and surveys; and in the use of Big data and citizen generated data and citizen science to produce official statistics.

4 In 1973 the IASS was founded and Italian official statisticians and academics also participated in its foundation

I remember that in the first years of life of the IASS over 100 Italians became members and a country representative was appointed. It was certainly an opportunity to organize the groups of Italians who intended to actively participate in the scientific meetings of the ISI and the IASS by proposing topics for the invited sessions and presenting papers. But also by carrying forward some important initiatives that I recall.

Under the impetus of survey statisticians, the Italian Statistical Society founded the SIS coordinating group on “Survey sampling methodology”, to contribute to the promotion and coordination of applied and methodological research on survey sampling.

The Survey Sampling Group decided to organize ITACOSM (Italian Conference on Survey Methodology) which is a bi-annual international conference, whose aim is promoting the scientific discussion on the developments of theory and application of survey sampling methodologies in the fields of economics, social and demographic sciences, of official statistics and in the studies on biological and environmental phenomena. The first edition of ITACOSM was held in Siena in 2009 and then the venue moved to Pisa in 2011, Milan in 2013, Rome in 2015, Bologna in 2017, Florence in 2019, Perugia in 2022, and Cosenza in 2023. In the first edition in Siena, the delegates were all Italians, apart from the 4 keynote speakers (Proff. Yves Tillé, Carl-Erik Sarndal, Yves Berger, and Tim Gregoire). In the last edition in Perugia, half of the 108 registered participants were not Italians. IASS has sponsored ITACOSM since the very beginning and the President or a member of the EC has always participated.

5 Concluding Remarks

Since the early 1800s, Italian statisticians have contributed a great deal to the development of survey statistics.

The latest developments concerned, as in many other countries, the use of Big data and citizen generated data and citizen science to produce official statistics. Istat will also develop these topics in the near future as stated during the webinar organized online on May 3, 2023, on "Big Data and new data sources to measure reality: A comparison on Trusted Smart Statistics". The introductory speech was carried out by Monica Pratesi, President of the IASS, and the webinar was followed by many Italian survey statisticians.

Acknowledgement. I thank Maria Giovanna Ranalli, Monica Pratesi and Luigi Fabbris for their support.

The Survey Statistician, 2023, Vol. 88, 72-74.

The IASS – 50 Years of Activity

Danny Pfeffermann[1,2]

[1] Department of Statistics, Hebrew University, Jerusalem, Israel
[2] Southampton Statistical Sciences Research Institute, University of Southampton, UK
IASS President 2013-2015
[1] msdanny@mail.huji.ac.il; [2] msdanny@soton.ac.uk

Abstract

In this short article I review the activities of the IASS, discuss the problem of its reducing membership and propose some possible directions for new developments in the future.

Keywords: Activities; Data science; IASS membership; Nonprobability samples

1 Introduction

The International Association of Survey Statisticians (IASS) is celebrating this year its 50 years jubilee, a milestone for celebration, reflecting on its big achievements so far, with a look to the future.

I believe that I joined the IASS already around 1980, shortly after completing my PhD. Why did I join? Probably because my PhD supervisor, the late Professor Gad Nathan, told me to do so. Mind you, I knew nothing about statistical organizations at that time. In 1985, I was elected as Fellow of the ISI. During the years 2001-2003, I chaired the programme committee of the IASS and 10 years later, during 2013-2015, I served as the President of the organization. So, it seems that I was quite an active member of the IASS in those years, which is probably why I was asked to contribute this short article that is due to appear in the July issue of *The Survey Statistician.*

2 The IASS mission and activities

The IASS was founded in 1973 "to promote the study and development of the theory and practice of sample surveys and censuses." How is this done? Mostly through the ISI meetings (the world statistical congress, WSC), held every two years; the IASS is one of the sections comprising the ISI. These meetings provide a forum for discussion of survey statistics. They include several specific sessions on recent advances and applications in survey and census methodologies. Other than the ISI meetings, the IASS helps to sponsor regional meetings and workshops devoted to specific aspects of surveys, and it publishes twice a year the journal *The Survey Statistician*, which is devoted to survey sampling and censuses and is distributed to all the IASS members. I always found the country reports in the journal, on activities performed by their national statistical institutes to be particularly useful.

I assume that other authors will cover the history and big achievements of the IASS since its foundation. Hence, in what follows, I like to discuss briefly some possible directions for new developments in the future.

3 Membership in the IASS

The IASS needs to increase its membership quite extensively. Here is a contingency table of the registry in February 2023, with the cells defined by continent of living and gender.

Region	Female	Male	Total
South/Central America	11	18	29
Oceania	5	18	23
North America	21	71	92
Europe	54	79	133
Asia	8	31	39
Africa	8	21	29
Total	107	238	345

When I started my presidency in 2013, the IASS had 445 individual members and 23 institutional members. Now we have 21 institutional members, but the numbers of members in the various categories have dropped quite significantly. The registration profile is also gloomy when looking at the distribution of the IASS members by age. Based on 284 members with known ages, out of 205 males, 106 are older than 65, 81 are at the age of 41-65 and only 18 are younger than 41. Out of 79 females, 30 are older than 65, 38 are at the age of 41-65, and only 11 are younger than 41. Thus, the majority of our members are in the older age groups.

I don't know how many survey statisticians there are worldwide, but it is obvious that only very few of them choose to become members of the IASS. We need a larger membership for a number of reasons. First, since the outcome of our work affects directly so many applications and decision-makings, we should have broad representation from all the various areas. A broad representation will stimulate more joint research and collaboration efforts. A larger membership will of course allow for greater financial flexibility, which in turn will facilitate more diverse activities.

The IASS should continue its efforts to encourage more survey statisticians to join the organization. No reason why only 35% of its members are females, and why continents other than North America and Europe are so poorly represented. The fact that young survey statisticians tend not to join the IASS is particularly worrying. If this trend continues, the prospect of the IASS to continue its mission and activities is in real risk.

PhD supervisors should encourage their students to join us. This is a first step to increase young member's registry. Social networks can be used as another platform to promote the IASS and motivate registration. A reviewer of this article proposed allocating a special section in *The Survey Statistician* for young survey statisticians from developing countries to publish reviewed scientific papers (possibly with discussion by senior members). These are all just examples of what can possibly be done to increase our membership.

4 Nonprobability sampling

In recent years there is growing research on the use of nonprobability samples. Such samples are not representative, and they require different kinds of inference on finite population parameters of interest, but they have their merits in terms of costs, logistic and possible reduction in nonresponse. See e.g., Beaumont and Rao (2021), published in *The Survey Statistician*. I think that the IASS should pay increased attention to this kind of samples, organize conferences and workshops and

encourage publications in *The Survey Statistician*, with emphasis on practical applications. IASS members and survey statisticians in general will undoubtedly benefit from this activity.

5 Survey sampling and data science

The last two decades have witnessed the rapid growth of data science. One of the facets of this growth is that there are people agitating that the existence of all sorts of “big data”, and the new advanced technologies that have been developed to handle them, will soon replace the use of sample surveys. In an article I published in 2015, I overviewed the problems with the use of big data for the production of official statistics but clearly, when such data sources are available, accessible and timely, they cannot and should not be ignored. I have no diploma in prophecy, but my own view is that survey samples will always be needed, at least in the foreseen future, so our profession is secured for many years to come.

I think that the IASS should play a leading role in the promotion of new theories and practices for the integration of classical survey sampling theory with data science, with the ultimate goal of improving the data and subsequent evidence-based decisions upon data obtained from only one of the sources. In simple words, how to benefit from both worlds.

6 Possible merge with the IAOS

Finally, I like to raise the possibility of merging with the International Association for Official Statistics (IAOS), another section of the ISI. I already raised this idea while I was president, some executive members of the IASS supported the idea, other objected, and IAOS executives with whom I discussed it were not very enthusiastic about it either, so I did not push it any further. I might be wrong, but I personally think that both organizations will benefit from such a merge, and if only because it will make the merged section the biggest or one of the biggest sections of the ISI. I propose therefore to examine the pros and cons of such a merge and if found worthy, negotiate it with IAOS representatives. The reviewer of this article proposed improving the interaction with the IAOS by organizing a joint conference. This is a nice idea, which could form a first step in a possible future merge.

7 Concluding remark

I think that the IASS is a very worthy and much needed organization, and I wish it to expand both in size and in its activities in the coming years.

References

Beaumont, J.F. and Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.

Pfeffermann, D. (2015). Methodological Issues and Challenges in the Production of Official Statistics. *The Journal of Survey Statistics and Methodology* (JSSAM), 3, 425–483.

The Survey Statistician, 2023, Vol. 88, 75–85.

Ask the Experts

Doubly and Multiply Robust Procedures for Missing Survey Data

Sixia Chen[1] **and David Haziza**[2]

[1]University of Oklahoma Health Sciences Center, U. S. A., Sixia-Chen@ouhsc.edu
[2]University of Ottawa, Canada, dhaziza@uottawa.ca

Abstract

Missing data are ubiquitous in surveys. Unadjusted estimators may be substantially biased as the set of respondents is generally a non-representative subset of the original sample. Item nonresponse, which most often treated by some form of imputation, may eliminate the potential nonresponse bias if the first moment of the imputation model is correctly specified. However, the resulting estimators may suffer from appreciable bias if the model is misspecified. In this paper, we review doubly and multiply robust imputation procedures. These procedures, that combine multiple nonresponse and imputation models, may provide some protection against model miss-specification.

Keywords: Deterministic imputation; imputation model; missing data; nonresponse model; random imputation.

1 Introduction

As response rates have declined sharply over the past two decades, reducing the nonresponse bias has become an important issue for survey statisticians. Unadjusted estimators tend to exhibit significant bias as the behaviour of the respondents typically differ from that of the nonrespondents. In the absence of non-sampling errors, bias is generally not an issue as customary point estimators (e.g., the Horvitz-Thompson estimator and calibration estimators) are design-unbiased or asymptotically design-unbiased. In this ideal setup, survey statisticians would typically opt for an estimator that exhibit a small variance. In the presence of missing values, bias is the main issue. Reducing the nonresponse bias as much as possible requires the availability of powerful auxiliary information. The nonresponse treatment stage involves a modeling task that puts an additional burden on the survey statistician's shoulders, as heavily biased estimators will lead to misleading inferences. In this article, we focus on item nonresponse, most often treated by some form of imputation, the first step of which is to postulate an imputation model describing the relationship between the survey variable Y requiring imputation and a set of fully observed variables $\mathbf{x}$. The modeling task involves the selection of variables that are predictive of the survey variable Y, and the specification of a suitable model for the relationship between Y and $\mathbf{x}$.

If the data are Missing At Random (MAR, Rubin, 1976), the estimators obtained after imputation will exhibit a negligible bias if the first moment of the imputation model is correctly specified. Otherwise, the bias may be significant.

This led researchers to develop imputation/estimation procedures that provide some robustness against model misspecification. This is where doubly and multiply robust procedures come into play. The concept of double robustness in the context of missing data is attributed to Robins et al. (1994) for their article published in the *Journal of the American Statistical Association*. However, it is worth pointing out that, in the same issue of the journal, Kott (1994) also independently introduced the concept of double robustness for missing survey data. In fact, the origin of doubly robust procedures can be traced back in the 1970s to the survey sampling literature on the generalized regression (GREG) estimator and, more generally, on model-assisted estimation procedures; see, e. g., Cassel et al. (1976), Särndal (1980), Särndal and Wright (1984) and Särndal et al. (1992). For instance, the GREG estimator of a population total, whose construction is assisted by a linear regression model, possesses the double robustness property: it is model-unbiased if the model is correctly specified, but remains design-consistent even if the model is misspecified.

In the context of missing data, doubly robust procedures combine two models. The first, called the imputation model or the outcome regression model, describes the relationship between the survey variable Y and a vector of explanatory variables. The second, called the nonresponse model or the propensity score model, describes the relationship between the response indicator R and a set of explanatory variables. If the data are MAR, doubly robust procedures remain consistent if either the nonresponse model or the imputation model is correctly specified. This is an attractive property closely related to the philosophy of model-assisted estimation in survey sampling. The literature on doubly robust procedures is very rich; see e. g., Robins et al. (1994), Scharfstein et al. (1999), Bang and Robins (2005), Haziza and Rao (2006), Tan (2006), Kang and Schafer (2008), Cao et al. (2009), Kim and Haziza (2014), Boistard et al. (2016) and Seaman and Vansteelandt (2018). However, doubly robust procedures have been criticized because the resulting estimators have been shown to exhibit poor performances if both models are (slightly) misspecified; e. g., see Kang and Schafer (2008).

Multiple robustness can be viewed as an extension of the concept of double robustness. Instead of postulating a single imputation model and a single nonresponse model, one rather postulates multiple imputation models and/or multiple nonresponse models. Each model may be based on a different link function and a different set of explanatory variables. An imputation procedure is called multiple robust if the resulting estimator remains consistent if anyone of the postulated models is correctly specified; see e. g., Han and Wang (2013), Chan and Yam (2014), Han, (2014; 2016), Chen and Haziza (2017), Duan and Yin (2017), Chen and Haziza (2019) and Han et al. (2019). Therefore, these procedures provide some type of insurance against a single misspecified model. Multiply robust procedures belong to the class of ensemble or aggregation procedures as the goal is to construct a set of imputed values that can be viewed as a suitable aggregate of the information contained in the multiple models.

2 Doubly robust procedures

Consider a finite population U of size N. Our goal is to estimate the population total of a survey variable Y, $t_y = \sum_{k \in U} y_k$. A sample S, of size n, is selected from U according to a sampling design with first-order inclusion probabilities π_k. Let R_k be a response indicator attached to unit k such that $R_k = 1$ if Y is observed and $R_k = 0$, otherwise. Let S_r be the set of respondents to item Y, of size n_r; that is the subset of S for which $R_k = 1$, and let $S_m = S - S_r$ be the set of nonrespondents. We assume that the data are MAR; i.e., the conditional distribution of Y given $\mathbf{x}$ observed among the respondents

is identical to the conditional distribution of Y given $\mathbf{x}$ observed among the nonrespondents, where $\mathbf{x}$ denotes a vector of fully observed variables. Under MAR, one can safely estimate the relationship between Y and $\mathbf{x}$ from the set of respondents S_r, and "extrapolate" from this relationship to construct a set of imputed values.

We assume that the relationship between R and $\mathbf{x}$ can be described by the following nonresponse model:

$$\mathbb{E}\left(R_k \mid \mathbf{x}_k\right) = p(\mathbf{x}_k; \boldsymbol{\alpha}), \tag{1}$$

where $p(\cdot; \boldsymbol{\alpha})$ is a prespecifed function and $\boldsymbol{\alpha}$ is a vector of unknown coefficients. We assume that the relationship between Y and $\mathbf{x}$ can be described by the following imputation model:

$$\mathbb{E}\left(y_k \mid \mathbf{x}_k\right) = m(\mathbf{x}_k; \boldsymbol{\beta}), \tag{2}$$

where $m(\cdot; \boldsymbol{\beta})$ is a prespecifed function and $\boldsymbol{\beta}$ is a vector of unknown coefficients. For simplicity, we assume that the first component of the $\mathbf{x}$-vector is 1 for all k and that $\mathbb{V}(y_k \mid \mathbf{x}_k) = \sigma^2$.

Doubly robust imputation can be described as follows:

(i) We obtain an estimator, $\widehat{\boldsymbol{\alpha}}$, of $\boldsymbol{\alpha}$ by solving, for example, the following estimating equations:

$$\sum_{k \in S} \pi_k^{-1} \frac{R_k - p(\mathbf{x}_k; \alpha)}{p(\mathbf{x}_k; \alpha)\left\{1 - p(\mathbf{x}_k; \alpha)\right\}} \frac{\partial p(\mathbf{x}_k; \alpha)}{\partial \alpha} = \mathbf{0}. \tag{3}$$

In the case of a logistic regression model, $p(\mathbf{x}_k; \alpha) = \exp\left(\mathbf{x}_k^\top \alpha\right)/\{1 + \exp\left(\mathbf{x}_k^\top \alpha\right)\}$, Expression (3) reduces to the customary estimating equations

$$\sum_{k \in S} \pi_k^{-1} \{R_k - p(\mathbf{x}_k; \alpha)\} \mathbf{x}_k = \mathbf{0}.$$

Let $p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})$ denote the resulting estimated response probability attached to unit k.

(ii) We obtain an estimator, $\widehat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$, by solving the estimating equations

$$\sum_{k \in S_r} \pi_k^{-1} \frac{1 - p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} \left\{y_k - m(\mathbf{x}_k; \boldsymbol{\beta})\right\} \mathbf{x}_k = \mathbf{0}.$$

Let $m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}})$ denote the predicted value attached to unit k. If $m(\mathbf{x}_k; \boldsymbol{\beta}) = \mathbf{x}_k^\top \boldsymbol{\beta}$, the estimator $\widehat{\boldsymbol{\beta}}$ reduces to the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{k \in S_r} \pi_k^{-1} \frac{1 - p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_r} \pi_k^{-1} \frac{1 - p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} \mathbf{x}_k y_k.$$

(iii) The imputed y-value for $k \in S_m$ is given by

$$y_k^* = m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}).$$

It follows that an estimator of t_y after imputation is given by

$$\widehat{t}_{y,DR} = \sum_{k \in S_r} \pi_k^{-1} y_k + \sum_{k \in S_m} \pi_k^{-1} m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}).$$

The estimator $\widehat{t}_{y,DR}$ is doubly robust in the sense that it remains consistent if either the nonresponse

model (1) or the imputation model (2) is correctly specified. To see why this is the case, note that $\widehat{t}_{y,DR}$ can be expressed in the following two forms:

$$\widehat{t}_{y,DR} = \widehat{t}_{y,F} - \sum_{k \in S_m} \pi_k^{-1} \left\{ y_k - m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}) \right\} \tag{4}$$

$$= \widehat{t}_{y,PSA} - \sum_{k \in S} \pi_k^{-1} \left(\frac{R_k}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} - 1 \right) m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}), \tag{5}$$

where $\widehat{t}_{y,F} = \sum_{k \in S} \pi_k^{-1} y_k$ denotes the (unfeasible) full sample estimator of t_y, and

$$\widehat{t}_{y,PSA} = \sum_{k \in S_r} \pi_k^{-1} \frac{y_k}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})}$$

corresponds to the propensity score estimator of t_y. If the imputation model is correctly specified, we have $\mathbb{E}\left\{ y_k - m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}) \right\} \approx 0$ and the second term on the right hand-side of (4) is, on average, approximately equal to 0. We are left with the full sample estimator $\widehat{t}_{y,F}$, which is consistent for t_y. Next, if the nonresponse model is correctly specified, we have $\mathbb{E}\left(\frac{R_k}{p(\mathbf{x}_k; \widehat{\boldsymbol{\alpha}})} - 1 \right) \approx 0$, and the second term on the right hand-side of (5) is approximately equal to 0. We are left with the propensity score adjusted estimator $\widehat{t}_{y,PSA}$, which is consistent for t_y since the nonresponse model is correctly specified.

The imputed values (2) belong to the class of deterministic imputation procedures. We can define a doubly robust random version (Haziza and Rao, 2006) as follows:

$$y_k^* = m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}) + e_k^*,$$

where e_k^* is selected at random with replacement from the set of standardized residuals observed among the respondents. That is,

$$e_k^* = e_\ell, \quad \ell \in S_r, \text{ with probability } \frac{\phi_\ell}{\sum_{t \in S_r} \phi_t},$$

where

$$\phi_\ell = \pi_\ell^{-1} \frac{1 - p(\mathbf{x}_\ell; \widehat{\boldsymbol{\alpha}})}{p(\mathbf{x}_\ell; \widehat{\boldsymbol{\alpha}})} \text{ and } e_\ell = y_k - m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}}).$$

Donor imputation is often used in practice as the imputed values are necessarily eligible values observed among the respondents, which is often deemed a desirable feature when, for instance, the variable requiring imputation is categorical. A doubly robust procedure random hot-deck imputation procedure can be described as follows. We first obtain the scores $m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}})$ and $p(\mathbf{x}_\ell; \widehat{\boldsymbol{\alpha}})$. Then, using a classification algorithm (e. g., the K-means algorithm), we create C homogeneous cells with respect to both $m(\mathbf{x}_k; \widehat{\boldsymbol{\beta}})$ and $p(\mathbf{x}_\ell; \widehat{\boldsymbol{\alpha}})$. Within each cell, a missing value is imputed using the y-value of a donor selected at random and with replacement from the set of donors belonging to the same cell.

3 Multiply robust imputation procedures

We consider two classes of parametric models: The first, $\mathcal{M}_1$, consists of H imputation models; i.e., $\mathcal{M}_1 = \left\{ m^{(h)}(\mathbf{x}_k^{(h)}; \boldsymbol{\beta}^{(h)}), h = 1, 2, \ldots, H \right\}$ and, the second, $\mathcal{M}_2$, consists of J nonresponse models; i.e., $\mathcal{M}_2 = \left\{ p^{(j)}(\mathbf{x}_k^{(j)}; \boldsymbol{\alpha}^{(j)}), j = 1, 2, \ldots, J \right\}$. The models in both classes may be based on different functionals and/or different sets of explanatory variables. Two methods for constructing a set of imputed values are: (i) Aggregation through calibration (Han and Wang, 2013; Han, 2014, 2016;

Chen and Haziza, 2017) and (ii) aggregation through refitting (Duan and Ying, 2017, Chen and Haziza, 2019). Although we focus on deterministic multiply robust imputation in the sequel, a random version as well as a random hot-deck version can be obtained using approaches similar to those described in Section 2 for doubly robust imputation.

Regardless of the aggregation approach, the first step is to fit each of the $H+J$ models in classes $\mathcal{M}_1$ and $\mathcal{M}_2$. For each $k \in S$, we can then construct 2 vectors: (i) The vector $\widehat{\mathbf{m}}$, of size H, given by $\widehat{\mathbf{m}} = \left(m^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)}), \cdots, m^{(H)}(\mathbf{x}_k^{(H)}; \widehat{\boldsymbol{\beta}}^{(H)})\right)^\top$. (ii) The vector $\widehat{\mathbf{p}}$, of size J, given by $\widehat{\mathbf{p}} = \left(p^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\alpha}}^{(1)}), \cdots, p^{(J)}(\mathbf{x}_k^{(J)}; \widehat{\boldsymbol{\alpha}}^{(J)})\right)^\top$. The estimators $\widehat{\boldsymbol{\beta}}^{(1)}, \cdots, \widehat{\boldsymbol{\beta}}^{(H)}$, $\widehat{\boldsymbol{\alpha}}^{(1)}, \cdots, \widehat{\boldsymbol{\alpha}}^{(J)}$, denote suitable estimators (e. g., maximum likelihood estimators or weighted least squares estimators) for their corresponding parameters.

3.1 Aggregation through calibration

Aggregation through calibration proceeds as follows:

(i) We start by obtaining a calibrated weighting system $\{w_1, w_2, \ldots, w_{n_r}\}$, where $w_k, k \in S_r$, is a scalar summary of the information contained in the H imputation models and the J nonresponse models. For simplicity, we consider the generalized chi-square distance (Deville and Särndal, 1992). We seek a weighting system $\{w_1, w_2, \ldots, w_{n_r}\}$ such that

$$\sum_{k \in S_r} \pi_k (w_k - \pi_k^{-1})^2/2$$

is minimum subject to the $H+J+1$ calibration constraints

$$\sum_{k \in S_r} w_k = \sum_{k \in S} \pi_k^{-1},$$

$$\sum_{k \in S_r} w_k m^{(h)}(\mathbf{x}_k^{(h)}; \widehat{\boldsymbol{\beta}}^{(h)}) = \sum_{k \in S} \pi_k^{-1} m^{(h)}(\mathbf{x}_k^{(h)}; \widehat{\boldsymbol{\beta}}^{(h)}), h = 1, \ldots, H,$$

and

$$\sum_{k \in S_r} w_k \frac{1}{p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)})} = \sum_{k \in S} \pi_k^{-1} \frac{1}{p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)})}, j = 1, \ldots J.$$

The resulting weights w_k are given by

$$w_k = \pi_k^{-1} \times (1 + \widehat{\boldsymbol{\lambda}}^\top \mathbf{h}_k), \tag{6}$$

where $\widehat{\boldsymbol{\lambda}}$ is a vector of estimated Lagrange multipliers of size $H+J+1$ and

$$\mathbf{h}_k = (1, \mathbf{h}_{1k}^\top, \mathbf{h}_{2k}^\top)^\top$$

with

$$\mathbf{h}_{1k} = \left(m^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)}) - \overline{m}^{(1)}, \ldots, m^{(H)}(\mathbf{x}_k^{(H)}; \widehat{\boldsymbol{\beta}}^{(H)}) - \overline{m}^{(H)}\right)^\top$$

and

$$\mathbf{h}_{2k} = \left(p^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\alpha}}^{(1)}) - \overline{p}^{(1)}, \ldots, p^{(J)}(\mathbf{x}_k^{(J)}; \widehat{\boldsymbol{\alpha}}^{(J)}) - \overline{p}^{(J)}\right)^\top,$$

with $\overline{m}^{(h)} = \sum_{k \in S} \pi_k^{-1} m^{(h)}(\mathbf{x}_k^{(h)}; \widehat{\boldsymbol{\beta}}^{(h)}) / \sum_{k \in S} \pi_k^{-1}$ and $\overline{p}^{(j)} = \sum_{k \in S} \pi_k^{-1} p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)})^{-1} / \sum_{k \in S} \pi_k^{-1}$. Distance functions other than the generalized chi-square distance can be used; see Chen and Haziza (2017). To better understand the rationale behind this type of aggregation, we define the standardized version of $\widehat{\boldsymbol{\lambda}}$ as $\widehat{\boldsymbol{\lambda}}^2 / \widehat{\boldsymbol{\lambda}}^\top \widehat{\boldsymbol{\lambda}}$, where $\widehat{\boldsymbol{\lambda}}^2 \equiv (\widehat{\lambda}_0^2, \cdots, \widehat{\lambda}_{J+H}^2)^\top$. It follows that the

standardized version of the term $\widehat{\boldsymbol{\lambda}}^\top \mathbf{h}_k$ on the right hand-side of (6) can be expressed as

$$\begin{aligned}\frac{\widehat{\boldsymbol{\lambda}}^2}{\widehat{\boldsymbol{\lambda}}^\top\widehat{\boldsymbol{\lambda}}}\mathbf{h}_k &= \delta_0 + \delta_1\left\{m^{(h)}(\mathbf{x}_k^{(h)};\widehat{\boldsymbol{\beta}}^{(H)})\right\} + \cdots + \delta_H\left\{m^{(H)}(\mathbf{x}_k^{(H)};\widehat{\boldsymbol{\beta}}^{(H)})\right\} \\ &\quad + \cdots + \delta_{H+1}\left\{p^{(1)}(\mathbf{x}_k^{(1)};\widehat{\boldsymbol{\alpha}}^{(1)}) - \overline{p}^{(1)}\right\} + \cdots + \delta_{H+J}\left\{p^{(J)}(\mathbf{x}_k^{(J)};\widehat{\boldsymbol{\alpha}}^{(J)}) - \overline{p}^{(J)}\right\}, \quad (7)\end{aligned}$$

where $\delta_h = \widehat{\lambda}_h^2/\sum_{h=0}^{J+H}\widehat{\lambda}_h^2, h = 0,\ldots,H+J$. The aggregation weights δ_h sum to 1, which makes (7) a convex combination of the individual predictions obtained from each of the $H+J$ models. Therefore, the calibration weight in (6) can be viewed as an aggregate score or a scalar summary of the information contained in the $H+J$ models. If one of the models in either class is correctly specified, we expect the associated aggregation weight δ_h to be large and the other weights to be small.

(ii) The imputed values y_k^* are obtained by fitting a weighted linear regression with Y as the dependent variable, and $\mathbf{h}_k$ as the vector of explanatory variables. The regression weights are given by $\phi_k = \pi_k^{-1}\left\{(1+\widehat{\lambda}_r^\top\mathbf{h}_k) - 1\right\}$. This leads to

$$y_k^* = \mathbf{h}_k^\top\widehat{\boldsymbol{\gamma}}, \quad k \in S_m,$$

where

$$\widehat{\boldsymbol{\gamma}} = \left(\sum_{k\in S_r}\phi_k\mathbf{h}_k\mathbf{h}_k^\top\right)^{-1}\left(\sum_{k\in S_r}\phi_k\mathbf{h}_k y_k\right).$$

It follows that an estimator of t_y after imputation is given by

$$\widehat{t}_{y,MR} = \sum_{k\in S_r}\pi_k^{-1}y_k + \sum_{k\in S_m}\pi_k^{-1}\mathbf{h}_k^\top\widehat{\boldsymbol{\gamma}}. \quad (8)$$

The estimator $\widehat{t}_{y,MR}$ given by (8) is multiply robust in the sense that it remains consistent if all but one of the $H+J$ models are misspecified.

3.2 Aggregation through refitting

Aggregation through refitting proceeds as follows:

(i) Fit a linear regression model based on $k \in S_r$ with Y as the dependent variable and $\widehat{\mathbf{m}}$ as the vector of explanatory variables. The vector of estimated regression coefficients is denoted as $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1,\cdots,\widehat{\eta}_H)^\top$. Define the standardized version of $\widehat{\boldsymbol{\eta}}$ as $\widehat{\boldsymbol{\eta}}^2/\widehat{\boldsymbol{\eta}}^\top\widehat{\boldsymbol{\eta}}$, where $\widehat{\boldsymbol{\eta}}^2 \equiv (\widehat{\eta}_1^2,\cdots,\widehat{\eta}_K^2)^\top$. The aggregate or compressed score attached to unit $k \in S_r$ is defined as

$$\widehat{m}_k = \sum_{h=1}^{H}\omega_h m^{(h)}(\mathbf{x}_k^{(h)};\widehat{\boldsymbol{\beta}}^{(h)}), \quad (9)$$

where $\omega_h = \widehat{\eta}_h^2/\sum_{h=1}^{H}\widehat{\eta}_h^2$. The aggregation weights ω_h sum to 1. Therefore, the aggregate score $\widehat{m}_k, k \in S_r$, can be viewed as a convex combination of the individual predictions obtained from each of the H imputation models.

(ii) Fit a linear regression model based on $k \in S$ with R as the dependent variable and $\widehat{\mathbf{p}}$ as the vector of explanatory variables. The vector of estimated regression coefficients is denoted as $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1,\cdots,\widehat{\tau}_J)^\top$. Define the standardized version of $\widehat{\boldsymbol{\tau}}$ as $\widehat{\boldsymbol{\tau}}^2/\widehat{\boldsymbol{\tau}}^\top\widehat{\boldsymbol{\tau}}$, where $\widehat{\boldsymbol{\tau}}^2 \equiv (\widehat{\tau}_1^2,\cdots,\widehat{\tau}_J^2)^\top$.

The aggregate score attached to unit $k \in S$ is defined as

$$\widehat{p}_k = \sum_{j=1}^{J} \phi_j p^{(j)}(\mathbf{x}_k^{(j)}; \widehat{\boldsymbol{\alpha}}^{(j)}), \tag{10}$$

where $\phi_j = \widehat{\tau}_j^2 / \sum_{j=1}^{J} \widehat{\tau}_j^2$. The aggregation weights ϕ_j sum to 1. Therefore, the aggregate score $\widehat{p}_k, k \in S$, can be viewed as a convex combination of the individual predictions obtained from each of the J nonresponse models. This ensures that the aggregate score $\widehat{p}_k$ lies between 0 and 1.

(iii) The imputed values $y_k^*, k \in S_m$ is given by

$$y_k^* = \mathbf{h}_k^\top \widehat{\boldsymbol{\gamma}}^*, \quad k \in S_m,$$

where $\mathbf{h}_k = (1, \widehat{m}_k)^\top$ and

$$\widehat{\boldsymbol{\gamma}}^* = \left(\sum_{k \in S_r} \pi_k^{-1} \frac{1 - \widehat{p}_k}{\widehat{p}_k} \mathbf{h}_k \mathbf{h}_k^\top \right)^{-1} \sum_{k \in S_r} \pi_k^{-1} \frac{1 - \widehat{p}_k}{\widehat{p}_k} \mathbf{h}_k y_k.$$

It follows that an estimator of t_y after imputation is given by

$$\widehat{t}_{y,MR} = \sum_{k \in S_r} \pi_k^{-1} y_k + \sum_{k \in S_m} \pi_k^{-1} \mathbf{h}_k^\top \widehat{\boldsymbol{\gamma}}^*. \tag{11}$$

The estimator $\widehat{t}_{y,MR}$ given by (11) is multiply robust in the sense that it remains consistent if all but one of the $H + J$ models are misspecified. This can be explained as follows: if the class $\mathcal{M}_1$ contains the true imputation model, say $m^{(1)}(\mathbf{x}_k^{(1)}; \boldsymbol{\beta}^{(1)})$, we expect the aggregation weight ω_1 associated with the prediction $m^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\beta}}^{(1)})$ to be close to 1, and the other aggregation weights $\omega_h, h = 2, \ldots, H$, to be close to 0. Similarly, if the class $\mathcal{M}_2$ contains the true nonresponse model, say $p^{(1)}(\mathbf{x}_k^{(1)}; \boldsymbol{\alpha}^{(1)})$, we expect the aggregation weight ϕ_1 associated with the prediction $p^{(1)}(\mathbf{x}_k^{(1)}; \widehat{\boldsymbol{\alpha}}^{(1)})$ to be close to 1, and the other aggregation weights $\phi_j, j = 2, \ldots, J$, to be close to 0. This is illustrated in Section 4.1

4 Empirical investigation

In this section, we present two limited empirical investigation: the first examines the distribution of the weights involved in the aggregation procedures, whereas the second compares the performance of several estimators in terms of bias and efficiency in the case of data Not Missing At Random (NMAR).

4.1 Distribution of the aggregation weights

We generated $1,000$ finite populations, each of size $N = 10,000$. In each population, we generated 4 explanatory variables X_1-X_4 independently from a standard normal distribution. The y-values were then generated according to $y_k = 1 + x_{1k} + x_{2k} + x_{3k} + x_{4k} + \epsilon_k$, $k = 1, 2, \ldots, N$, where the ϵ_k's were independently generated from a standard normal distribution. In each population, we selected a sample S, of size $n = 800$, according to inclusion probability proportional-to-size systematic sampling with size variable $s_k = 0.5 v_k + 1$, where v_k was generated from a standard chi-square distribution with one degree of freedom. In each sample the response indicators R_k were independently generated from a Bernoulli distribution with probability $logit(p(\mathbf{x}_k; \alpha)) = 0.5 + x_{1k} + x_{2k} + x_{3k} + x_{4k}$. This led to an overall response rate of about 56%.

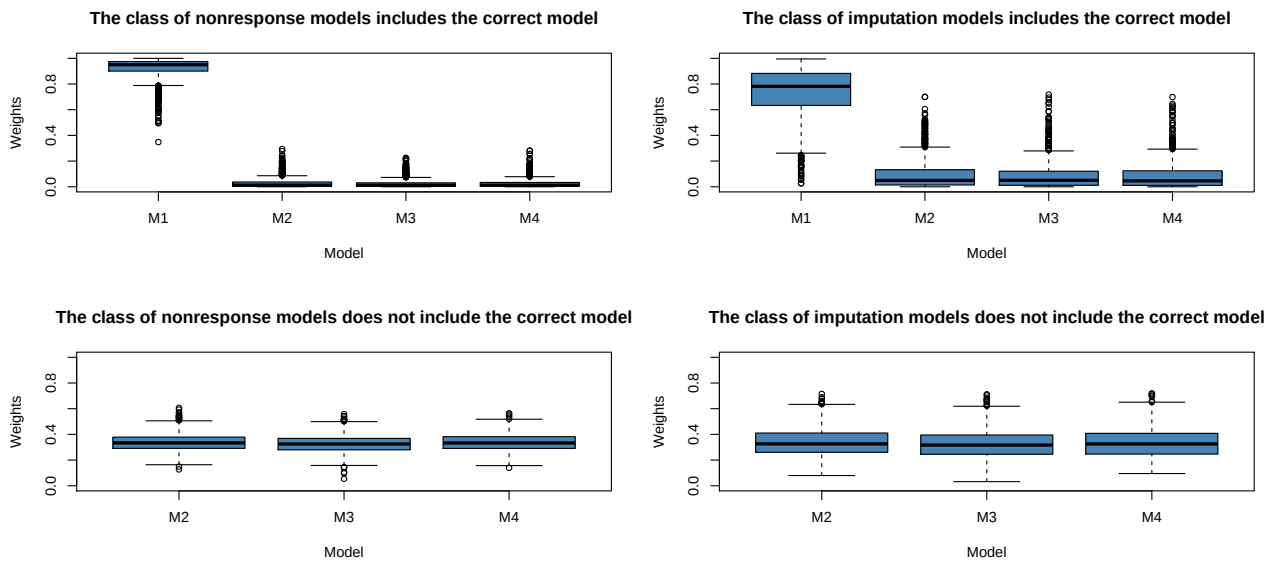


Figure 1: Distribution of the aggregation weights δ_h for the aggregation through calibration

The class of imputation models, $\mathcal{M}_1$, consisted of 4 imputation models: the correct imputation model based on X_1-$X4$ (M1); an incorrect linear regression model based on X_2 only (M2); an incorrect linear regression model based on X_3 only (M3); an incorrect linear regression model based on x_4 only (M4). The class of nonresponse models, $\mathcal{M}_2$, also consisted of 4 nonresponse models: the correct nonresponse model based on X_1-X_4 (M1); an incorrect logistic regression model based on X_2 only (M2); an incorrect logistic regression model based on X_3 only (M3); an incorrect logistic regression model based X_4 only (M4). In each class, each of the 4 models was fitted and the predictions were aggregated using both aggregation through calibration (see Section 3.1) and aggregation through refitting (see Section 3.2). For the aggregation through calibration procedure, we computed, in each sample, the aggregation weights δ_h in (7). For the aggregation through refitting procedure, we computed, in each sample, the aggregation weights ω_h in (9) and the aggregation weights ϕ_j in (10).

Figure 1 and Figure 2 display the distribution of the aggregation weights for the aggregation through calibration and the aggregation through refitting, respectively. When the class $\mathcal{M}_1$ or $\mathcal{M}_2$ included the correct model (M1), we note that both aggregation procedures put most of the weight on the correct model (M1). The incorrect models received a much smaller weight. This suggests that both aggregation procedures perform some type of implicit of model selection. When the classes $\mathcal{M}_1$ and $\mathcal{M}_2$ did not include the correct model, each of the models (M2)-(M4) contributed almost equally to the resulting predictions. In other words, a prediction was essentially defined as the average of the predictions generated from each of the models.

4.2 Data Not Missing At Random

We evaluated the performance of several estimators in terms of bias and efficiency in the context of NMAR. We used a simulation setup similar to that of Chen and Haziza (2021). We generated $B = 1,000$ finite populations, each of size $N = 10,000$. In each population, we generated 4 explanatory variables X_1-X_4 independently from a standard normal distribution. The y-values were then generated according to $y_k = 210 + 27.4x_{1k} + 13.7(x_{2k} + x_{3k} + x_{4k}) + \epsilon_k,\ k = 1, 2, \ldots, N$, where the ϵ_k's were independently generated from a standard normal distribution. In each population, we selected a sample S, of size $n = 800$, according to the same sampling design described in Section 4.1. In each sample, the response indicators R_k were independently generated from a

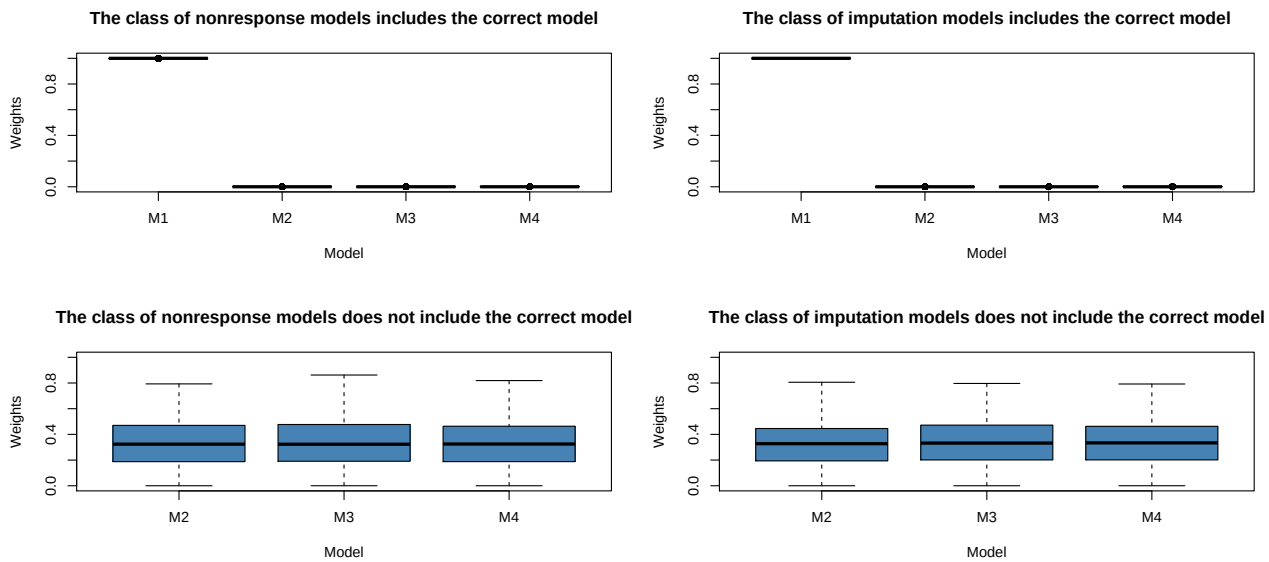


Figure 2: Distribution of the aggregation weights ω_h and ϕ_j for the aggregation through refitting

Bernoulli distribution with probability $logit(p_k(\alpha)) = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 y_i^{1/4}$ with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (-2.4, -1, 0.5, -0.25, -0.1, 0.5)$, which corresponds to a response rate of about 40% and $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (-1.3, -1, 0.5, -0.25, -0.1, 0.5)$, which corresponds to a response rate of about 60%. We assumed that only the transformed variables Z_1-Z_4 of X_1-X_4, were available to the imputer, where $z_{1k} = \exp(x_{1k}/2)$, $z_{2k} = x_{2k}\{1 + \exp(x_{1k})\}^{-1} + 10$, $z_{3k} = (x_{1k}x_{3k}/25 + 0.6)^3$, and $z_{4k} = (x_{2k} + x_{4k} + 20)^2$. In other words, the imputer did not have access to the variables X_1-X_4. Kang and Schafer (2008) used a similar setup.

We were interested in estimating the finite population mean of Y. In each sample, we computed the following estimators of the mean: (1). The complete data estimator that corresponds to the weighted mean of the sample y-values (COM); (2). The estimator based on doubly robust imputation, where the nonresponse model was a logistic regression model based on Z_1-Z_4 and the imputation model was a linear regression model based on Z_1-Z_4 (DR); (3). The estimator based on nearest-neighbor imputation using Z_1-Z_4 as matching variables (NN); (4). Five multiply robust estimators based on aggregation through calibration with the pseudo-empirical likelihood distance function. (MRC1)– Both the nonresponse model and the imputation model were based on Z_1-Z_4. (MRC2)– Both models in (MRC1) and their two-way, three-way, and four-way interaction terms; (MRC3)– Both models in (MRC1) and the additional models with $|Z|_1^{1/2}$, $|Z|_2^{1/2}$, $|Z|_3^{1/2}$, $|Z|_4^{1/2}$, and their two-way, three-way, and four-way interaction terms; (MRC4)– Both models in (MRC1) and the additional models with $\log|Z_1|$, $\log|Z_2|$, $\log|Z_3|$, $\log|Z_4|$ as explanatory variables, and their two-way, three-way, and four-way interaction terms; (MRC5)–Based on all the models used in (MRC1) to (MRC4); (5). Four multiply robust robust estimators based on aggregation through refitting: (MRP2)–Using the same models as in (MRC2); (MRP3)– Using the same models as in (MRC3); (MRP4)– Using the same models as in (MRC4); (MRP5)– Using the same models as in (MRC5).

For each estimator, we computed the following Monte Carlo measures: bias, standard error and root mean squared error. The results are shown in Table 1. As expected, the complete data estimator COM exhibited negligible bias and was the most efficient. Both the estimator DR and NN showed appreciable bias. Except for MRP2, the MRC and MRP estimators performed much better than DR and NN. The MRC estimators performed generally better than their MRP counterparts. This can be explained by the fact that the calibration procedure used in the aggregation through calibration provides some robustness against the presence of small estimated response probabilities.

Table 1: Monte Carlo Bias (BIAS), Standard Error (SE), and Root Mean Squared Error (RRMSE) for different estimation procedures.

	Response rate–40%			Response rate–60%		
Procedure	Bias	SE	RMSE	Bias	SE	RMSE
COM	- 0.05	1.39	1.39	0.01	1.34	1.34
DR	-12.69	44.49	46.26	-6.71	40.90	41.44
NN	-9.83	1.97	10.02	-5.30	1.52	5.52
MRC1	-2.70	1.95	3.33	-1.84	1.57	2.42
MRC2	-0.69	1.90	2.02	-0.69	1.57	1.72
MRC3	-0.83	1.65	1.84	-0.57	1.45	1.56
MRC4	-1.27	1.58	2.02	-0.67	1.40	1.55
MRC5	-1.02	2.26	2.48	-0.66	1.47	1.61
MRP2	-6.49	34.95	35.55	-3.98	17.51	17.95
MRP3	-1.62	4.27	4.56	-1.16	3.05	3.26
MRP4	-1.28	1.61	2.05	-0.72	1.47	1.64
MRP5	-1.08	2.16	2.42	-0.80	1.91	2.07

References

Bang, H. and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Boistard, H., Chauvet, G. and Haziza, D. (2016). Doubly robust inference for the distribution function in the presence of missing survey data. *Scandinavian Journal of Statistics* **43**, 683—699.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley New York.

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science* **29**, 380—396.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439—453.

Chen, S. and Haziza, D. (2019). On the nonparametric multiple imputation with multiply robustness. *Statistica Sinica* **29**, 2035–2053.

Chen, S. and D. Haziza, D. (2021). A review of multiply robust estimation with missing data. *In Modern Statistical Methods for Health Research*, 103–118.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–734.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Duan, X. and Yin, G. (2017). Ensemble approached to estimation of the population mean with missing responses. *Scandinavian Journal of Statistics* **44**, 899–917.

Han, P. (2014). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference* **148**, 101–110.

Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika* **103**, 683–700.

Han, P. and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika* **100**, 417–430.

Han, P., Kong, L., Zhao, J. and Zhou, X. (2019). A general framework for quantile estimation with incomplete data. *Journal of the Royal Statistical Society B* **81**, 305–333.

Haziza, D. and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32**, 53–64.

Kang, J. D. Y. and Schafer, J. L. (2008). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **24**, 375–394.

Kott, P. S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association* **89**, 693–696.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficient when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling.* Springer-Verlag.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion and rejoinder). *Journal of the American Statistical Association* **94**, 1096–1146.

Seaman, S. R. and Vansteelandt, S. (2018). Introduction to double robust methods for incomplete data. *Statistical Science* **33**, 184–197.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101**, 1619–1637.

The Survey Statistician, 2023, Vol. 88, 86–93.

New and Emerging Methods

Calibration Techniques for Model-based Prediction and Doubly Robust Estimation

Changbao Wu

University of Waterloo, Canada, cbwu@uwaterloo.ca

Abstract

We present a brief overview of calibration techniques for model-assisted estimation for probability survey samples and their extensions for model-based prediction and doubly robust estimation to missing data problems, causal inference, and analysis of non-probability samples. The focuses are to provide a clear description of the setting for each of these areas and on how doubly robust estimators are constructed either through a set of calibration equations or using model-calibrated empirical likelihood methods. Theoretical details are left to additional references.

Keywords: Empirical likelihood; Double robustness; Inverse probability weighting; Model-assisted estimation; Model-calibration.

1 Introduction

Survey samples have an important feature of representing a finite target population. Statistical tools for dealing with descriptive finite population parameters are often discrete in nature, such as series summations and double summations. There has been a separation between survey sampling and the so-called mainstream statistics in terms of tools and methodologies, highlighted by the extensive use of parametric or semi-parametric models, the likelihood principle, Bayesian pedagogies, etc., in other fields of statistics but not or less so in survey sampling. The field of survey sampling often lags behind on development of innovative general statistical tools.

There have been examples, however, where a method was first developed or rooted in survey sampling and later became widely used in other fields of statistics. The most prominent example is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Narain, 1951), which is popularly termed as the "*inverse probability weighted*" (IPW) estimator and is a fundamental tool for propensity score based methods in missing data analysis and causal inference. Another less known example is the doubly robust estimator, also popularized in missing data and causal inference literature starting from the 1990s. It is rooted in model-assisted estimation methods first developed in survey sampling going back to the 1970s. The generalized difference estimator of the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ of the study variable y, where N is the population size, as discussed in Cassel et al. (1976) is given by

$$\hat{\mu}_{yGD} = \frac{1}{N}\Big\{\sum_{i\in\mathcal{S}} \frac{y_i - c_i}{\pi_i} + \sum_{i=1}^{N} c_i\Big\}, \tag{1}$$

where $\mathcal{S}$ is a probability sample of size n, the π_i's are the first order inclusion probabilities, and $\{c_1, c_2, \ldots, c_N\}$ is an arbitrary sequence of known numbers. The estimator $\hat{\mu}_{yGD}$ is exactly unbiased for μ_y under the probability sampling design p for any given sequence c_i, and is also model-unbiased if we choose $c_i = m_i = E_\xi(y_i \mid \boldsymbol{x}_i)$ under the assumed model ξ on y given $\boldsymbol{x}$. The estimator $\hat{\mu}_{yGD}$ with the choice $c_i = m_i$ is the same as the doubly robust estimator in the missing data and causal inference literature where π_i becomes the propensity score and m_i is the mean function of the outcome regression. Both π_i and m_i require an assumed model to be estimated, and the estimator remains valid if one of the models is correctly specified. The doubly robust estimator is also called the "*augmented inverse probability weighted*" (AIPW) estimator in the literature.

Calibration methods are also first developed in survey sampling and later find general uses in other areas. While the popularity of calibration methods is often credited to the highly cited JASA paper by Deville and Särndal (1992), the original idea of calibration estimation goes back to Deming and Stephan (1940) on raking ratio estimators. The model-calibration approach proposed by Wu and Sitter (2001) serves as the basis for the discussions presented in the rest of the paper on model-based prediction and doubly robust estimation.

2 Calibration methods for probability survey samples

The fundamental tool for design-based approach to survey sampling is the Horvitz-Thompson estimator for the finite population total $T_y = \sum_{i=1}^{N} y_i$, which is given by $\hat{T}_{yHT} = \sum_{i\in\mathcal{S}} d_i y_i$, where $d_i = 1/\pi_i$ are the basic design weights. Most surveys collect information on a vector of auxiliary variables, $\boldsymbol{x}$, leading to a survey dataset $\{(y_i, \boldsymbol{x}_i, d_i), i \in \mathcal{S}\}$. The initial motivation of calibration estimators is to use the known population totals of the auxiliary variables, $T_{\mathbf{x}} = \sum_{i=1}^{N} \boldsymbol{x}_i$, to achieve the so-called *internal consistency* by using calibrated weights w_i instead of d_i such that

$$\sum_{i\in\mathcal{S}} w_i \boldsymbol{x}_i = T_{\mathbf{x}}\,. \tag{2}$$

Equations (2) are referred to as the calibration equations or benchmark constraints. Deville and Särndal (1992) formulated the general calibration methods as a constrained minimization problem where the calibration weights w_i are obtained by minimizing a distance measure $D(\boldsymbol{d}, \boldsymbol{w})$ between $\boldsymbol{d} = (d_1, \ldots, d_n)$ and $\boldsymbol{w} = (w_1, \ldots, w_n)$ subject to constraints (2). Deville and Särndal (1992) argued intuitively that the calibration estimator $\hat{T}_{yC} = \sum_{i\in\mathcal{S}} w_i y_i$ should be more efficient than $\hat{T}_{yHT}$ since "... *weights that perform well for the auxiliary variable also should perform well for the study variable*".

The calibration estimator $\hat{T}_{yC} = \sum_{i\in\mathcal{S}} w_i y_i$ is indeed a model-assisted estimator with the same spirit of "double robustness" under a linear regression model with the mean function $E_\xi(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where E_ξ denotes the expectation with respect to the model ξ and $\boldsymbol{\beta}$ is the vector of regression coefficients. Under the constrained minimization procedure of Deville and Särndal (1992), the estimator $\hat{T}_{yC}$ is design-consistent regardless of any models. It is also an unbiased model-based prediction estimator under the linear regression model ξ since $E_\xi(\hat{T}_{yC} - T_y) = 0$.

The calibration estimator $\hat{T}_{yC} = \sum_{i\in\mathcal{S}} w_i y_i$ with the constraints (2) is no longer model-unbiased under any nonlinear models. Wu and Sitter (2001) considered a semiparametric model with a general mean function $E_\xi(y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ and proposed a model-calibration approach through the use of

the constraint

$$\sum_{i\in\mathcal{S}} w_i \mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})\,, \tag{3}$$

where $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ under the assumed model. There are three basic features of the model-calibration estimator $\hat{T}_{yMC} = \sum_{i\in\mathcal{S}} w_i y_i$ with the model-calibration constraint (3): (i) it is design-consistent irrespective of the model; (ii) it is an approximately model-unbiased prediction estimator under the assumed model; and (iii) the use of the estimated model parameters $\hat{\boldsymbol{\beta}}$ in (3) has no impact on the asymptotic variance of $\hat{T}_{yMC}$ under the survey design. Nonparametric models can also be used to construct model-calibration estimators (Montanari and Ranalli, 2005). The model-calibration constraint requires the "population control" $\sum_{i=1}^{N} \mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})$ to be known, which typically requires complete auxiliary information $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ to be available under a nonlinear model for y given $\boldsymbol{x}$.

Calibration methods can be formulated under the framework of pseudo empirical likelihood (PEL) where the distance measure $D(\boldsymbol{d}, \boldsymbol{w})$ is replaced by the pseudo empirical log-likelihood function of Chen and Sitter (1999) defined as

$$\ell_{PEL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} d_i \log(p_i)\,, \tag{4}$$

where $\boldsymbol{p} = (p_1, \ldots, p_n)$ satisfying $p_i > 0$ and the normalization constraint

$$\sum_{i\in\mathcal{S}} p_i = 1\,, \tag{5}$$

and the calibration weights are given by $\boldsymbol{w} = N\boldsymbol{p}$. The PEL approach with calibration equations has a major advantage of constructing better behaved PEL ratio confidence intervals (Wu and Rao, 2006). The PEL function $\ell_{PEL}(\boldsymbol{p})$ is defined explicitly through the design weights d_i. An alternative approach is to incorporate the survey weights through an additional constraint and use the standard empirical likelihood (EL) of Owen (1988) for the constrained maximization. Let

$$\ell_{EL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}} \log(p_i)\,. \tag{6}$$

The maximum EL estimator of the population mean μ_y is given by $\hat{\mu}_{yEL} = \sum_{i\in\mathcal{S}} \hat{p}_i y_i$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_n)$ maximizes $\ell_{EL}(\boldsymbol{p})$ subject to the normalization constraint (5) and other suitably chosen constraints. The estimator $\hat{\mu}_{yEL}$ is design-consistent if the following constraint is included:

$$\sum_{i\in\mathcal{S}} p_i \pi_i = \frac{n}{N}\,. \tag{7}$$

Note that constraint (7) is a sample version of the population moment condition $N^{-1}\sum_{i=1}^{N} \pi_i = n/N$ under survey designs with fixed sample size n. Alternative versions of (7) are used by Kim (2009) and by Oguz-Alper and Berger (2016), among others. Estimator $\hat{\mu}_{yEL}$ is also approximately model-unbiased under the assumed semiparametric model if we include the model-calibration constraint

$$\sum_{i\in\mathcal{S}} p_i \mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N} \mu(\boldsymbol{x}_i;\hat{\boldsymbol{\beta}})\,. \tag{8}$$

The standard EL formulation through constrained maximization of $\ell_{EL}(\boldsymbol{p})$ subject to (5), (7) and (8) brings a unified framework for model-assisted estimation with probability survey samples and doubly robust estimators in other areas, as discussed in Section 3 below. The quantities on the right hand

side of equations (7) and (8) are "population controls" and need to be replaced by suitable estimates depending on the setting of the problem, as discussed in Section 3.

3 Calibration approach to propensity score based estimation

In this section, we describe suitable formulations of EL-based inference for missing data problems, causal inference, and estimation with non-probability samples to construct doubly robust estimators through calibration techniques. The focus is on similarities of these problems and their connections to the calibration methods presented in Section 2.

3.1 Missing data

Let $\mathcal{S}$ be a set of n subjects with independent and identically distributed observations from an underlying infinite population. The vector of covariates $\boldsymbol{x}$ is fully observed but the study variable y is subject to missingness. Let $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. Let $\mathcal{S}_R = \{i \mid i \in \mathcal{S} \text{ and } \delta_i = 1\}$ be the set of respondents with observed y and $\mathcal{S}_M = \{i \mid i \in \mathcal{S} \text{ and } \delta_i = 0\}$ be the set of nonrespondents with missing y. The observed data can be represented by $\{(\delta_i, \delta_i y_i, \boldsymbol{x}_i), i \in \mathcal{S}\}$.

Propensity scores, defined as $\pi_i = P(\delta_i = 1 \mid y_i, \boldsymbol{x}_i)$, play an important role for missing data analysis. Under the missing at random assumption where $\pi_i = P(\delta_i = 1 \mid \boldsymbol{x}_i)$, the π_i's can be estimated based on an assumed parametric model on δ given $\boldsymbol{x}$, denoted as model q, using the observed dataset $\{(\delta_i, \boldsymbol{x}_i), i \in \mathcal{S}\}$. For instance, one can use a logistic regression model where $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = 1 - [1 + \exp(\boldsymbol{x}_i'\boldsymbol{\alpha})]^{-1}$ and estimate the model parameters $\boldsymbol{\alpha}$ using maximum likelihood.

The IPW estimator of the population mean μ_y is given by $\hat{\mu}_{yIPW} = n^{-1} \sum_{i \in \mathcal{S}_R} y_i / \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$, where the estimator $\hat{\boldsymbol{\alpha}}$ is obtained by a suitable method such as maximum likelihood and the IPW estimator is consistent. With an assumed parametric form $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$, the propensity score model parameters $\boldsymbol{\alpha}$ can be estimated using a calibration method, and the resulting IPW estimator $\hat{\mu}_{yIPW}$ is doubly robust if the outcome regression model ξ on y given $\boldsymbol{x}$ is linear. The calibration estimator $\hat{\boldsymbol{\alpha}}$ is defined as the solution to the calibration equations

$$\sum_{i \in \mathcal{S}_R} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i \in \mathcal{S}} \boldsymbol{x}_i \,. \tag{9}$$

The double robustness property of $\hat{\mu}_{yIPW}$ is justified based on the following two arguments. First, the equation system (9) is "unbiased" with respect to the propensity score model q in the sense that $E_q\{\sum_{i \in \mathcal{S}_R} \boldsymbol{x}_i / \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) - \sum_{i \in \mathcal{S}} \boldsymbol{x}_i \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} = \boldsymbol{0}$, and the resulting calibration estimator $\hat{\boldsymbol{\alpha}}$ is consistent for $\boldsymbol{\alpha}$. Second, the sample means $n^{-1} \sum_{i \in \mathcal{S}} \boldsymbol{x}_i$ is a valid approximation to the "population controls" of the variables $\boldsymbol{x}$ because $\mathcal{S}$ is an iid sample. In practice, the calibration estimator $\hat{\boldsymbol{\alpha}}$ obtained as the solution to (9) tends to be less stable as compared to the maximum likelihood estimator; see, for instance, Chen et al. (2020) for a discussion under the context of non-probability samples.

The EL-based methods for achieving double robustness through model-calibration is a more desirable approach and is applicable to linear or nonlinear outcome regression models with a mean function $\mu(\boldsymbol{x}, \boldsymbol{\beta})$. It involves modifications to the three crucial components: the EL function, the constraint on propensity scores, and the model-calibration constraint on the outcome regression. Let $m = \sum_{i \in \mathcal{S}} \delta_i$ be the number of units with observed y_i. Let $\boldsymbol{p} = (p_1, \ldots, p_m)$ and $\ell_{EL}(\boldsymbol{p}) = \sum_{i \in \mathcal{S}_R} \log(p_i)$. The maximum EL estimator of μ_y is computed as $\hat{\mu}_{yEL} = \sum_{i \in \mathcal{S}_R} \hat{p}_i y_i$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_m)$ maximizes $\ell_{EL}(\boldsymbol{p})$ subject to the normalization constraint $\sum_{i \in \mathcal{S}_R} p_i = 1$, the constraint for the propensity score model $\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = E_q(\delta_i \mid \boldsymbol{x}_i)$,

$$\sum_{i \in \mathcal{S}_R} p_i \, \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \frac{m}{n} \,, \tag{10}$$

and the constraint for the outcome regression model $\mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = E_\xi(y_i \mid \boldsymbol{x}_i)$,

$$\sum_{i \in \mathcal{S}_R} p_i \, \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i \in \mathcal{S}} \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) \,. \tag{11}$$

The m used in equation (10) may be replaced by $\sum_{i \in \mathcal{S}} \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$. For computational simplicity, the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in equations (10) and (11) can be replaced by suitable estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, and the resulting estimator $\hat{\mu}_{yEL} = \sum_{i \in \mathcal{S}_R} \hat{p}_i y_i$ remains doubly robust.

3.2 Causal inference

Estimation of the Average Treatment Effect (ATE) by comparing the responses of the treatment group to the ones for the control group is a fundamental problem in causal inference. Let $\mathcal{S}$ be the set of initial n subjects randomly selected from the target population, with measures on baseline variables $\boldsymbol{x}$ for each subject. Let T be the treatment assignment indictor with $T_i = 1$ if subject i is assigned to the treatment group and $T_i = 0$ if i is assigned to the control group. Let $\mathcal{S}_1$ and $\mathcal{S}_0$ be the set of subjects in the treatment group and in the control group, with sizes n_1 and n_0, respectively. We have $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_0$ and $n = n_1 + n_0$. Let y_1 and y_0 be, respectively, the study variable under the treatment and the control. We have a unique two-sample setting with two datasets $\{(y_{1i}, T_i = 1, \boldsymbol{x}_i), i \in \mathcal{S}_1\}$ and $\{(y_{0i}, T_i = 0, \boldsymbol{x}_i), i \in \mathcal{S}_0\}$. The ATE is the parameter of interest and is defined as $\theta = \mu_1 - \mu_0$ where μ_1 and μ_0 are, respectively, the population means of the study variable under the treatment and under the control. We assume that T_i is conditionally independent of y_{1i} and y_{0i} given $\boldsymbol{x}_i$.

It is possible to construct a doubly robust estimator for each of μ_1 and μ_0 separately, and estimate θ by $\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_0$, using a parallel procedure for the missing data problem described in Section 3.1 for obtaining $\hat{\mu}_1$ and $\hat{\mu}_0$. Huang et al. (2023) used a two-sample EL formulation and dealt with θ directly for EL-ratio confidence intervals. Let $\pi_i = P(T_i = 1 \mid \boldsymbol{x}_i)$ be the propensity score for treatment assignments, with an assumed parametric form $\pi_i = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$; let $\mu_j(\boldsymbol{x}_i, \boldsymbol{\beta}_j) = E_\xi(y_{ji} \mid \boldsymbol{x}_i)$ be the mean functions of the response variable y_j for the two groups $j = 1, 0$ under two assumed outcome regression models. Let $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jn_j})$, $j = 1, 0$. The two-sample EL function is given by

$$\ell(\boldsymbol{p}_1, \boldsymbol{p}_0) = \sum_{i \in \mathcal{S}_1} \log(p_{1i}) + \sum_{i \in \mathcal{S}_0} \log(p_{0i}) \,. \tag{12}$$

The maximum EL estimator of θ is computed as $\hat{\theta}_{EL} = \sum_{i \in \mathcal{S}_1} \hat{p}_{1i} y_{1i} - \sum_{i \in \mathcal{S}_0} \hat{p}_{0i} y_{0i}$, where $\hat{\boldsymbol{p}}_j = (\hat{p}_{j1}, \ldots, \hat{p}_{jn_j})$, $j = 1, 0$ maximize $\ell(\boldsymbol{p}_1, \boldsymbol{p}_0)$ subject to the normalization constraints $\sum_{i \in \mathcal{S}_1} p_{1i} = 1$ and $\sum_{i \in \mathcal{S}_0} p_{0i} = 1$, the model-calibration constraints induced by the propensity scores,

$$\sum_{i \in \mathcal{S}_1} p_{1i} \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \frac{n_1}{n} \,, \quad \sum_{i \in \mathcal{S}_0} p_{0i} [1 - \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})] = \frac{n_0}{n} \,, \tag{13}$$

and the model-calibration constraints from the two outcome regression with respect to y_1 and y_0 conditional on $\boldsymbol{x}$,

$$\sum_{i \in \mathcal{S}_1} p_{1i} \mu_1(\boldsymbol{x}_i, \boldsymbol{\beta}_1) = \frac{1}{n} \sum_{i \in \mathcal{S}} \mu_1(\boldsymbol{x}_i, \boldsymbol{\beta}_1) \,, \quad \sum_{i \in \mathcal{S}_0} p_{0i} \mu_0(\boldsymbol{x}_i, \boldsymbol{\beta}_0) = \frac{1}{n} \sum_{i \in \mathcal{S}} \mu_0(\boldsymbol{x}_i, \boldsymbol{\beta}_0) \,. \tag{14}$$

The model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ used in constraints (13) and (14) can be replaced by suitable estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$. The constraint for the parameter of interest, $\theta = \mu_1 - \mu_0$, which is part of the system for computing the EL ratio function, is given by

$$\sum_{i \in \mathcal{S}_1} p_{1i} y_{1i} - \sum_{i \in \mathcal{S}_0} p_{0i} y_{0i} = \theta \,. \tag{15}$$

The two-sample EL formulation with a single parameter of interest θ imposes some computational challenges for the constrained maximization problem. Huang et al. (2023) contain further discussions.

3.3 Non-probability samples

One of the basic features of probability samples is that the sample inclusion probabilities are known under the given sampling design. Statistical analysis with non-probability samples requires assumptions about and modelling on the unknown sample selection/inclusion process, which further requires auxiliary information on the target population. A popular setup widely used in the recent literature involves a reference probability sample containing auxiliary information from the same target population; see, for instance, Chen et al. (2020) and references therein. Let $\mathcal{S}_A$ be the set of n_A units for the non-probability sample and $\mathcal{S}_B$ be the set of n_B units for the reference probability sample, both from the same target population of size N. The two sample datasets are represented by $\{(y_i, \boldsymbol{x}_i), i \in \mathcal{S}_A\}$ and $\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$, where the d_i^B are the survey weights for the reference probability sample.

Let $R_i = 1$ if $i \in \mathcal{S}_A$ and $R_i = 0$ otherwise, $i = 1, 2, \ldots, N$. Assume that R_i and y_i are independent given $\boldsymbol{x}_i$. A crucial step in analyzing the non-probability sample dataset is the modelling on the propensity scores, also called the participation probabilities by some authors, i.e., $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i)$, $i = 1, 2, \ldots, N$. The participation probability π_i^A is defined for all units in the target population, and it is immediately clear that estimation of the π_i^A's requires information on $\boldsymbol{x}$ from the entire target population as well as an assumed model, even though the final IPW estimator of $\mu_y = N^{-1}\sum_{i=1}^N y_i$, computed as $\hat{\mu}_{yIPW} = N^{-1}\sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^A$, only requires the estimated π_i^A for units in $\mathcal{S}_A$.

Let the form of $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\alpha})$ be specified from a parametric model q on $(R_i \mid \boldsymbol{x}_i)$. A pseudo maximum likelihood estimator of $\boldsymbol{\alpha}$ was described in Chen et al. (2020). A calibration estimator $\hat{\boldsymbol{\alpha}}$ can also be obtained as the solution to the calibration equations

$$\sum_{i\in\mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\alpha})} = \sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i \,. \tag{16}$$

The right hand side of (16) is an estimate for the population controls $\sum_{i=1}^N \boldsymbol{x}_i$ using the reference probability sample $\mathcal{S}_B$. Consistency of $\hat{\boldsymbol{\alpha}}$ follows from the result that $E_{qp}\{\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i/\pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) - \sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i\} = \boldsymbol{0}$ under the joint randomization of the model, q, for sample participation and the probability sampling design, p, for the reference sample. The IPW estimator $\hat{\mu}_{yIPW} = N^{-1}\sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^A$, with the calibration estimator $\hat{\boldsymbol{\alpha}}$ used in $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\alpha}})$, is doubly robust if the outcome regression model ξ for $(y_i \mid \boldsymbol{x}_i)$ is linear since $E_{\xi p}\{\hat{\mu}_{yIPW} - \mu_y\} \doteq 0$ under the linear mean function $E_\xi(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$.

Chen et al. (2022) presented the PEL approach to doubly robust estimation with non-probability samples with a linear or nonlinear outcome regression model $E_\xi(y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$. Doubly robust estimation can also be achieved through the standard EL. Let $\ell_{EL}(\boldsymbol{p}) = \sum_{i\in\mathcal{S}_A} \log(p_i)$, where $\boldsymbol{p} = (p_1, \ldots, p_{n_A})$. The maximum EL estimator of μ_y is computed as $\hat{\mu}_{yEL} = \sum_{i\in\mathcal{S}_A} \hat{p}_i y_i$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_{n_A})$ maximizes $\ell_{EL}(\boldsymbol{p})$ subject to the normalization constraint $\sum_{i\in\mathcal{S}_A} p_i = 1$, the constraint for the participation probabilities,

$$\sum_{i\in\mathcal{S}_A} p_i \pi(\boldsymbol{x}_i, \boldsymbol{\alpha}) = \frac{n_A}{N} \,, \tag{17}$$

and the model-calibration constraint for the outcome regression model,

$$\sum_{i\in\mathcal{S}_A} p_i \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) \,. \tag{18}$$

The model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ used in (17) and (18) can be replaced by suitable estimates, and the

population size N can be replaced by $\hat{N} = \sum_{i \in \mathcal{S}_B} d_i^B$. The resulting estimator $\hat{\mu}_{yEL}$ is doubly robust as defined in Chen et al. (2020) where, in addition to model q for the sample participation and the model ξ for outcome regression, the probability sampling design p is part of the joint randomization framework.

4 Concluding remarks

Maximum Likelihood (ML) and Least Square (LS) are two fundamental principles for statistical inference. Calibration techniques have shown potential to be a general statistical tool, especially in the modern era for combining data from different sources as well as information from different models. The concept of model-calibration has found applications in a wide range of problems in recent years and has demonstrated certain optimality and robustness for best use of auxiliary information through an assumed model; see, for instance, Wu (2003) and Zhang et al. (2022), among others. The constrained minimization of a distance measure as described in Deville and Särndal (1992) provides a natural connection to the constrained maximization of the empirical likelihood function, which has been used in many areas of statistics. Calibration techniques for model-based prediction and doubly robust estimation have been shown to be useful for problems described in this short article and their potential for other problems and extensions to the so-called multiply robust estimation (Han and Wang, 2013) deserves further exploration.

References

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, **63**, 615–620.

Chen, J., and Sitter, R. R. (1999) A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, **9**, 385–406.

Chen, Y., Li, P., and Wu, C. (2020) Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, **115**, 2011–2021.

Chen, Y., Li, P., Rao, J. N. K., and Wu, C. (2022) Pseudo empirical likelihood methods for non-probability survey samples. *The Canadian Journal of Statistics*, **50**, 1166–1185.

Deming, W. E., and Stephan, F. F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, **11**, 427–444.

Deville, J., and Särndal, C. -E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Han, P., and Wang, L. (2013) Estimation with missing data: beyond double robustness. *Biometrika*, **100**, 417–430.

Horvitz, D. G., and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Huang, J., Wu, C., and Zeng, L. (2023) Empirical likelihood methods for causal inference. Working paper, Department of Statistics and Actuarial Science, University of Waterloo.

Kim, J. K. (2009) Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, **19**, 145–157.

Montanari, G. E., and Ranalli, M. G. (2005) Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, **100**, 1429–1442.

Narain, R. D. (1951) On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–174.

Oguz-Alper, M. and Berger, Y. G. (2016) Modelling complex survey data with population level information: An empirical likelihood approach. *Biometrika*, **103**, 447–459.

Owen, A. B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Wu, C. (2003) Optimal calibration estimators in survey sampling. *Biometrika*, **90**, 937–951.

Wu, C., and Rao, J. N. K. (2006) Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, **34**, 359–375.

Wu, C., and Sitter, R. R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185–193.

Zhang, S., Han, P., and Wu, C. (2022) Calibration techniques encompassing survey sampling, missing data analysis and causal inference. *International Statistical Review*, accepted for publication.

The Survey Statistician, 2023, Vol. 88, 94–95.

Book and Software Review

Book review: "Sampling: Design and Analysis, Third Edition" by Sharon L. Lohr

Camelia Goga[1]

[1]Laboratoire de Mathématiques de Besançon, Université de Franche-Comté
Besançon, France, camelia.goga@univ-fcomte.fr

Abstract

The book *"Sampling: Design and Analysis, Third Edition"* by Sharon L. Lohr is an outstanding work in the field of survey sampling. This third edition is accompanied by two additional books for readers who wish to implement the survey sampling techniques presented in the book using either R or SAS software. Like the previous two editions, this book is primarily aimed at students and instructors.

Keywords: categorical data analysis, linearization techniques, nonprobability samples, R and SAS companion books, sampling designs.

The book *"Sampling: Design and Analysis, Third Edition"* by Sharon L. Lohr was published by Chapman and Hall/CRS Press in 2022 (Lohr, 2022a). This new edition is structured in 16 chapters, for more than 600 pages, and covers a broad spectrum of survey theory concepts, the whole supported by numerous examples from social sciences, public opinion research, public health, business, agriculture, and ecology. Chapters 1 to 6 are dedicated to the most commonly used probability sampling designs (simple random sampling without replacement, stratified sampling, unequal probability sampling designs, cluster sampling design) to estimate totals and means by using the Horvitz-Thompson, ratio and regression estimators. These chapters also contain sections dedicated to the model-based versus design-based estimation points of view, with discussions of recent research in this field. Chapters 7 to 16 are dedicated to advanced topics in survey sampling theory usually related to complex surveys, such as nonresponse, linearization techniques, categorical data analysis, two-phases sampling designs, rare populations and small area estimation. A chapter on nonprobability samples is included as well in this new edition.

Throughout this book, focus is given on the sampling phase, which is considered by the author to be the most important step in the survey process. The reader is guided in this process by means of many examples on real survey data, inevitably involving graphical analysis of survey data, which can be a real challenge in practice.

As the book is primarily aimed at students and instructors, the book contains more than 550 exercises from which more than 150 exercises are new for this third edition.

As in the previous editions, these exercises are structured in four types intended to meet the objectives of the book: *Introductory*, *Working with survey data*, *Working with theory* and *Project and Activities*. As a teacher of statistics and survey theory myself, I can testify of their utility and great interest.

The novelty of this new edition is the addition of two free downloadable books: *"R Companion for Sampling: Design and Analysis, Third Edition"* by Yan Lu and Sharon L. Lohr (Yan & Lohr, 2022) and *"SAS Software Companion for Sampling: Design and Analysis, Third Edition"* by Sharon L. Lohr (Lohr, 2022b). These two additional books are intended to be a guide for a novice reader in the field of surveys but also in the use of R and SAS software. The companion books have exactly the same structure. They start with an introductory chapter which gives basic instructions to get started with R and SAS (e. g. for R, installing the software and specific packages, reading and saving statistical datasets as well as conducting first statistical analyses and drawing plots). The books then provide respectively R and SAS implementation of the exercises proposed in the book *"Sampling: Design and Analysis, Third Edition"* (Chapters 1 to 11). For the R companion, the R packages `survey` (Lumley, 2020) for computing statistics from complex surveys, `sampling` (Tillé & Matei, 2021) for drawing complex samples and `SDARessources` (Lu & Lohr, 2021) for datasets from the *"Sampling: Design and Analysis, Third Edition"* book are mainly used; some functions of these packages are described in detail. For the SAS part, the focus is on the procedures: `surveyselect`, `surveymeans`, `surveyfreq`, `surveyreg` and `surveylogistic`. The implemented examples are provided with code, annotated output, and helpful tips. Both companions end with a chapter entitled *"Additional Topics for Survey Data Analysis"* containing implementations for some advanced methodology contained in Chapters 12 to 16 from the *"Sampling: Design and Analysis, Third Edition"* book as well as a *"Data Set Descriptions"* in Appendix. The SAS companion contains a specific Appendix, *"Jackknife Macros"*, dedicated to jackknife methods with survey data.

Concluding, the Lohr's book *"Sampling: Design and Analysis, Third Edition"* will be again a reference book in the field of survey sampling as the first two editions. The two companion books for carrying out surveys and estimates with R and SAS software fill a gap in the specialist literature. Students, instructors and anyone wishing to train in survey techniques will find in this book the necessary methodology, as well as many examples of sample surveys on real data and how to implement them using R and SAS software in its two companion books.

References

Lohr, S. L. (2022a). *Sampling: Design and Analysis, Third Edition*. New York: Chapman and Hall/CRC Press.

Lohr, S. L. (2022b). *SAS Companion for Sampling: Design and Analysis, Third Edition*. New York: Chapman and Hall/CRC Press.

Lu, Y. & Lohr, S. L. (2021). SDAResources: Datasets and Functions for 'Sampling: Design and Analysis, 3rd Edition'. R package version 0.1.1, `https://CRAN.R-project.org/package=SDAResources`.

Lumley, T. (2020). *survey: analysis of complex survey samples*. R package version 4.0, `https://CRAN.R-project.org/package=survey`.

Tillé, Y. & Matei, A. (2021). *sampling: Survey Sampling*. R package version 2.9, `https://CRAN.R-project.org/package=sampling`.

Yan, L. & Lohr, S. L. (2022). *R Companion for Sampling: Design and Analysis, Third Edition*. New York: Chapman and Hall/CRC Press.

The Survey Statistician, 2023, Vol. 88, 96-104.

Book and Software Review

The survey Package for R, 15 Years on

Thomas Lumley[1]

[1] University of Auckland, t.lumley@auckland.ac.nz

Abstract

In 2008, issue 57 of *The Survey Statistician* published an article about the **survey** package for R, which at that point had been under development for about five years. The current version of the package is 4.2; the version in 2008 was 3.6-12. In this article I will discuss the changes since 2008, and major changes planned for the near future.

Keywords: software, two-phase designs, domain estimation, regression, plausible values.

1 Introduction

The **survey** package has three main goals:

- Integrate standard design-based analyses into R
- Allow non-specialist users to display, analyse, and model complex surveys with only the necessary changes from their usual data analysis practice
- Provide an implementation platform for novel analysis methods.

Since 2008 there has been substantial progress on all these goals. The package is now quite widely used in teaching, research, and at some official statistics agencies. The main download site reports an average of about 2000 downloads per day, and 131 other packages list survey as a dependency.

It's probably no longer necessary in 2023 to introduce readers to R, the free statistical computing environment. It is worth quoting from 2008:

The flexibility of R comes at a price in performance: it is often slower and usually requires more memory than competing systems. This disadvantage is becoming progressively less important as computers improve.

The trend has continued; it is entirely possible to handle national survey data sets on commodity servers, and moderate-sized surveys such as NHANES (see https://www.cdc.gov/nchs/nhanes/index.htm) can be analysed interactively on standard laptops. Speed has not been a priority for implementation, however I am happy to learn about real examples where the package is genuinely too slow and have worked to optimise the code in such cases. A notable example early in development was designs where a large stratum is sampled with certainty, as in some business surveys.

While the package, like R, remains based on command-line code, there is now a basic graphical user interface from the iNZight project (https://inzight.nz). A full description is outside the scope of this article, but iNZight was developed for teaching introductory statistics and for supporting data

exploration and analysis. The system now allows survey designs to be specified for data files. When a design is specified, appropriate methods for analysis and graphics are transparently used with no extra effort from the user. The visual interface allows for only two stages of sampling, with stratification at the first stage, and sampling with or without replacement.

The survey package does not provide much in the way of output formatting. This is deliberate; R has other packages for constructing and formatting tables, notably the new **gt** package (Iannone et al, 2023).

Basic structure of the package

Analysis functions in the **survey** package take two standard inputs: a model formula describing the variables to be used, and a survey design object containing the data and meta-data. The design objects are created by the functions `svydesign` (when strata, clusters, etc are supplied) and `svrepdesign` (when replicate weights are supplied). Replicate weights can also be created for a `svydesign` object. Arbitrarily many stages of stratified cluster sampling are possible, with or without replacement. There is also some support for PPS sampling and for arbitrary designs specified via a pairwise probability matrix.

The design objects are not simply data sets: subsetting a design object gives a design subpopulation object that will produce valid domain estimates. Raking, calibration, and trimming of weights are also available for design objects.

When data are inconveniently large to reside in memory, they can now be kept in a database. The `svydesign` and `svrepdesign` functions accept a database table as a data source and produce survey design objects that load only the necessary data into memory for each analysis.

2. New features since 2008

Comparisons between domains

The package has always supported domain estimation, but did not initially provide variance estimates for contrasts *between* domains (except through regression or ratio estimation). When using replicate weights it is easy to estimate these variances by carrying the replicate estimates through any calculation, and this was added not long after 2008. With linearisation it is somewhat more complicated, but since version 4.0 of the package, analysis functions can return a matrix of influence functions that will be carried through subsequent calculations. This approach means new functions – built into the package or written by users – can automatically take advantage of the between-domain comparisons.

Here are two examples comparing different levels of school, in a built-in dataset on California standardised assessments. The first compares the mean size of high schools and elementary schools, the second compares the agreement between two yes/no progress indicators, measured with Cohen's kappa.

```
> library(survey)
> data(api)
> dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
> means<- svyby(~enroll, by=~stype, design=dclus1, svymean, covmat=TRUE)
> means
  stype    enroll        se
E     E  432.8542  16.51599
H     H 1130.2857 357.12197
M     M  897.7200  99.53188
> svycontrast(means, c(H=1, E=-1))
         contrast     SE
contrast   697.43 346.47
> kappas<- svyby(~comp.imp+sch.wide, by=~stype, design=dclus1, svykappa,
covmat=TRUE)
```

```
> kappas
  stype     kappa  se.kappa
E     E 0.4827586 0.0716370
H     H 0.4285714 0.1638083
M     M 0.7491639 0.1235138
> svycontrast(kappas, c(M=1,E=-1))
         contrast     SE
contrast  0.26641 0.1575
```

Rank tests

My initial view on rank tests for survey data was that they were not well motivated: they could not be exact and would rely on the same asymptotics as a t-test. Users, however, were interested in having rank tests available, and expressed some surprise that they had not been implemented, so Alastair Scott and I worked out how to define them (Lumley & Scott, 2013). The resulting tests use the values of the estimated population cumulative distribution function as scaled ranks, and the population or superpopulation null hypothesis of the test (though not, of course, the power) is the same regardless of the sampling design. The most useful of these are probably the Wilcoxon test and the tests for specific quantiles, but users can define their own rank transformations. As is well known, the Wilcoxon test can also be derived as a score test in a proportional odds model, but this approach requires more computational effort and seems more difficult to generalise.

Weighted version of the $G^{\rho,\gamma}$ family of logrank tests for survival data have also been implemented, based on ideas of Rader (2014) but with some speed and memory optimisations.

Two-phase samples

Sampling from existing databases has become increasingly important in health research, either when new variables are being measured on existing cohorts, or when subsamples are taken for validation from electronic health record databases. Shepherd et al (2022) gives a modern example of a multi-wave twophase validation study. Two-phase sample objects can also be used to implement calibration for non-response in a single-phase sample.

The `twophase` function defines a two-phase sample object, by providing the sampling design at each phase. The resulting object can be used in all analyses, and supports calibration of phase 2 to phase 1 as well as calibration of phase 1 to the population. This example shows a two-phase case-control design as described by Breslow and Chatterjee (1999). The first phase is actually a pair of clinical trials, treated here as a simple random sample; the second phase is stratified on outcome and tumour histology. In the second analysis, the design is post-stratified on tumour stage, improving precision for a number of the parameters.

```
> data(nwtco, package="survival")
> nwtco$incc2<-
as.logical(with(nwtco, ifelse(rel | instit==2,1,rbinom(nrow(nwtco),1,.1))))
> dccs2<-
twophase(id=list(~seqno,~seqno),strata=list(NULL,~interaction(rel,instit)),
    data=nwtco, subset=~incc2)
> summary(svyglm(rel~factor(stage)*factor(histol),design=dccs2,
family=quasibinomial()))
```

```
Call:
svyglm(formula = rel ~ factor(stage) * factor(histol), design = dccs2,
    family = quasibinomial())

Survey design:
twophase2(id = id, strata = strata, probs = probs, fpc = fpc,
    subset = subset, data = data)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.6802     0.1264 -21.206  < 2e-16 ***
factor(stage)2                  0.6994     0.2004   3.490 0.000502 ***
factor(stage)3                  1.0275     0.2208   4.653 3.67e-06 ***
factor(stage)4                  0.7804     0.2449   3.187 0.001477 **
factor(histol)2                 1.2147     0.3246   3.743 0.000192 ***
factor(stage)2:factor(histol)2  0.2073     0.4589   0.452 0.651547
factor(stage)3:factor(histol)2  0.4942     0.4343   1.138 0.255383
factor(stage)4:factor(histol)2  1.0384     0.6137   1.692 0.090918 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.000901)

Number of Fisher Scoring iterations: 4
> gccs8<-calibrate(dccs2, phase=2, formula=~interaction(rel,stage,instit))
> summary(svyglm(rel~factor(stage)*factor(histol),design=gccs8,
family=quasibinomial()))

Call:
svyglm(formula = rel ~ factor(stage) * factor(histol), design = gccs8,
    family = quasibinomial())

Survey design:
calibrate(dccs2, phase = 2, formula = ~interaction(rel, stage,
    instit))

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -2.6836     0.1089 -24.647  < 2e-16 ***
factor(stage)2                  0.7706     0.1478   5.213 2.22e-07 ***
factor(stage)3                  0.7707     0.1543   4.995 6.85e-07 ***
factor(stage)4                  1.0658     0.1768   6.028 2.27e-09 ***
factor(histol)2                 1.2167     0.3226   3.771 0.000171 ***
factor(stage)2:factor(histol)2  0.1647     0.4474   0.368 0.712865
factor(stage)3:factor(histol)2  0.7037     0.4346   1.619 0.105717
factor(stage)4:factor(histol)2  0.9331     0.5760   1.620 0.105501
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.000901)

Number of Fisher Scoring iterations: 4
```

Regression models

Even in 2008, **survey** supported generalised linear models and the Cox proportional hazards model. Additions since then include loglinear models for contingency tables, the proportional odds model and other cumulative link models, and accelerated failure models for survival data. A companion package, **svyVGAM** provides an interface to the wide range of models in the **VGAM** package (Yee, 2023; Lumley 2020), including multinomial regression, negative binomial, and zero-inflated and zero-truncated Poisson.

```
> library(svyVGAM)
> dclus2<-svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)
> dclus2<-update(dclus2, mealcat=cut(meals,c(0,25,50,75,100)))
> svy_vglm(mealcat~avg.ed+mobility+ell, design=dclus2,
family=multinomial(refLevel=1))
2 - level Cluster Sampling design
With (38, 116) clusters.
update(dclus2, mealcat = cut(meals, c(0, 25, 50, 75, 100)))

Call:
vglm(formula = formula, family = family, data = surveydata, weights =
.survey.prob.weights)


Coefficients:
(Intercept):1 (Intercept):2 (Intercept):3      avg.ed:1      avg.ed:2
avg.ed:3
  16.94881271   17.10446320   13.33436684   -5.80250194   -7.26575425   -
9.08907835
   mobility:1    mobility:2    mobility:3         ell:1         ell:2
ell:3
  -0.00363934    0.16130315    0.21750911    0.10998773    0.17127858
0.38340792

Degrees of Freedom: 348 Total; 336 Residual
Residual deviance: 4368.787
Log-likelihood: -2184.394

This is a multinomial logit model with 4 levels
> svy_vglm(ordered(mealcat)~avg.ed+mobility+ell, design=dclus2,
family=propodds())
2 - level Cluster Sampling design
With (38, 116) clusters.
update(dclus2, mealcat = cut(meals, c(0, 25, 50, 75, 100)))

Call:
vglm(formula = formula, family = family, data = surveydata, weights =
.survey.prob.weights)


Coefficients:
(Intercept):1 (Intercept):2 (Intercept):3        avg.ed      mobility
ell
    7.6020176     5.0890255     1.4022044    -3.3569363     0.1117306
0.1446080

Degrees of Freedom: 348 Total; 342 Residual
Residual deviance: 5091.878
Log-likelihood: -2545.939
```

The package implements research on ‘working likelihood’ analyses of generalised linear models and the Cox model. This comes in two parts. First, the familiar Rao-Scott tests for multiway tables have been extended to these regression models. Second, there are now design-based analogues of AIC and BIC for model selection (Lumley & Scott, 2015) in generalised linear models.

```
> model0<-svyglm(I(sch.wide=="Yes")~ell+meals+mobility, design=dclus2,
family=quasibinomial())
> model1<-svyglm(I(sch.wide=="Yes")~ell+meals+mobility+as.numeric(stype),
+      design=dclus2, family=quasibinomial())
> model2<-svyglm(I(sch.wide=="Yes")~ell+meals+mobility+stype, design=dclus2,
family=quasibinomial())
> anova(model2)
```

```
Anova table:  (Rao-Scott LRT)
svyglm(formula = I(sch.wide == "Yes") ~ ell, design = dclus2,
    family = quasibinomial())
           stats     DEff       df ddf        p
ell       1.1259  0.76870  1.00000  38 0.236339
meals     4.8189  1.24181  1.00000  37 0.058326 .
mobility  0.3712  1.42335  1.00000  36 0.608418
stype    52.4054  2.43494  2.00000  34 0.001341 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model0,model2)
Working (Rao-Scott+F) LRT for stype
 in svyglm(formula = I(sch.wide == "Yes") ~ ell + meals + mobility +
    stype, design = dclus2, family = quasibinomial())
Working 2logLR =  21.52228 p= 0.0013407
(scale factors:  1.7 0.3 );  denominator df= 34
> anova(model1, model2)
Working (Rao-Scott+F) LRT for stype - as.numeric(stype)
 in svyglm(formula = I(sch.wide == "Yes") ~ ell + meals + mobility +
    stype, design = dclus2, family = quasibinomial())
Working 2logLR =  25.10744 p= 1.816e-05
df=1;  denominator df= 34
```

Predictive margins (Korn & Graubard, 1999) are also available for generalised linear models. In this example an extract from NHANES is used to estimate the prevalence of high cholesterol by race/ethnicity, standardised for age and sex. These are then transformed to prevalence ratios (relative risks) by svycontrast.

```
> data(nhanes)
> nhanes_design <- svydesign(id=~SDMVPSU, strata=~SDMVSTRA, weights=~WTMEC2YR,
nest=TRUE,data=nhanes)
> agesexmodel<-svyglm(HI_CHOL~agecat+RIAGENDR,
design=nhanes_design,family=quasibinomial)
> means<-svypredmeans(adjustmodel=agesexmodel, groupfactor= ~factor(race))
> means
      mean     SE
2 0.114596 0.0065
3 0.084718 0.0106
1 0.123081 0.0058
4 0.108770 0.0304
> ## relative risks compared to non-Hispanic white
> svycontrast(means,quote(`1`/`2`))
         nlcon     SE
contrast 1.074 0.0722
> svycontrast(means,quote(`3`/`2`))
           nlcon     SE
contrast 0.73928 0.0923
```

Multiple imputation and plausible values

Constructing multiple imputations is a specialised task for which there is other software, but with the accompanying **mitools** package the **survey** package does support the analysis of multiply-imputed data and of data with 'plausible values', as in some educational surveys. From a computational viewpoint the distinction between the two is that imputation provides multiple complete datasets whereas plausible values provide multiple alternative columns in a single dataset. In both settings, the results of multiple analyses are combined using Rubin's rules.

For plausible values, the inputs consist of an *action* to be performed on each plausible value, and a set of formulas specifying which columns go into which variables in the action. Here is an example using the New Zealand subset of the PISA 2012 educational survey (OECD, 2013).

```
> library(mitools)
> data(pisamaths, package="mitools")
> des<-svydesign(id=~SCHOOLID+STIDSTD, strata=~STRATUM, nest=TRUE,
+     weights=~W_FSCHWT+condwt, data=pisamaths)
> results<-withPV(list(maths~PV1MATH+PV2MATH+PV3MATH+PV4MATH+PV5MATH),
+    data=des,
+    action=quote(svyglm(maths~ST04Q01*(PCGIRLS+SMRATIO)+MATHEFF+OPENPS,
design=des))
+    )
> MIcombine(results)
Multiple imputation results:
      withPV.survey.design(list(maths ~ PV1MATH + PV2MATH + PV3MATH +
    PV4MATH + PV5MATH), data = des, action = quote(svyglm(maths ~
    ST04Q01 * (PCGIRLS + SMRATIO) + MATHEFF + OPENPS, design = des)))
      MIcombine.default(results)
                          results         se
(Intercept)          4.729436e+02 20.7907335
ST04Q01Male          5.268461e+01 24.1976184
PCGIRLS              5.974293e+01 17.6683218
SMRATIO              3.552268e-02  0.1233799
MATHEFF              4.736517e+01  3.0904746
OPENPS               1.317289e+01  3.0565353
ST04Q01Male:PCGIRLS -1.109811e+02 32.8851005
ST04Q01Male:SMRATIO  4.391909e-03  0.1358470
```

the action is to fit a design-weighted linear model for maths performance with predictors gender, percentage of girls in the school, student:teacher ratio, and two attitude questions. The response variable maths is not a variable in the data set; instead, the first argument of withPV says that the action should be performed five times, with each of the actual maths score variables substituted for maths.

For multiple imputation, the function `imputationList` wraps a set of imputed complete data sets so they can be used to create a packaged list of design objects that can then be used for analyses. Again, `MIcombine` is used to combine the sets of results.

Your own extensions

Users will inevitably need some estimators that are not already implemented. Two tools for extending the package are `withReplicates` and `svycontrast`.

When the estimator can be expressed as an explicit function of quantities that can already be estimated, svycontrast will compute standard errors using either the delta method (with symbolic differentiation) or replicate weights. A couple of examples have already been seen above. Here is a slightly more complicated example. For each school in a sample, we have the number of enrolled students, and we wish to calculate the fraction of students in elementary, middle, and high schools. The code first computes the estimated population total, then divides the domain totals by it.

```
> totals<-svyby(~enroll,~stype,design=dclus1,svytotal,covmat=TRUE)
> totals<-svycontrast(totals, list(total=quote(E+M+H)), add=TRUE)
> svycontrast(totals, list(quote(E/total), quote(M/total), quote(H/total)))
      nlcon     SE
[1,] 0.6196 0.0657
[2,] 0.2231 0.0412
[3,] 0.1573 0.0431
```

The `withReplicates` function is useful when you already have code to compute point estimates, but not standard errors. This could be because you have written the code, or it could be because there is existing code in some other package that accepts precision weights or frequency weights and gives correct point estimates. The `withReplicates` function provides a simple interface to run this existing code using replicate weights and compute a variance estimate. For example, here we fit a median regression using code that expects precision weights, and extract the coefficient

estimates. Re-running this code with 100 sets of bootstrap replicate weights allows for design-based standard error computations.

```
> library(quantreg)
> data(api)
> ## one-stage cluster sample
> dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
> ## convert to bootstrap
> bclus1<-as.svrepdesign(dclus1,type="bootstrap", replicates=100)
> ## median regression
> withReplicates(bclus1, quote(coef(rq(api00~api99, tau=0.5,
weights=.weights))))
               theta      SE
(Intercept) 87.78505 16.6070
api99        0.91589  0.0237
```

To provide more examples of how to add new estimators to the package, I wrote a post about different ways to implement zero-inflated Poisson regression (Lumley, 2015).

The future

There are two major upgrades planned in the short term. The first is to speed up the basic multistage variance computation using C++ code contributed by Ben Schneider (2022). The relative speed gains are greatest for small data sets, but the package will be faster for surveys of all sizes. The second major upgrade is to add an interface to small-area estimation code developed by Richard Li, Jon Wakefield and co-workers (Li et al, 2022).

I also plan to add explicit support for multi-phase and multi-frame samples. Multi-phase sampling is of increasing interest in health research, with validation subsampling of large existing databases, and there is no real barrier to extending from the current two phases to three or more. Multi-frame sampling has well-developed methods and the main task in implementation is developing an appropriate user interface.

Mixed models are an area of interest for future development. It is surprisingly difficult to estimate these models under complex sampling, especially when the model structure is not aligned with the sampling structure, but I plan to implement methods based on pairwise likelihood (Yi et al 2016; Huang, 2019).

Other developments in the future will depend in part on what users ask for. Many changes to the package over the years have been responses to user feedback. Sometimes this simply meant fixing bugs; other times it needed new coding or even new methods research.

I will end by noting that the **survey** package does not have any dedicated external or internal funding for either development or maintenance, though at some points it has been supported indirectly by statistical methods research funding (from the Marsden Fund of the Royal Society of New Zealand). It is not clear that this situation is sustainable indefinitely.

References

Breslow N.E. and Chatterjee N. (1999), Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis, *Applied Statistics,* 48:457-68.

Graubard B., Korn E. (1999), Predictive Margins with Survey Data, *Biometrics,* 55:652-659.

Huang X. (2019), *Mixed Models for Complex Survey Data* PhD thesis, University of Auckland.

Iannone R., Cheng J., Schloerke B., Hughes E., Lauer A. and Seo J. (2023), *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.9.0. https://CRAN.R-project.org/package=gt

Li Z.R., Martin B.D., Hsiao Y., Godwin J., Paige J., Gao P., Wakefield J., Clark S.J., Fuglstad G-A, and Riebler A. (2022), *SUMMER: Small-Area-Estimation Unit/Area Models and Methods for Estimation in R*. R package version 1.3.0. https://CRAN.R-project.org/package=SUMMER

Lumley T. (2015), *Zero-inflated Poisson from complex samples.* https://notstatschat.rbind.io/2015/05/26/zero-inflated-poisson-from-complex-samples/

Lumley T. (2020), *MOAR survey regression models,* https://notstatschat.rbind.io/2020/09/24/moar-survey-regression-models/

Lumley T., and Scott A.J. (2013), Two-sample rank tests under complex sampling, *Biometrika*, 100 (4), 831-842.

Lumley T., and Scott A.J. (2015), AIC and BIC for modelling with complex survey data, *Journal of Survey Statistics and Methodology,* 3 (1): 1-18.

OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing.

Rader KA (2014), *Methods for Analyzing Survival and Binary Data in Complex Surveys.* Doctoral dissertation, Harvard University. http://nrs.harvard.edu/urn-3:HUL.InstRepos:12274283

Schneider B (2022), *Practical Significance: More on Speeding up the Survey Package: Adding the New C++ Functions to the Package.* https://www.practicalsignificance.com/posts/adding-rcpp-to-the-survey-package/

Shepherd B.E., Han K., Chen T., Bian A., Pugh S., Duda S.N., Lumley T., Heerman W.J., and Shaw P.A. (2022), Multiwave validation sampling for error-prone electronic health records, *Biometrics*. doi: 10.1111/biom.13713. Epub ahead of print. PMID: 35775996

Yee, T. (2023), *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-8. https://CRAN.R-project.org/package=VGAM

Yi G.Y., Rao J.N.K., and Li H. (2016), A weighted composite likelihood approach for analysis of survey data under two-level models, *Statistica Sinica,* 26(2) 569-587.

ARGENTINA

Reporting: **Verónica Beritich**

INDEC presents the new update of the Integrated System of Social Statistics

On January 5, 2023, the National Institute of Statistics and Censuses (INDEC) presented an update of the Integrated System of Social Statistics (SIES). It is a dynamic platform that provides statistical information on current and future well-being in our country from a multidimensional perspective that includes material conditions, quality of life, and sustainability over time.

With respect to the initial version, published on July 6, 2022, the "Summary" section is added, which includes the temporal evolution of each dimension at the country level, and a comparative analysis of the well-being indicators per jurisdiction within the "Here and Now" subsection. Four new context indicators are also included in the Income and Employment dimensions.

Through its interactive visualization, the SIES presents a federal approach to the state of the welfare situation from the capabilities approach in Argentina. In this way, the possibility of selecting 86 indicators and obtaining personalized charts and data files in CSV or XLS format is provided.

The visualization system, developed by the team of the Directorate of Sectoral Statistics dependent on the National Directorate of Social and Population Statistics, is in line with the multidimensional approach of measuring well-being [1] which, at the same time, is articulated with the Sustainable Development Goals of the United Nations Organization.

Within the new Summary section, in the "Recent evaluation" option, the data of the last 5 years are considered. There is a comparison between each end of the series, and three colors are used in the graphic to reveal whether their situations have improved (green), worsened (red), or has not changed (yellow) along the time. For these calculations, a hypothesis test (t-test) is used to evaluate the difference of means when data comes from a survey; and an interval of +/- 0.5% of percentage variation is established for data coming from records and censuses.

In "Temporal evolution", a disaggregated visualization of the indicators is offered, grouped into the dimensions of housing, employment, education and skills, state of health, civic commitment, and human capital, which continues with the system of colors and labels proposed in "Recent evaluation".

Finally, in "Here and now" the distribution of the indicators is presented per jurisdiction based on the median absolute deviation of all jurisdictions (see more in SIES>About SIES>General considerations), by means of an interactive bubble chart.

[1] Stiglitz, J., Fitoussi, J.P. and Durand, M. (2018). Beyond GDP. Measuring What Counts for Economic and Social Performance, https://doi.org/10.1787/9789264307292-en.

The Integrated System of Social Statistics can be found at https://shiny.indec.gob.ar/sies

General information can be found at www.indec.gob.ar.

For further information, please contact https://www.indec.gob.ar/indec/web/Institucional-Indec-Contacto.

AUSTRALIA

Reporting: **Emma Farrell**

Introducing a new statistical data collection mode - Video Assisted Live Interviewing

Video-Assisted Live Interviewing (VALI) is data collection conducted online using a video conferencing platform. Following extensive success with video-interviewing for questionnaire testing, the Australian Bureau of Statistics (ABS) is now pursuing VALI for quantitative data collection. As a replacement for face-to-face household survey interviews, the mode reduces nonresponse in a range of situations (e.g., pandemic related reasons). It also shows potential to improve data collection efficiency, reduce costs, enhance interviewer safety, and reduces respondent burden.

Comprehensive research was undertaken to develop the video interviewing process prior to offering this option to live survey respondents. This included conducting a feasibility study of mode preferences using a commercial online panel, sentiment testing with respondents following traditional ABS face to face interviews, and multiple usability testing rounds with ABS field interviewers and various respondent cohorts. The live pilot study involved offering the mode in the later stages of a large health survey, and VALI interviews have been successfully completed for over 400 respondents. Debriefings were conducted with respondents and interviewers at the beginning of the pilot.

Feedback from tests and debriefings are that respondents:

- are positive about the mode, and many had used video calls previously for other purposes.
- appreciated being interviewed from a private location where they felt most comfortable. For example, in a bedroom or study when otherwise they would have met with an interviewer in a main room of the house.
- liked and felt more at ease with the physical separation that video provided when talking about sensitive content.
- appreciated being able to see survey prompt cards online during the interview, and would have liked more to be shown on the cards to more easily comprehend complex questions.

The ABS is now collaborating with the Social Research Centre to conduct an experiment comparing data quality between VALI, online and telephone collection modes.

Contact: Emma Farrell (emma.farrell@abs.gov.au)

CANADA

Reporting: **Jean-François Naud**

Canadian Covid-19 Antibody and Health Survey – Cycle 2

Statistics Canada, in partnership with the Public Health Agency of Canada and Canada's COVID-19 Immunity Task Force, conducted the second cycle of the Canadian COVID-19 Antibody and Health Survey (CCAHS-2) to better understand the spread of SARS-CoV-2, the virus that causes COVID-19, and the longer-term impacts of COVID-19 in Canadian adults. The survey data was collected in three waves from April to August 2022, from a sample of more than 100,000 Canadians aged 18+. Selected individuals were asked to complete a short questionnaire and to provide a dry blood sample (DBS) to be sent by mail to Statistics Canada. Respondents from Waves 2 and 3 were also asked to provide a saliva sample and send it along with the DBS. The saliva samples were used to detect

current and recent infections through a polymerase chain reaction (PCR) test, while the DBS were used to measure immunity by detecting antibodies from past infection or vaccination.

The laboratory tests measure the concentration of three antibodies. For a sample to be considered positive from either vaccination or infection, at least two of the three antibodies must be detected in the blood sample. For a sample to be considered positive from a past infection, the nucleocapsid antibody must be one of the two positive antibodies. The CCAHS-2 was collected during a period when a very high proportion of the population had been vaccinated, and other studies showed that vaccinated individuals generate fewer nucleocapsid antibodies following an infection. Therefore, determining whether an individual had been previously infected or not was not as clear as it was for CCAHS-1, which was collected at the start of the vaccination roll-out.

To take this uncertainty into account, instead of using the Rogan-Gladen estimator, a modelled probability of having been infected was used. The model was created in partnership with the University of Ottawa and Sinai Health, which were the institutions responsible for the CCAHS-2 DBS laboratory testing. The probabilities were modelled using data from a baseline cohort followed from a point in time after the vaccine roll-out through to the initial onset of the Omicron variant. From these probabilities, weighted estimates of the prevalence of previously infected persons were produced.

Furthermore, given the increased uncertainty caused by the high proportion of vaccinated persons and different conditions in the two labs, some systematic differences in the measures from each were observed. These differences were perfectly acceptable from a microbiology point of view but could have impacted statistical estimates from the survey. A cross-over study was done to put the lab measures to the same scale. A random sample of three hundred specimens collected by CCAHS-2 that had been tested by one lab were sent to the other for re-testing and vice-versa. The measured concentrations from these six hundred specimens were used to create a regression model, from which adjusted concentrations were derived. The probability models described above for the nucleocapsid were derived based on these adjusted concentrations.

For more information about the CCAHS-2 methodology, please contact jean-francois.naud@statcan.gc.ca

CROATIA

Reporting: **Ksenija Dumičić**

The Croatian Bureau of Statistics completed the first “EU survey on gender-based violence against women and other forms of inter-personal violence” (EU-GBV)

The Croatian Bureau of Statistics (CBS) launched the “EU survey on gender-based violence against women and other forms of inter-personal violence” (EU-GBV), which is being conducted for the first time in the Republic of Croatia, with the goal to collect statistical data that will answer the question to what extent residents feel safe in the environment in which they live and work. These data will help in providing the necessary information for creating social and security policies, in various scientific analyses and international comparisons, and in informing the general public about the state of personal security.

In April 2023, the CBS completed, in cooperation with the research agency Ipsos, West Adria office from Zagreb, the EU-granted survey EU-GBV, which followed the Eurostat harmonized methodology, as given at https://ec.europa.eu/eurostat/documents/3859598/13484289/KS-GQ-21-009-EN-N.pdf/1478786c-5fb3-fe31-d759-7bbe0e9066ad?t=1633004533458.

Prior to that, the pilot survey took place from 1 October 2017 to 30 September 2019, as described at

https://dzs.gov.hr/UserDocsImages/dokumenti/Zavr%C5%A1eni%20projekti%20financirani%20iz%20EU.pdf (p 91-92).

The national EU-GBV survey was conducted from 1 December 2020 to 30 April 2023, https://dzs.gov.hr/vijesti/pocinje-provedba-ankete-o-sigurnosti/1281. It used slightly adjusted survey questions, https://dzs.gov.hr/UserDocsImages/dokumenti/Dokumenti/Projekti%20u%20tijeku%20financirani%20iz%20EU.pdf (p 16.).

The national sample survey was planned in two steps. In the first step, until 31 October 2022, self-administered electronically (online) filled questionnaires were applied. For those that did not fill them, in the second step, from 1 November to 15 February 2023, Ipsos applied CATI.

The sampling frame included individuals aged 18 to 74 in private HH and private apartments according to Census 2021, as found on 31 August 2022. A sample of 42,977 people was selected from the frame (Census 2021), and 22,695 of them were matched with the phone book by address and surname (for fixed phones) or by address, surname, and first name (for mobile phones). Eight strata were defined by statistical regions from the year 2021 (NUTS-2) and by the designation of whether it is an urban settlement or a settlement belonging to the other category.

Stratified systematic sampling with the implicit gender and age-based strata, and the sample allocation proportional to the number of persons in the stratum based on Census 2021 was designed. Only the sample size in the biggest stratum for “City of Zagreb – other” was increased. Documents available at: https://dzs.gov.hr/UserDocsImages/dokumenti/Dokumenti/Projekti%20u%20tijeku%20financirani%20iz%20EU.pdf (p 16.); https://dzs.gov.hr/vijesti/pocinje-provedba-ankete-o-sigurnosti/1281.

All 22,695 respondents received a letter with a link to the questionnaire and an invitation to complete it online, and 2,300 of them did so within the deadline set for the first step. After that, it was switched to CATI, but the option was left for people to fill out the questionnaire online themselves. In the end, 4,112 questionnaires were filled out online, and 2,059 by telephone (CATI).

DENMARK

Reporting: **Martine Friisenbach and Lotte Yssing Jacobsen**

A qualitative look at the respondent journey

Over the past 10 years, more and more respondents complete surveys via their smartphones. At the same time, the response rate on surveys has also been slowly decreasing. In 2022 roughly 80 pct. of surveys were completed via smartphone and the average response rate on surveys were approximately 45 pct. using mix-mode CAWI and CATI. We therefore wanted to take initiatives to improve the respondent experience and raise the response rates. From our prior studies of the invitation letter and the use of incentives using split sample testing on different surveys, we gained valuable knowledge about our communication with the respondents. However, we wanted more knowledge about the perspective of the individual respondents and what works and does not work in the current context. Therefore, we decided to work with a private company that conducts in-depth interviews.

Test set-up

The respondent journey from the first contact with the respondent to the completion of a survey and how we “leave” the respondent was described in detail. This gave us the opportunity to identify the direct contact points we have with the respondents and where we can improve the respondents’ experience.

The key points of contact are:

- The invitation (digital post or letter in mail)
- The respondent starting the survey (via direct link)
- The respondent starting the survey via special webpage
- The questionnaire – layout and functionality
- Landing page and how we thank them for their time

A total of 18 respondents were interviewed. Main recruitment criteria were age, family status, education and presumed IT-capabilities.

Test results

As expected, the smartphone is the preferred device to participate in a survey. Respondents aged 75+ also preferred the smartphone, whereas we had expected this age group to use the PC to a greater extent. One respondent aged 75+ commented that she had not used a mouse for a PC in more than 10 years. The preferred device is in general determined by the respondents' expected length of time to participate in the survey. The longer the questionnaire, the more likely the use of an iPad or PC.

Invitations to the survey must be short and easy to read so the respondents can decode the information quickly. This also applies to the layout of the questionnaire; the easier the respondents can get an overview of the question(s) presented, the more likely they are to continue to answer the survey. The focus among the younger respondents is to feel effective while answering. For the older respondents it is important to make them feel that it is easy to answer the questions and progress in the survey. The younger respondents are highly motivated by the subject whereas the older respondents are more likely to participate by the feeling of helping the survey sender.

Actions taken

As a result of the test, we have made changes in the letter of invitation, the questionnaire layout, the closing words after last question is answered and the landing page. The test emphasized the 'mobile first' mindset, which entails adaptions of questions, shortening of introduction to questions and the general functionality of the questionnaire. We have still in favour to see the effect on response rates which can be a bit difficult to measure since due to other changing circumstances.

FIJI

Reporting: **M.G.M. Khan**

Balance of Payment

The Balance of Payment (BOP) unit is working closely with the Reserve Bank of Fiji (RBF) in reviewing and updating the outdated Overseas Exchange Transaction (OET) codes. More codes were proposed to the team for implementation in order to meet the data needs of the external sector statistics compilation.

In addition, the BOP unit is updating the International Investment Survey frame to improve the recording of liabilities in the financial accounts and to reduce/minimize the errors and omissions in the balance of payment statistics. A technical mission is planned later this year by the International Monetary Fund (IMF) to review all accounts and provide further support and expert advice for improving the accounts.

The unit is also planning to implement the latest International Visitors Survey (IVS) Report for the years 2016 to 2019 released by the Ministry of Tourism and Civil Aviation in the Fiji's Earning's from Tourism. The IVS reports are used to calculate the Per-diem rates for the tourism earnings release.

Contact persons: Mr. Shonal Deo shonal.deo@statsfiji.gov.fj and Ms. Torika Ketenilagi tketenilagi@statsfiji.gov.fj

Civil Registration & Vital Statistics (CRVS) Inequality Assessment in Fiji

The objective of the assessment was to ensure no one is left behind; further investigation is needed to understand whose vital events are least likely to be registered. Fiji is one of the first countries, to our knowledge, to embark on an in-depth quantitative inequality assessment, examining differentials by sex, age, ethnicity and mother's marital status. Assessment results will be used to inform future research and policy interventions and to bridge gaps in registration between different populations in the country.

The work was primarily led by the Fiji Bureau of Statistics in coordination with other key stakeholders, including the Ministry of Justice and the Ministry of Health and Medical Services and other key stakeholders with technical support from UNESCAP and International Consultant Ms. Renee Sorchik.

The report is expected to be released by the end of Quarter 2, 2023.

Contact person: Ms. Amelia Tungi ameliat@statsfiji.gov.fj and Mr. Meli Nadakuca mnadakuca@statsfiji.gov.fj

Developing Vital Statistics Indicators and Assessing Completeness and Inequalities in the Registration of Births and Deaths

Over the past year, the Fiji Bureau of Statistics worked with the Ministry of Health and Medical Services and the Ministry of Justice to produce a Vital Statistics Report for Fiji covering the period 2016 to 2021.

This project was financially and technically supported by Vital Strategies (Bloomberg Philanthropies – Data for Health Initiative) and the University of New South Wales, Senior Lecturer Dr. Christine Linhart.

The report contains important information regarding levels and trends in fertility and mortality across the country, including cause-specific mortality, life expectancy, and excess mortality. This report is expected to be released by the end of Quarter 2, 2023.

Contact persons: Ms. Amelia Tungi ameliat@statsfiji.gov.fj and Mr. Meli Nadakuca mnadakuca@statsfiji.gov.fj

Household Surveys

Over the years, the Household Survey Division (HSD) has embarked on new innovations of data collection such as the transition from PAPI to CAPI in the 2017 Population and Housing Census. The division embarked on its first survey using a CATI (Computer Assisted Telephone Interview) system for the Pre-Screening phase of the Secured Transaction Reform Impact Evaluation (STRIE) funded by the Asian Development Bank focused on Small Micro Medium businesses. The STRIE survey is scheduled to conclude by May 2023. With the support of the United Nations Office for Drugs and Crime (UNODC), the division also conducted the first National Trafficking in Persons Prevalence Survey and was involved with the data processing and analysis using the Network Scale-Up Method (NSUM). Lastly, as per national surveys every five years, the division is preparing the Household Labour Force Survey or the Employment Unemployment Survey (EUS) 2023-24, the first EUS conducted using CAPI with technical support from the International Labour Organisation (ILO), with collection from August 2023 to July 2024.

Contact person: Ms. Salanieta Soli salanietas@statsfiji.gov.fj

FRANCE

Reporting: **Philippe Brion**

The French Health Barometer and its adaptation during the Covid-19 pandemic

The French Health Barometer is a repeated health telephone survey that has been conducted in the general adult French population since 1992 by Santé publique France, the national public health agency, allowing trends in health risk behaviours to be measured.

This survey covers various topics, including: tobacco, alcohol and drug consumption, vaccination, sexual practices, nutrition, physical activity, mental health, etc. The survey design is a random digit dialing sample of landline and cell phones. Interviews are conducted using computer-assisted telephone interviewing, of people living in France, non-institutionalized, and who speak French. Participation is anonymous and voluntary.

More elements on the Health Barometer may be found (in French) at:

https://www.santepubliquefrance.fr/etudes-et-enquetes/barometres-de-sante-publique-france#block-65435

In 2020, the survey had to adapt after the Covid-19 pandemic hit. First, the survey collection relied on interviewers who were gathered at call centres: lockdown made impossible this way of working and the survey collection had to stop. Second, the pursued objectives could not remain the same in this extra-ordinary period, especially regarding health attitudes in a pandemic context. It was then decided to distinguish data collected before and after the first French lockdown, which took place from March 17th to May 10th. The data collected from January 8th to March 16th (before the lockdown) were used to produce the usual indicators. Due to the unexpected halt to the data collection, the sample size was reduced, leading these indicators to be produced only at a national level, and not at a regional level as initially intended. In a second step, a new survey was launched on June 4th, just after the lockdown was finished and featuring a different questionnaire, to obtain timely information on the impact of the pandemic on the health behaviours and on mental health, and also to gather information on the spread of the epidemic itself. This was made possible by the quick development of solutions for the interviewers to work from home, as gathering in confined spaces was not yet permitted.

More elements on this survey may be found (in French) at:

https://www.santepubliquefrance.fr/etudes-et-enquetes/barometres-de-sante-publique-france/barometre-sante-2020

Contact: Noémie Soullier (Noemie.SOULLIER@santepubliquefrance.fr)

KENYA

Reporting: **David I. Ojakaa**

Kenya Demographic and Health Survey

This report recapitulates the technical processes of the 2022 Kenya Demographic and Health Survey (KDHS), a key and quinquennially recurring sample survey in Kenya as in many developing countries. The intent is to shine the spotlight on innovations/new approaches. The survey neared completion with the release of the key indicators report (KIR) at the end of January 2023; the complete document will be shared later in 2023. Notable transitions were implemented in the survey

processes of sample design, questionnaire development, anthropometric measurement, training, fieldwork, and data processing.

The sampling frame for the survey, baptized the Kenya household master sampling frame (KHMSF) in the current phase, was developed from the 129,067 enumeration areas (EAs) of the 2019 Kenya Population and Housing census. Out of these, 10,000 EAs were selected for the KHMSF. The EAs were then transformed into clusters through household listing and geo-referencing. In a subsequent step, the 47 Kenya counties were stratified into urban and rural areas to yield 92 strata, Nairobi and Mombasa counties both being urban. A two-stage stratified design was applied; in the first step a total of 1,692 clusters being selected through equal probability sampling. In the second stage 25 households were systematically selected in each sampled cluster. This yielded a total of 32,156 interviews among women aged 15-49 years (a quantum leap of 281% from the 2008 survey, and a stabilizing 3.5% increase from the previous, 2014, survey), and 14,453 with men in the 15-54 age range.

To collect the data, eight questionnaires were used: the full and short household questionnaire; the full and short woman's questionnaire; the man's questionnaire; the full and short biomarker questionnaire; the fieldworker questionnaire. The separation into full and short questionnaires aimed at reducing the duration of fieldwork, as well as interviewer and respondent fatigue. To determine nutritional status using anthropometric measures, the weights and heights of children under age five, women aged 15–49, and men aged 15–54 were taken using precision Seca digital scales and Shorr boards respectively.

Training, to ensure data quality, consisted of three steps – training of trainers (TOTs), the pre-test, and field-staff training. This process resulted in a total of 45 trainers and 314 personnel participating in the master, and pre-test cum fieldworker sessions respectively. Collection of data was accomplished by 48 teams categorized mainly by local languages. Every team comprised a supervisor, biomarker technician, three female interviewers, a male interviewer, and a driver. Data collection involved computer-assisted personal interviewing (CAPI), specifically Android computer tablets programmed with CSpro software.

Note: The views expressed here are those of the author solely and not of KNBS nor of the DHS program.

More information on the survey can be obtained from: directorgeneral@knbs.or.ke; archive@dhsprogram.com. The key indicator report (KIR) can be accessed through the following link: https://www.knbs.or.ke/download/2022-kdhs-key-indicators-report/

THE NETHERLANDS

Reporting: **Deirdre Giesen**

Two programs to increase effectiveness and efficiency

Statistics Netherlands has set up two programs that aim to make statistical processing more effective and efficient. Their common goal is to free up space for innovation. EBN2.x is in the economic division, where EBN is the Dutch acronym for the Division of Economic and Business statistics and National account. KERS is in the social statistics division, where KERS stands for Chain Efficiency Registers Socioeconomic and spatial statistics. The programs involve using office-wide standardized tooling and working methods to enable reusing data, tools, and methods across production systems.

Two of the principles of the renewal program EBN2.x are: we share all our data from the start, and we centrally manage all our (population) frameworks, which are the basis of our statistics. For KERS the aim is to free up time (by realizing efficiency gains) in the processing of registers. Based on best

practices, 'KERS principles' and a 'KERS standard process' promote more flexible and agile processing, more sharing of data, knowledge, ICT and methods. Of the 134 registers that are in scope of the program, 34 are already in progress of implementing the new way of working. Initial results show that the statistical departments are enthusiastic about the new way of working and there will be a gain in maintainability.

For advice and guidance, each of KERS and EBN 2.x has a core team including tool specialists, methodologists, business analysts and representatives from the statistical departments. The duration of both EBN 2.x and KERS is until the end of 2025. For more information on EBN2.x please contact program manager Anita Vaasen amvj.vaasen-otten@cbs.nl ; for KERS you can contact program manager Elia Bleuten e.bleuten@cbs.nl

Official statistics on mobility trends in the Netherlands for small domains using multilevel time series models

The longstanding Dutch Travel Survey (DTS) aims to produce reliable estimates on mobility of the Dutch population on an annual frequency. A multilevel time series model serves to estimate mobility trends at several aggregation levels. The models account for discontinuities induced by three survey redesigns, outliers due to less reliable outcomes in one particular year and the effect of the COVID-19 crisis on mobility. The input for the model is a set of direct annual estimates with their standard errors for the period 1999–2021 for about 700 domains cross-classified using gender, age, transportation purpose and transportation mode. As a result, the model can be considered as a multivariate time series extension of the Fay-Herriot model. The model structure is predominantly based on the required output tables, which implies that temporal and cross-sectional components are included at different aggregation levels. To reduce the risk of overfitting, many effects including discontinuities and COVID-19 effects are modelled as random effects. Using Laplace and Horseshoe distributions, a regularization method employing non-normally distributed random effects both suppresses noisy model coefficients and allows large effects sufficiently supported by the data.

Appropriate transformations for the direct estimates and generalized variance functions to smooth the standard errors of the direct estimates are developed for better model fits. The models are fitted in an hierarchical Bayesian framework using MCMC simulations. Smooth trend estimates are computed at the most detailed domain level. Predictions at higher aggregation levels obtained by aggregation of the most detailed domain predictions result in a numerically consistent set of trend estimates for all target variables, that have been published recently by Statistics Netherlands as official statistics.

Contact persons: Harm Jan Boonstra (hbta@cbs.nl) and Jan van den Brakel jbrl@cbs.nl

Boonstra, H.J. and J.A. van den Brakel. (2022), Multilevel time series models for small area estimation at different frequencies and domain levels. Annals of Applied Statistics. Vol. 16, No. 4, pp 2314-2338

Boonstra, H.J., J.A. van den Brakel and S. Das (2021). Multilevel time series modeling of mobility trends. Journal of the Royal Statistical Society A series. Vol 184, pp. 985-1007

PERU

Reporting: **Leonor Laguna**

Implementation of a Microdata System

El Instituto Nacional de Estadística e Informática (INEI) is in charge of all the statistics produced in Peru. The INEI is currently organizing a System of Microdata for the promotion and diffusion of the research that is carried out.

This system offers both the database and the relevant documentation of the surveys and census carried out by the INEI in recent years.

One of the advantages of this system is to facilitate the research, identification and recovery of the information and documentation of the surveys and census that the INEI carries out. The users can also obtain the information and documentation of the surveys and census in popularly used formats and of wide publication in the market (SPSS, Microsoft Excel, Acrobat Reader).

You may enter to the system of data by following this link: https://proyectos.inei.gob.pe/microdatos/.

The system of data is open to the public in general.

POLAND

Reporting: **Tomasz Żądło**

Poland ranked 2nd in the Open Data Inventory ranking

The Open Data Inventory assesses the coverage and openness of official statistics. In the current ranking (https://odin.opendatawatch.com/Report/rankings) updated March 9, 2023, Poland was ranked 2nd among 192 countries after Singapore and before Finland.

Currently, Statistics Poland gives the open access to the following databases:

- Regional Atlas - a map module that allows the spatial visualization of data concerning regional and local economy (available in English: http://swaid.stat.gov.pl/EN/SitePagesDBW/AtlasRegionow.aspx),

- Local Data Bank - Poland's largest database of information on the economy, society, and environment (available in English: https://bdl.stat.gov.pl/bdl/start),

- Macroeconomic Data Bank - a statistical database providing access to a long-time series of basic macroeconomic indicators (available in English: https://bdm.stat.gov.pl),

- Data Bank Poland - a repository that collects historical data from the Polish official statistics system. The time series, depending on the category, begins in 1946 and ends in 1999 (available in Polish: https://bdp.stat.gov.pl/),

- Polish organisations and institutions abroad and Polish diaspora organisations and institutions database (available in English as xlsx file at the bottom of the page: https://stat.gov.pl/en/topics/population/poles-and-polish-community-abroad/the-polish-organisations-and-institutions-abroad-and-polish-diaspora-organisations-and-institutions-database,1,2.html),

- Decompositions – a database which presents methodology and analyses results that aim to identify regional discrepancies in the degree of economic development, dividing them into subcomponents crucial for socio-economic policy intervention (available in Polish but automatic Google Translation of the website works properly: https://dekompozycje.stat.gov.pl/),

- Database Demographics – a source of statistics about population status and structure, natural movement, and migration (available in Polish: https://demografia.stat.gov.pl),

- Knowledge Databases – it focuses on giving detailed information about 31 different topics, such as Prices, Demography, Education, Public finances, Business and consumer tendency, Science and technology, Labour Market (available in English: https://dbw.stat.gov.pl/en),

- Foreign trade – it provides information on international trade in goods by countries and goods, as well as international trade in services by trading partners (available in English: http://swaid.stat.gov.pl/EN/SitePagesDBW/HandelZagraniczny.aspx),

- Sustainable Development Goals – information on over 120 indicators concerning implementation of the Sustainable Development Goals in Poland (available in English: https://sdg.gov.pl/en/),

- Public Service Monitoring System – it provides local government units, businesses, and the society with the information necessary to comprehensively evaluate services provided at the local level (available in Polish but automatic Google Translation of the website works properly: https://smup.gov.pl/),

- Statistical Handbook of Local Government – a system created by public statistics to meet the growing information needs of local government units corresponding to their tasks (available in Polish but automatic Google Translation of the website works properly: https://svs.stat.gov.pl/),

- Strateg – a system, updated at least once a quarter, supporting the process of monitoring development and evaluating the effects of actions taken to strengthen social cohesion (available in English: https://strateg.stat.gov.pl).

SWITZERLAND

Reporting: **Georg Lutz and Alina Matei**

linkhub.ch: a project for promoting data linking

The linkhub.ch project aim is twofold: first, to provide the creation of a legal and institutional environment that supports academic and administrative research based on data linking while respecting data security, and second, to ensure dissemination of knowledge about data linking. Created in 2019, the linkhub.ch project is the result of collaboration between several institutions (FORS, TREE - University of Berne, Swiss Network on Fiscal Federalism - University of Basel, Swiss RDL - University of Berne, NCCR on the move - University of Geneva). The project is held by FORS, the Swiss Centre of Expertise in the Social Sciences, hosted at the University of Lausanne and funded by the Swiss National Science Foundation.

A data link usually requires two different data sources. Linking data requires specific procedures because: first, linking datasets is not possible without identifying information, and second, the sensitivity of data and, hence, the potential harm may increase once data is linked.

The Swiss Federal Statistical Office (SFSO) carries out the linking of datasets with at least one dataset coming from the Swiss Federal Administration. Researchers can submit demands for data linking to SFSO upon some conditions: they must sign a data contract that limits the usage of data to a maximum of five years, and data has to be destroyed after the end of a project.

The Federal Statistics Act provides the legal basis for the linkage of data for statistical purposes. However, the linkage of research data and private data is not regulated as for the administrative one. There are no standards or even principles governing how such data should be used for research. Moreover, Switzerland does not have a general strategy to promote and facilitate open data access for research. In its report from 2020, the linkhub.ch project proposes options and suggestions for changes to the legal framework that would facilitate access to data and data linkage. Since the Federal Statistical Act already has a provision on linking data, the current proposal of the linkhub.ch project is to encourage the establishment of a legal framework to make data available to third parties for research purposes in a broader sense, which currently fall outside the Statistical System.

Given that the secondary use of data is becoming an increasingly important topic in many areas, e.g., also around the discussion at the European Union level around data spaces, Switzerland is now also moving towards creating a new legal framework law for the secondary use of data in order to establish a clear legal basis how to use, link and access data from different sources. After approval of the national parliament to work on such a law in June 2023, the federal administration will start the preparatory work in autumn 2023. Linkhub.ch will actively participate in this preparatory phase in order to highlight the importance of the law and the needs of the research community to have good access to existing data within an established legal and institutional framework.

More information: https://linkhub.ch/about/

Contact : https://linkhub.ch/contact/

URUGUAY

Reporting: **Diego Aboal**

Innovations in Uruguay's 2023 Census

Uruguay is performing the 2023 Population, Households and Housing Census (2023 PHHC). It presents innovative features for the Latin America and Caribbean region.

Firstly, the completely digital 2023 PHHC census will take place in two phases: a web phase where households will be able to self-register, and a face-to-face phase with electronic recording of data through 8-inch tablets. Unlike previous censuses, for the 2023 PHHC self-registration is intended as a crucial census strategy and not only as a marginal or recovery strategy. In the region, the only known successful web census, in terms of coverage, is the 2022 Argentine PHHC.

Secondly, in parallel, a census based on administrative records will be conducted. It is the first time that a country, in the region, has carried out such a pilot. It will allow comparisons between a traditional census and a register-based census for some basic variables. In this manner, gaps could be assessed taken into consideration the national plan for moving to a complete register- based census in the future. During the last years, Uruguay has been improving the quality and integration of administrative records. In this sense, Nordic countries are examples that are being followed by Uruguay.

Thirdly, the Uruguayan census is making strong use of administrative records in the pre-census phase. Early results indicate that the correction to office work carried out on satellite images of the census areas will be less than 1%. On the other hand, the count of addresses in the field is yielding similar numbers to the previous count carried out with administrative data based on the number of connections to the electricity service. In Uruguay, electricity coverage reaches 98% of homes and can be georeferenced appropriately. The convenience of such traditional field count in the pre-census phase is an open question. In future it will probably be more efficient to use administrative data during such phase.

The high coverage of the electricity service and the possibility of managing the consumption meters of customers with high precision permits the geo-location of the census data collected through the web. Households responding to the 2023 PHHC via Internet must login with their customer number and the number of their electricity meter. Incentives have been designed for the population to respond through this mechanism with the aim of lowering costs and data collection times: those who respond to the web questionnaire will be able to participate in a lottery for one year of free electricity consumption.

Finally, the Uruguayan census is innovating in another dimension: leveraging technology to flatten the pyramid usually observed in field operations with respect to the information flow to and from the

field. In fact, one of the limitations of the traditional pyramidal structure of census supervision is the distortion of information received by census enumerators and from them up to the survey chiefs. Usually, the vast majority of supervisors are staff with no previous survey experience, so those messages and information flowing through those channels present limitations. Technology can help to overcome this problem.

In the case of Uruguay, a planning and control centre has been established. This center is in permanent contact with the census enumerators and supervisors of the first level in the field directly (all tablets have access to electronic messaging services and have connection to mobile telephone lines) and it is the first reference for conceptual doubts, data collection and other field incidents. The heads of the field work and their assistants are the coordinators of those centers. The centers have access through computer systems to various tools to evacuate the above-mentioned doubts and solve problems in a short time. In addition to being able to monitor the progress of the survey. These centers receive incidents quickly (without distortions from intermediaries), can adequately systematize the information and can give rapid, unified and validated responses by the survey managers to all field staff, thereby improving the quality of the flow of information allowing more timely decisions to be made.

This report was prepared in collaboration with Leonardo Cuello, Lucía Pérez and Federico Segui.

General Information can be found at: www.ine.gub.uy

For further information, please contact daboal@ine.gub.uy

UNITED STATES

Reporting: **Andreea L. Erciulescu**

Toward a Vision for a New Data Infrastructure for Federal Statistics in the 21st Century

Ad hoc committees appointed by the National Academies of Sciences, Engineering, and Medicine are developing a vision for the new national data infrastructure to help inform decision makers on matters regarding economy, society, and life in general. For this, virtual public workshops are being held to gather information from key stakeholders and external experts, which is then being disseminated in public reports. The first two reports, on the overall vision for a new national data infrastructure and on using multiple data sources, have already been released. A third report will be produced, on data privacy and confidentiality. More information on the overall project can be found at the following link: Toward a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century | National Academies.

Upcoming Conferences and Workshops

Conferences on survey statistics and related areas

WSC 2023

ISI2023 The *64th ISI World Statistics Congress* will be face-to-face and held in Ottawa, Canada on July 16 - 20, 2023. https://www.isi2023.org/conferences/ottawa-2023/

OTTAWA 2023
64TH WORLD STATISTICS CONGRESS

The program of the upcoming 64th ISI WSC is now available here:
https://www.isi2023.org/conferences/15/programme/

IASS 2023 General Assembly

The IASS 2023 General Assembly will be held on Wednesday, 19th July 2023 at 12:10-13:50 pm (EDT - Eastern Daylight Time) during the WSC in Ottawa in ***hybrid* format**.
Join Zoom Meeting: https://zoom.us/j/97827663892.

The proposed agenda for the meeting is:

1. Welcome and Opening
2. President's Communications
3. IASS annual report 2022
4. President's report
5. TSS editor's report
6. Any other business

If you have any other items for discussion, please send an email to natalie.shlomo@manchester.ac.uk.

https://wiki.helsinki.fi/display/BNU/BANOCOSS2023

European Establishment Statistics Workshop 2023

The 2023 European Establishment Statistics Workshop - EESW23 – will be held at Statistics Portugal, in Lisbon, on 20-22 September, 2023. The deadline for abstract submission is past. However, there are **three short courses** that cover a selection of topics of high current relevance to establishment statistics methodologists and practitioners, and delivered by renowned international experts in their fields:

New developments in business data collection methodology, by Sally-Anne Aubrey-Smith, Ger Snijkers and Paulo Saraiva

Business network analysis, by Carolina Mattsson

Quality of multisource statistics, by Arnout van Delden and Sander Scholtus

Lunch and refreshments are included in course fees, and a certificate of attendance can be provided. For fees and registration to a short course, please follow this link

More information: sites.google.com/enbes.org/home/home/news-and-events/eesw23

Connecting Innovations in Data Science, Survey Research, and the Social Sciences

BigSurv23 The **3rd** international conference on **Big Data Meets Survey Science**

will be held on **October 26-29, 2023**,

at Universidad San Francisco de Quito in **Ecuador**.

The call for abstracts and session proposals is now closed.
Additional information can be found on the BigSurv23 website at https://www.bigsurv.org/

In Other Journals

Journal of Survey Statistics and Methodology

Volume 11, Issue 1, February 2023

https://academic.oup.com/jssam/issue/11/1

Survey Methodology

Mail Communications and Survey Response: A Test of Social Exchange Versus Pre-Suasion Theory for Improving Response Rates and Data Quality
Pierce Greenberg and Don Dillman

An Experimental Comparison of Three Strategies for Converting Mail Respondents in a Probability-Based Mixed-Mode Panel to Internet Respondents
David Bretschi, Ines Schaurer, and Don A. Dillman

Can Appended Auxiliary Data be Used to Tailor the Offered Response Mode in Cross-Sectional Studies? Evidence from An Address-Based Sample
Michael T. Jackson, Rebecca L. Medway, and Mahi W. Megra

Sequential and Concurrent Internet-Telephone Mixed-Mode Designs in Sexual Health Behavior Research
Stéphane Legleye and Géraldine Charrance

Predicting Nonresponse in Future Waves of a Probability-Based Mixed-Mode Panel with Machine Learning
Christoph Kern, Bernd Weiß, and Jan-Philipp Kolb

An Experimental Evaluation of Two Approaches for Improving Response to Household Screening Efforts in National Mail/Web Surveys
James Wagner, Brady T. West, Mick P. Couper, Shiyu Zhang, Rebecca Gatward, Raphael Nishimura, and Htay-Wah Saw

Survey Statistics

Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: An Empirical Comparison
Mehdi Dagdoug, Camelia Goga, and David Haziza

A Comparative Study of Imputation Methods for Multivariate Ordinal Data
Chayut Wongkamthong and Olanrewaju Akande

Adapting Nearest Neighbor for Multiple Imputation: Advantages, Challenges, and Drawbacks
Rebecca Roberts Andridge and Katherine Jenny Thompson

A Rescaling Bootstrap Approach for Imputed Survey Data
Zeinab Mashreghi and Huiqi Deng

Applications

Multiple Imputation with Massive Data: An Application to the Panel Study of Income Dynamics
Yajuan Si, Steve Heeringa, David Johnson, Roderick J. A. Little, Wenshuo Liu, Fabian Pfeffer, and Trivellore Raghunathan

Volume 11, Issue 2, April 2023

https://academic.oup.com/jssam/issue/11/2

Survey Methodology

How Does Back Translation Fare Against Team Translation? An Experimental Case Study in the Language Combination English–German
Dorothée Behr and Michael Braun

Multi-Project Assessments of Sample Quality in Cross-National Surveys: The Role of Weights in Applying External and Internal Measures of Sample Bias
Piotr Jabkowski, Piotr Cichocki, and Marta Kołczynska

Effects of Address Coverage Enhancement on Estimates from Address-Based Sampling Studies
Michael Jones, J. Michael Brick, and Wendy Van de Kerckhove

Deriving Priors for Bayesian Prediction of Daily Response Propensity in Responsive Survey Design: Historical Data Analysis Versus Literature Review
Brady T. West, James Wagner, Stephanie Coffey, and Michael R. Elliott

Survey Statistics

Bootstrap Estimation of the Conditional Bias for Measuring Influence in Complex Surveys
Jean-François Beaumont, Cynthia Bocci, and Michel St-Louis

Rank-Based Inference for Survey Sampling Data
Akim Adekpedjou and Huybrechts F. Bindele

Inference from Nonrandom Samples Using Bayesian Machine Learning
Yutao Liu, Andrew Gelman, and Qixuan Chen

Calibrated Multilevel Regression with Poststratification for the Analysis of SMS Survey Data
Jonathan Gellar, Constance Delannoy, Erin Lipman, Shirley Jeoffreys-Leach, Bobby Berkowitz, Grant J. Robertson, and Sarah M. Hughes

Fully Bayesian Estimation Under Dependent and Informative Cluster Sampling
Luis G León-Novelo and Terrance D Savitsky

Survey Statistics

Corrigendum to: Fully Bayesian Estimation Under Dependent and Informative Cluster Sampling
Luis G León-Novelo and Terrance D Savitsky

Journal of Official Statistics

Volume 39 (2023): Issue 1 (March 2023)

https://sciendo.com/issue/JOS/39/1

Characteristics of Respondents to Web-Based or Traditional Interviews in Mixed-Mode Surveys. Evidence from the Italian Permanent Population Census
Elena Grimaccia, Alessia Naccarato, Gerardo Gallo, Novella Cecconi and Alessandro Fratoni

A Multivariate Regression Estimator of Levels and Change for Surveys Over Time
Anne Konrad and Yves Berger

Investigating an Alternative for Estimation from a Nonprobability Sample: Matching plus Calibration
Zhan Liu and Richard Valliant

Using Eye-Tracking Methodology to Study Grid Question Designs in Web Surveys
Cornelia E. Neuert, Joss Roßmann and Henning Silber

A Statistical Comparison of Call Volume Uniformity Due to Mailing Strategy
Andrew M. Raim, Elizabeth Nichols and Thomas Mathew

A Two-Stage Bennet Decomposition of the Change in the Weighted Arithmetic Mean
Thomas von Brasch, Håkon Grini, Magnus Berglund Johnsen and Trond Christian Vigtel

Volume 39 (2023): Issue 2 (June 2023)

https://sciendo.com/issue/JOS/39/2

Effects of Changing Modes on Item Nonresponse in Panel Surveys
Oliver Lipps, Marieke Voorpostel and Gian-Andrea Monsch

Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach
Daniela Marella

Design and Sample Size Determination for Experiments on Nonresponse Followup using a Sequential Regression Model
Andrew M. Raim, Thomas Mathew, Kimberly F. Sellers, Renee Ellis and Mikelyn Meyers

Estimating Intra-Regional Inequality with an Application to German Spatial Planning Regions
Marina Runge

Constructing Building Price Index Using Administrative Data
Masahiro Higo, Yumi Saita, Chihiro Shimizu and Yuta Tachi

From Quarterly to Monthly Turnover Figures Using Nowcasting Methods
Daan Zult, Sabine Krieg, Bernd Schouten, Pim Ouwehand and Jan van den Brakel

Survey Practice

Vol. 16, Issue 1, 2023

https://www.surveypractice.org/issue/6753

Articles

The Shy Respondent and Propensity to Participate in Surveys: A Proof-of-Concept Study
John Boyle, James Dayton, Randy ZuWallack, Ronaldo Iachan

Null Effects of Framing Welcoming Ordinances
David Doherty, Dana Garbarski, Pablo Guzman Rivera

How Weighting by Past Vote Can Improve Estimates of Voting Intentions
Darren Pennay, Sebastian Misson, Dina Neiger, Paul J Lavrakas

Adapting Clinical Instruments for Population Mental Health Surveillance: Should an Explicit "Don't Know" Response Option Be Given?
Rachel Suss, Tashema Bholanath, Tenzin Yangchen, Dongchung Amber, Levanon Seligson, Christina C. Norman, Sarah E. Dumas

Mail to One or Mail to All? An Experiment (Sub)Sampling Drop Point Units in a Self-Administered Address-Based Sampling Frame Survey
Taylor Lewis, Joseph McMichael, Charlotte Looby

Survey Research Methods

Journal of the European Survey Research Association

Vol 17 No 1 (2023)

https://ojs.ub.uni-konstanz.de/srm/issue/view/233

Articles

Puzzling Answers to Crosswise Questions: Examining Overall Prevalence Rates, Response Order Effects, and Learning Effects
Sandra Walzenbach, Thomas Hinz

Respondents for Nearly Three Decades: How Do Loyal Sample Members Differ From Others?
Nicole D. James

Memory Effects: A Comparison Across Question Types
Tobias Rettig, Annelies G. Blom, Jan Karem Höhne

Religious Involvement Across Europe: Examining its Measurement Comparability
Alisa Remizova, Eldad Davidov, Maksim Rudnev

Ambiguity in the Item Wording, Ambiguity in the Respondents' Comprehension? An Experiment on the 'Immigrants/Foreign Workers' Social Distance Item in Values Surveys
Riccardo Ladini, Ferruccio Biolcati

Using Cognitive Interviews to Evaluate and Improve a Danish Translation of a Compiled Questionnaire on Existential and Spiritual Constructs
Tobias Anker Stripp, Dorte Toudal Viftrup, Ricko Damberg Nissen, Sonja Wehberg, Jens Sondergaard, Niels Christian Hvidt

Hard-to-Survey and Negligible? The Institutionalized Population in Europe
Jan-Lucas Schanze

Other Journals

- **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/uasa20
- **Statistics in Transition**
  - https://sit.stat.gov.pl

Welcome New Members!

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

Title	First name	Surname	Country
MR.	Peter	Buwembo	India
PROF	Mahmoud	Torabi	Canada
DR.	Dina	Neiger	Australia
DRS	Luciana	Crosilla	Italy
MS	Camilla	Salvatore	Italy
MR.	Darryl	Creel	United States
PROF	David	Haziza	Canada
DR.	Vilma	Nekrašaitė-Liegė	Lithuania
MR.	Mawdo	Gibba	Gambia
DR.	Jae Kwang	Kim	United States
MR.	Rees	Morrison	United States
PROF	Enrico	Fabrizi	Italy
MS	An-Chiao	Liu	The Netherlands
MR.	José	Zea	Colombia
MRS	Aulia Dini	Rafsanjani	Indonesia
DR.	Eva	Elvers	Sweden
DR.	Daniela	Cialfi	Italy
DR.	Monika	Galambosne Tiszberger	Hungary
DR.	Renata	Benda-Prokeinova	Slovakia
DR.	Bernard	Baffour-Awuah	Australia
PROF. DR.	Reza C.	Daniels	South Africa
PROF	Nicola	Salvati	Italy

IASS Executive Committee Members

Executive officers (2022 – 2024)

President:	Monica Pratesi (Italy)	monica.pratesi@unipi.it
President-elect:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
Vice-Presidents:		
Scientific Secretary:	M. Giovanna Ranalli (Italy)	maria.ranalli@unipg.it
VP Finance	Jairo Arrow (South Africa)	jairo.arrow@gmail.com
Liaising with ISI EC and ISI PO plus administrative matters	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
Chair of the Cochran-Hansen Prize Committee and IASS representative on the ISI Awards Committee:	Nikos Tzavidis (UK)	n.tzavidis@soton.ac.uk
IASS representatives on the World Statistics Congress Scientific Programme Committee:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
IASS representative on the World Statistics Congress short course committee:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
IASS representative on the ISI publications committee	M. Giovanna Ranalli (Italy)	maria.ranalli@unipg.it
IASS Webinars Representatives 2021-2023	Andrea da Silva (Brazil)	andrea.silva@ibge.gov.br
Ex Officio Member:	Ada van Krimpen	an.vankrimpen@cbs.nl

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

IASS LinkedIn Account https://www.linkedin.com/company/international-association-of-survey-statisticians-iass

Institutional Members

International organisations:

- Eurostat (European Statistical Office)

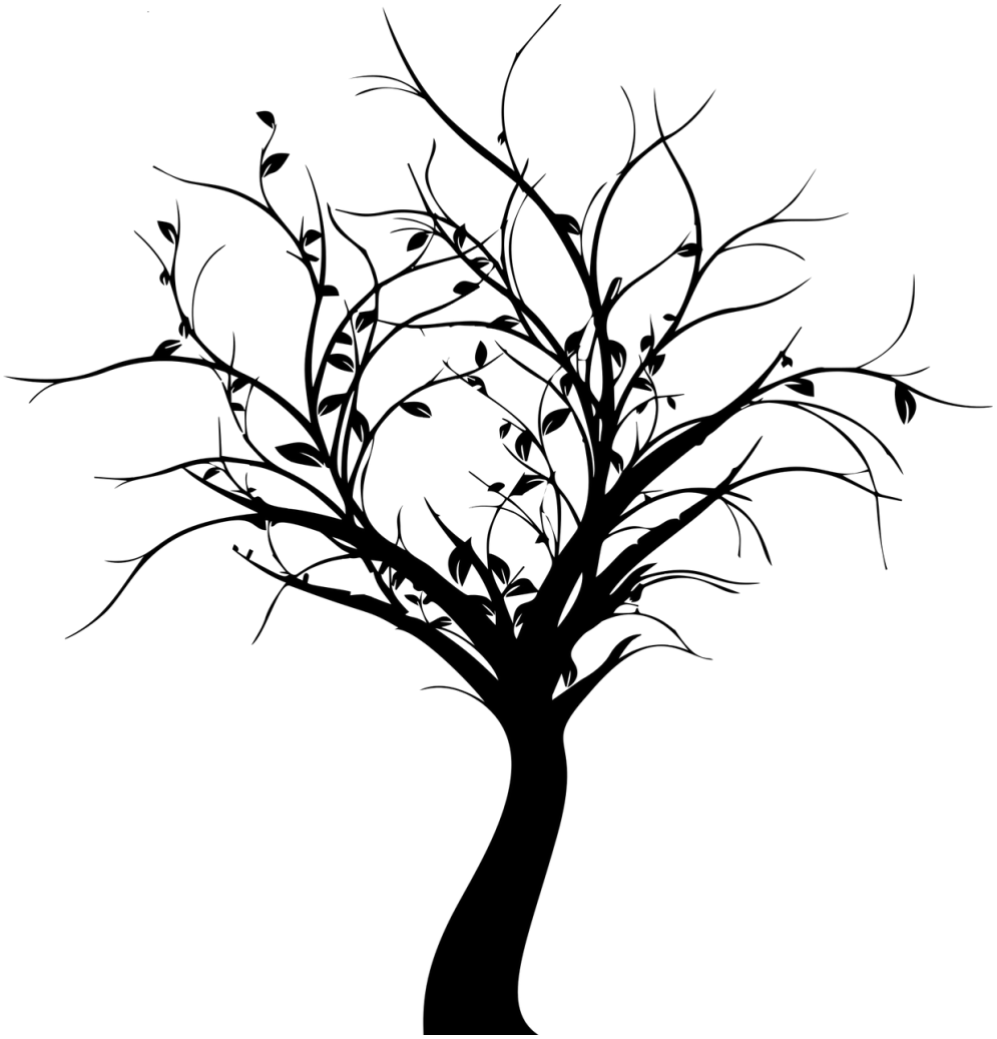
National statistical offices:

- Australian Bureau of Statistics, Australia
- Inst Brasileno de Geografia y Estatistica, Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistiches Bundesamt (Destatis), Germany
- International Rel. & Statistical Coordination, Israel
- ISTAT, Italy
- Dept. of Economics and Management, Italy
- Statistics Korea (KOSTAT), Korea, Republic of
- EC Eurostat - Unit 01: External & Inter., Luxembourg
- Dir.dos Serviços de Estatística e Census, Macao, SAR China
- Statistics Mauritius, Mauritius
- INEGI, Mexico
- Statistics New Zealand, New Zealand
- Statistics Norway, Norway
- Inst. Nacional de Estatística (INE), Portugal
- Statistics Sweden, Sweden
- National Agriculture Statistics Service, United States
- WESTAT Inc., United States
- National Center of Health Statistics, United States

Private companies:

- Westat, United States

Read *the Survey Statistician* online!



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