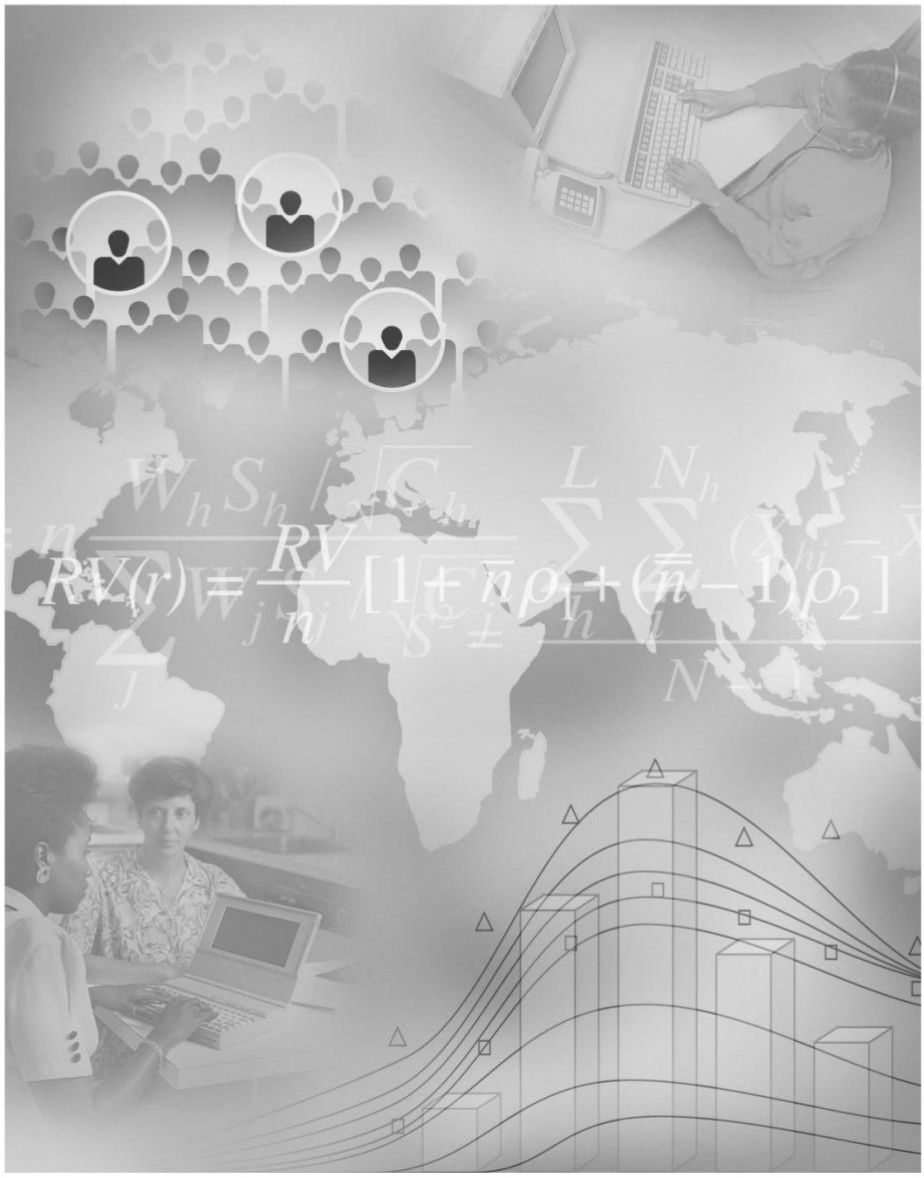
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

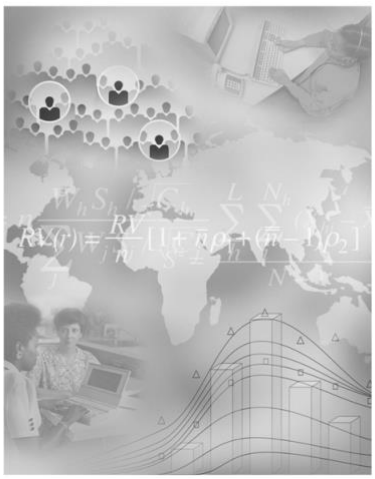
No. 89 **January 2024**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

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



The Survey Statistician No. 89, January 2024

The Survey Statistician is published twice a year by the International Association of Survey Statisticians and 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. 8 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 Alina Matei (alina.matei@unine.ch).

ISSN 2521-991X

In this Issue

Letter from the Editors

Dear Readers,

May the New Year bring you more peace and health!

The current issue of TSS includes information about the latest events of the last half-year: World Statistics Congress 2023 (WSC2023) and workshops having international significance; the recent awards; as well as traditional and new TSS sections.

Two award recipients are honoured in this issue: Professor Richard Valliant who received the Joseph Waksberg award by *Survey Methodology* journal for 2024; and Professor C.R. Rao who was awarded the 2023 International Prize in Statistics by the ISI. Unfortunately, he could not rejoice long with it as he passed away at almost 103 years old. The issue includes a tribute to him written by renown professors who were among his former students and who knew him well and had the chance to work with him. These are Arni S.R. Srivastava Rao, T. Krishna Kumar and Pramod Kumar Pathak. Their paper describes the scientific life of Professor Rao who was key in the history of statistics.

This issue continues with the two papers devoted to the 50th anniversary of the IASS. First, Carl-Erik Särndal recalls a number of ideas, concepts and lines of thought that influenced theory and practice in survey statistics. Then, Annamaria Bianchi presents testimonials of 15 former IASS presidents.

Our Association is very lively, as a young generation continues the traditions started by elders and predecessors. To reflect this, *The Survey Statistician* is introducing a new section called *Early Career Survey Statistician*. It will be edited by Dr. Veronica Ballerini from the University of Florence, Italy.

The 2021-2023 IASS Executive Committee (EC) of IASS completed its work. We thank President Monica Pratesi (University of Pisa, ISTAT) and IASS Scientific Secretary Giovanna Ranalli (University of Perugia) for their regular *Letters* to *TSS* and interesting articles organized in the *New and Emerging Methods* section. We wish them success in their activities.

The IASS Executive Committee (EC) for 2023-2025 was elected before the WSC 2023. You can see the list of EC members at the end of this issue. The new IASS EC president is Natalie Shlomo, Professor of Social Statistics at the University of Manchester, UK. The scientific secretary is Dr. Annamaria Bianchi from the University of Bergamo, Italy. They start to contribute to the current issue in those capacities. The co-editor of *TSS* Eric Rancourt was elected vice-president and so this issue will be his last one as an editor. He passes on this editorial role to Alina Matei, Titular Professor of Statistics at the University of Neuchâtel, Switzerland. Alina is currently the *Book and Software Review* section editor. Her responsibilities go to Dr. Gaia Bertarelli from Ca' Foscari University of Venice.

The *Ask the experts, New and Emerging Methods, Book and Software Review* sections all happen to contain articles written by Italian statisticians. Taking also into account two Italian section editors and their inputs to this *TSS* issue, one could almost call it an Italian issue. This comment is simply to illustrate the highly and widely developed interest in research in survey statistics in Italy and to wish the Italian statisticians (and all others!) to follow this successful path in the future.

Finally, should you wish to contribute to any section of *TSS,* please write to the editors or to the section editors about your proposals, and you will be welcomed.

Danutė Krapavickaitė and **Eric Rancourt**

TSS editors

Letter from the President

Dear IASS Members,

My first letter as President of the IASS appeared in the August 2023 newsletter here: [July-August Newsletter] where I introduced the new 2023-2025 Executive Committee members and thanked the outgoing 2021-2023 Executive Committee members, particularly the outgoing President Monica Pratesi, for their dedication in promoting IASS and expanding its activities. I also thanked the ISI Permanent Office for their continuing support in maintaining IASS operations and administration.

The 2023-2025 Executive Committee will be continuing with the successful and impactful current activities: the monthly webinar series, the monthly newsletters (with a reading of the month) and the twice-yearly The Survey Statistician, the Hukum Chandra prize in 2024 and the Cochran-Hansen prize in 2025, financial support for survey statistics-related conferences and finally maintaining our website and increasing social media presence. We will also continue to promote the IASS and expand our membership, as well as enlist more institutional membership, through the help of our over 65 IASS country representatives.

There are also exciting new initiatives to watch out for:

- There will be a new section in *The Survey Statistician* dedicated to the *Early Career Survey Statistician* under the leadership of the Editor-in-Chief, Danutė Krapavickaitė and the new Section Editor: Veronica Ballerini. We are also exploring the idea of dedicating a section of The Survey Statistician to one specific topic in survey statistics with contributions from the county representatives.
- We are increasing our presence at more regional workshops, for example we delivered a presentation at the online Asia-Pacific Stats Café held on October 30th, 2025, and sponsored a session titled: Machine Learning Methods in Survey Statistics, at the Second Workshop on Methodologies for Official Statistics held at ISTAT in Rome December 6th - 7th, 2025.
- We aim to host a bespoke IASS satellite conference prior to the 2025 65th WSC taking place in The Hague, July 13th - 17th, 2025. The satellite conference will be held in Manchester July 9th - 11th, 2025 so please save these dates. The conference will include all areas of research and development related to survey statistics and survey methodology.

A new Strategy Document 2023-2025 will be out soon for consultation to be ratified at the July 17th, 2024, IASS General Assembly (GA) at 2 pm ECT to be held virtually. More details will follow. Please put this date in your calendar to attend the GA.

In this January 2024 issue of *The Survey Statistician*, we are continuing to celebrate our 50 year jubilee of the IASS with more articles and input from past IASS Presidents. This is a momentous occasion. It shows the ever- increasing importance of survey statistics, particularly as it evolves to include new forms of data and administrative data into our statistical systems. Research in the areas of data integration, compensating for measurement and coverage errors, introducing new tools offered by machine learning and AI, all demonstrate more than ever the need for high-quality unbiased survey data collections to be used as reference samples. These samples can assess the quality and compensate for errors in new sources of data, as well as continue to collect data that are not available in administrative data or other sources of data.

We are now at the crucial point of the calendar year where we ask all IASS members to renew their IASS membership. We will also carry out a recruitment drive through the help of our country representatives, focusing on inviting institutional members as well. Please spread the news of IASS to all your networks so that we can increase our membership and allow funding for all the important activities that we do to benefit survey statisticians around the world. The website for membership of

the IASS is here: [Join IASS]. The membership fees are 50 euros/25 euros a year depending on your country of residence. Please download our brochure here: [Brochure] to help recruit new members from your networks.

Please get in touch with me at natalie.shlomo@manchester.ac.uk if you have ideas, suggestions, comments, improvements, or criticisms, regarding IASS activities and strategic priorities.

With best wishes,

Natalie Shlomo

President IASS, 2023-2025

Report from the Scientific Secretary

I have been appointed Scientific Secretary of IASS during the first meeting of the newly elected IASS Executive Committee (EC) in July during the WSC in Ottawa. I am very grateful to the members of the EC for their trust, and I am indebted to Maria Giovanna Ranalli for her legacy on this role.

As my first duty, I had to choose a topic for the **New and emerging methods** section of this issue of The Survey Statistician. I decided to give space to the use of big data in the context of small area estimation. This is a rather recent topic, that I believe needs more exploration in order to fully develop the potential of using big data sources for enhancing estimation at small domain level. I am very grateful to Professors Stefano Marchetti (University of Pisa) and Francesco Schirripa-Spagnolo (University of Pisa) for having accepted my invitation to write a paper on **Social big data to enhance small area estimates**. This contribution provides a discussion about the opportunity to use big data in small area estimation. The Authors also show an application where data mined from Twitter are used to improve small area estimates of consumption expenditures for leisure at local level in Italy.

To celebrate the 50th jubilee of the IASS, I have also prepared the contribution **Past Presidents Testimonials on IASS** where past Presidents of the IASS share their experience and involvement with the Association as well as their view on its future. I started this work with our Past President Monica Pratesi and continued it with our current President Natalie Shlomo. We were able to collect testimonials from fifteen past Presidents. I would like to thank all of them for their contribution to this section.

The organization of the monthly **Webinar series** has continued, and we are particularly thankful to Andrés Gutiérrez for his engagement in organizing the webinars. We have now reached Webinar number 35 and we are happy to have made it a monthly event that has attracted an audience of over two **hundred registered participants**. Please, visit the IASS events page for upcoming webinars: http://isi-iass.org/home/events/. Also visit the webinar section of our website http://isi-iass.org/home/webinars/ for slides of past webinars, and that of ISI https://isi-web.org/courses-webinars-workshops?type=2&field_type_courses=All for recorded past webinars. Contact Andrés (andres.gutierrez@cepal.org) if you have suggestions for topics and/or speakers for an upcoming webinar. Those webinars held in the last six months of 2023 have covered the role of the Inter-Secretariat Working Group on Household Surveys by Haoyi Chen, the Production of Small Area Estimates for Labor Market Indicators in Latin America by Andrés Gutiérrez, the Weight Share Method when using a Continuous Sampling Frame by Guillaume Chauvet, Missing Voices in Household Survey Data by Cheryl Doss, Recent Developments in Unit-Level Models for Small Area Estimation by Emily Berg and Enhancing the Credibility of Survey Data: Old Tricks and New Techniques in Improving the Respondent Experience by Barbara Rater.

The IASS EC supported three conferences/workshops in 2023: the 8th ITAlian COnference on Survey Methodology – ITACOSM2023 – "New Challenges for sample surveys: innovation through tradition" held at the University of Calabria, Italy, on June 7-9; the Sixth Baltic-Nordic Conference on Survey Statistics - BaNoCoSS-2023 - held on August 21-25 in Helsinki, Finland; and A Day Survey Researchers Conference on the theme “Good design and proper conduct of surveys – Key determinants of reliable data for planning”, held on July 13 at the Department of Statistics, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria. The new **Call for Conference Support for 2024** is now out. We will consider providing financial support for workshops, conferences and similar events that promote the study and development of the theory and practice of statistical surveys and censuses and associated subjects, and foster interest in these subjects among statisticians, organizations, governments, and the general public in different countries of the world. To apply for

support, please send a document project, including a budget sheet, to IASS President Natalie Shlomo (natalie.shlomo@manchester.ac.uk) and in cc to IASS President-Elect Partha Lahiri (plahiri@umd.edu) **before June 30th**. More details can be found here: http://isi-iass.org/home/wp-content/uploads/Call-for-Conference-Support-2024.pdf.

Turning to our social media activity, during these past months I have continued posting about webinars, conferences, books, articles, prizes, the newsletter release and its contents, and the recent recruitment drive. The number of followers to the pages are increasing: Since July, X followers increased from 365 to 392, LinkedIn followers from 1075 to 1434 and Facebook followers from 133 to 178. Thus, please, follow the updates on the life of IASS via **social media** and reading our **monthly Newsletter**. Other than webinars, information on conferences, on the recipients of awards and on call for nominations, it now features a **reading of the month** section in which we suggest monographs, special issues or edited books on topics of interest to the members of IASS. Please, feel free to contact us for news and information to be added in the Newsletter by the 15th of each month.

Annamaria Bianchi
annamaria.bianchi@unibg.it
IASS Scientific Secretary 2023-2025

Promoting good survey theory and practice around the world

The International Association of Survey Statisticians (IASS) aims to promote the study and development of the theory and practice of sample surveys and censuses.

It also aims to increase the interest in surveys and censuses among statisticians, governments and the public in the different countries of the world.

Dear statistician,

Join the IASS and become a member of an association of 350+ survey statisticians!

To become a member (never before a member of the IASS nor of the ISI):

You have to register first at Login (isiwebshop.org), an activation link will be sent to the email address you entered. Note that you must activate the account by selecting the activation link in the email you receive. If you don't see the email in your Inbox, please check your SPAM box or your other email boxes. After activation, you can log in to https://www.isiwebshop.org/ using the username (email address) and the password you entered during registration.

Current members can renew their membership/s and update their contact details via the webshop. The username is the email address you used for your contact details and the password is your membership number, that is if you haven't customized it.

Should you require any assistance in becoming a member, renewing your membership/s, or forgotten your membership number please send an email to isimembership@cbs.nl.

News and Announcements

The 2024 Waksberg Award goes to Richard Valliant

Richard Valliant is the recipient of the 2024 Waksberg Award. Professor Valliant who is Research Professor Emeritus at the University of Michigan as well as at the Joint Program in Survey Methodology of Maryland will deliver the Waksberg address at Statistics Canada's Methodology Symposium in the fall of 2024 and write a paper for the December issue for Survey Methodology.

Richard Valliant wrote several books, including *Finite Population Sampling and Inference: A Prediction Approach* in 2000 with A. H. Dorfman and R. M. Royall and he produced numerous articles contributing to advance survey sampling. He was an Associate editor in *Survey Methodology* and has also refereed many papers.

The Waksberg Award is given each year to an eminent survey statistician who has made a contribution in an area of theory and / or practice in relation to the work of Joseph Waskberg who was an icon in survey sampling. Joseph Waksberg worked at the U.S. Census Bureau and Westat where he was chairman of the board for over 25 years. The winner of the Waksberg award is selected to write a paper on developments related to the work of Waksberg and receives an honorarium from Westat.

The selection committee appointed by *Survey Methodology* and the American Statistical Association was composed of M. Giovanna Rannali (Chair), Jae-Kwang Kim, Kristen Olsen and Denise Silva.

More information on the Award can be found at:
https://www.isi-web.org/article/richard-valliant-wins-2024-waksberg-award

World Statistics Congress 2023 in Ottawa, a great success

The 64th World Statistics Congress of the International Statistical Institute (ISI) took place from July 16th to 20th in Ottawa, Canada. The event attracted much of the statistics and data world's attention. Opening remarks were provided by an Ottawa Canadian Parliamentarian, The Honorable Jenna Sudds as well as by the Chief Statistician of Canada, Anil Arora. ISI President Stephen Penneck officially opened the congress.

The World Statistics Congress 2023 (WSC23) attracted close to 1600 people from 110 countries and included 310 invited paper, contributed and poster sessions that included a total of 967 presentations. As part of the WSC23, there were short courses attended by 76 people and booth from many organizations.

During the WSC23, the ISI recognized many people through a number of awards. Highest in this list was the International Prize in Statistics that was awarded to Dr. C.R. Rao for his immense contribution to our field. His son Veerendra Rao was present at the event to proudly accept the prize. At the time of the WSC23, Dr. Rao was 102 years old and regretfully passed away the following month.

The WSC23 took place in Ottawa 60 years after it had first been held there in 1963. During the event, delegates had the privilege to welcome Dr. Ivan P. Fellegi, Chief Statistician Emeritus from Canada who was present at the ISI congress in 1963! The WSC23 provided a forum for practitioners, academics and business statistician to nurture and expand their network. While many were longtime ISI members having participated in many congresses, many new faces, students took part of the event.

As part of the WSC, 35 delegates also had the chance to take part in a city tour of Ottawa in which they were provided historical information about the Ottawa and Canada. Delegates appreciated the tour and more generally had very good words about their overall experience of Ottawa and the WSC23.

Following five days of presentations, meetings, networking and friends making, the WSC23 was concluded and Canadian Senator, the Honourable Colin Deacon provided closing remarks.

Eric Rancourt

IASS Highlights from the 64th ISI World Statistics Congress, Ottawa, Canada, July 16th – July 20th, 2023

The International Association of Survey Statisticians (IASS) is one of seven associations of The International Statistical Institute (ISI) and therefore has an important role in the organization of the ISI World Statistics Congresses.

The IASS aims to promote the study and development of the theory and practice of sample surveys and censuses around the world. Besides statistical and methodological research and development in surveys and censuses, including small area estimation, there is now large demand for timely, more detailed, less burdensome and costly statistics, with wider coverage. Therefore, survey statisticians are also working in innovation, for example investigating new sources of data, such as Administrative Data and Big Data (mobile phones, social media interactions, electronic commercial transactions, sensor networks, smart meters, GPS tracking devices, satellite imagery), that can potentially be included in statistical systems. New forms of data also require developing new tools, such as

algorithms arising from Machine Learning and AI, for enabling analyses and predictions. Moreover, given these new forms of data, there is more urgency on developing statistical methods to ensure the validity, measurement and representativeness of these data, typically carried out through the use of reference data coming from high-quality and representative survey data.

At the 64th WSC, the IASS sponsored 14 Invited Papers Sessions (IPS) covering both traditional statistical research in survey statistics and new developments related to new forms of data and innovation. The IPS were:

- Better ways to address nonresponse in social surveys
- Longitudinal observation of human populations
- Measurement Error Modelling: Advances and Applications
- Recent Advances in Data Anonymization and Data Modelling for Education and Poverty Research
- Unlocking Microdata: Experience from International Organizations
- Inference under Informative Sample Designs
- Harnessing data by citizens for public policy and SDG monitoring: a conceptual framework and what is next?
- Monitoring Progress towards Sustainable Development Goals with Small Area Estimation over Space and Time
- The use of alternative data through modelling in Official Statistics
- Innovative statistical methods for large-scale surveys
- Developments in Small Area Statistics Leveraging Non-Random Sampling
- Sample surveys in the era of Big Data and Machine Learning
- Advanced inference on mixed effects models for SAE
- Data Integration in Survey Sampling

There were also three Special Invited Paper Sessions (SIPS) sponsored by the IASS:

- SIPS Waksberg Award with a presentation by the 2023 Waksberg Award Winner, Prof. Ray Chambers. The title of Ray's talk was 'The Missing Information Principle - A Paradigm For Analysis Of Messy Sample Survey Data'. In this talk, Ray discussed alternative approaches to inference from multiple data sources that can compensate for coverage errors, measurement errors and response errors in the statistical modelling process embedded in a Missing Information Principle (MIP) framework.

- SIPS IASS President's Invited Speaker session was delivered by Prof. Fulvia Mecatti. The title of her talk was: "Bridging Big Data and Sampling Methodology: A Sampling Statistician Perspective". In her talk, Fulvia focused on survey sampling methods and big data for the production and analysis of survey data. In particular, the talk focused on the multifaceted relationship connecting Big Data to statistical and to sampling methodology.

- SIPS Cochran-Hansen Prize session with presentations by our two Cochran-Hansen prize winners: Alejandra Arias-Salazar (Costa Rica) with her paper: "Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach" and Ziqing Dong (China) with his paper: "Linearization and Variance Estimation of the Bonferroni Inequality Index". The session was chaired by Nikos Tzavidis, chair of the Cochran-Hansen prize committee, with the help of the committee members: Bernard Baffour and Maria Giovanna Ranalli. The Cochran-Hansen prize winners were also awarded with a certificate presented at the ISI and Associations Awards Ceremony.

The General Assembly of the IASS was also held at the 64th WSC and it was a great opportunity to meet face-to-face with IASS members attending the WSC and to promote discussion on the state of our society. The presentation included an overview of IASS activities and the 2022 IASS budget. The final 2022 report and budget were approved by the IASS members. It was also an opportunity to thank the outgoing Executive Committee (Monica Pratesi, Nikos Tzavidis, Jairo Arrow and Andrea Diniz da Silva) as well as welcome the incoming Executive Committee (Natalie Shlomo, Partha Lahiri, Annamaria Bianchi, Jiraphan Suntornchost and Andres Gutierrez Rojas).

One feature of the WSC is that ISI Associations hold a satellite conference adjacent to the WSC. The IASS however has generally taken a different approach over the years in that it supports survey statistics-related conferences around the world. This past year, IASS supported conferences and workshops in Italy, Pakistan, Nigeria and Finland. For the 65th WSC to be held in the Hague on July 13th – 17th, 2025, the IASS Executive Committee is supporting the idea of holding an IASS satellite conference (possibly combined with the annual Small Area Estimation conference) in the city of Manchester, United Kingdom on July 9th – 11th, 2025. Please save the dates!

Natalie Shlomo

Annamaria Bianchi

ENBES Workshop, 20-22 September 2023

The eighth *European Establishment Statistics Workshop* (EESW23) was held in Lisbon, hosted by Statistics Portugal. The workshop was spread over three days and preceded by a day with short courses. The programme, papers and an extended report can be found at https://sites.google.com/enbes.org/home**.** The leading theme was "traditional and new data sources for establishment statistics". The workshop is designed to promote discussion on the topics presented.

The first day of the workshop began with a session on innovation in data collection. We discussed the use of cognitive testing on an international scale. While it is very useful, time constraints sometimes prevent us from applying cognitive testing. We discussed the use of paradata dashboards for monitoring responses, which is essential for enhancing the quality of the collected data. Finally, there was an application of system-to-system data collection in an agricultural survey as a form of automatic data collection that helps us to with the challenge to collecting more data in a faster and more secure manner while keeping respondents or data providers motivated.

The next session was dedicated to redesigning and optimizing sample designs. The first paper was on the UK Business Enterprises Research and Development survey, whose sample was increased approximately tenfold in order to cover a wider range of businesses and improve quality. It also changed to electronic data collection with improved guidance to businesses. The second paper was on the Producer Price index design: a two-phase probabilistic sample has been applied in the UK and was considered for Ireland. The authors discussed the change of method for calculating sample variances (model-based) and the way to move from the current towards the new sample design.

The final presentation session that day covered different approaches on linking households and establishments, looking into household consumption and income. A wide variety of sources was used in the three studies: survey data, bank transaction data and card issuers, tax authority business data, administrative micro-data on personal expenses and social security register data. Some common objectives were: Improving the quality of the statistics, promoting replacement of survey variables with administrative data, increasing the timeliness and level of detail of statistics, and reducing costs and the statistical burden on respondents.

On the first day we also had a general ENBES meeting on the history of ENBES and on how we can best continue, and we had a session with three small discussion groups: on improving response rates, statistics for rare populations, and making business surveys resilient to future shocks.

The second day started with a session on the use of alternative data sources to achieve earlier estimates in short-term statistics. The papers explored different ways in which new data sources can be incorporated through modelling approaches, including use of machine learning (ML). One paper considered the use of aggregate credit card data either as an early estimate or as a replacement for retail turnover data for the smaller retail. Another paper considered using scanner data, though accessing the data is a serious challenge. The last paper considered the use of e-invoices in Italy for small and medium-sized businesses, but with some challenges with GDPR.

Then followed a poster session. Topics were a new method using real time shipping data to produce timelier data, outlier treatment in seasonal adjustment handling the effects of COVID-19, new data infrastructure with analytical capabilities and an agile and multisource approach to business surveys.

The following session looked into the use of new data sources in business statistics. The first paper focused on using multilingual web scraping to generate a business register. The second paper showed how we can define a new classification of enterprises by mixing survey data with information from websites. In the third paper administrative data was used to improve the accuracy of monthly construction estimates using ML. The session concluded that in official statistics, transparency and the ability to explain how decisions are made are critical for accountability and trust of novel data sources and techniques.

The fourth session was on analysis and quality evaluation of complex business statistics. Three very different examples of complexity were given. The first paper proposed new ways to measure differences in trade values as measure by the exporting versus the importing country. The second paper explained a bootstrap method to estimate variances of an estimator that combined administrative data with a survey using unequal sampling probabilities. The final paper gave a short overview of recent research on networks of buyer-supplier ties among businesses.

The final session of day two was dedicated to respondent management. It began with the use of permanent random numbers when drawing samples in order to estimate changes over time, to balance response burden and to account for stratum changers. Further, we discussed changes in data collection for the IAB Establishment Panel in times of the COVID-19 pandemic showing results of a multimodal survey since face-to-face interviews face were no longer possible. The final paper proposed ways to reduce the number of companies that receive multiple surveys. It was focused on the smaller businesses, but we also discussed options for larger businesses.

The first session on day three was on infrastructure for data integration. The first paper showed ways to achieve consistency across different statistics by using manual and automatic multisource editing at Statistics Netherlands. We discussed that in the long run it might be desirable to integrate the data collection in one large survey. Next, a multimode data collection system for tourism statistics was presented, supporting XML, electronic forms and paper and pencil. Special attention was needed for data confidentiality so different users could access the data. The final paper analysed investments of Italian enterprises, using a system of integrated registers (ISSR) developed by Istat. The ISSR allows Istat to do all kinds of new analyses without the need for dedicated surveys.

Then we dealt with measuring and documenting process and output quality of multisource statistics. It started with a method to automatically detect errors in registered NACE codes, using a ML model with scraped website texts. It requires extensive website information and innovative methods for data processing. The second paper presented a generalized way to produce process and output quality estimates from Istat's ISSR without defining the exact output in advance. The third paper illustrated the use of different small area estimators for the production of land use statistics by exploiting surveys and auxiliary sources. We discussed that not only the variance but also the bias of the different estimators should be evaluated.

In the closing session, there was some discussion of developments in establishment statistics. Since we have many sources available nowadays, are we going to ask respondents for their data, or do we ask them to verify the data that we have? Do we want to focus on which sources can be used for our regular output, or are we also going to explore new outputs? Finally, if our output is based on non-probability samples, how are we going to evaluate output quality? We might need to draw audit samples for quality estimation specifically for that purpose.

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

Arnout van Delden and **Paul Smith** on behalf of ENBES.

Second Workshop on Methodologies for Official Statistics, Rome, December 6-7, 2023

The Italian National Statistical Institute (Istat) organized and hosted the Second Italian Workshop on Methodologies for Official Statistics, that was held in Rome on 6 and 7 December, 2023 at Istat premises. The workshop was by invitation, and it was possible to follow it online as well.The event had the objective to gather researchers on statistical methodologies applied in the official statistics context, in order to promote the exchange of ideas and best practices. The workshop was mainly focused on understanding the impact on methodologies and quality of new data sources (big data and in general non- probability data) for the official statistics production.

The workshop was organized into four sessions, each one consisting of a master class and invited speakers from many universities and statistical agencies across the US, Canada and Europe.

The workshop opened with a welcome and opening session with addresses by Francesco Maria Chelli (Istat President, Italy), Massimo Fedeli (Head of Department for Development of Methods and Technologies for Production and Dissemination of Statistical Information, Istat) and Monica Pratesi (Head of Department for Statistical Production, Istat).

During the first morning of the workshop, the first session was entitled “Methodologies and Designs for Multi-source Processes with Non-probability Data”. It entailed a master class given by Changbao Wu (University of Waterloo, Canada) on “Challenges and Strategies in Dealing with Non-probability Samples”. This was followed by three invited talks by Danila Filipponi (Istat, Italy), Aurélien Lavergne (INSEE, France), J. David Drown (U.S. Census Bureau, U.S.A.), and a discussion by Maria Giovanna Ranalli (University of Perugia, Italy).

The second session on the first day was entitled “Innovative Data for Official Statistics: Methodological Challenges”. It started with a master class given by Stefano M. Iacus (Harvard University and IQSS, U.S.A.) on “Limits and Challenges of Incorporating Innovative Data in Official Statistics”. This was followed by three invited talks by Claudia De Vitiis (Istat, Italy), Fabrizio De Fausti (Istat, Italy), and Maurizio Naldi (LUMSA University, Italy). The session concluded with a discussion by Li-Chun Zhang (Statistisk Sentralbyrå, Norway and University of Southampton, U.K.).

The second day started in the morning with the third session titled “Quality for Non-traditional Sources” with a master class given by Fabio Ricciato (Eurostat) on “Quality for Innovative Data Sources: Progress, Challenges and Directions of Work for the European Statistical System”. This was followed by three invited talks by Gabriele Ascari (Istat), Magdalena Six and Alexander Kowarik (Statistik Austria, Austria), and Anthony Dawson and Calvin O’Brien (Ireland Central Statistical Office - An Phríomh-Oifig Staidrimh, Eire). The discussant for the session was Natalie Shlomo (University of Manchester, U.K., IASS President).

The final session on the second day was sponsored and organized in collaboration with the International Association of Survey Statisticians (IASS). The title of the session was “Machine Learning Methods in Survey Statistics”. The session started with an introduction by the IASS President Natalie Shlomo (University of Manchester, U.K.) entitled “Overview of Machine Learning in Survey Research”. This was followed by three invited talks by David Haziza (University of Ottawa, Canada) on “Machine Learning Procedures for the Treatment of Unit Non-response in Surveys”, Marco Di Zio (Istat, Italy) on “State of Play and Perspectives on Machine Learning at Istat”, and Petrus J.H. Daas (Centraal Bureau voor de Statistiek, The Netherlands) on “Machine Learning in Official Statistics: towards Statistical Based Machine Learning”. Below are photos of the Italian representatives of the IASS along with the IASS President and the invited speaker David Haziza.

The details of the programme, all abstracts and presentations are available here: https://www.istat.it/en/archivio/288564 [istat.it]

Photos from the International Association of Survey Statisticians (IASS) sponsored session: "Machine Learning Methods in Survey Statistics" at the Second Italian Workshop on Methodologies for Official Statistics, Rome, December 6-7, 2023

The closing session entailed an adress by Orietta Luzi (Head of Directorate for Methodology and Statistical Process Design, Istat, Italy) thanking the speakers and audience participation.

The event was a great success with around 70 people in presence in the audience and other people following online. Papers presented will be published in a Proceedings produced by Istat at the beginning of 2024.

Reported by: **Annamaria Bianchi**, IASS Executive Committee 2023-2025

The Survey Statistician, 2024, Vol. 89, 16–28.

A Tribute to the Centenarian Statistician C. R. Rao

Arni S. R. Srinivasa Rao[1,2], **T. Krishna Kumar**[3] **and Pramod Kumar Pathak**[4]

[1] Laboratory for Theory and Mathematical Modeling, Medical College of Georgia, Augusta, U. S. A., arrao@augusta.edu

[2] Department of Mathematics, Augusta University, Georgia, U. S. A.

[3] Rockville Analytics, Rockville, MD., U. S. A., tkkumar@gmail.com

[4] Michigan State University, East Lansing, MI., U. S. A. pathakp@msu.edu

Abstract

Professor Calyampudi Radhakrishna Rao, (fondly known as Dr. Rao to his students and CR to his professional colleagues in US) was born on September 10, 1920 and died on August 22, 2023, at his daughter's place in Buffalo, New York just 18 days before his 103rd birthday. He was a legendary statistician, born along with modern mathematical statistics. He had the unique distinction of being one of three distinguished centenarian statisticians who lived in the century of the birth of the discipline and the century of its maturity, he having contributed immensely for its maturity. C. R. Rao adored statistics and adorns statistics. Here is a tribute to him.

Keywords: Cramer-Rao inequality, Rao-Blackwellization, Rao Distance, Information Geometry, Econometrics, Score Test

1 Introduction

The statistics legend C. R. Rao (Calyampudi Radhakrishna Rao) who lived a glorious life of 102 years passed away on August 22, 2023, at his daughter's home in Buffalo, New York. He was born in India on September 10, 1920.

Dr. Rao was always at the frontiers of statistics knowledge, and often those frontiers chased him as he stretched those frontiers. He communicated a paper to the *Proceedings of the National Academy of Sciences* in his 100th year (2020) (Baisuo Jin, C. R. Rao, Yuehua Wu, and Li Hou, "Estimation and model selection in general spatial dynamic panel data models", *Proceedings of the US National Academy of Sciences*). He visualized the role of artificial Intelligence and machine intelligence as far back as 1969 (C. R. Rao, Computers and the Future of Human Society, Statistical Learning Society) that are just now being explored. Rao co-edited a volume of *Handbook of Statistics*, Vol. 49, titled *Artificial Intelligence* for Elsevier in 2023 (with S. G. Krantz and Arni Rao (the co-author of this tribute)). C. R. Rao was the series editor of *Handbook of Statistics* series for more than three decades and coedited several volumes.

Dr. Rao co-authored a book on artificial intelligence in his 100th year (2020) with Pereira and Oliveira (Pereira, Basilio de Braganca, C. R. Rao, and Fabio Borges de Oliveira, *Statistical Learning Using Neural Networks*, C. R. C. Press, Chapman, and Hall). We three, the authors of this article, are privileged to be associated with him as his students at Indian Statistical Institute (I. S. I.) or collaborators. Speaking of his teaching, he would ask his students, by choosing one student at random in each class, to go to the blackboard and derive from a practical example a very useful statistical result-such as the best estimator of a parameter, and in a later class prove the same result using the Rao-Blackwell theorem. In the afternoon the same result or similar ones are worked out with a real life problem with data. Quite often he used an interesting example with a joke and with his characteristic chuckle and smile. He made his students, several hundreds of them, feel nostalgic in later years about the teaching at the Research and Training School of the Indian Statistical Institute that he oversaw. See obituaries on C. R. Rao published in *Nature* (2023 by Peddada and Khattree), in *Science* (2023 by Banks and Clarke), and in the *IMS Bulletin* for his role in elevating the Indian Statistical Institute and the official Statistical Systems in India.

C. R. Rao was born around the same time the fundamental concepts of sufficiency, efficiency, and likelihood were introduced by Fisher. Twenty years later Rao studied statistics under the tutelage of Professor P. C. Mahalanobis with the dictum that "statistics has a purpose", and exploited these concepts of sufficiency, efficiency, and likelihood and used them for good purpose, thus making breakthroughs in statistics. He had keen interest in all branches of knowledge where statistical thinking is useful. This is the hallmark of the Indian school of statistics developed by Mahalanobis and Rao over fifty years (1930-1980). Statistics was developed in India as a core subject surrounded by various branches of science, with the Banyan tree serving as the logo. Dr. Rao developed statistical theories and methods from practical problems and applied them to a variety of disciplines such as anthropometry, biometry, psychometry, econometrics, engineering, and environmental science. That is why the citation of the US Medal of Science to Dr. Rao, awarded by President Bush in 2002, stated "..... for his pioneering contributions to the foundations of statistical theory and multivariate statistical methodology and their applications, enriching the physical biological, mathematical, economic and engineering sciences ...". Fisher founded the Biometric Society in 1947 (when Rao was his Ph. D. student) with five regional Chapters, comprising of Britain, France, Australia, United States, and India. The Indian Region of the Biometric Society was established with P. C. Mahalanobis, who received the coveted Weldon medal for his work on biometry from the University of Oxford in 1944 as its President, and C. R. Rao as the Secretary. Rao would later become the President and Honorary Life Member of the Biometric Society.

C. R. Rao was a brilliant student of mathematics since the beginning (*Significance* 2020 and *Significance*, 2012). He had received very good training in high school to master's in mathematics at Andhra University, India. According to the MR database of the American Mathematical Society, C. R. Rao published 19 articles during 1941-1945 from Kolkata, India before he went to pursue a Ph. D. in Cambridge, England. Before his groundbreaking article of 1945 titled "Information and the accuracy attainable in the estimation of statistical parameters" which was published in the *Bulletin of Calcutta Mathematical Society*, Rao worked on topics like general probability, combinatorial designs, characterization of probability distributions, problems in number theory, etc., C. R. Rao completed Ph. D. in 1948 from Cambridge under the supervision of R. A. Fisher and returned to Indian Statistical Institute, Kolkata as a Professor, a position created to him by P. C. Mahalanobis, the founder Director of the Institute. Rao lead teaching and research activities at I. S. I. for three decades. Speaking to an audience at I. S. I., R. A. Fisher remarked that at one time most of the professionally trained statisticians in the world were from India and they were all trained by Rao. After retiring from I. S. I., Rao moved to US and spent three more decades leading teaching and research in statistics, and multivariate statistics, in particular at the University of Pittsburgh and the Pennsylvania State University.

2 The impact Rao's Work on Various Branches of Sciences and Engineering

We start by describing Rao's path-breaking contribution of introducing the concept of differential geometry in statistics and deriving the all too powerful result-the Cramer-Rao inequality.

2.1 Rao's work prior to 1948

Problems arising out of geometrical thinking have been of interest to Rao since the beginning. For example, in the article in 1941 published in *The Mathematics Student*, using a method to compute the volume of a prismoid in $n-$space, Rao showed how the geometrical probability functions can be treated as useful tools. In the article published in 1944 in *Science and Culture*, Rao developed a method of analysis of experimental block design where an experiment in randomized blocks is not independent of each other. In a series of articles in 1945 published in *Sankhya* and *Science and Culture*, Rao considered Markov's theorem on matrices and generalized it to make it flexible to test linear hypotheses. He also considered the equality of the means of a $p-$variate normal population whose covariance matrices are unknown. Rao alone and also along with R. C. Bose and S. Chowla in a series of articles published in 1945 in journals: *Bulletin of Calcutta Mathematical Society, Proceedings of National Academy Sciences, India Section A, Proceedings of National Academy Sciences, India, Proceedings of Lahore Philosophical Society*, etc., studied a variety of problems arising in Galois fields, theory

of congruences, quadratic functions of the type $x^2 + ax + b \,(\text{mod}\, p)$, where p is an odd prime, a and b are arbitrary integers.

For some integer c, the integral order of $x^2 + ax + b \,(\text{mod}\, p)$ is defined as the least positive integer n such that

$$x^n \cong c \left(\text{mod}\ p, x^2 + ax + b\right). \tag{1}$$

By the age of 25, Rao published more than 18 articles including the monumental article mentioned at the beginning of this tribute that helped create a solid foundation for modern statistical science through Cramér–Rao inequality, and Rao–Blackwell theorem. This article also created a new subject area called information geometry. We will describe the basics of the 1945 *Bulletin of Calcutta Mathematical Society* article of Rao and some necessary technical elements of it in the following paragraphs. Let

$$\sum_{n=0}^{\infty} a_n \quad \text{and} \quad \sum_{n=0}^{\infty} b_n$$

be two convergent series such that

$$\sum_{n=0}^{\infty} a_n \geq \sum_{n=0}^{\infty} b_n,$$

then

$$\sum_{n=0}^{\infty} a_n \log \frac{b_n}{a_n} \leq 0. \tag{2}$$

In the inequality (2), a lot of information is stored, and based on the values of the sequence of terms of a_n and b_n, the future events can be predicted. The questions like how much uncertainty can be born in the future based on random variables defined in a space or the level of uncertainty that can be allowed based on the sequence of random variables within a space can be explained using the above kind of inequality or in its continuous equivalent. Fisher information measure, say, $\mathcal{F}$ on a parameter θ contained in a random variable X (in some space S) having continuous probability density $\phi(.,\theta)$ with $\sigma-$finite measure ν is defined by

$$\mathcal{F}(\theta) = \int_{-\infty}^{\infty} \frac{\partial^2 \text{log}\phi(X,\theta)}{\partial\theta^2} \phi(X,\theta) dx. \tag{3}$$

The probability density $\phi(X,\theta)$ is assumed to be differentiable with respect to θ for any

measurable set $C \subset S$, and

$$E\left(\frac{\partial \log \phi}{\partial \theta}\right)^2 = \int_{-\infty}^{\infty} \frac{\partial^2 \log \phi(X,\theta)}{\partial \theta^2} \phi(X,\theta) dx.$$

C. R. Rao in his 1945 ground-breaking article asked the question "What do we mean and know by the information on an unknown parameter θ?" and systematically answered it. He explains that the extent to which uncertainty regarding the unknown value of θ is reduced as a consequence of a prior observed value of X.

Suppose n sample observations $x_1, ..., x_n$ with a parameter θ are estimated by the function $t = f(x_1, ..., x_n)$, then Rao showed that

$$\text{Var}(t) \not< \frac{1}{I}, \tag{4}$$

where $\text{Var}(t)$ is the variance of t, and $I = \text{Var}\left(\frac{1}{\phi}\frac{d\phi}{d\theta}\right) = \text{E}\left(-\frac{\partial^2 \log \phi}{\partial \theta^2}\right)$. Rao further considered information matrix with several unknown parameters $\theta_1, ..., \theta_q$ with probability density function $\phi(X, \theta_1, ..., \theta_q)$. Suppose $t_1, ..., t_q$ be the estimates of $\theta_1, ..., \theta_q$ with the joint distribution $\Phi(t_1, ..., t_q; \theta_1, ..., \theta_q)$, then the information matrix on $\theta_1, ..., \theta_q$ due to $t_1, ..., t_q$ is defined by Rao as $\|F_{ij}\|$, where

$$F_{ij} = E\left[-\frac{\partial^2 \log \Phi}{\partial \theta_i \partial \theta_j}\right]. \tag{5}$$

Further, Rao considered distance between two populations (which is now popularly known as Rao distance) with two sets of parameters $\theta_1, ..., \theta_q$ and $\theta_1 + \delta\theta_1, ..., \theta_q + \delta\theta_q$. Using Riemanian geometry, Rao showed that the geodesic distance between these two populations as

$$ds^2 = \sum\sum g_{ij} \delta\theta_i \delta\theta_j, \tag{6}$$

where

$$g_{ij} = E\left[\left(\frac{1}{\phi}\frac{d\phi}{d\theta_i}\right)\left(\frac{1}{\phi}\frac{d\phi}{d\theta_j}\right)\right]. \tag{7}$$

Arni Rao and Steven G. Krantz extended the Rao distances to complex planes by using conformal mapping principles of complex planes. In a series of collaborative articles (Arni Rao, *Handbook of Statistics*: *Geometry and Statistics*, Vol. 46, pp: 401-464, Steven Krantz and Arni Rao, *Handbook of Statistics*: *Geometry and Statistics*, Vol. 46, pp: 3-19, Arni Rao and Steven Krantz, *Handbook of Statistics: Information Geometry*, Vol. 45, pp: 43-56, Arni Rao and Steven Krantz, *Cell Patterns*, 2020) they showed that such extensions have implications for virtual tourism technologies.

2.2 Rao's work after 1948

During 1948-2023, C. R. Rao developed individually and also through collaborative efforts, various statistical methods, and theories, which include linear statistical inference, multivariate analysis, combinatorial designs, information geometry, orthogonal arrays, M-estimation, second-order efficiency, characterization of distributions, Minimum Norm Quadratic Unbiased Estimation (MINQUE) theory, random coefficient modeling, and matrix theory and their applications in statistical inference (for example, see, Efron B. et al., *Significance,* 2020, Andrews G. E. et al. *Notices of the AMS,* 2022, Khattree, R. et al. *Bulletin of the IMS,* 2023, etc.).

C. R. Rao published more than 15 books and over 475 research articles throughout his research career spanning over 80 years. His work impacted research in other fields such as computer science, especially in signal processing, physics, information geometry, biology, anthropology, economics, etc. He was conferred an Honorary membership of IEEE for his pioneering contribution to signal processing through his Cramer-Rao inequality. Apart from receiving various prizes and medals Rao was honored with 40 honorary doctorates from various countries. He was also elected as the President of several statistical societies. .He was elected as a member of several national academies of science, including the US Academy of Science, Indian Academy of Science, the Third World Academy of Science.

Cramer-Rao inequality is used as an industry standard for noise reduction in communications and signal processing for which IEEE honored Rao with a distinguished Honorary Membership of IEEE during the platinum jubilee year of its Signal Processing Society. The orthogonal designs developed by Rao constitute the engineering designs to enhance productivity promoted by Taguchi. Cramer-Rao bound is used as Quantum Cramer-Rao bound in quantum physics. Almost all existing theories of physics, and some new ones, were derived from a principle of reaching a suitably defined Cramer-Rao bound (Frieden (2004)). Rao's score test, Fisher-Rao distance, MINQUE theory, his concept of second order efficiency, etc. are examples of his contributions of lasting value.

3 The impact of Rao's work on Econometrics and Other Social Sciences

3.1 Rao's Contributions to Econometrics

In this section we examine, in a bird's eye view, the work Rao did in the field of social sciences and the impact it had on those disciplines. Mahalanobis and Rao initiated econometrics research in India by organizing a two-day session of the international econometric society meetings in India in December of 1951, the same year Mahalanobis was elected a Fellow of the International Econometric Society in recognition of

his work on large scale sample surveys and National Income Statistics for the United Nations Statistical Commission. A decade later in 1960 Rao organized a special session in the Indian Science Congress on the applications of mathematics and statistics to economics. The same year he and N. S. Iyengar organized the first Indian Econometric Conference at the Indian Statistical Institute. Rao was the Founder President of the Indian Econometric Society (1970-75), and he was elected the Fellow of the International Econometric Society in 1972. Rao organized a Workshop on the Database of the Indian economy and helped Mahalanobis establish the Central Statistical Organization and State Statistical Bureaus for providing reliable economic statistics needed for national planning. For the first time in the history of the world, statistics and operations research were employed, through the active participation of an academic institute, to prepare economic plans for development affecting the lives of millions of people. Rao played a significant role in this effort, along with Mahalanobis.

Econometrics deals with both observed variables (price index, current income, etc.) and unobserved conceptual and or subjectively defined variables (expected price, lifetime income, etc.). Econometrics appears mainly in three forms: i) as a system of independent equations in which both errors in variables and errors in equations appear (1926-1944), ii) as a system of independent equation with errors in equations but not in variables (1944-) and iii) as a system of equations in observables and non-observables (factor analysis). Rao made fundamental contributions to statistical methodology needed in each of these three different approaches. Ragnar Frisch, who founded and pioneered the development of econometrics as a new discipline, conceived it as a confluence of mathematics, economics, and statistics whose objective it was to understand and tame the economic environment. A paper by Frisch and Mudgett published in 1931 (Ragnar Frisch and Bruce D. Mudgett "Statistical Correlation and the Theory of Cluster Types", *Journal of American Statistical Association*) is an excellent exposition of his intuitive conceptualization of statistical modeling in economics, on how to use sample information to impose parametric restrictions to identify and estimate linear models. Visualization of clustering of sample observations could suggest what parametric restrictions characterize that cluster. It is this duality between the parametric and sample spaces, given an assumed probability distribution (and the likelihood function), that characterized the inferential work of Rao using differential geometric concepts. Thus, Rao was both a Fisherian and a Frishian.

Frisch interpreted economic variables as observed and unobserved, and as endogenous, and exogenous. Endogenous variables are those that are explained by the economic model of the economic system, and exogenous variables are variables that are determined outside the economic model and affect the variables in the economic system but are not affected by them. An econometric model is a simultaneous system of as many equations as there are endogenous variables that impose constraints on the

free movement of variables in a finite-dimensional space. Thus this system gives rise to the correlation between variables on restricted hyperplanes or hyper surfaces rise to the correlation between variables and the confluence of the variables on restricted hyperplanes or hyper surfaces.

In connection with the above errors in variables model, used as an exploratory model to discover economic laws, Frisch posed the following question to Neyman at the 1936 European meetings of the Econometric Society: if two endogenous variables are linearly related to an exogenous variable measured with error by two equations with errors in equations, under what distributional assumptions on the errors in variables and errors in the two equations does there exist a linear regression relation between the two endogenous variables. The motivation for this question was to determine in a very simple two equation model under what conditions a structural equation (with a possible economic interpretation) exists between two endogenous variables with one exogenous variable. A complete answer to this question was provided by Rao just six years later in 1943 in his Master's thesis (as cited in C. R. Rao "Note on a problem of Ragnar Frisch", *Econometrica*, Vol. 15, 1947). In 2021 as a birth centennial tribute to Rao, Prakasa Rao and Kumar extended this result of Rao (*Journal of Quantitative Economics*, Vol. 19, 2021) by finding conditions under which two endogenous variables linearly related to two exogenous variables, measured with errors (and hence stochastic) would admit two independent linear regressions between the two endogenous variables. This is a typical model of a partial market equilibrium for a commodity as a system of two simultaneous equations, one a demand equation and the other a supply equation. Thus, Rao can be regarded as one of the pioneers in the development of econometric theory formulated by Frisch with both errors in variables and errors in equations. While this line of research proposed by Frisch with errors in variables was abandoned by econometricians, Rao followed this line of work and developed a new field within statistics with the name characterization of probability distributions. The application of this method to econometrics, however, remained underexplored.

The Cowles Commission for Research in Economics developed econometric theory dealing with identification and consistent estimation of a system of linear econometric relations employing likelihood and quasi-likelihood methods. In doing so it used models with no errors in variables and depended on the prevailing theories of statistical inference in linear models, not recognizing that most of that work was done by one person, C. R. Rao. The econometricians' concept of identification is nothing but a special case of the estimability concept developed by Rao (when the sample size is very large or when we are dealing with the entire population rather than a finite sample). Likewise, the efficient estimation of the unknown parameters in econometrics is nothing but obtaining consistent estimators whose variances are as close to the Cramer-Rao variance bound as possible. Cramer-Rao bound and Rao-Blackwellization are important

tools for this purpose. At a time when econometricians were analysing identification and estimation, model by model, Thomas Rothenberg ("Efficient Estimation with a Priori Information", *Cowles Commission Monograph*, No. 23, 1973) took a more general approach. He showed that in a very broad class of models, local identification depends on the information matrix: if it is non- singular in a neighborhood of the true parameter value, the model is identified in that neighborhood. Thus he related identification to estimability concept of Rao. The problem of estimating the parameters of simultaneous equation systems in econometrics is shown by Rothenberg to be a special case of the general theory of statistical inference developed by Fisher and Rao, the basic tool being Cramer-Rao inequality. Rao-Blackwellization is very rarely used in econometrics and there is a lot of scope for further research in this area (For Rao-Blackwellization in econometrics see H. D. Vinod, "Rao-Blackwellization using Bootstraps for pre-test and instrumental variables estimator", *Journal of Quantitative Economics*, Vol. 20. 2022, Supplement, C. R. Rao Birth Centennial Issue.

Rao promoted applications of new statistical methods in econometrics by co-editing *Handbook of Statistics* on Econometrics with Maddala and Vinod (G. S. Maddala, C. R. Rao, and H. D. Vinod, *Handbook of Statistics*, Vol. 11, *Econometrics*, 1993, Amsterdam, North-Holland; G. S. Maddala, C. R. Rao, *Handbook of Statistics*, Vol. 14, *Finance*, 1996, Amsterdam, North-Holland; G. S. Maddala and C. R. Rao, *Handbook of Statistics*, Vol. 15, *Robust Inference*, 1997). For C. R. Rao's contribution to the application of big data analysis in Finance see T. J. Rao ("Influence of C. R. Rao in financial statistics", *Macroeconomics and Finance in Emerging Market Economies*, 2023).

In his *Econometric Theory* interview with Bera, Rao remarked that he was unhappy with economists not refining the measurements of variables, and in their not finding new relevant variables to improve prediction. Thus, he was suggesting the importance of errors in variables models and in devising methods to recognize the possibility of omitting relevant variables that are correlated with the included variables. It is easy to reformulate a model with omission of variables that are correlated with included variables as a model with included variables but with random coefficients (C. R. Rao, "The theory of Least Squares when the parameters are stochastic with application to the analysis of growth curves", *Biometrika*, Vol. 52, No. 3/4, 1965). Rao extended the standard linear models to linear random coefficient models. Thus, factor analysis models and models with random coefficients, and characterization of distributions, all developed by Rao, have a great future in economics. The method of instrumental variables plays an important role in obtaining consistent estimators. What is the best choice of an instrumental variable? This choice can benefit immensely from a differential geometric approach pioneered by Rao applied to the empirical likelihood function (Marriott, Paul, and Mark Salmon, *Application of differential geometry to econometrics*, Cambridge University Press, 2020).

For a more detailed description on Rao's work in econometrics one may see other references(G. S. Maddala, P. C. B. Phillip, and T. N. Srinivasan, *Advances in Econometrics and Quantitative Economics*, Wiley-Blackwell, 1995; Bera, Anil and Aman Ullah, "Rao's Score Test in Econometrics", *Journal of Quantitative Economics*, Vol. 7, No. 1: T. K. Kumar, H. D. Vinod, and S. Deman, "Dr. C. R. Rao's contributions to economic science", Proceedings of the Indian Academy of Sciences *Proceedings – Mathematical Sciences*, 130, 45 (2020); B. L. S. Prakasa Rao and H. D. Vinod (Editors), *Journal of Quantitative Economics*, (2022), Dr. C. R. Rao Birth Centenary Issue.

3.2 Rao's Contributions to other Social Sciences

Griliches showed that it was a mistake to abandon the errors in variables models just around the time that economic theory required their use. He formulated errors in variables model, with both observed and unobserved variables, as factor analysis models (Zwi Griliches, "Errors in variables and other unobservables" Henry Schultz Lecture, 1974, Econometric Society, *Econometrica*, Vol. 42, No. 6). This is a statistical model which was analysed from a multivariate statistics perspective by Rao and Slater twenty five years earlier in 1949 ("Multivariate analysis applied to differences between neurotic groups", *The British Journal of Mathematical and Statistical Psychology,* Vol. 2, Issue 1). Factor analysis models and the methods developed by Rao and Bartlett are now widely used in economics, sociology, psychology, and educational research. For an interface between economics, statistics, psychology, and sociology dealing with factor analysis one may see Grliches' paper cited above.

4 Rao's Influence on Survey Sampling Theory and Practice

One of the authors (Pathak) started his professional career with interest in survey sampling theory, literally and figuratively at the footsteps of C. R. Rao. His office was just opposite Rao's, and the typical routine used to be whenever Pathak had an exciting new result he would show it to Rao, and Rao would comment: "Interesting result but not of publishable significance". One of us, Kumar, used to see Rao go for tea and snacks in the afternoon mostly with two close friends S. Raja Rao and S. J. Poti. We also used to see D. B. Lahiri and M. N. Murthy of National Sample Survey Organization visit Rao at his office for consultations. Pramod recollects, that there were other instances when C. R. Rao would make significant revisions in his work before publication, e. g. Pramod's paper on the efficiency of Des Raj and Murthy's estimators in one such case. Salem H. Khamis, Des Raj, D. B. Lahiri, and M. N. Murthy who worked in close collaboration with Rao and significantly contributed to the theory and practice of sample surveys. As a student of Mahalanobis, Rao's initial work on survey sampling was with the anthropometric survey in 1941 in which he developed the random permutation model for sampling (T.

J. Rao, "C. R. Rao's Influence on theory and practice of sample surveys", *Journal of Indian Society of Probability and Statistics*, 2020). Rao was very much involved with the National Sample Survey Organization in the sampling designs, setting up sample survey questionnaires, and in their analyses. It was V. P. Godambe's paper ("A Unified Theory of Sampling From Finite Populations", *Journal of the Royal Statistical Society*. *Series B* (Methodological), Vol. 17, No. 2 (1955), and his visiting position as a Senior Research Fellow at I. S. I. during 1958-1959 that stimulated the interest of Rao, Basu, and Pathak to apply the theory of statistical inference to sample survey methods. It was S. Raja Rao of the National Sample Survey Organization in Kolkata, a close friend of C. R. Rao, who brought Rao's attention to the importance of the concept of sufficiency in sample surveys. This concept of sufficiency and the associated Rao-Blackwellization were followed up by Basu and Pathak (D. Basu, "On sampling with and without replacement", *Sankhya*, Vol. 20, 1958, and P. K. Pathak, "On simple random sampling with replacement", *Sankhya*, Vol. 24, Ser A, 1962, and P. K. Pathak, "Sufficiency in sampling theory", *Annals of Mathematical Statistics*, Vol. 35, 1964). Hartley and Rao noted that the non-existence of minimum variance unbiased estimator was due to the class of estimators considered by Godambe, that depended on identification labels. They showed that once that dependence is dropped appropriately, Minimum Variance Unbiased Estimator (MVUE) exists (H. O. Hartley and J. N. K. Rao "A new estimation theory for sample surveys", *Biometrika*, Vol 58, No. 3, 1968).

Some of these developments resulted in pointing out serious problems when we make inferences with survey sampling with finite populations, problems such as the non-existence of MVUE, the likelihood surface being flat in the direction of some parameters, whether one should look for Rao-Blackwellization with a bad sample or a bad sample design or look for a better samples or better sampling designs. The non-existence of MVUE was also addressed by advancing various admissibility criteria without any practical significance. This prompted Rao in 1968 to request J. N. K. Rao, who was then visiting I. S. I., to deliver a few lectures on inference issues in sample surveys. After attending those lectures Rao prepared two papers addressing these issues (C. R. Rao, "Some aspects of statistical inference in problems of sampling from finite populations, (with discussion and reply by author)", In: V. P. Godambe and D. A. Sprott (eds.) *Foundations of Statistical Inference*, 1971, Holt, Rinehart and Winston of Canada; and Rao, C. R. "Some Problems of Sample Surveys", *Advances in Applied Probability*, Vol. 7, Sept. 1975). Rao reformulated the issues using a basic probabilistic structure to the inference problems along the lines of Kolmogrov's measure theoretic formulation. Noting that the question posed and the answer obtained in statistical inference depend on the sample space, parameter space, and the probability distribution assumed for the generation of the observations, Rao concluded: However, many controversies can be avoided if the basic issues are clearly stated and statisticians do not insist on a

monolithic structure for all problems of statistical inference. Much damage has been done by fashions and slogans in statistics introduced by theoretical statisticians who have no experience of handling live data and extracting information from them. Jointly with Mitra, Mathai and Ramamurthy, Rao edited the *Tables for Statistical Work* (C. R. Rao, S. K. Mitra, A. Mathai and K. G. Ramamurthy (1975), *Formulae and Tables for Statistical Work*, 1966, Statistical Publishing Society, Calcutta). This book included a brief review of basic sample survey estimates and their standard errors, in addition to random sample numbers and random permutations. These chapters of the tables are often consulted by students, scholars and survey practitioners. Here we also suggest referring to two volumes of the *Handbook of Statistics* that Rao edited one with P. R. Krishnaiah in 1988 (*Sampling*, Volume 6) and the second with D. Pfeffermann in 2009 (*Sample surveys: Design, Methods and Applications*, Volume 29A).

5 Conclusions

In conclusion, we believe, the teaching and research tradition set by the legend C. R. Rao continues worldwide and his students, textbook users, collaborators, and friends take forward the rich culture of theoretical and applied research. The C. R. Rao Advanced Institute in Hyderabad, India started to conduct high-class original work in all the research areas that C. R. Rao inspired. We wish good luck to all the scientists. Finally, it gives us a great feeling to know the legend personally who lived a cheerful and inspiring life till the end.

Acknowledgements

We thank the editor-in-chief Dr. Danutė Krapavickaitė for inviting us to write about Dr. C. R. Rao and for editorial comments and edits. We thank Dr. J. N. K. Rao and Dr. T. J. Rao for helpful discussions and edits on a previous draft.

The Survey Statistician, 2024, Vol. 89, 29-36

The first fifty years of the IASS, some thoughts

Carl-Erik Särndal

Professor emeritus, Statistics Sweden

Abstract

This article recalls and re-examines a number of ideas, concepts and lines of thought that influenced theory and practice in survey statistics, in particular during the fifty years of the IASS.

Keywords: randomization theory, types of inference, rigorous statistical treatment, modeling, theorization, the nineteen-seventies.

1 Introduction

The International Association of Survey Statisticians (IASS), created in 1973, marked its first quarter-century by publishing a Jubilee Commemorative Volume, subtitled Landmark Papers in Survey Statistics. I am reminded of it, in commenting now on two quarter-centuries in the life of the IASS. I shall refer to it as The Association.

The jubilee volume was an unusual and thought-provoking initiative of The Association. It paints a portrait of survey statistics, as it was seen twenty-five years ago. It bears testimony to an era in survey science.

This volume remains today as a witness to a period of growth in survey science, a document of a certain value in the history of the discipline. Its content reflects a subjectivity that “a selection of the best” will invariably bring. Many other excellent articles were published over the years.

The purpose of the present article is not to review the nine-teen selected articles. This may be of considerable interest, but is not the objective here. Nor is it to trace the steps in the progress in survey science that those articles brought.

The article offers a perspective on survey science taking the volume’s excellent preface as the point of departure. I note how it introduces and justifies the selected articles in terms of statistical ideas, concepts and expressions that were in vogue at the time, often mentioned in the literature of not so very long ago, but which may since have fallen more or less into desuetude. This article is thus an essay on ideas that influenced the discipline. It expresses the author’s personal opinions and impressions. It makes no claims to a complete coverage of the field, makes no attempt to write the history of the field. The result is of some educational value for a younger generation.

A creative period in survey science began in the late 1960's. It questioned the established state of the discipline. New ideas and approaches had a profound impact. I like to think that the birth of The Association in 1973 was to some degree a reaction to the new directions. The Association was a welcome addition to the scope and activities of the International Statistical Institute (ISI).

2 The selection

As the preface mentions, The Association had trusted a committee with the difficult task of selecting "landmark papers" from the vast stream of literature in survey statistics, beginning early in the twentieth century. The resulting volume presents nine-teen full-length papers with publication dates between 1934 and 1989. The stated objective was to choose among articles from roughly the last fifty years, emphasizing the significance of the contribution, rather than quality of exposition or direct usefulness.

Eleven out of the nine-teen had appeared in the years from 1969, an unbalanced selection, from a time perspective. Recent work tends to come more readily to mind. But it also reflects the fact that the times from around 1970 brought significant and in a sense revolutionary progress.

All nine-teen articles in the volume are framed, to varying degrees, in mathematical formulation and language. Although not a display of "hard mathematics", this nevertheless suggests that formal expression helps to bring about "a seminal contribution", which requires not only the recognition of a practically important survey question, but also a convincing mathematical formulation, treatment and resolution. It is the kind of article that is capable of generating a stream of further contributions.

3 One exceptional name

Altogether twenty-nine names get credit for authorship or co-authorship in this collection of seminal contributions. One name stands out, as first co-author of three of the chosen nine-teen articles. Dated 1943, 1961 and 1983, they point to a remarkable forty-year activity span. The name is Morris H. Hansen (1910-1990). He deserves to be recognized also because he was the first president of The Association.

While at the United States Census Bureau, 1935-1968, Morris Hansen was highly influential in the discipline, one of the first to develop methods for statistical sampling. He made important contributions in many areas of surveys and censuses. He was one of the principal builders of the "probability sampling paradigm"; he vigorously defended it, when needed. He and his colleagues pioneered in viewing survey quality under a broad umbrella, covering survey errors of different kinds.

4 The rise and the high point of randomization theory

The preface notes that "the randomization theory of sampling … proposed a logic of inference based on confidence intervals". This grew out of the important theoretical advances in the 1930's by J. Neyman and others. Design unbiased estimation was a key feature in this logic.

The decades that followed saw randomization inference developed to perfection, in other words, the inference built on the randomization feature of the probability sampling design. Known selection probabilities of the identifiable population units was the key to this. The selection could be stratified, in two or more stages, in two or more phases, by probability proportional to size, and yet others. A vast literature from that era bears witness to a panoply of methods; although in essence just variations of one single method: probability sampling. Randomization inference became popularly known as design-based inference.

Probability sampling with design-based inference progressed rapidly and convincingly, approved by the general public and authorities. It became a recognized tool, capable of handling a great variety of situations. Some form of probability sampling could usually be designed and carried out to deliver accurate information needed about some aspect of society.

Prior to 1970 there was no need to qualify inferences about the finite population - confidence intervals and other forms – as “design-based”. It was the golden rule. But in short time after 1970 came a need for making distinctions.

Survey statistics was seen by some as a field closed in itself, unaware or ignorant of reasoning and methods in “general statistical theory”. The topic was typically taught as “sampling” in university courses, if at all offered. It had a reputation of a rather special, somewhat marginal, field of study within statistical science. Some survey statisticians felt this distinctly, and hoped for a change.

In those days, some senior stake holders no doubt considered survey statistics as an essentially complete and saturated field. Probability sampling was the uncontested methodology. The national statistical agencies applied it. The probability sampling paradigm had tremendous power behind it. It was a formidable task to challenge this bulwark.

Survey statistics and “sampling” had built its reputation on the privileged setting of the finite universe, composed of *N* identifiable objects, or units, ideally listed in a frame, perhaps together with known properties linked to individual units, such as membership in population groups of potential interest.

Survey statistics is a scientific field in its own right. As such, it uses concepts and ideas of statistical science. This became much more evident from the late 1960’s and on, when general statistical theory came to influence survey statistics as rarely before.

5 A period of change

In the new scientific discourse, “models” and “modeling” became key concepts. Was “modeling” something new? Not at all, as some defenders of “the traditional thinking” liked to remind. Modeling was in fact present much earlier. For example, planners of a new survey could rightfully claim that their choice of an unusual probability sampling plan, say, a complex stratification, or a complex sampling design in several stages, was the result of a conscious – although perhaps not declared – modeling effort, in the interest of accurate estimation, unbiased by virtue of the randomization theory.

However, from around 1970, models became more apparent and explicitly declared, both in regard to the sampling design and in the construction of estimators.

Ample warnings were voiced. To “rely on models” – possibly wrongful or misleading – might expose the estimates to severe bias. Despite a privileged position of design-based inference, the new forms of reasoning not only survived; they thrived and set their important mark on developments in survey statistics in the decades until today.

6 Theoreticians

The preface notes that in the 1950’s and 1960’s “theoreticians addressed the foundations of randomization inference”, in attempts to “integrate randomization inference into mainstream statistical inference”.

Who are “the theoreticians”? What is their role in this practical field of survey statistics? Were theoreticians less prominent, less influential, in those early path-breaking decades of survey sampling?

“Theoretician” may refer to someone highly knowledgeable in “general statistical theory”, especially advanced estimation theory, someone with a pronounced mathematical/statistical training and orientation, but with possibly little experience of the practical work in a national statistical agency.

Important conferences were devoted to survey statistics, especially to “its foundations”. There was introspection, attempts to place of survey statistics into “the mainstream” of statistical science. Two such occasions, in 1968 and in 1977, were at the University of North Carolina at Chapel Hill. A 1970 symposium at the University of Waterloo on “the foundations of statistical inference” had an important portion on survey statistics theory.

However, the preface strikes a somewhat negative note: some of the resulting theoretical work "seemed too abstract to practitioners and may have resulted in the perceived divide between theory and practice". Nonetheless, the decade that saw the creation of The Association was one of lively exchange and debate; a march in new directions.

7 A new era

From around 1970, new theories emerged. Terms were coined for proposed new types of inference for the finite population. They were integrated into the scientific language. Survey statisticians began to communicate in a language and with a terminology hardly needed before. New terms enriched the scientific language, to inform readers of articles and participants at conferences on the nature of a contribution.

"Superpopulation" was an exotic new term for an imagined infinite universe with certain assumed features, and assumed to have generated the actual finite population, from which sample selection then took place.

As the preface further notes, "theoreticians indicated that a modeling approach could be adapted to complex finite population structures and sampling schemes". A few visionary statisticians – theoreticians, in the eyes of some – had, in the years around 1970, the audacity to use postulated assumptions – more or less trustworthy relationships among variables – as the basis for addressing the substantive issues: the sample selection, the properties of estimators, such as their unbiasedness, variance and mean square error.

Model-based theory of inference for a finite population saw the light of day, holding that inference could well be based entirely on a modeled relationship between variables, notably that of the auxiliary variables with the survey variable(s).

In its pure form, this theory is model dependent. The validity of the resulting estimates depends on "the truth" of the model that the survey statistician ventured to assume. It was both a vulnerable theory, because the truth of the model can never be taken for granted, and a revolutionary theory, because it challenged the classical randomization theory, which had taken pride in delivering trustworthy inferences without any assumptions, valid whatever the form of the finite population.

Not surprisingly, the new ideas were at first controversial, albeit received by some as refreshing and vitalizing. Defenders of the classical design-based school received the new theory with a good deal of suspicion. Practitioners were at first hesitant to use methods that appeal more or less directly to "modeling".

Then came the theory called "model assisted (design-based) survey sampling". Capitalizing on advanced forms of modeled relationship among variables, it nevertheless preserved the precious design-based nature of the inferences. Model assisted design-based theory and methods became widely accepted and used in national statistical offices, especially fruitful in countries, such as Scandinavia, where reliable registers give ample supply of explanatory variables – those called "auxiliary" – for the model fit. The theory of calibration estimation of recent decades is a further outgrowth of this thinking.

8 Non-sampling aspects

The preface notes, somewhat apologetically: "Although non-sampling aspects of our subject, such as response errors, editing and imputation, are well recognized as of prime importance in the practice of survey work they have not always received the rigorous statistical treatment of topics such as sample design and estimation." A couple of the nine-teen articles do deal with non-sampling aspects.

A classical distinction in survey statistics was that between sampling error and non-sampling error. One can claim that a disproportionate part of research and published work focused on the former type, on methods to reduce that error with the aid of efficient sample design and advanced estimation theory. Much of this theoretical work was set in ideal conditions: the absence of nonresponse, measurement error and other imperfections labelled as non-sampling error. As some critics

maintained, it was a focus on “finding a (marginally) better estimator” under ideal conditions, with results often of limited use in practice; a display of “rigorous statistical treatment” in unrealistic settings.

9 Rigorous statistical treatment

When the preface uses this concept, it sounds as a self-evident obligation, a high ideal, for survey statistics to live up to. What does it require, today or in the future? How rigorous must the treatment be? Is it a question of mathematical rigor, or some other kind? For example, are all of the competing theories in survey science commensurate with the concept?

In one interpretation, the concept asks for a discourse where powerful theoretical tools can be brought to bear on the practical question, in a formal language and structure - as the word “rigorous” begs - rather than just a fleeting verbal discussion.

Also, “rigorous statistical treatment” obliges producers of statistics to keep users informed, in appropriate statistical measures, on the reliability, trustworthiness, and probable error of estimates. A part of this should be probability statements on accuracy - such as 95% confidence intervals - interpretable in one of the acclaimed theory frameworks, design-based, model-based, Bayesian or yet other.

But, as the preface hints, to genuinely accomplish this in the presence of the various non-sampling errors proved difficult. It is, somewhat paradoxically, an unresolved dilemma for the discipline.

Research in recent decades did try to make up for a perceived lack of “rigorous statistical treatment” with respect to non-sampling errors, but without any complete or decisive result.

In particular, much attention was devoted to one of those imperfections, the rapidly growing problem of survey nonresponse, and the bias it leads to in the estimates. An improved understanding of respondent motivation and behavior did help to reduce nonresponse at the data collection stage. At the estimation stage, advanced nonresponse bias adjustment methods helped to improve the quality of statistics produced.

Was the “rigorous treatment” of non-sampling error too much to ask? Will users do without this assurance in the future, and just accept declarations that “the numbers were produced with the best methodology we know”, without any further assurance of closeness of estimates to the truth.

The roots of the failure are traceable to the rigor imposed by randomization theory: its implicit obligation of one hundred percent response rate, and from precisely those in the designated probability sample. Unless “precisely those” are obtained, the theory is strictly speaking transgressed and compromised; estimates are biased.

It is a vulnerable theory, hardly made for the tough conditions of today’s survey climate. That the theory worked well, with some amendments, for as long as it did may surprise us now. Certainly, in the theory’s youth, some eighty or more years ago, nonresponse was low, or negligible, hardly worth worrying about. But today, the theory sidesteps a reality that surveys now face.

10 Theorization

The preface sees a “rigorous statistical treatment” as desirable, yet regrets that some published theoretical work may be seen as abstract, causing a “divide between theory and practice”. Realistic and convincing theorization can indeed pave the way for further advances, on non-sampling error as on sampling error.

How should research in survey science accomplish a proper balance between purely theoretical progress and a practical utility? It is a difficult question. Theorization can certainly be pursued for its own sake, without much chance of being applicable in practice, yet qualify as high caliber research from a technical point of view, as evidenced by hundreds of articles published in the last eighty years, many them “run-of-the-mill” papers, as will happen in what philosophers of science call “a period of normal science”.

The degree of theorization depends on the nature of the topic at hand. Some topics in survey statistics seem to strike researchers as more inviting, and more directly suited, for theorization and a certain abstraction. Small area estimation is of this kind. Other topics, although important for practice, seem to resist, or discourage, a desirable theorization. Non-sampling errors tended to be of this kind.

However, much valuable work on new important themes was addressed in the literature with little mathematical formalization, in a discourse framed in theory and concepts proper to other sciences, notably the behavioral sciences.

11 Subpopulations

"Small area estimation" became a prominent topic in recent decades, in a sense a unique topic. Such estimation has always resided in the realm of survey statistics, but only in the 1970's did the width of the question catch the attention of the theoreticians and other stakeholders in survey statistics.

An immediate challenge lies in the title: The words "small area" warn about a possible shortage of data from within the area itself, under any realistic cost frame for the entire survey. To overcome this dilemma by the development of new theory was the answer. Advanced theorization came quickly and is continuing, a beneficial playground for "the theoreticians".

The topic was a welcome and attractive avenue for survey science, which at times seemed rather short of challenges that could inspire both advanced theorization and deliver results of prime importance for practice.

"Small area" proliferates the idea of "a set of identifiable objects" as a target of inference. The set is not only finite, but smallish finite. "Small area" is to be understood in a wide sense, as the estimation for subpopulations, also called domains, contained within the greater population concept. Focus is usually not on one single domain, but rather on estimation for many, all contained in the large entire finite population, also a target of inference in the same survey.

Domains of interest are often placed in the geographical or administrative context of a country. Who can deny the need for accurate information for smallish but politically critical regions of a country? More generally, domains of interest can be any subgrouping of the well-defined population at large.

12 Recent trends and future prospects

Certain topics of importance did not set a clear mark on the jubilee volume of twenty-five years ago, understandably, since much of the work on those is relatively recent. Among them are *survey quality*, *total survey error*, *survey cost*, and *mixed data inputs* for statistics production. They deserve to be mentioned in these concluding sections.

Survey conditions changed, "for the worse" in the view of some, during the second twenty-five year period of The Association. The demise of the probability sampling paradigm has far-reaching consequences that will take time for the national statistical agencies to comply with and adjust to.

One must recognize that research in survey statistics is steeped in two different lines of scientific discourse. One is held in more or less formal mathematical language. The other progresses rather as a verbal discourse, framed in concepts and ideas proper to "survey generalists". To say that "the theoreticians" shape only the former type of discourse is not correct. The second type also draws on theory, but coming more likely from the behavioral sciences. Survey statistics is indeed interdisciplinary in character.

Behind this division lies the different educational backgrounds of survey statisticians. One stream is trained to work comfortably in a formalized mathematical idiom. Another is trained in concepts and ideas from fields with a less formalized structure. Both kinds contribute. The Association embraces both categories and encourages a fruitful co-operation between the two.

Survey Quality and Total Survey Error are areas that attracted considerable attention in the survey statistics literature of recent years. The latter has taken on the role of a conceptual framework, a

central organizing structure of the field of survey methodology. As such, it has filled an important need.

The elusive concept “quality”, especially “quality of survey statistics”, came to the fore and inspired much thought and drew much attention. Statisticians asked themselves: Survey quality, what is this? Quality, as subsequently argued, is a multi-faceted concept. Several national statistics offices spent considerable time to identify and elaborate their own vision of the essential dimensions, as many as six or more, of survey quality, so as to back up their mandate to provide valid numbers for the nation. Although no doubt helpful to some, those dimensions remain little more than “just names”, of limited value to many. They seem to escape attempts at synthesis, proper measurement and “rigorous statistical treatment”.

13 Challenges

One can claim that survey science is a science that is not free to act on its own behalf. It is always driven by extraneous conditions, notably the cost consideration. Statistics should be accurate and timely and relevant, but must not cost too much to produce. Quality statistics is a goal, naturally, but always subject to survey cost.

The jubilee volume does not dwell directly on the critical role of survey cost. Nevertheless, cost continues to be a driving force for survey theory and development.

The cost aspect is not new. It was important already a hundred years ago, when theoretical progress and empirical evidence finally convinced the statistical community that a complete enumeration of the country´s entire population was not necessary. National statistics of excellent – or at least sufficient – accuracy were obtainable at a reasonable cost with “just a sample”, more specifically a probability sample, just a modest fraction of the population. Public trust in such “low cost but accurate statistics” gradually developed.

For a long time, the statistical profession took pride in and thrived on this trust, on behalf of authorities and the general public, until, some decades later, high cost, high non-response and other non-sampling survey errors darkened the outlook.

High nonresponse rates in recent times made data collection from the units in the designated probability sample cumbersome, time consuming and delaying. Multiple attempts at contact with those particular units drove survey cost up. And even after a costly effort, remaining nonresponse bias plagues the estimates and has to be “adjusted for”, by a variety of suggested methods.

“The high cost syndrome” is to a degree responsible for the demise of the probability sampling paradigm. Meanwhile, data input from less expensive “alternative data inputs”, or “mixed data inputs”, have become ingredients that one cannot afford to disregard in building future theories for survey science, for “statistics at a reasonable cost”.

If probability sampling shall vanish from the survey science scene, the nonresponse problem in its original sense – the failure to get response from precisely those in the designated probability sample – will also disappear. On a positive note, this will liberate the discipline from burdensome or embarrassing chains. But how well the future data inputs shall represent the finite population is a matter that has to be addressed through other terms and concepts.

An open question is the validity of the statistics produced; how do we guarantee it? A “rigorous statistical treatment” seems mandatory. But what this attractive but elusive concept will require in the future is not clear. It is evolving over time, in tune with an evolving society and changing survey conditions. Nevertheless, the concept is an indispensable guideline for the discipline and the scientific community.

“Theorization” will take on new forms and directions. The sharp distinction between sampling error and non-sampling is expected to lose its contours. The fixation on theory development within the probability sampling paradigm is likely to disappear. Powerful new determinants will come from non-statistical considerations, such as survey cost.

Some prestigious sciences – such as physics – are capable of presenting to the world an admirable image of cumulative progress: On the basis of truths so far established, we find out more and more, we establish piece by piece further insights into a fixed – but highly complex – component of the universe, such as the atom. The universe that survey science is addressing is not unalterable, but a changing and evolving world.

14 Conclusion

The Association's Jubilee Volume, from about twenty-five years ago, was my guiding light in these notes. I commented on issues and lines of development that seemed important. Times are changing. If The Association were to now paint a portrait of the field that we call survey statistics, what would a new commemorative volume contain? Which contributions stand out today as the landmarks and breakthroughs that will guide the discipline into the future?

The author acknowledges the contribution of two anonymous reviewers of this article.

The Survey Statistician, 2024, Vol. 89, 37-46

Past Presidents Testimonials on IASS

Annamaria Bianchi[1]

[1] University of Bergamo, Italy, annamaria.bianchi@unibg.it

Abstract

To celebrate the 50th birthday of the International Association of Survey Statisticians (IASS), this paper reports past Presidents' personal experiences with the Association, offering an interesting angle that allows to read the history of IASS through the lives of its Presidents. Their view on the future of the Association is also provided. After a brief introduction, testimonials from fifteen past Presidents are reported.

Keywords: International Association of Survey Statisticians, History.

1 Introduction

The IASS is celebrating its 50th anniversary! Since its foundation in 1973, many changes have happened in survey research and many more are still to come. Such changes are reflected in the long history of the Association, under the lead of its Presidents. Their personal experience with the Association is intertwined with the survey research landscape and influenced the evolution of the Association itself.

It is thus interesting to hear of past Presidents' personal experiences with the Association and their view on the future of the IASS to adapt to new scenarios. To this purpose, we asked them the following three questions:

1) How and why did you get involved in the IASS?
2) Has the involvement in the IASS brought you any benefits in your career?
3) Which suggestions would you give to the IASS to make a strategy for the future?

We collected answers from fifteen past Presidents. They are reported in Section 2.

2 Testimonials

2.1 Monica Pratesi – Past President 2021-2023

1) My involvement has been a natural process as a PhD student in Applied Statistics, and then junior researcher at the University of Florence, Italy. As a PhD student I had the possibility to attend lectures from legendary survey statisticians, as Leslie Kish, and to be in contact with many researchers and professors, members of the IASS.

Since the 90s I have been participating in the ISI congresses and activities, Luigi Biggeri and Andrea Giommi had told me about the International Association of Survey Statisticians, I attended the meetings and then joined the IASS.

2) IASS is a network of researchers and as all networks gives you many advantages. Sometimes scientists are considered as isolated individuals, ignoring the network effects of collaboration. Involvement in the IASS has definitely launched me in an open network of collaboration. I do not consider only the network effects of productivity defined as paper counts, but I am convinced that the discussion during the scientific activities of the IASS (national and international conferences, sessions, seminars) has surely improved the prominence and the impact of my publications and research. All this has certainly allowed me to have an adequate CV in the selection and appointment in Italian Universities and various bodies of international NGOs, the European and Italian official statistical system and to be elected president of the IASS for the period 2021-2023 and, then, to be Director of the Department for Statistical Production of the National Institute of Italian Statistics (Istat) for the period 2022-2023.

3) My wish is that all objectives of the IASS 2022-23 strategic plan be achieved. I would stress the role of the representatives of the various countries to reinforce local participation and at the same time would continue and strengthen the organization of online activities. In my vision IASS is an important form of social capital and online webinars/activities could make the IASS a more "equally" distributed social capital that shapes scientific results and easily supports the dissemination and delivery of research on survey statistics.

Also, the creation of strong and successful partnerships with other scientific associations and between researchers and stakeholders will increase the value of IASS social capital.

2.2 Denise Silva – Past President 2019-2021

1) My first contact with IASS was many years ago, in 1993, when I attended the 49th ISI World Statistics Congress in Florence. At that time, I was studying at the University of Southampton and my PhD supervisor, Prof. Fred Smith, was IASS Vice-President. On that occasion, I learned about the many experts and young professionals in our area who were members of the IASS community. The Association was known for promoting capacity building and for providing precious networking opportunities. After that, I made an effort to attend, not only the subsequent ISI World Statistics Congress, but also IASS sponsored conferences, such as IASS WSC satellite meetings, joint IASS-IAOS events and Small Area Estimation Conferences. More and more, I felt part of this community, a knowledge-sharing environment that fosters collaboration and empathy.

2) The IASS played a vital role on my professional development as an open window to the international statistical scenery and a pathway to learn and to exchange information on survey methods and practice. Through IASS forums and short courses, I met the authors whose papers I read, or would read, and had the chance to present my work. I also helped organise workshops and conferences in South America with IASS support and invited IASS colleagues. Moreover, I had the honour to contribute to IASS in more than one executive role where I learned about the functioning of international organisations, IASS and ISI, and improved my teamwork abilities in an international context.

3) Coming from a country that still faces social and digital inequality, I have always appreciated the importance IASS gives to bridging gaps and creating opportunities for learning and networking that stimulated the development of young statisticians around the world. In the last years, IASS has been able to provide a lively and welcoming virtual environment for the development of survey statistics. I hope IASS will promote initiatives to develop survey statistics communities in developing countries. Also, as part of the ISI family, I believe that activities to encourage collaboration between ISI associations and its various communities would be very welcome.

2.3 Peter Lynn – Past President 2017-2019

As a young researcher starting out on my career in the late 1980s, my mentor Denise Lievesley suggested I should join the IASS. I enjoyed reading "The Survey Statistician" and eventually contributed to it with an article on substitution and with country reports in my role as country representative. Denise and Roger Jowell, Director of SCPR, where I worked at the time, regularly attended the ISI Sessions (as the WSC was known at that time) and came back with inspiring tales of foreign lands (and slide shows!) and of the work being done in statistical organisations around the world. From 1993 (Firenze) I started going regularly myself, always presenting a paper, listening to others, and joining the IASS assembly. It was inspiring to meet leading survey statisticians from many countries and to feel their enthusiasm for spreading knowledge and nurturing future generations. I think some of that enthusiasm may have rubbed off on me. During my time as IASS President (2017-2019) I was very aware that IASS membership was an aging cohort and that more needed to be done to ensure involvement of younger generations. We have made progress in that direction, including by supporting attendance at many international conferences and workshops. For the future, we should be alert to the need to communicate well the strengths of surveys and the important role they still have in society, to continue to share and spread good survey practice, and to encourage talented young people to work in this exciting field.

2.4 Steeve Heeringa – Past President 2015-2017

1) Starting in 1975, I trained under Leslie Kish and Irene Hess at the University of Michigan Institute for Social Research. In 1978, I joined Leslie and Irene as a member of the ISR faculty. Both Leslie and Irene were heavily involved in the early years of the IASS and regularly attended the biennial World Statistics Conferences (WSC) organized by the ISI. Based on their experiences as members of the IASS and those of my colleagues, Graham Kalton and Colm O'Muircheartaigh, I joined the IASS in the early 1980s. However, almost two decades passed before I attended my first WSC (Berlin, 2003). My first formal role was to join Graham and Colm in teaching the *Survey Sampling* short course at the 2003 (Berlin), 2005 (Sydney), 2007 (Lisbon) and 2009 (Durban) WSCs. During this same period, I also served as the Editor of the *Survey Statistician* (2004-2006) and was elected to the position of Scientific Secretary for 2007-2009. I organized the IASS short course program for the 2009 WSC and the ISI short course programs for the 2011 (Dublin) and 2013 (Hong Kong) conferences. I was elected IASS Vice-President/President Elect for 2011-2013 and as was the procedure at the time transitioned to the President role for 2013-2015.

2) For me, the greatest benefit of IASS membership and participation in the biennial ISI meetings has been the opportunity to meet and work with international colleagues at all levels—from leaders of national statistical offices to survey statisticians and methodologists who do the important day-to-day work of our international statistical community. Through Leslie's, Graham's and Colm's influences, early on I was engaged in teaching international students in our Summer Program and consulting globally with government agencies and non-profit organizations. However, it was my membership and engagement in the IASS and ISI that have really exposed me to the global statistical community and the challenges and opportunities that we share.

3) More than any other national or regional professional organization, IASS (and the ISI) serves a unique and important role in building and maintaining our international network of survey statisticians. Success in this role requires regular, strong communication both internally among the organizations' members but also externally with national statistical offices and decision-making bodies at all levels. I find that the recent IASS initiative to host webinars on important statistical topics is an excellent way to bring our community together and also reach new, younger members who otherwise may not realize the benefits of IASS membership. The webinars and conference support provided by the IASS also advance the important education/ capacity building goals of our organization.

2.5 Raymond Chambers – Past President 2011-2013

1) Looking through my IASS related papers, and associated emails, I guess that I really started to get involved with both the IASS and the ISI about the time that I moved from Southampton to Wollongong. But my awareness of IASS and willingness to get involved in its activities was really set during my ten year stint in Southampton, where I was lucky enough to be part of a group that included Tim Holt, Fred Smith, Chris Skinner, Danny Pfeffermann and Pedro and Denise Silva, all of whom were actively involved in international statistical outreach and as a consequence were involved with both ISI and IASS to a greater or lesser extent. It was their commitment to this work that ensured that when describing to a friend why I agreed to stand as President-Elect of IASS in 2008 I said that it was my professional responsibility to do so.

2) It has brought many benefits to me personally because it has brought me into contact with so many survey statisticians around the world, through participation in short courses and conferences, organising IASS events, and steering the careers of young survey statisticians through prizes and travel awards. It has also allowed me to have an impact on the IASS through guiding it, together with my fantastic Executive Director, Catherine Meunier, in its transition from being an organisation somewhat outside the ISI family with an office in Libourne to one that is now completely integrated into the bigger world of the ISI in the Netherlands. Whether any of this brought benefits to my career is somewhat moot. I made the transition from a senior government statistics position to a senior academic one in the early 90s. My career after that (and especially after my move to Southampton) was definitely international in its outlook and IASS was definitely part of that. But whether it was internationally focussed because of IASS is much less certain. Southampton was international. Full stop.

3) I would strongly suggest that you read the last two paragraphs of my June 2013 letter to IASS members as outgoing President. The future that I described then is here now, and maybe is already part of the recent past. The IASS has implemented a number of strategic initiatives (I am thinking of its very successful webinar series) over the last few years. These clearly need to be continued, as well as the initiatives that support and progress the careers of younger survey statisticians. However, there is still the question of how the organisation adapts to a much more fluid data capture and analysis environment, particularly involving methodologies focussed on integrating data from registers, convenience samples and scientific surveys. Whether we like it or not, survey statisticians are now required to be data scientists, able to communicate new information as well as uncertainty about this information using sources that can be very much more complex than were described in the sampling texts of the last century. I believe the time has come for us to become less precise and more useful. Perhaps an important new strategic objective will be to identify best practices in this new environment and then to actively seek to popularise their use (online training courses?) by IASS members and more generally by people in government, business and academia who are responsible for data acquisition and analysis.

2.6 Susan Linacre – Past President 2009-2011

1) I was encouraged to join the IASS by my colleagues at the Australian Bureau of Statistics, a number of whom were already members. Someone senior organised the process, although funding of membership fees was an individual thing for each of us. I chose to join as I felt that joining a society with a common interest in developing our knowledge of survey design and estimation, as well as related issues, would be good for our work, for my career, and would be interesting from a networking perspective.

2) Involvement in the IASS has been beneficial in terms of statistical knowledge gained, networks formed and the experience of working with others on an international basis, including providing help to statisticians from developing countries. The statistical confidence that this has helped me build has been beneficial to my development and to my career.

3) As I have not been involved with the IASS in recent years, I am not well placed to provide suggestions for the future. One thing I used to wonder about, that seems likely to remain relevant, is whether the title of IASS still reflects the key interests of its members, or would be members. I have not had dealings with the IASS in recent years, but the integration of survey and non-survey data, and effective use of non-survey data seem to bring many big issues for today's methodologists. Maybe the common issue now revolves around statistical inference from incomplete data.

2.7 Pedro Silva – Past President 2007-2009

1) I attended a 'Workshop on Survey Sampling in Developing Countries', held just before the Tokyo ISI Congress, in 1987. My participation was made possible by a grant to cover registration, travel, hotel, and meals provided by the IASS, thanks to donations secured to support this event by the IASS leadership. Prof. Tim Holt was my contact, and I cannot thank him and the IASS enough for that fantastic opportunity. It changed my life for good. After attending the event and the ISI Congress in Tokyo I found my community in the IASS and the ISI, and never left. I got involved because I was always made to feel welcome, met many colleagues who worked on topics that I was working on and/or interested in, and was supported in developing my career as a professional in this area. Such contacts were vital to help me deliver ideas and advance solutions for many challenges faced as part of my work on the Methodology Division in Brazil's national statistics institute.

2) I benefited immensely from joining the IASS and attending its sponsored sessions and short courses at the ISI World Statistics Congresses. The short courses allowed me to develop skills and expertise in areas which are not covered by regular university or academic education. In addition, I could learn from and take part in discussions with the best in the field. Later I was asked to work on Association functions, such as joining and then chairing scientific programme committees, where I developed collaboration and leadership skills in an international setting.

3) I have valued immensely the opportunity to attend an ISI Congress and related activities as my first ever trip abroad. Therefore, I hope that the IASS could engage in organising events where similar opportunities can be offered to the young members, particularly those from the developing world.

I also hope that it can encourage its members to continue sharing their knowledge by developing courses, publications and other learning materials that can reach the widest possible audiences around the world.

2.8 Gordon Brackstone – Past President 2005-2007

THE IASS reaches 50!

The formation of IASS within the ISI umbrella in 1973 coincided with my own early days as a survey statistician at Statistics Canada, so to join was natural, perhaps influenced by the fact that my supervisor then was among the IASS founders.

The term "survey" was always broadly interpreted to include sample surveys, censuses, and the use of existing data sources, often administrative records, for statistical purposes. The new Association aimed to serve the interests of those concerned with the design and implementation of such surveys, particularly but not exclusively in government statistical agencies and in the survey research organisations that supported them, as well as the interests of those in academia researching and teaching statistical methods with application to surveys.

A decade later the proposal to create an association for Official Statistics under the ISI umbrella did raise some concerns about potential overlap between these two sections, given IASS's presence in the realm of official statistics. These were addressed and IAOS was formed in 1985. Since then the two associations have co-operated and flourished in parallel including many joint sessions at ISI meetings (WSCs).

In more recent years a crucial issue for IASS and its members has been the future role of the traditional "survey" under the outside pressures of "big data" from without and declining response rates from within. Improved timeliness is also expected. These pressures have, of course, been widely recognised and their implications for methodology and inference are being addressed by statistical Associations and Agencies across the world. Though our broad definition of "survey" could stretch to accommodate innovative usages of existing data sources as well as new creative methods of data acquisition, it remains to be seen whether the statisticians who end up designing and implementing these new data systems will still see themselves as "survey statisticians".

Having not been active in the profession for the past decade or so I am reluctant to comment on the future lest it be mistaken for wisdom.

2.9 Luigi Biggeri – Past President 2003-2005

1) In 1973 I decided to participate in the 39th congress of the ISI since many invited and contributed paper sessions concerned the organization of statistical surveys, sample surveys and the evaluation of the quality of statistical information, research topics which I had been dealing with for some years. Slobodan S. Zarkovic and Leslie Kish had told me about the International Association of Survey Statisticians being established and given my interest, I attended the meetings and immediately joined the IASS since its foundation.

2) Involvement in the IASS has definitely benefited my career. By participating in all subsequent scientific activities of the IASS I have benefited from the teaching of many legendary survey statisticians, improving my research in the sector and thus increasing my reputation both nationally and internationally. The friendship with various talented young survey statisticians has allowed me to call them to hold seminars and short lecture cycles at the PhD in applied statistics of the statistics department of the University of Florence, which has thus become one of the most appreciated PhDs. All this has certainly allowed me to have an adequate CV in the selection and appointment in various bodies of the Italian official statistical system and to be elected president of the IASS for the period 2001-2003 and, finally, to be nominated president of the National Institute of Italian Statistics (Istat) for the period 2001-2009.

3) I have read and greatly appreciated the IASS strategic plan for the period 2022-2023. The wish is that all objectives can be achieved. From my point of view, I would stress the role of the representatives of the various countries more, by asking them to organize semi-annual or annual telematic meetings (or webinars) of the IASS members to discuss an activity plan for the development of the Association in the country.

2.10 Xavier Charoy – Past President 2001-2003

1) I was involved in the IASS in 1985. I was then working in INSEE, dealing with international cooperation. Some years before, I had worked with Gérard Théodore, who was one of the founders of IASS and the first Executive Director of the Association, and with Jean-Louis Bodin who was the second one. The latter then asked me to become the third Executive Director.

I could not assume the function after 1988, when I was appointed overseas and then joined FAO.

When back in INSEE headquarters in 1993, I had no special activity with IASS until 1997, when I was appointed as Chairman of the Nominations Committee. I was then elected as President-elect in 1999 and became President in 2001.

2) My involvement in the IASS did not bring me any benefit in my career. It was accepted and encouraged by my hierarchy, nothing more.

3) I am now 84. I retired 23 years ago. I still had a few contacts with NSIs, mainly in Africa, sometimes through international organisations (AFRISTAT, FAO), often on questions of statistical training with the three (now four) French speaking statistical schools in Africa, but never on survey subjects. I am

unfortunately unable to make any suggestion for the future of IASS. Younger and still active persons might be able to do it, not me. Sorry.

2.11 Kirk Wolter – Past President 1999-2001

1) All of us "stand on the shoulders of giants", and my giants were Wayne Fuller, Barbara Bailar, and many others. As a PhD student at Iowa State University, I became aware of the important paper Wayne delivered at the first meeting of the IASS in Vienna, 1973. Later, as a young professional employee at the Unites States Census Bureau, I helped Barbara prepare for the IASS meetings in Manila, 1979 and Buenos Aires, 1981. With the twin goals of advancing my knowledge of survey statistics and learning about statistical practices in other countries, and to model the work of my giants, I naturally sought to become a member of IASS and to participate in future conferences and other association business. My first meeting was Madrid, 1983.

2) IASS involvement brought wonderful friendships with statisticians from around the world, with people I may otherwise never have met. Through IASS and the network of friendships fostered at association conferences and business meetings, I learned about survey methods and procedures in other countries and cultures, which has enriched my own practice of survey statistics in the United States.

3) There are big questions ahead, such as

- What is the future of survey research?
- What role should IASS play going forward?

I would offer the following suggestions to current and future leaders of the Association:

- Remain committed to both the theory and practice of survey research. Avoid imbalance in which one side or the other dominates.
- Remain open to new methods and provide a forum and other opportunities for development, dissemination, and discussion of new methods and for evaluation of their statistical properties. The profession faces a time of substantial change potentially elevating the importance of methods such as multiple modes of response, probability and nonprobability sampling, integration of survey data with business or administrative data, web data collection, nontraditional data, and natural language processing of textual information collected in surveys.
- Grapple with the tradeoffs inherent in the budget and technical challenges of the future.
- Attend to questionnaire design. Emphasis in real surveys increasingly shifts from methods of sampling and estimation to evaluation of nonsampling errors and total survey error, and to the validity of survey responses.
- Encourage collaboration between survey statisticians and scientists in other disciplines. Working together, we can discover, test and implement methods and technologies that enable collection of valid data that bear on the critical problems facing humanity in the decades ahead.

2.12 Nanjamma Chinnappa – Past President 1997-1999

1) I became a member of the IASS in the late 1970s, when I started working as a Survey Statistician at Statistics Canada, Ottawa. As a survey Statistician in India and Canada, I was aware of the good work being done by the IASS in promoting the study and development of the theory and practice of sample surveys and censuses, and in publicizing research and good practices in Survey Statistics among statisticians, governments and the public all over the world. I therefore decided to become a member of the IASS and get involved in its efforts.

My official roles in the IASS were as

- President of the IASS, 1997-1999
- Chair of the Programme Committee for the scientific programme of the IASS at the 49th session of the International Statistical Institute, 1993
- Member of the IASS Council, 1991-1997.

As the President of the IASS in 1997-1999 I implemented the following initiatives:

a) The establishment of the Cochran-Hansen prize for promising young Survey Statisticians from developing countries in 1999, in celebration of its 25th anniversary.
b) Publication of 25 Years of the History of the IASS, which served as a useful document of the journey of the institution.
c) IASS Jubilee Commemorative Volume – Landmark papers in Survey Statistics (2001).

2) I believe it has, because I came in contact with survey statisticians around the world when I attended IASS meetings, especially in my official roles. That helped me in gaining more visibility in the field and discussing and contributing to the efficiency and practicability of survey methods and practices.

3) I whole-heartedly endorse the IASS Strategy Document 2022-2023. I would add

- Initiating and supporting regular regional meetings or seminars devoted to specific aspects of surveys.
- Distributing its publications to statistical libraries to be suggested by country representatives in developing countries.
- Allowing members from developing countries to subscribe to Survey Methodology journals at reduced rates.

2.13 Dennis Trewin – Past President 1995-1997

1) My boss at the time (Ken Foreman) was part of the Committee that established the IASS at the 1973 ISI Session. He encouraged me to join which I did.

2) Yes, it did. It enabled to establish relationships with many of the great survey statisticians. Leslie Kish and Ivan Fellegi are two examples. It also enabled me to keep abreast of developments of direct relevance to my own organisation through Conferences, publications and Newsletters. I will just mention one example. In Australia, we knew little about CATI until I was able to attend several CATI presentations at the 1983 ISI Session.

3) I think the main role of the IASS is to facilitate communication between its members and will remain so in the future. Traditionally, this has been through Conferences and their proceedings, publications like the Journal of Official Statistics and The Survey Statistician. Webinars have commenced in more recent years but it is important that they are recorded to enable those members who cannot attend the Webinar to also benefit. They are generally held at a time that is unfriendly for Australia. No doubt new technologies will provide additional options. However, it is extremely important to understand what works best for developing countries and to respond accordingly. Should we set up a network of experienced mentors to assist developing countries? I would certainly be willing to assist in this way.

2.14 Graham Kalton – Past President 1991-1993

1) The first ISI meeting that I attended was the 1973 Congress in Vienna, where I presented an invited paper in a session on variance estimation and design effects organized by Leslie Kish. The IASS was established at that Congress, and the conference program included a sizable number of sessions and many impressive presentations on survey statistics. I was attracted by the formation of an international community of survey statisticians that covered both theory and practice of the subject. I joined IASS and I resolved to try to attend future Congresses.

2) My many contacts with survey statisticians around the world have broadened my outlook on the applications of survey statistics in different environments and on the variety of techniques being applied.

In addition, I have had a long-standing interest in training survey statisticians in countries that have less survey experience. On behalf of IASS, from 1991 to 2005 I co-presented short courses on “Survey Sampling in Developing Countries” at eight ISI sessions, six jointly with Colm O’

Muircheartaigh and two with Steve Heeringa. Notably, international organizations and others often funded the participation of statisticians from developing countries for these courses and for attendance at the ISI sessions.

I met Pali Lehohla, then the Statistician General of South Africa, at the 2005 ISI session in Sydney. We discussed the benefits of running a training program in survey methods in Africa for statisticians from a group of neighbouring countries. Pali followed up on our conversation and he arranged four courses in survey sampling and survey methods to be given in South Africa for statisticians from the Southern African Development Community (SADC) countries. I helped structure the program and taught on it for a couple of years. I found all these training activities highly rewarding.

3) The membership of the IASS is disappointingly small given the considerable number of statisticians engaged in survey research around the world. To attract new members, I think that the IASS needs to expand its offering to its members, recognizing that few survey statisticians can afford to attend the ISI biennial sessions. I applaud the recent IASS initiatives in sponsoring conferences on special topics in different regions of the world and in supporting a series of web-based presentations on methodological issues in survey statistics. However, more needs to be done. One possibility is that the IASS could play a role in stimulating web-based training in survey methods, both statistical and methodological, perhaps seeking international funding to support the activity.

The IASS has had fruitful collaborations with the IAOS. The IASS could also usefully explore collaborations with the subsections for survey research that now operate in many national professional statistical associations.

2.15 Colm O'Muircheartaigh – Past President 1987-1989

1) Leslie Kish (University of Michigan, USA) was a great champion of IASS. I had met Leslie when I attended his Sampling Program for Foreign Statisticians at the University of Michigan in 1968, and subsequently taught at the program from 1975 onwards. He encouraged me to submit a paper to the 1977 ISI Meeting in New Delhi; my paper was part of an IASS session on sampling in developing countries. Morris Hansen, who was then President of IASS, met with all the speakers at IASS sessions and encouraged us to join. Morris and Leslie (both US) were very active in those early days of IASS. Leslie Kish, in particular, used his international work to recruit statisticians from many countries.

2) Yes, indeed. Through IASS I met survey statisticians from all over the world and developed friendships and collaborations with many of them.

IASS was important in the development of the World Fertility Survey (WFS), encouraged methodological research and engaged with members of the national statistical agencies where the WFS data collections were carried out. IASS organized many sessions over the years dealing with the work of WFS.

Through an IASS paper that I gave at the 1981 sessions in Buenos Aires, I met a group of statisticians from the US Census Bureau, in particular Barbara Bailar and Kirk Wolter, both of whom later made important contributions to IASS. As a result, I developed a working relationship with the Bureau, and have continued to consult with them since that time.

It was also through IASS that I met an important group of (then) young Italian statisticians, of whom Luigi Biggeri was a key figure from the earliest stages. This group did important work in China (for FAO), and I was delighted to play a part in that work. These connections also led to my being invited to teach from time to time at the consortium of Italian universities with graduate programs in survey and social statistics (Firenze, Padova, Bologna, Perugia, Siena). These universities played a significant part in the development of international collaborations and in IASS itself. I would recognize particularly Luigi Fabbris (Padova), Carlo Filippucci (Bologna), Giuseppe Cicchitelli (Perugia) and Andrea Giommi (Firenze).

INSEE (France) played a fundamental role, providing the secretariat for IASS for a great number of years. Gerard Theodore and Xavier Charoy were both excellent Executive Directors and Anne-Marie Vespa provided outstanding administrative support while IASS was based in Paris.

3) IASS is probably the most important network of statisticians working in the public policy arena in the social sciences and has the potential to disseminate high quality methodology across the world.

To me, the key is to create a network within which survey statisticians can develop relationships across borders and across sub-disciplines, providing support to each other and developing, testing, and implementing methodologies. The IASS program should also have an educational component, providing materials helpful to statisticians who are working in environments where specialized training in surveys is not locally available. While the work of IASS may increasingly take place online, I feel that continuing to provide regular in-person opportunities will be important in building and maintaining the inter-personal relationships that provide a lasting foundation for collaboration. As the cost of attending the World Statistics Congress increases, there should be particular emphasis on having smaller-scale in-person research and training meetings in diverse regions.

Acknowledgement

The idea of this paper started with the past Executive Committee. I would like to particularly thank Monica Pratesi that helped me in the conception and implementation of this paper.

The Survey Statistician, 2024, Vol. 89, 47-58

Ask the Experts

What is the state of play on statistical matching with a focus on auxiliary information, complex survey designs and quality issues?

Marcello D'Orazio, Marco Di Zio and Mauro Scanu[1]

[1] Istituto Nazionale di Statistica – Istat, madorazi@istat.it, dizio@istat.it, scanu@istat.it

Abstract

The statistical matching problem consists in fusing information from two data sources that are representative of the same population but contain observations on disjoint sets of units. The lack of joint information on variables observed distinctly in the two data sources induces a source of uncertainty that usually statistics does not tackle directly, under the status of unidentifiability of the model given the data at hand. This paper gives an updated account of what has been proposed in order to deal with this problem.

Keywords: identifiability, uncertainty, imputation, data fusion.

1 Introduction

The widespread use of data integration for statistical purposes gave rise to new challenges: statistical matching aims at overcoming one of these challenges for which there is not an immediate answer under the statistical point of view. The statistical matching problem consists in making "estimates" on parameters of the joint distribution of Y and Z when Y and Z are observed in two distinct data sources (*A* and *B* respectively), and the sets of units on which Y and Z are observed, although representative of the same population, are disjoint. Hence, it is not possible to connect records through the use of identifiers or exact or probabilistic record linkage.

We avoid the description of well know methods, mostly already covered in D'Orazio et al. (2006a) and references therein. In this paper, we try to give the state of play on statistical matching on some specific issues. First of all, it should be clear once and for all that the use of the data in *A* and *B* is not enough for the two purposes for which statistical matching has been considered: i) a "fused" complete but synthetic data set on which whatever statistical analyses involving Y and Z could be performed (micro approach); ii) estimation of specific parameters on the joint Y and Z distribution (macro approach). Hence, a discussion on how to incorporate additional information in terms of data sources or constraints is given in Section 2. Secondly, most data sets are drawn according to complex survey designs, and Section 3 covers this issue. Section 4 illustrates different areas on which statistical matching applications have been conducted. Quality of the results obtained by statistical matching is an essential point that sometimes seems to be neglected in real applications: we describe the state of play on quality measures in Section 5. Some conclusions and hints on areas of research are finally discussed in Section 6.

2 Use of auxiliary information

The statistical matching (SM henceforth) problem is naturally affected by an identifiability problem in the sense that data in *A* and *B* are not enough to estimate the parameters of the joint distribution of Y and Z. To fill this gap, assumptions as the independence of Y and Z given the matching variables $\boldsymbol{X}$ (conditional independence, CI henceforth) is needed and it is explicitly or implicitly used in the base SM procedures. This assumption severely limits SM procedures applicability. One way to move away from this assumption and thus improve the conclusions is that of using auxiliary information, particularly on the variables Y, Z or $(Y, Z|\boldsymbol{X})$. Such information can take several forms: it can be a set of micro data, information on parameters or aggregate values of the variables under observation, or refer to logical and statistical constraints on the variables (partial information) see D'Orazio et al. (2006b).

2.1 Exploitation of additional data sources

An auxiliary data set *C* can be used for matching purposes if it is a representative sample of the target population, otherwise the inference will be affected by a bias. Several methods have been proposed in the literature: parametric, nonparametric and mixed, see D'Orazio et al. (2006a) for a review. In the presence of a micro data set, the approaches described in D'Orazio et al. (2006a) follow the idea of creating a data set by appending file *A*, *B* and the auxiliary file *C* and treating it as a statistical inference problem in the presence of missing data. This approach is studied and further explored in a Bayesian context by Fosdick et al. (2016). They use a data augmentation algorithm that intrinsically produces multiply imputed values. The simulations show that the size of the auxiliary file *C* essentially represents the degree of confidence with respect to the auxiliary information. Although this result was expected, it is important to remember that every time a file *C* is used, its observed size naturally becomes our degree of confidence about its quality.

The representativeness of *C* can be a limiting assumption as well. In fact, most of the times *C* is an outdated sample, or a file composed of proxy variables, i.e., it is composed of information related to the variables under investigation - and thus it is important to take *C* into account in order to avoid the CI assumption - but characterized by a sort of proxy information. How to deal with this issue is an important topic still under investigation. In Moretti and Shlomo (2023) and Fosdick et al. (2016) there are two ways of approaching the non-representativeness of the auxiliary file *C*. In the first, there is the research and use of further additional information, while the second is characterized by a further but weaker assumption about the representativeness of *C*.

Moretti and Shlomo (2023) propose calibrating the prediction regression model of a mixed approach (predictive mean matching) to known marginal totals of the variables $(\boldsymbol{X}, Y, Z)$ to make the estimation of parameters robust to misspecification of the model. They empirically show that this approach improves the results of the matched file.

In Fosdick et al. (2016), an SM method is proposed for the case where the sample *C* is not representative for the joint distribution of $(\boldsymbol{X}, Y, Z)$ but is representative for its conditional distributions $Y|\boldsymbol{X}$ and $Z|\boldsymbol{X}$. The algorithm essentially consists of:

1. estimating the conditional distributions $Y|Z, \boldsymbol{X}$ and $Z|Y, \boldsymbol{X}$ from *C*,
2. obtaining a synthetic auxiliary file *C** by
    a) generating the observations from $Y, \boldsymbol{X}$ and $Z, \boldsymbol{X}$ observed in *A* and *B* respectively (e.g., by duplicating or sampling records with replacement from *A* and *B*),
    b) imputing the missing values of Y and Z in the respective subsets of *C** using the conditional distributions estimated in the first step given the observations generated by step 2a.

This algorithm produces an auxiliary file *C** that preserves the marginal distributions observed in *A* and *B* and the conditional distributions observed in *C*.

Even in this case, however, an assumption is made, namely that the conditional distributions observed in *C* are representative of those in the target population. To measure the validity of this assumption, the authors propose an empirical evaluation. They suggest comparing the marginal distributions of Y and Z obtained in the synthetic file *C** with those observed in *A* and *B*. A high discrepancy suggests that the conditional distributions of *C* are not representative and therefore this approach should be avoided.

Though developed in a Bayesian context, this approach can be an interesting proposal for dealing with non-representative auxiliary information in different inferential contexts.

2.2 Auxiliary information in terms of constraints

Another type of auxiliary information often available in official statistics concerns the use of logical or statistical constraints, known in the field of editing and imputation as edit rules (soft and hard edit rules). Hard constraints (hard edits) refer to relationships between variables that must be necessarily fulfilled by the values of each observation, for instance, babies cannot have an academic degree and the total costs of a company are greater than or equal to the amount spent on purchases. Soft constraints (soft edits), on the other hand, identify abnormal although possible behaviors, e.g., the ratio of purchases to sales is generally within an interval $[l, u]$, see De Waal et al. (2011). Auxiliary information in terms of constraints was firstly studied in D'Orazio et al. (2006a and 2006b), later with discretized continuous variables by Conti, Marella and Scanu (2016 and 2017). An interesting and extensive recent study with continuous variables is in Claramunt-González et al. (2023).

The introduction of hard constraints on the Y, Z variables naturally makes the conditional independence model unfit and impossible. However, the use of hard constraints is not immediate and needs further investigation. In D'Orazio et al. (2006a), constraints are introduced at the model estimation stage, i.e., the estimable parameters that determine the region of uncertainty is bounded by the introduced constraints. Claramunt-González et al. (2023) focuses on a mixed method, that is in fact a predictive mean matching, and the constraints are used not in the parameter estimation step, but in the stage of donor imputation, namely, donors are chosen taking into account the hard and soft constraints. As noted in D'Orazio et al. (2006a) and Claramunt-González et al. (2023), there is a general improvement in the matching results. However, some constraints may be more or less useful. In particular, the introduction of hard constraints in the imputation phase can lead to having empty imputation cells or cells with a low number of donors. So, the introduction of a constraint on $Y, \boldsymbol{X}$ or $Z, \boldsymbol{X}$ should be well thought because it does not make direct changes on the conditional independence model but may introduce problems for the imputation phase. It is also interesting to note that Claramunt-González et al. (2023) studied the use of constraints in the case of additional information on $(\boldsymbol{X}, Y, Z)$ that allows estimating a model that is not based on CI.

3 Approaches accounting for complex sample survey design

A large part of SM methods proposed in literature are designed to integrate random samples consisting of independent and identically distributed (iid) observations. This assumption is seldom valid in official statistics where the available data come from complex probabilistic surveys that commonly include stratification and clustering; these complex selection mechanisms consist of two or more stages that typically invalidate the independence assumption (units belonging to a cluster show a degree of homogeneity) and often result in unequal weighting of the final in-sample units (Base weights, which are the reciprocal of first order inclusion probabilities, are corrected to compensate for unit nonresponse and for coverage problems, so the final survey weights are often the outcome of calibration or post-stratification procedures). In this framework, the target of inference are finite population quantities and the approach to inference is typically *design-based* or *model-assisted design-based* (Särndal et al., 1992). Therefore, the application of SM methods has the

objective of estimating the correlation coefficient between Y or Z or the contingency table crossing Y and Z at finite population level, or to create a synthetic representative sample that can be used for the subsequent analyses (under the paradigm of design-based inference). In this context, when SM aims at estimating model parameters (e.g., the correlation between Y and Z), it should be considered that the model assumed for the data in the sample is often not the same as that in the population and the sampling design is said to be *nonignorable* or *informative* (cf. Opsomer, 2009).

The SM methods accounting explicitly for the sampling design (and survey weights) are quite limited. In the past, two main different approaches were suggested: (i) Rubin's file concatenation (Rubin, 1986) and, (ii) Renssen's matching by weight calibration (Renssen, 1998).

Rubin's file concatenation consists in appending the two data sources and re-calculating the sampling weights in order to achieve representativeness of the target population. As noted by Ballin et al. (2008 and 2009) the recalculation of weights is not straightforward and requires knowledge of information on several aspects (sampling frame, design variables, etc.) typically available solely within the statistical agency that administers both the surveys. In addition, the approach does not consider unit nonresponse and corresponding weights' correction. In any case, the re-calculation of the weights does not solve the problem of lack of joint information regarding Y and Z. In practice, two imputations steps are still required (Y in subsample *B* and of Z in subsample *A*). For all these reasons the approach is seldom applied (see e.g., Ballin et al., 2008; Torelli et al., 2009).

The matching by weights' calibration proposed by Renssen (1998) is primarily intended to estimate the contingency table crossing Y and Z when Y and Z are categorical target variables. The method has the advantage of providing an estimated table whose marginal distributions are fully coherent with those estimated from the starting data sets via the Horvitz-Thompson (HT) estimator ($\widehat{N}_j = \sum_{k\in A} \widetilde{w}_A I(y_k = j);\ j = 1, \dots, J$). However, Renssen pinpoints that for coherence purposes, before estimating the Y and Z cross-table, it is necessary to align the marginal/joint distributions of the matching variables $\boldsymbol{X}$ in order to return the same known totals; this latter task requires two weights' calibration steps. Renssen's approach can exploit additional auxiliary information coming from a third sample *C* that observes both Y and Z (and possibly also $\boldsymbol{X}$). This method permits also the creation of a synthetic sample by means of a two-step procedure (resembling predictive mean matching). In fact, the estimation of the Y and Z cross-table requires the adoption of linear models (*linear probability models* in the case of categorical variables), whose predictions can be used as input of nearest neighbor hotdeck to impute in the recipient the target variable observed in the donor sample (see e.g., Donatiello et al., 2022). D'Orazio et al. (2010) extend Renssen's idea by replacing weights' calibration with the procedure suggested by Wu (2004).

Renssen's procedure is the most popular approach to handle survey weights in the two samples. It belongs to the larger class of SM methods that use survey weights in the matching step. A seminal proposal in this sense is that of Barr and Turner (1981) that consists in creating a synthetic sample using a constrained nearest neighbor hotdeck where the weights assigned to matched units in the synthetic sample are obtained by solving an optimization problem that guarantees reproducing the same estimated total amount of Y (Z) obtained by applying the traditional HT estimator in *A* (*B*) (this procedure assumes CI on Y and Z given $\boldsymbol{X}$ and that the starting weights in *A* and *B* return the same estimated population size). Renssen notes that this result can be achieved with much less computational effort by applying his procedure under CI (in absence of an additional auxiliary sample *C*), after the initial weights' calibration aimed at aligning the totals of the matching variables. D'Orazio (2015) suggests a slight modification of random hotdeck to use donors' survey weights in the random draw of a donor: this approach follows the ideas of weighted random hotdeck (cf. Andridge and Little, 2010).

Recently Kim et al. (2016) proposed a fractional imputation method aimed at creating a synthetic sample where the survey weights are used in the different steps of the imputation procedure.

Jauslin and Tillé (2023) follow Renssen's ideas to develop a nonparametric procedure; it at first applies weights' calibration to harmonize the totals of the matching variables, and then imputes the

recipient data set by using a nearest neighbor hot-deck approach that is constrained to use a donor only once and to return estimated totals of the imputed variable equal to those estimated from the donor sample. The optimization problem is solved by means of an algorithm commonly used in balanced sampling.

Schifeling et al. (2019) use SM to assess how measurement errors affect observation of a target variable (Z) in one survey when the same variable is observed free of errors (Y) in another survey. They suggest adoption of design-based inference to calculate the estimates of the cells of the contingency table crossing $\boldsymbol{X}$ and Y and the corresponding sampling variance. Then this information, coupled with assumed measurement error models (that avoid CI), becomes the input of a Bayesian approach that ends with an estimated posterior distribution of the true values and of model parameters.

Marella and Pfeffermann (2019) propose a unified framework for making inference on model parameters in the SM case when dealing with samples from complex sample surveys (informative samples). In particular, they define a sample likelihood that enables the estimation of the target population distributions and subsequently to impute the missing values. The authors give the conditions under which the sample models are identifiable and estimable from the starting data.

4 Some applications

SM methods have been and are applied in different domains: investigation of poverty and well-being, education statistics, travel and transportation statistics, agriculture statistics, etc. Although SM applications are mainly tailored to integrate data of probabilistic surveys, in some cases they are applied also to integrate probabilistic and non-probabilistic sources (administrative registers and more in general big data). Obviously, it is not possible to keep track of all the various applications of SM methods in the various domains and with different data settings. For this reason, in this section we limit our attention to some relevant applications with data stemming from probabilistic surveys referring to the same target population.

A huge number of papers apply SM methods to get insight on people's well-being (see e.g. Leulescu and Agafitei, 2013; Donatiello et al., 2016; Bernini et al., 2021). A large contribution to this objective is given by studies investigating the relationship between people's income and consumption, due to the well-known difficulties in collecting detailed information on both these items in the same survey. In the European Statistical System (ESS) this objective has been pursued by integrating the EU-Statistics on Income and Living Conditions (SILC) and the Household Budget Survey (HBS), in most of the cases by creating a synthetic sample that serves as basis for an in-depth investigation. Tonkin and Webber (2013) compare nonparametric methods and mixed SM methods. Donatiello et al. (2014, 2016 and 2022) investigate hotdeck imputation and warn about the consequences of assuming CI when matching the data of these two surveys. They show that CI approximately holds when considering a proxy of income or consumption in the matching process. As an alternative, they avoid CI by carrying out an assessment of uncertainty. Conti, Marella and Neri (2017) use Italian surveys to assess uncertainty due to the SM framework by including some constraints on the joint distribution of income and consumption. More recently, Donatiello et al. (2022) adopt Renssen's approach to derive a synthetic sample through a two-step procedure that uses a proxy of consumption in the matching process. The paper shows how crucial it is to think about SM when designing both the surveys, by harmonizing ex-ante the definitions and the classifications used for the common variables and by collecting the information (proxy of consumption in SILC) needed to make CI valid. In addition, the paper stresses that the nice feature of Renssen's approach of ensuring that the consumption imputed in the SILC survey maintains a marginal distribution aligned to that estimated in the HBS (donor) is crucial in official statistics, where the synthetic sample should provide estimates coherent with those obtained from the starting surveys. For the same reason, Rios-Avila (2015) suggests the use of a weight-splitting strategy to better comply with the constrained SM rationale and corresponding advantages. Ucar and Betti (2016) consider also the longitudinal dimension of the SILC survey and impute in it the consumption expenditure variable observed in the

cross-sectional HBS data; they also use a two-step Renssen procedure. Another application to Turkey data performs a constrained SM using propensity score ranking (Albayrak and Masterson, 2017). Decoster et al. (2020) compare different methods to impute consumption in a dataset on income in Belgium and create a data set for microsimulation purposes. Lopez-Laborda et al. (2020) suggest fitting Engel curves in a parametric SM approach with the objective of imputing consumption in SILC.

Recently attention moved from the joint distribution of income and consumption to a wider picture by including also wealth data. This is a very relevant domain of study, as it permits a thorough investigation of multidimensional poverty. A seminal paper on this topic is from Tedeschi and Pisano (2013) that investigate how to integrate the Survey on Household Income and Wealth (SHIW) carried out by the Bank of Italy with data on consumption available from the Italian HBS survey carried out by the Italian National Statistical Institute (Istat). Given the importance of the topic, recently Eurostat and OECD decided to join efforts and carry out an extensive SM exercise involving data from different countries within and outside EU (Balestra and Ohler, 2023), with the objective of measuring the joint distribution of household income, consumption and wealth at the micro level. In the same direction goes the work of Tram and Osier (2023) who want to explore multidimensional poverty in Luxembourg by applying a two-step approach that performs multiple imputation via Bayesian Bootstrap predictive mean matching.

Other survey data that are often involved in SM applications are those related to time use (TU), a very important topic, in particular when jointly investigated together with data collected in Labour Force Surveys (LFS). Gazzelloni et al. (2008) present a hot-deck application with Italian LFS and TU data; Ghahroodi (2023) considers Iranian data and suggests fitting tailored models for TU data (conditional predictive Dirichlet distribution or conditional predictive multinomial distribution) in the first step of an SM mixed approach. Zacharias et al. (2014) investigate the relationship between TU data and consumption expenditures by applying SM based on propensity scores. To investigate the measure of time and consumption poverty at microdata level Rios-Aviola (2016 and 2020) applies SM to integrate TU and living conditions data for different African countries.

Dalla Chiara et al. (2029) suggest an application of SM based on propensity scores for creating a synthetic dataset by integrating SILC, HBS, TU survey and data on household conditions and social capital; the fused file allows investigating households' living conditions in Italy.

Wiest et al. (2019) use SM to investigate effects of educational participation on well-being in later life. Bernini et al. (2021) aim at analyzing how happiness affects expenditure behaviour in different urbanized areas in Italy. Hossain et al. (2022) integrate data from the household travel survey with different specialized "satellite" surveys to assess the impact of COVID-19 on passenger travel demand in the Greater Toronto Area.

Torelli et al. (2009) and Ballin et al. (2009) explore the application of SM to the Farm Structure Survey and Farm Accountancy Data Network survey carried out on Italian farms, with the additional difficulty of managing dependent surveys. In the same framework, D'Orazio and Catanese (2016) use SM to assess the revenues and economic growth of farms producing renewable energies.

It is worth noting that a large part of the SM applications have a very challenging objective, namely the creation of a synthetic sample that serves as basis for in-depth analyses. Regardless of the SM method being applied and its complexity, a critical reading shows that several applications seem unaware of the assumptions underlying integration, in particular that of independence between the target variables conditional on the selected matching variables. Sometimes also the applications that are aware of CI and claim that it is valid, often ignore its consequences on extended analyses carried out on the synthetic sample. In fact, while CI permits to reliably explore the association/correlation between Y and the imputed Z, the same assumption may not lead to valid results when studying for instance the relationship between W, a different variable observed in the recipient file, and the imputed Z. In general, managing implications of the CI assumption can become quite difficult when SM is applied to integrate three or more data sources. For this reason, studies having very ambitious

objectives that require the integration of several samples should proceed very carefully and should dedicate much effort to understand whether the objectives can be reliably pursued given the available data and the underlying assumptions in the integration process.

5 Quality issues

As it is clear from the last considerations in Section 3, the big question in an SM problem is the assessment of the quality of the results. De Waal (2015) states that this is a primary field of research for SM, identifying two key issues: how to better extract information from the available data, and what kind of additional information could support SM?

As far as the first question is concerned, an update on the possible different estimators or imputation procedures has already been given in the previous sections. Hence, we focus here on the modelling issues that, frequently, imply the use of specific methods. As already said, a multivariate model that includes (Y, Z) is unidentifiable for the data at hand for SM. Since the publication of D'Orazio et al. (2006a), this additional source of uncertainty (i.e., uncertainty due to the lack of joint observations on Y and Z) has become more important in the evaluations of SM results than the uncertainty due to sampling (that can be always investigated by means of the usual tools, as for instance coefficients of variation). A thorough discussion on uncertainty in statistical matching is given in Conti, Marella and Scanu (2017). Up to now, this kind of uncertainty has been treated in the following ways.

1. It was resolved by assuming, possibly in an explicit way, specific models that are identifiable for the data. Much has been already said on the CI (mostly assumed subconsciously) and, as already remarked, we consider it important to be conscious of that assumption and to report it explicitly, if taken into account. This assumption seems appropriate just in those applications (see for instance Donatiello et al., 2022) that make use of at least one matching variable that is (very) highly correlated with either Y or Z, so that Y and Z become almost independent given the matching variables (something that can be imposed by construction when matching is planned while organizing the observation of the source files *A* and *B*). The CI cannot be tested by the data at hand.
2. A different identifiable model has been suggested by Kim et al. (2016). This model assumes that matching variables can be decomposed in two groups, say $\boldsymbol{X} = (V, W)$, and that V is an "instrumental variable" for Y, i.e., V is conditionally independent of Z given W and Y but V is correlated with Y given W. Under this model, the authors suggest the use of parametric fractional imputation (PFI, Kim et al., 2016). Also in this case, there is not a test that can validate the assumed model. Anyway, the authors state that "a sufficient condition for model identifiability is the existence of an instrumental variable in the model." Furthermore "The proposed methodology is applicable without the instrumental variable assumption, as long as the model is identified." Their estimation approach, based on the use of the EM algorithm, does not necessarily converge if the model is unidentifiable. They consequently claim: "In practice, one can treat the specified model as identified if the EM sequence converges." This seems the most interesting and intriguing aspect of this approach, that can justify the use of model assumptions even if untestable.
3. If no identifiable model can be constructed, a (possibly) large set of models are indistinguishable by the data at hand. D'Orazio et al. (2006a) use the notion of "likelihood ridge" as the set of all the equally likely maximum likelihood estimates of some parameters in order to represent the uncertainty on some parameters of the (Y, Z) distribution given the data at hand. For specific parameters, the width of the interval or space of all the equally plausible solutions quantifies how uncertain these parameters are given the available data on *A* and *B*. Results on Y and Z correlation coefficients are discussed in D'Orazio et al. (2006b), frequencies of contingency tables in D'Orazio et al. (2006a), for ordered categorical variables in Marella et al. (2013) and for generic empirical distributions on Y and Z in Conti et al. (2016).

The second question raised by De Waal (2015) was on what additional information can be considered in order to improve the quality of SM results. Much has already been described in Section 2 as far as additional data sources are considered. Here we focus on the effects of the use of constraints in terms of uncertainty. Conti et al. (2016) adopt as SM estimate for the distribution of (Y, Z) given $\boldsymbol{X}$ the central distribution among the ones in the estimated likelihood ridge given the data at hand. Even if the likelihood ridge is rather well known in general (consider for instance the Fréchet bounds for categorical variables and the parameter under the CI as a midpoint), this computation of the likelihood ridge's central distribution becomes cumbersome when constraints are imposed. They consequently define a very general estimator and derive its asymptotic properties as well as the width of the likelihood ridge in order to derive tests on the likelihood ridge's sparseness around the estimated distribution. This is just one of the papers that make use of the likelihood ridge width as a measure of the SM uncertainty due to the lack of joint observations on Y and Z.

A specific measure of the sparseness of the uncertainty set of distributions when the variables are categorical is given by the Fréchet bounds (D'Orazio et al., 2006a). Fosdick et al. (2016) compute the Fréchet bounds of Y and Z given $\boldsymbol{X}$ in order to verify, in a simulated context, the goodness of their estimator (described in Section 2.1) under the presence of different kinds of additional files *C*, and examine if these bounds are as tight as possible.

A simulated set up is the context where quality measures can be defined and applied, exploiting all their potentiality, given that the actual parameters to be estimated are known in advance. For instance, this happens in the already cited paper by Claramunt-González et al. (2023) where a multivariate mixed method for SM for the estimation of the correlation between Y and Z is proposed: in that context, quality has been assessed by i) computing the estimator bias (which can be calculated due to knowledge on the actual parameters) in a multiple imputation case (allowing variance estimation) and ii) transforming estimates by means of Fisher *z*-transformation (in order to ensure that the resulting transformed estimates are generated by a normal distribution). Besides using the mean squared error as a measure of performance of an estimator, the same authors identify also a measure for the imputed data set, checking whether the individual imputations are "correct within a $100 \times \tau$ per cent". For instance, when the Z observations z_a (which are known in advance in file *A* in a simulation study) are imputed by an SM procedure, it is verified whether each imputation lies in the interval with extremes $(1-\tau)z_a$ and $(1+\tau)z_a$. The authors note that this is a way to derive the so-called "matching noise" (see Conti et al., 2010): given that the objective of the matching noise is to describe the distance between the actual data generation process and the imputation process, and that this computation can be cumbersome for some estimators as the mixed ones, the identification of such an empirical computation of the matching noise is a nice trick to take into account. The computation of the fraction of "correct within a $100 \times \tau$ per cent" imputed values should be considered as a quality measure for the statistically matched file. In fact, they say in the paper that: "we do not intend to release any statistically matched datasets." We agree with their approach. Anytime a statistically matched data set is created and released for whatever unplanned statistical data production, it could happen that the chosen but unplanned Y and Z are connected in ways that that are not taken into account in the SM method, e.g., even by hard edit rules (see Section 2.2 and discussion therein). The introduction of constraints, as already discussed in D'Orazio et al. (2006a), dramatically improves quality of results obtained by SM. In fact, the uncertainty set of equally plausible estimated distributions for the data set at hand changes significantly, and the use of hard constraints (as introduced for instance in Section 2.2) excludes the conditional independence model among the distributions that can contribute to the uncertainty space, moving the "conditional" uncertainty space towards the actual but unknown distribution. Claramunt-González et al. (2023), as noted in Section 2.2, suggest to include not only hard constraints in an SM problem, but also soft ones. As the authors note, these constraints generally need the help of a third complete data set that allows one to fix the characteristics of soft rules that allow to isolate those values that are unlikely, even if possible. The introduction of these rules should help an SM procedure in reducing the uncertainty space. However, it is yet not clear how effective they are, and this could be, in our opinion, an interesting area of research.

As far as the computation of uncertainty is concerned when dealing with samples drawn according to complex survey designs, D'Orazio (2015) includes survey weights in estimating the uncertainty in the case of categorical Y and Z for the cells in the contingency table $Y \times Z$ in a standard matching framework. In this case, before the assessment for coherence purposes, it is suggested to align the marginal/joint distribution of the chosen matching variables, e.g., by using the IPF algorithm (starting from v.1.3.0 of the R package StatMatch, D'Orazio 2022). In addition, robust estimation methods are introduced to handle the problem of statistical zeroes.

6 Conclusions

Although appealing, SM can be tricky and hides features that can be dangerous for the credibility of the results. As remarked in all the sections, a clear assessment and declaration of all the assumptions underlying the specific statistical matching application is absolutely necessary. In particular, the micro approach is the one that could be most harming, and could lead to the "dog food problem" (see Claramunt-González et al., 2023 and references therein). Even in case of additional information and/or specific assumptions, an evaluation of uncertainty should be given and the reasons which lead the analyst to choose just one of the equally plausible estimates for either the micro or the macro approach should be clearly stated.

Furthermore, there are some approaches that need more attention and additional research, also in an applied setting. We mention just two of them. The first considers the use of models that include instrumental variables and make use of PFI (Kim et al., 2016): this approach touches all the main statistical matching issues, such as the presence of additional data sources and the use of complex survey designs for *A* and *B*, while paying attention to model identifiability. The second consists of a micro approach in which uncertainty is taken into account in the imputation process (imprecise imputation, see Endres et al., 2019), and that could be worthwhile also outside the usual statistical matching framework.

References

Albayrak, O. and Masterson T. (2017). Quality of statistical match of Household Budget Survey and SILC for Turkey. Levy Economics Institute of Bard College, Working Paper No. **885**.

Andridge, R.R. and Little, R.J.A. (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, **78**, 40-64.

Balestra, C. and Ohler, F. (2023). Measuring the joint distribution of household income, consumption and wealth at the micro level. Methodological issues and experimental results. Edition 2023. European Union/OECD, Statistical working papers.

Ballin, M., Di Zio, M., D'Orazio, M., Scanu, M. and Torelli, N. (2008), File concatenation of survey data: a computer intensive approach to sampling weights estimation. *Rivista di Statistica Ufficiale*, **2-3**, 5-12.

Ballin, M., D'Orazio, M., Di Zio, M., Scanu, M. and Torelli, N. (2009), Statistical matching of two surveys with a common subset. Univ. di Trieste, Dip. Scienze Economiche e Statistiche, Working Paper, **124**.

Barr, R.S. and Turner, J.S. (1981). Microdata file merging through large-scale network technology. *Mathematical Programming Study*, **15**, 1-22.

Bernini, C., Emili, S. and Galli, F. (2021). Does urbanization matter in the expenditure happiness nexus?. *Pap Reg Sci*., 1-26.

Claramunt-González, J., Van Delden, A. and De Waal, T. (2023). Assessment of the effect of constraints in a new multivariate mixed method for statistical matching. *Computational Statistics and Data Analysis*, **177.**

Conti, P.L., Marella, D. and Neri, A. (2017). Statistical matching and uncertainty analysis in combining household income and expenditure data. *Statistical Methods & Applications*, **26**, 485-505.

Conti, P.L., Marella, D. and Scanu, M. (2010). Evaluation of matching noise for imputation techniques based on the local linear regression estimator. *Computational Statistics and Data Analysis*, **53**, 354-365. DOI 10.1016/j.csda.2008.07.041.

Conti, P.L., Marella, D. and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association*, **111**, 1715-1725. DOI 10.1080/01621459.2015.11128.

Conti, P.L., Marella, D. and Scanu, M. (2017). How far from identifiability? A systematic overview of the statistical matching problem in a nonparametric framework. *Communications in Statistics Theory and Methods*, **46**, 967-994.

Dalla Chiara, E., Menon, M. and Perali, F. (2019). An integrated database to measure living standards. *Journal of Official Statistics*, **35**, 531-576.

Decoster, A., De Rock, B., De Swerdt, K., Loughrey, J., O'Donoghue, C. and Verwerft, D. (2020). Comparative analysis of different techniques to impute expenditures into an income data set, *International Journal of Microsimulation*, **13**, 70-94.

De Waal, T. (2015). General approaches for consistent estimation based on administrative data and surveys, Discussion paper, Statistics Netherlands.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. John Wiley & Sons.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M. and Spaziani, M. (2014). Statistical matching of income and consumption expenditures. *International Journal of Economic Sc.*, **3**.

Donatiello, G., D'Orazio, M., Frattarola, D., Rizzi, A., Scanu, M. and Spaziani, M. (2016). The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics. DGINS - Conference of the Directors General of the National Statistical Institutes, 26-27 September 2016, Vienna.

Donatiello, G., D'Orazio, M., Frattarola, D. and Spaziani, M. (2022). The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching. *Rivista di Statistica Ufficiale - Review of Official Statistics*, **3**, 77-109.

D'Orazio, M. (2015). Integration and imputation of survey data in R: the StatMatch package. *Romanian Statistical Review*, **2**, 57-68.

D'Orazio, M. (2022) StatMatch: statistical matching or data fusion. R package version 1.4.1, https://CRAN.R-project.org/package=StatMatch.

D'Orazio, M. and Catanese, E. (2016). Evaluating revenues and economic growth for farms producing renewable energies: an investigation based on integration of FSS and EOAH 2013 survey data. Proceedings of the Seventh International Conference on Agricultural Statistics – ICAS VII, Rome 26-28 October 2016, 938-945 (DOI: 10.1481/icasVII.2016.e26c).

D'Orazio, M., Di Zio, M. and Scanu, M. (2006a). *Statistical Matching: Theory and Practice*. Wiley, Chichester.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006b). Statistical matching for categorical data: displaying uncertainty and using logical constraints. *Journal of Official Statistics*, **22**, 137-157.

D'Orazio, M., Di Zio, M., and Scanu, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. *Proceedings of the* "XLV Riunione Scientifica" of the Italian Statistical Society (SIS), Padova, 16-18 June 2010.

Endres, E., Fink, P. and Augustin, T. (2019). Imprecise imputation: a nonparametric micro approach reflecting the natural uncertainty of statistical matching with categorical data. *Journal of Official Statistics*, **35**, 599-624.

Fosdick, B.K., De Yoreo, M. and Reiter, J.P. (2016). Categorical data fusion using auxiliary information. *The Annals of Applied Statistics*, **10**, 1907-1929.

Gazzelloni, S., Romano, M.C., Corsetti, G., Di Zio, M., D'Orazio, M., Pintaldi, F., Scanu, M. and Torelli, N. (2008). Time Use and Labour Force: a proposal to integrate the data through statistical matching. In: (Romano, M. C. ed.) I tempi della vita quotidiana: un approccio multidisciplinare all'analisi dell'uso del tempo, Argomenti N. **32**, Istat, 375-403.

Ghahroodi, Z.H. (2023). Statistical matching of sample survey data: application to integrate Iranian time use and labour force surveys. *Statistical Methods & Applications*, **32**, 1023-1051.

Hossain, S., Loa, P., Wang, K., Mashrur, S.M., Dianat, A. and Habib, K.N. (2022). Comprehensive data fusion to evaluate the impacts of covid-19 on passenger travel demands: application of a core-satellite data collection paradigm. Available at SSRN: https://ssrn.com/abstract=4181189 or http://dx.doi.org/10.2139/ssrn.4181189.

Jausling, R. and Tillé, Y. (2023). An efficient approach for statistical matching of survey data trough calibration, optimal transport and balanced sampling. *Journal of Statistical Planning and Inference*, **225**, 121-131.

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

Lopez-Laborda, J., Marin, C. and Onrubia, J. (2020). Estimating Engel curves: A new way to improve the SILC-HBS matching process using GLM methods. *Journal of Applied Statistics*, **48**, 3233-3250.

Leulescu, A. and Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Publications Office of the European Union, Methodologies & Working papers.

Marella, D., Conti, P.L. and Scanu, M. (2013). Uncertainty analysis for statistical matching of ordered categorical variables. *Computational Statistics and data Analysis*, **68**, 311-325. DOI 10.1016/j.csda.2013.07.004.

Marella, D. and Pfeffermann, D. (2019). Accounting for non-ignorable sampling and non-response in statistical matching. *International Statistical Review*, **91**, 269-293.

Moretti, A. and Shlomo, N. (2023). Improving statistical matching when auxiliary information is available. *Journal of Survey Statistics and Methodology*, **11**, 619-642.

Opsomer, J.D. (2009). Introduction to part 4. Alternative approaches to inference from survey data. In Pfeffermann D. and C.R. Rao (Eds) *Sample Surveys: Inference and Analysis*, **29B**. Elsevier, Amsterdam.

Renssen, R.H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology*, **24**, 171-183.

Rios-Avila, F. (2015). Quality of match for statistical matches using the consumer expenditure survey 2011 and annual social economic supplement 2011. Levy Economics Institute of Bard College, Working Paper No. **830**.

Rios-Avila, F. (2016). Quality of match for statistical matches used in the development of the Levy institute measure of time and consumption poverty (limtcp) for Ghana and Tanzania. Levy Economics Institute of Bard College, Working Paper No. **873.**

Rios-Avila, F. (2020). Quality of match for statistical matches used in the development of the Levy institute measure of time and consumption poverty (limtcp) for Ethiopia and South Africa. Levy Economics Institute of Bard College, Working Paper No. **970**.

Rubin, D.B. (1986). Statistical matching using sample concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat*., **4**, 87-94.

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

Schifeling, T., Reiter, J.P. and De Yoreo, M. (2019). Data fusion for correcting measurement errors. *Journal of Survey Statistics and Methodology*, **7**, 175-200.

Tedeschi, S. and Pisano, E. (2013). Data fusion between Bank of Italy-SHIW and ISTAT-HBS, MPRA Paper No. **51253**.

Tonkin, R. and Webber, D. (2013). Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation. 2013 Edition. European Commission Statistical Working Papers.

Torelli N., Ballin, M., D'Orazio, M., Di Zio, M., Scanu, M. and Corsetti, G. (2009). Statistical matching of two surveys with a non-randomly selected common subset. In: Eurostat, Insights on Data Integration Methodologies. Office for Official Publications of the European Communities, Luxembourg, 68-79, isbn: 9789279123061

Tram, T.T.H. and Osier, G. (2023). Identifying the disadvantaged in Luxembourg – Measuring multidimensional poverty by statistical matching. Economie et Statistiques, Working papers du STATEC, N. **133.**

Ucar, B. and Betti, G. (2016). Longitudinal statistical matching: transferring consumption expenditure from HBS to SILC panel survey. Univ. di Siena, Quaderni del Dipartimento di Economia Politica e Statistica, N. **739**.

Wiest, M., Kutscher, T., Willeke, J., Merkel, J., Hoffmann, M., Kaufmann-Kuchta, K. and Widany, S. (2019). The potential of statistical matching for the analysis of wider benefits of learning in later life. *European Journal for Research on the Education and Learning of Adults*, **10**, 291-306.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Can. J. Stat.*, **32**, 1-12.

Zacharias, A., Masterson, T. and Memis, E. (2014). Time deficits and poverty, the Levy Institute measure of time and consumption poverty for Turkey. UNDP & Levy Economics Institute of Bard College, Research Project Report.

The Survey Statistician, 2024, Vol. 89, 59–67.

Social big data to enhance small area estimates

Stefano Marchetti[1] **and Francesco Schirripa Spagnolo**[2]

[1]University of Pisa, Italy, stefano.marchetti@unipi.it
[2]University of Pisa, Italy, francesco.schirripa@unipi.it

Abstract

The paper discusses the challenge and the opportunity of the use of big data in small area estimation applications. Big data come in an unstructured way, and they are self-selected. Therefore, they have to be handled with care, and they need adequate statistical methods to produce sound statistics. Together with the discussion, we present an application where data mined from Twitter (now changed to X) are used to improve small area estimates of consumption expenditure for leisure at the local level in Italy.

Keywords: Twitter, area-level models, leisure consumption

1 Introduction

In recent decades, there has been a growing demand for small area official statistics for decision-making at the local level. By small area, we mean subdomains of a population (such as geographical areas or socio-economic groups) where direct estimates – based on area units only – from surveys typically do not offer accurate estimations. To overcome this problem, survey statisticians rely on a model-based approach and use statistical models to *borrow strength* across areas. Rao and Molina (2015) and Pratesi (2016) provide a comprehensive account of model-based approaches for Small Area Estimation (SAE).

At the same time, as a result of technological innovations and the growth of the Internet and the Web, the availability of new kinds of unstructured and heterogeneous data originating from ICT systems, the so-called big data, is increasing at an unprecedented rate. Examples of big data sources are GPS data, mobile phone data, internet searches, and social networking. Many of these data can be viewed as proxies of social behaviour along various dimensions. For instance, data coming from social networks, blogs, or web search keywords can trace desires, opinions, and feelings; records of mobile phone calls and GPS trajectories can trace the movement of individuals (Marchetti et al., 2015). Consequently, a growing number of analysts and academics have looked into the benefits of utilizing big data in socioeconomic studies.

For what concerns the use of big data in SAE there is a need to point out the dimension of big data in terms of the number of units in the small areas. We start by recalling the definitions of big data, which are many. A popular definition is: "*big data is the information asset characterized by such a high Volume, Velocity and Variety to require specific technology and analytical methods for its transformation into Value*" (De Mauro et al., 2016). It is also known as the four V's definition (Volume, Velocity, Variety, Value). Other definitions add more V's, like Veracity, Variability, etc., up to 42 V's in Farooqi et al. (2019).

Assuming that a high volume of data is fundamental to dealing with big data, this does not automatically imply that we have the availability of a large sample size at the small area level. Indeed, we can distinguish two kinds of big data: i) "horizontal" and ii) "vertical".

The first kind of big data is characterized by many observations for each unit, e.g. vehicles with GPS track can produce data for longitude, latitude and altitude variables every 5", for about 520 thousand observations in one month for each variable for each unit. Nevertheless, having many observations for each unit does not imply having observed many units, e.g. we can have a few vehicles with GPS track at the small area level, like in Marchetti et al. (2015).

The latter are big data characterized by a large number of units at the small area level. Of course, we can have big data that are both vertical and horizontal.

Moreover, the high spatial granularity of big data and their availability at an unprecedented temporal detail may enable us to use them to infer in near real-time socio-economic characteristics for an entire nation as well as for disaggregated geographical domains, which are important for timely, evidence-based policies. The integration of big data in the SAE framework represents a valuable resource to improve the accuracy of local estimates.

In the last decade, the use of big data in the SAE framework has been increasingly explored by researchers, with the aim of estimating well-being and other socioeconomic indicators, such as poverty indicators, for unplanned domains such as provinces in Italy (NUTS 3 in the Eurostat nomenclature), as their knowledge may be useful in better planning local policies and distributing welfare resources.

Model-based SAE can be divided into two approaches: area-level and unit-level models (Rao and Molina, 2015). The first uses aggregated area-level data in the regression model. The latter uses unit-level data (microdata), where a model fitted on sample data is then projected on the population data. Area-level models represent a natural way to include big data that can be aggregated at the area-level (Porter et al., 2014; Marchetti et al., 2015, 2016; Schmid et al., 2017); on the contrary, to the best of our knowledge, the use of unit level models by including big data has been only recently investigated (Pratesi et al., 2022).

We can identify three possible approaches for the use of big data in the SAE framework (Marchetti et al., 2015; Pratesi and Schirripa Spagnolo, 2023):

1) Local indicators are created by using big data sources, and they are then compared with the results obtained from SAE techniques.

2) Big data are used to generate new covariates for small area models

3) Survey data are used to check and remove the self-selection bias of the values of the indicators obtained using big data.

The first approach has been shown by Marchetti et al. (2015), where an entropy-based mobility variability index has been compared to poverty incidence at the province (small area) level in Tuscany, showing the potential of big data to catch the direction of poverty incidence. The idea of this first approach is to use big data to estimate a target parameter or an index strictly related to that target,

and then validate the results using traditional high-quality surveys. If results based on big data prove to be reliable over time, then they can be used to anticipate results from surveys, which are typically available several times after the big data.

At the moment, the second approach is the most explored. In particular, under this approach, Porter et al. (2014) used Google Trends data as covariates in a standard spatial Fay–Herriot (FH) model (as in Pratesi et al., 2009) to estimate the relative changes in rates of household Spanish-speaking in the United States. Marchetti et al. (2015), presented an application of the modified FH model (Fay and Herriot, 1979) proposed by Ybarra and Lohr (2008) to estimate poverty indicators for local areas in Tuscany using big data on mobility as covariates. In particular, the authors used mobility indexes based on different car journeys between locations automatically tracked with a GPS device. Similarly, Marchetti et al. (2016) used data coming from the social network Twitter to predict the share of food consumption expenditure of Italian households at the provincial level. In particular, they included as a covariate in the FH model an indicator, called iHappy, obtained from the analysis of Twitter (now named X) data measuring happy tweets to the total of tweets at the provincial level. They showed that this indicator has a good predictive power for food consumption expenditure and can be used as a proxy to measure households' living conditions. Another interesting application developed following the second approach described above was proposed by Schmid et al. (2017), who used mobile phone data to estimate subnational estimates of the share of illiterate individuals by gender at the local (*commune*) level in Senegal.

When we deal with horizontal big data - the big data size at the small area level is small - if we want to use big data as auxiliary variables in the SAE models, we can use the modified version of the FH model proposed by Ybarra and Lohr (2008), which allows for the sampling error in the auxiliary variables.

On the other hand, when we deal with vertical big data, the survey error at the area-level can be negligible and a standard FH model can be appropriate if we want to use them as auxiliary variables in the small area model (Marchetti et al., 2016).

For what concern the third approach, in the last year, many scholars have focused on the issue of selection bias that arises when big data are used. In particular, according to the third approach mentioned above, integrating data from a probability survey and a non-probability source is used to make valid inferences (for a review on this topic see Yang and Kim, 2020; Lohr and Raghunathan, 2017). In the SAE framework, this problem has been addressed by Pratesi et al. (2022), who proposed a method based on the integration of a probability and a nonprobability sample to reduce the selection bias associated with the big data source when the aim is to predict statistics related to enterprises at the local level. In their work, the authors assumed that the variable of interest is only in the non-probability (big data) data source.

In this paper, we show the potential of big data coming from the social network Twitter to improve small area estimates based on an area-level FH model (see details on Rao and Molina, 2015). We use the second approach for this application among the three possible approaches suggested above.

2 Potential of social big data to improve small area estimate, an application

In this section, we use a social index based on text analysis of georeferenced tweets from Twitter to improve the small area estimates of consumption expenditure for leisure at the province level in Italy in 2017.

2.1 Data

In this application, we use the Italian Household Budget Survey (HBS) 2017, aggregated data coming from the tax and population registers and the iHappy index based on big data from Twitter.

In Italy, the HBS is the main source of information concerning consumption expenditure. The HBS has a stratified two-stage design, which allows for reliable estimates at the regional level (NUTS 2). The sample size is about 17000 households and 40000 persons. The sample size at the province level varies between 20 and 1036, with a median of 125. Therefore, in most of the provinces direct estimation leads to unreliable estimates. From the HBS we estimate the consumption expenditure for leisure activities, which is a driver of subjective well-being (Noll and Weick, 2015), and it is our target variable from which we want to obtain a mean estimate at the province level in Italy.

Model-based small area methods require the use of auxiliary variables which are related to the target variable. We found that province-level variables coming from the Italian tax register are suitable for our purpose. From this source, we have available the following variables: per capita tax, per capita income from real estate, per capita income from labour, and proportion of taxpayers. From the population register we obtain the mean age at the province level.

We also considered as a potential source of auxiliary information big data obtained from Twitter in 2017. In particular, we use the iHappy index at the province level available from the Opinion Analytics platform Voices from the Blogs (available here `http://media2.corriere.it/corriere/pdf/2018/cultura/ihappy2017-18-def.pdf?fbclid=IwAR2dUuMDUxXEJICRi6Owpw4ICpFGi3jTLlyOZ-y1YxXj1tAaC6GIAfj2bn4`)[1]. This index referring to the year 2017 was obtained from more than 52 million tweets posted on a daily basis in all the Italian provinces. A text analysis of the tweets classifies them into two categories: "happy" and "unhappy" (Curing et al., 2015). The iHappy index is the percentage ratio of the number of happy tweets to the sum of happy and unhappy tweets. The iHappy index for Italy in 2017 is 55.1%. The happiest province is Genova (north-west of Italy) with 59.5%, while the unhappiest is Aosta with (also in the north-west) 49.9%.

2.2 Small area estimation method

Given the availability of data aggregated at the provincial level we use an area-level model, as described in Rao and Molina (2015). Let θ_i be the target parameter (mean or total of a target variable) for area i, and let $\hat{\theta}_i$ be its direct estimator, then $\hat{\theta}_i = \theta + \varepsilon$, where under random sampling we usually assume $\varepsilon \sim N(0, \psi_i)$, and ψ_i is the variance of $\hat{\theta}_i$. Let $\boldsymbol{x}_i$ be a vector of auxiliary variables for area i, then we assume $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i$, where $\boldsymbol{\beta}$ is a vector of regression coefficients and $u_i \sim N(0, \sigma_u^2)$. The FH model is then

$$\hat{\theta}_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + u_i + \varepsilon_i,$$

where u_i represents the area i random effect, which is independent from ε_i.

The best linear unbiased predictor (BLUP) of θ_i is $\tilde{\theta}_i = \hat{\theta}_i\gamma_i + \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}(1-\gamma_i)$, where $\gamma_i = \sigma_u^2/(\sigma_u^2+\psi_i)$ and $\tilde{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, while σ_u^2 and ψ_i are unknown. Usually, ψ_i is considered known, and the smoothed estimator of ψ_i is treated as if it is the true sampling variance. The BLUP is a convex combination of the direct estimator $\hat{\theta}_i$ and of the predicted value $\boldsymbol{x}_i^T\boldsymbol{\beta}$, with weights γ_i and $1-\gamma_1$, where γ_i is the relative sizes of the model error variance σ_u^2 and the sampling error variance ψ_i.

Using maximum likelihood estimation or restricted maximum likelihood estimation we can obtain es-

[1] Link verified on December 11th 2023

timates of σ_u^2 and β, so to obtain the empirical best linear unbiased predictor (EBLUP):

$$\hat{\theta}_i^{FH} = \psi_i\hat{\gamma}_i + \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_i(1-\hat{\gamma}_i), \quad \hat{\gamma}_i = \frac{\hat{\sigma_u^2}}{\hat{\sigma_u^2}+\psi_i}.$$

The mean squared error (MSE) of the BLUP is $MSE(\tilde{\theta}_i) = E[(\tilde{\theta}_i - \theta_i)^2] = \psi_i\gamma_i$. Therefore γ_i measures the reduction of the variability of the BLUP with respect to the variability of the direct estimator. The MSE of the EBLUP is $MSE(\hat{\theta}_i^{FH}) = \psi_i\hat{\gamma}_i + \boldsymbol{x}_i^T V(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i(1-\hat{\gamma}_i)^2 + \psi_i^2(\psi_i+\sigma_u^2)^{-3}V(\hat{\sigma}_u^2)$. Analytical estimation of the MSE of the EBLUP is possible, details in Rao and Molina (2015).

Estimates in our application have been obtained using the package `emdi` (Kreutzmann et al., 2019) in the `R` environment (R Core Team, 2023).

2.3 Estimate of the mean consumption expenditure for leisure at the provincial level with and without big data covariates

We show how the use of the iHappy index as a covariate in the FH model in addition to other register-based covariates can improve the efficiency of estimates in many areas. Our target is the mean consumption expenditure for leisure at the province level in Italy. As discussed before, the sample size of the HBS does not allow for reliable estimates at such a level of aggregation. Therefore, we resort to SAE models, and in particular to the FH area-level model described above because we have access to aggregated data only.

First, we obtain direct estimates of the mean consumption expenditure at the province level (θ_i) for all the 107 Italian provinces using micro-data from the HBS 2017 edition[2]. Let y_{ij} be the expenditure for leisure for household j in province i, and let w_{ij} be the associated survey weight, adjusted for non-response and measurement error, and calibrated at the provincial level. Then a direct estimator of θ_i is $\hat{\theta}_i = \sum_{j=1}^{n_i} y_{ij}w_{ij} / \sum_{j=1}^{n_i} w_{ij}$. According to Statistics Canada, there are no restrictions to publish estimates with a coefficient of variation (CV) less than 16%, other national statistical offices use different thresholds (Eurostat, 2013). Using a CV less than 16% as a reference, in our case only 3 provinces out of 107 have such a CV, making evident the need to resort to SAE methods.

We adopt two different FH models to improve the efficiency of direct estimates, one without and one with the iHappy index (x_1). As auxiliary variables in the FH model without iHappy, we use the province mean age (x_2) and the per capita tax (x_3).

The two FH models successfully increase the efficiency of direct estimates. In table 1 we show the number of provinces for which the CV is less or equal to 16% and for which is greater than 16%, i.e. the number of provinces for which the estimates are considered reliable or not. As already noted, only three direct estimates can be considered reliable. On the contrary, the small area estimates are reliable for all 107 provinces.

Table 1: Number of provinces by coefficient of variation (CV)

Estimator	CV $\leq 16\%$	CV $> 16\%$
Direct	3	104
FH without iHappy	107	0
FH with iHappy	107	0

Given that both the FH models work well, we observed that the estimates based on the FH model that include the iHappy index are for 87 out of 107 provinces a little bit more efficient than the estimates

[2]These data are made available to the authors under the European Project MAKSWELL, grant agreement 770643

obtained without the iHappy index. In table 2 we show the ratio between the estimated MSE of small area estimates obtained with and without the iHappy index. A value smaller than one means that small area estimates obtained with iHappy are more efficient than those obtained without.

Table 2: Summary of the ratio between estimated MSEs of small area estimates obtained with and without the iHappy index

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.969	0.985	0.990	0.994	0.996	1.068

The gain in efficiency is very limited, but is however present.

The estimates (with iHappy) of the mean consumption expenditure for leisure at the provincial level in Italy in 2017 are mapped in figure 1. The values are expressed in euro per month, and represent the average household expenditure for leisure in a month. The mean consumption expenditure for leisure at national level in Italy is 42.61 euro per month. At province level the expenditure for leisure ranges from 1.76 per month in Nuoro, the central part of Sardinia (the Italian island on the 40°N parallel) to 85.00 euro per month in Milan, north-west of Italy. Milan is an outlier, with a very high expenditure level. The highest one without considering Milan is 67.42 euro per month in Bolzano, in the north-east of Italy, on the border with Austria. A similar level of consumption expenditure (67.22 euro) is in Monza-Brianza, which borders the Milano province. From figure 1 is also evident the socio-economic north-south divide present in Italy, with the highest values of consumption expenditure for leisure in the northern provinces and the lowest values in the southern provinces.

3 Concluding remarks

This article summarises the possible use of big data in small area estimation. Big data are a valuable source of information and knowledge that come in an unstructured way. After appropriate elaboration, they can be used, among others, in the framework of small area estimation, mainly in three ways: i) to create indexes that predict similar indicators from reliable data sources, such as surveys, ii) as auxiliary variables in small area estimation methods, iii) to estimate small area target parameters that are adjusted for self-selection and measurement error using survey sample. All these three methods have been explored in the literature, however, the approach ii) is the most used, and it is used in the application shown in this article. Similarly to previous works, we use an index obtained from Twitter data as a covariate in a small area model to estimate the consumption expenditure for leisure in Italy at the provincial (small area) level in 2017. Even if the application can be considered as a bit more than an example, it shows the potential of the use of big data in SAE.

Big data represents a challenge and an opportunity, in the field of small area estimation and in many other fields of statistics. In the last years, many efforts have been made to use them to obtain sound statistics. However, it is important to remark that it is often difficult to have access to big data, that are mainly available to big tech companies, banks, big retail shops, etc. Nevertheless, the possibility of scraping data from the web makes it possible to have access to a wide range of big data, that can be used by researchers to explore the potential of this new source of information.

Acknowledgment

This paper is supported by the Ministry of University and Research (MUR) as part of the FSE REACT-EU - PON 2014-2020 “Research and Innovation” resources - Innovation Action - DM MUR 1062/2021 - Title of the Research: “Statistical Machine Learning nelle Indagini Campionarie”.

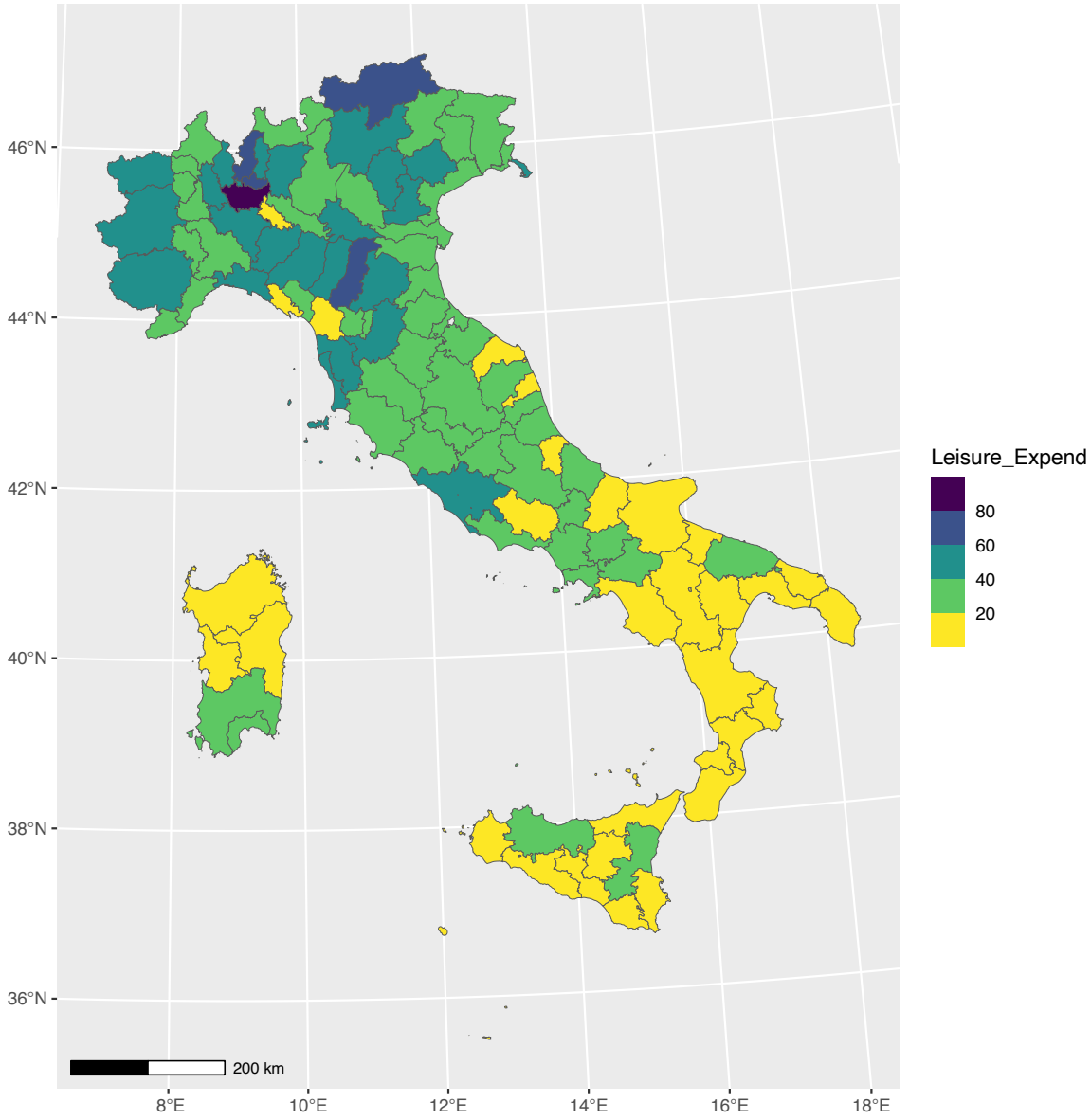


Figure 1: Estimates of consumption expenditure on leisure at provincial level in Italy, 2017

References

Curing, L., S. Iacus, and L. Canova (2015). Measuring idiosyncratic happiness through the analysis of twitter: An application to the italian case. *Social Indicators Research 121*(2), 525–542.

De Mauro, A., M. Greco, and M. Grimaldi (2016, 03). A formal definition of big data based on its essential features. *Library Review 65*, 122–135.

Eurostat (2013). *Handbook on precision requirements and variance estimation for ESS households surveys*. Eurostat.

Farooqi, M. M., M. A. Shah, A. Wahid, A. Akhunzada, F. Khan, N. ul Amin, and I. Ali (2019). Big data in healthcare: A survey. In F. Khan, M. A. Jan, and M. Alam (Eds.), *Applications of Intelligent Technologies in Healthcare*, pp. 143–152. Springer International Publishing.

Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association 74*(366a), 269–277.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software 91*(7), 1–33.

Lohr, S. L. and T. E. Raghunathan (2017). Combining survey data with other data sources. *Statistical Science 32*(2), 293–312.

Marchetti, S., C. Giusti, and M. Pratesi (2016). The use of Twitter data to improve small area estimates of households'share of food consumption expenditure in Italy. *AStA Wirtschafts- und Sozialstatistisches Archiv 10*(2), 79–93.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics 31*(2), 263–281.

Noll, H.-H. and S. Weick (2015). Consumption expenditures and subjective well-being: empirical evidence from germany. *International Review of Economics 62*(2), 101–119.

Porter, A. T., S. H. Holan, C. K. Wikle, and N. Cressie (2014). Spatial Fay–Herriot models for small area estimation with functional covariates. *Spatial Statistics 10*, 27–42.

Pratesi, M. (Ed.) (2016). *Analysis of Poverty Data by Small Area Estimation*. New Yok: John Wiley & Sons.

Pratesi, M., N. Salvati, et al. (2009). Small area estimation in the presence of correlated random area effects. *Journal of Official Statistics 25*(1), 37.

Pratesi, M. and F. Schirripa Spagnolo (2023). Small area methodology for measuring poverty at a local level. In J. Silber (Ed.), *Research Handbook on Measuring Poverty and Deprivation*, pp. 129–140. Edward Elgar Publishing.

Pratesi, M., F. Schirripa Spagnolo, G. Bertarelli, S. Marchetti, M. Scannapieco, N. Salvati, and D. Summa (2022). Inference for big data assisted by small area methods: an application to OBEC (on-line based enterprise characteristics). In A. Balzanella, M. Bini, , C. Cavicchia, and R. Verde (Eds.), *Book of short papers SIS 2022*, pp. 305–311. Pearson.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. N. and I. Molina (2015). *Small area estimation*. New Yok: John Wiley & Sons.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. *Journal of the Royal Statistical Society Series A: Statistics in Society 180*(4), 1163–1190.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 1–26.

Ybarra, L. M. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika 95*(4), 919–931.

The Survey Statistician, 2024, Vol. 89, 68–73.

EARLY CAREER SURVEY STATISTICIAN

The growth of new researchers in the era of new data sources

Veronica Ballerini[1] **and Lisa Braito**[2]

[1]University of Florence, Italy, veronica.ballerini@unifi.it
[2]University of Florence, Italy, lisa.braito@unifi.it

Abstract

In this article, we introduce a new section of The Survey Statistician, the "Early Career Survey Statistician." We report our personal experience in the field of survey statistics as early career researchers and, inspired by the recent workshop on methodologies for official statistics held in Rome last December, we make a digression into the new challenges in the era of innovative data.

Keywords: Early Career Researchers, Innovative data, Machine Learning, Nonprobability samples

1 Introduction

In the last issue of *The Survey Statistician*, Prof. Danny Pfefferman raised his concern about the involvement of "young" statisticians in the activities of the IASS. In the same article, he reported a suggestion from a reviewer that did not go unheeded: allocating a special section in TSS for young survey statisticians. Said and done, this is the first introductory number of a new section of TSS, the "Early Career Survey Statistician" (ECSS). The ECSS welcomes original research works of junior researchers, summaries of their research, review papers on survey statistics-related topics, reviews of events on survey statistics, and innovations introduced by junior researchers at the statistical offices. An "early career survey statistician" is a person with up to 5 years of employment or within the fifth year since the achievement of their Ph. D., who is researching in the field of survey statistics. To continue citing the recent article by Pfefferman in TSS, "the outcome of our [*survey statisticians'*] work affects directly so many applications and decision makings." In the era of novel data sources and huge data availability, this is truer than ever. Indeed, the research work of survey statisticians is instrumental to the proper secondary use of big or complex data and nonprobabilistic samples in general, and it is essential to contribute to making the estimates obtained reliable. Every time and in every context, change has always been embraced by the "youth"; we should also ride the change in the survey statistics field.

I (Veronica) met many early career researchers like myself in this year's events on survey statistics; the majority of us come from mixed backgrounds and our research crosses different statistical fields. Such a transdisciplinary attitude might be a positive aspect. Another of my main research interests is causal inference for clinical and observational studies, which is a broad field perceiving the opportunities offered by data integration and facing its challenges at the same time.

Lisa, who is a Ph. D. student in my own Department, and I have come across such challenges; for this reason, we are broadly reviewing the state of the art of what concerns new data sources and methods, and their opportunities. These have been topics also at the core of the "2nd Workshop on Methodologies for Official Statistics," held at the Italian National Statistical Institute (Istat) in Rome last December, from which we partially draw inspiration in this article. All presentations are available at https://www.istat.it/en/archivio/288564. We share some insights here, hoping to stimulate other early career researchers to get involved in such new survey statistics challenges and calling for contributions for the next issues of this new TSS section.

2 Challenges is Opportunities in Survey Statistics

If one were to identify the most significant drawback in the production of reliable statistics, it would be poor timeliness. Hence, the (almost) real-time production of "big" data is very attractive. However, it is important to quickly elaborate trustworthy data to prevent users from blindly relying on raw big data. Because "big" is not enough for reliable inference: "[...] classical statistical ideas continue to have a crucial role to play in keeping data analysis honest, efficient, and effective. [...] huge new computing resources do not put an end to the need for careful modelling, for honest assessment of uncertainty, or for good experimental design" (Lockhart, 2018). With the words of Monica Pratesi at the aforementioned workshop, "uncertainty is here to stay; so, we need inference".

Innovative data need to be processed and analyzed using innovative methods. Thus, there is a need for bridging traditional survey statistics with machine learning methods, which are more suitable to deal with such nontraditional data.

2.1 Innovative data for survey statistics

Satellite data, remote sensing, mobile network data, mobile sensor data, social media platforms, web scraping, Google trends, web surveys, and scanner data: all of these, and even more, belong to the class of "innovative data". As pointed out by Stefano Iacus in his master class on "Limits and challenges of incorporating innovative data in official statistics", their main competitive advantages are the fine geographic and temporal granularity, the profound timeliness, and the large coverage they can reach.

Let us consider the illustrative case brought forth by Iacus and his research conducted in collaboration with the European Comission, focusing on the complexities of migration as a phenomenon. Migration presents inherent characteristics that pose challenges when relying solely on traditional data sources for analysis. Surveys, for instance, suffer from high costs, infrequent data collection, and limited coverage, making them inadequate tools to comprehensively analyze this dynamic and multifaceted phenomenon. Given the rapid evolution of migration, the aforementioned limitations in timeliness and granularity associated with traditional sources, such as surveys, contribute to their diminished reliability. Moreover, the elusive nature of the target population further reduces the capture probability of the units within these surveys. To tackle these limitations, Iacus and his collaborators explored the combined use of innovative and traditional data. For insights on this topic see, e. g., Carammia, Iacus and Wilkin (2022) and Spyratos et al. (2020).

Another context where innovative data might be useful is what is called "data equity", namely the need for representative data and disaggregated statistics about those population groups that may be discriminated by the data production process. For instance, sometimes the linking procedure may introduce biases because linkage errors can disproportionately affect members of population subgroups due to, e. g., spelling and/or typing errors (National Academies of Sciences, Engineering, and Medicine, 2023). Alternative data sources are crucial in advancing data equity by identifying data

gaps or misrepresentations. They contribute by offering insights into population subgroups that are often under-represented in traditional surveys, such as individuals experiencing homelessness or residing in institutions like nursing homes. Additionally, these sources facilitate the generation of statistics that are disaggregated by key characteristics like race, ethnicity, education, disability status, and other factors of interest. This inclusive approach helps address disparities and ensures a more comprehensive understanding of diverse demographic groups.

Despite the undeniable opportunities, the utilization of non-traditional data sources for statistical purposes also presents several challenges. Before methodological considerations, let us underline that a relevant matter pertains to the management and processing of these data. Notably, new data sources lack stability, as they may cease to exist over time based on the discretion of the private entities that possess them. Data quality, data management and processing are open research fields, especially for NSIs.

Foremost among the methodological challenges is the issue of data linkage, which is further complicated by concerns surrounding data protection and privacy. Then, transparency, in general, remains a significant concern when dealing with such data sources, as the data production process is often unclear, necessitating reverse engineering efforts to integrate this data coherently with traditional sources. Additionally, various types of biases must be carefully addressed, such as selection bias stemming from the nonprobabilistic nature of these data.

Traditionally, a nonprobability sample is a sample with an unknown participation mechanism and an unknown sampled population; nonprobability samples that have been deeply investigated in the last decades are not only web surveys, volunteer surveys, administrative data, but also probability samples that encounter issues such as very low nonresponse rates and nonignorable nonresponse. The issues related to the nonprobability samples are intrinsic in their definition. Nowadays, the set of nonprobability samples comprises also the realm of big data. Whereas the literature about the integration of probability and nonprobability samples in a traditional framework is rich (among others, see Wu (2022) and his master class presentation slides during the workshop in Rome), with emerging new data sources and reshaped views of traditional data sources, data integration and data harmonization have become a very broad area that calls for continued research. Survey statisticians have started making efforts to integrate traditional literature on combining probability samples and innovative data (Yang and Kim, 2020). On the one hand, when the nonprobability sample has a large sample size and a probability sample including the response variable is available, it is possible to exploit the auxiliary information in the big data to improve the efficiency of the estimators of interest (Kim and Tam, 2021; Yang and Ding, 2020). On the other hand, when the big data includes the response variable, research has been done in the direction of leveraging probability samples to correct for selection bias improving robust mass imputation methods; for instance, see Yang, Kim and Hwang (2021). Lastly, a problem that may arise is linked to the large availability of variables in the big data sample; in this case, irrelevant auxiliary variables can introduce large variability in the estimation. Research is moving towards variable selection approaches tackling data integration and estimation rather than prediction; see, e. g., Chen, Valliant and Elliott (2018) and Yang, Kim and Song (2019).

2.2 Machine learning in survey statistics

The paradigm shift occurring in the field of survey statistics involves not only data sources but also methods for inference. Within this context, the shortcomings (and opportunities) we previously discussed about innovative data are similarly applicable to the new methodologies. The final session of the workshop, organized in collaboration with IASS, discussed these topics in detail, focusing on machine learning methods (ML) in survey statistics. In the last decade, the discussion concerning ML in survey statistics has been a hot topic. As pointed out by Puts and Daas (2021), ML methods

provide several advantages, such as better scalability, less sensitivity to outliers and erroneous data, and the ability to capture non-linear relationships. However, the authors also reflect upon challenges and limitations arising from the application of ML methods in survey statistics. Among the challenges, accessibility and clarity have to be taken into account; especially in the field of survey statistics, it is important to define and fully understand the process by which results are obtained. This is a problem for some ML algorithms that result in black boxes, and it touches on the topic of explainability and the development of explainable AI. Furthermore, we need to pay attention to accuracy and reliability. In the employment of ML methods, the spotlight is often far from the uncertainty assessment and the estimators' robustness itself; these concerns are at the core of survey and official statistics instead.

It is noteworthy that the use of ML methods in survey statistics is two-fold. On the one hand, there is the application of classical prediction tasks to problems of survey methodology (optimizing data collection, adaptive survey designs, predicting nonresponse break off in online web surveys) and survey statistics (imputation, classification, data integration, automatic coding, anomaly detection, forecasting), as also mentioned by IASS president Natalie Shlomo. The advantages of ML in this context have been already exploited by NSIs, especially for what concern imputation, data quality assurance, survey sampling, document classification issues, time series forecasting, topic modeling, sentiment analysis, and geospatial analysis. For examples of works on these topics, see Beck, Dumpert, Feuerhake (2018); Buskirk, Bear and Bareham (2018); Burkirk et al. (2018); Kern et al. (2023), Kern, Klausch and Kreuter (2019); see also UNECE (2021). However, the application of these methods is still conceived as "experimental" statistics, in the sense that they are not consistently integrated in the NSIs systems.

On the other hand, "ML in survey statistics" also refers to the development of novel intersections between ML and survey statistics methodologies. This area represents an ongoing research field with ample scope for exploration and innovation. Examples of research areas at this intersection are the issues of sampling the population to obtain representative training sets, using stratification in the context of ML, reducing spurious correlations and assessing causal relationships, correcting the bias caused by the ML model, dealing with concept drift (Puts and Daas, 2021). To have some insights on research studies going in this direction see Breidt and Opsomer (2017); Chen and Haziza (2019); Dagdoug, Goga, and Haziza (2023a, 2023b); see also Buskirk and Kirchner (2020).

3 Concluding remarks

Encouraging the active participation of early career statisticians in research and collaboration endeavors will not only nurture their professional growth but also foster the development of novel methodologies and approaches that address the evolving challenges and opportunities in survey statistics. The interdisciplinary background we may have could be a potential strength, not only in the research work per se, but also in the creation of synergies among statisticians, data scientists, computer scientists, and operational and applied researchers, which is crucial for advancing the field of survey statistics in the era of data-driven decision-making. Inspired by the topics of the recent workshop on methodologies for official statistics held in Rome last December, we overviewed some of the hot topics in modern survey statistics, namely the use of innovative data sources and the role of machine learning methods in this field, to stimulate other early career statisticians to contribute to this field.

References

Beck, M., Dumpert, F. and Feuerhake, J. (2018) Machine Learning in Official Statistics. *arXiv preprint*, https://arxiv.org/abs/1812.10422; https://DOI:10.18356/9789210011143.

Breidt, F. J., and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32 (2) 190 - 205, May 2017. https://doi.org/10.1214/16-STS589

Buskirk, T. D., Bear, T., and Bareham, J. (2018). Machine made sampling designs: applying machine learning methods for generating stratified sampling designs. *Paper presented at the BigSurv18 Conference*, Barcelona, Spain (25-27 October 2018).

Buskirk, T. D., and Kirchner, A. (2021). Why machines matter for survey and social science researchers: Exploring applications of machine learning methods for design, data collection, and analysis. *Big data meets survey science: A collection of innovative methods*, Eds. Hill, C. A.; Biemer, P. P.; Buskirk, T. D.; Japc, L.; Kirshner, A.; Kolenikov, S.; Lyberg, L. E. John Wiley & Sons, 9-62. https://DOI:10.1002/9781118976357.

Buskirk, T. D., Kirchner, A., Eck, A., and Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1).

Carammia, M., Iacus, S. M., and Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, 12(1), 1457. https://doi.org/10.1038/s41598-022-05241-8

Chen, J. K. T., Valliant, R., and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44, 117–144.

Chen, S., and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: A critical review. *International Statistical Review*, 87, S192-S218. https://doi.org/10.1111/insr.12305

Dagdoug, M., Goga, C. and Haziza, D. (2023a). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 542, 1234-1251.

Dagdoug, M., Goga, C. and Haziza, D. (2023b). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Journal of Survey Statistics and Methodology* 11, 141-188.

Kern, C., Eckman, S., Beck, J., Chew, R., Ma, B., and Kreuter, F. (2023). Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. *arXiv preprint* arXiv:2311.14212, https://doi.org/10.48550/arXiv.2311.14212.

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey research methods*, 13(1), 73-93. https://doi.org/10.18148/srm/2019.v1i1.7395

Kim, J. K. and Tam, S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382-401. https://doi.org/10.1111/insr.12434

Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46(1), 4-9. https://doi.org/10.1002/cjs.11350

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Eds. S. L. Lohr, D. H. Weinberg, K. Marton. The National Academies Press. https://doi.org/10.17226/26804

Pfeffermann, D. (2023). The IASS–50 Years of Activity. *The Survey Statistician*, 88, 72-74.

Puts, M. J. H. and Daas, P. J. H. (2021). Machine Learning from the Perspective of Official Statistics. *The Survey Statistician*, 84, 12-17.

Spyratos, S., Vespe, M., Natale, F., Iacus, S. M., and Santamaria, C. (2020). Explaining the travelling behaviour of migrants using Facebook audience estimates. *Plos ONE*, 15(9), e0238947.

UNECE (2021). *Machine Learning for Official Statistics*. United Nations publication, Geneva. https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*. 48(2), 283-311. http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm

Yang, S. and Ding, P. (2019) Combining Multiple Observational Data Sources to Estimate Causal Effects, *Journal of the American Statistical Association*, 115:531, 1540-1554, doi:10.1080/01621459.2019.1609973

Yang, S., Kim, J. K., and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47(1), 29-58.

Yang, S., Kim, J. K., and Song, R. (2019). Doubly robust inference when combining probability and nonprobability samples with high-dimensional data. *Journal of the Royal Statistical Society*, Series B, 82, 445–465.

Yang, S., and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

The Survey Statistician, 2024, Vol. 89, 74–75.

Book and Software Review

A Course on Small Area Estimation and Mixed Models Methods Theory and Applications in R

Caterina Giusti[1]

[1]Department of Economics and Management, University of Pisa, Italy.
caterina.giusti@unipi.it

Abstract

The book *A Course on Small Area Estimation and Mixed Models Methods. Theory and Applications in R* by Domingo Morales, María Dolores Esteban, Agustín Pérez and Tomáš Hobza was published by Springer in 2021 in the series Statistics for Social and Behavioral Sciences. The book will easily become a reference manual for researchers working in Universities and statistical offices and for PhD students who aim at studying small area estimation methodologies and mixed models from both a theoretical and applied perspective. In its 21 chapters the book covers some of the main models of small area estimation with plenty of details in the main mathematical developments, and with applications to synthetic socioeconomic indicators using R code lines.

Keywords: Small area models, random effect models, synthetic data, R coding.

In the last decades, Small Area Estimation (SAE) models have received a growing attention in the scientific literature, with a corresponding increasing number of methodological articles and manuals. However, these materials are often too advanced for researchers approaching the topic for the first time. Moreover, SAE models are relevant not only from a methodological perspective, but also for their application to obtain reliable estimates for unplanned domains in sample surveys. Therefore, training in SAE should focus not only on the mathematical developments of the estimators, but also on the relevant aspects of their application, including open software codes.

The book *A Course on Small Area Estimation and Mixed Models Methods. Theory and Applications in R* by Domingo Morales, María Dolores Esteban, Agustín Pérez and Tomáš Hobza addresses the basic aspects of the theory and practice of SAE for readers who don't need to be expert in sampling, statistical modeling, or programming languages. Specifically, as stated by the authors, the book aims at being useful to researchers from universities and statistical offices, and to doctoral students. For this reason each chapter, dedicated to a specific SAE model, reports the main mathematical developments with plenty of details, and examples of application of the models to synthetic data using R codes. The coding is as simple as possible, so that the reader can easily identify the corresponding mathematical formulas. The synthetic data files and the codes are available by chapter on a dedicated website.

The data consist in two files containing unit-level data from a labor force survey and a from a living conditions survey, together with two corresponding files with aggregated data at the domain level to be used as covariate information in the models.

The content of the book is organized into 21 self-contained chapters. The first six chapters are dedicated to the introduction of the main basic estimators and models, the following nine chapters to models focusing on SAE unit-level models, and the last six chapters to SAE area-level models.

Specifically, Chapter 1 *Small Area Estimation* introduces the basic elements of SAE and mixed models and describes the data files used in the examples with R. Chapters 2 *Design-Based Direct Estimation* and 3 *Design-Based Indirect Estimation* introduce the most popular design-based estimators of domain means and totals and describe some resampling procedures for estimating their mean squared errors. Chapter 4 *Prediction Theory* introduces the prediction theory in finite populations and the Best Linear Unbiased Predictor (BLUP) of a linear parameter and the corresponding prediction variance. Chapter 5 *Linear Models* and Chapter 6 *Linear Mixed Models* present these models in the framework of SAE.

Chapter 7 *Nested Error Regression Models* introduces the basic unit-level model in SAE, the nested error regression model. The EBLUP of domain linear parameters under this model is derived in Chapter 8 *EBLUPs Under Nested Error Regression Models*, while the corresponding MSE in Chapter 9 *Mean Squared Error of EBLUPs*. Chapter 10 *EBPs Under Nested Error Regression Models* refers to the estimation of non linear parameters (e.g. poverty indicators as the poverty incidence), introducing the Empirical Best Predictors (EBPs) and the corresponding MSE parametric bootstrap estimators. Chapter 11 *EBLUPs Under Two-Fold Nested Error Regression Models* and Chapter 12 *EBPs Under Two-Fold Nested Error Regression Models* are dedicated to the two-fold nested error regression model, taking into account the variability between domains and between subdomains inside each domain. Chapter 13 *Random Regression Coefficient Models* introduces random regression coefficient models that can be used when a more flexible model specification is needed. Finally, Chapter 14 *EBPs Under Unit-Level Logit Mixed Models* and Chapter 15 *EBPs Under Unit-Level Two-Fold Logit Mixed Models* introduce two specific unit-level mixed models belonging to the class of generalized linear mixed models.

Chapter 16 *Fay-Herriot Models* introduces the basic area-level model for SAE, with the EBLUPs of domain means and the corresponding MSE estimators. Chapter 17 *Area-Level Temporal Linear Mixed Models* and Chapter 18 *Area-Level Spatio-Temporal Linear Mixed Models* generalize the basic area-level model by considering structures of temporal and/or spatial correlation. Further extensions are introduced in Chapter 19 *Area-Level Bivariate Linear Mixed Models*, Chapter 20 *Area-Level Poisson Mixed Models* and Chapter 21 *Area-Level Temporal Poisson Mixed Models*.

In short, this book is accessible to a wide audience, including students, survey statisticians, practitioners and researchers interested in the use of SAE.

The Survey Statistician, 2024, Vol. 89, 76–85.

Book and Software Review

SAE models for indicators in the unit interval: The tipsae R package and its Shiny interface

Silvia De Nicolò[1] **and Aldo Gardini**[2]

[1]University of Bologna, Italy, silvia.denicolo@unibo.it
[2]University of Bologna, Italy, aldo.gardini@unibo.it

Abstract

The **tipsae** R package is a dedicated tool for mapping proportions and indicators defined on the unit interval, e.g., common poverty and inequality measures, in the framework of small area estimation. It implements Beta-based Bayesian Hierarchical models at the area-level through Stan language. A set of diagnostics, exploratory analysis and complementary tools, such as benchmarking and variance smoothing methods, complete the package. To enhance accessibility for practitioners, we have developed a user-friendly Shiny application alongside the R package. The purpose of this paper is to illustrate its potential by describing the workflow of a typical small area analysis and by providing the underlying R code. The current version 0.0.18 is avalaible on https://cran.r-project.org/web/packages/tipsae/index.html

Keywords: Bayesian Inference, Beta Regression Models, Small Area Estimation, Shiny, Stan.

1 Introduction and software review

Timely and reliable statistical estimates defined for granular geographical levels or specific socio-demographic groups are increasingly in demand. Many of those require an extensive exploitation of survey data; nonetheless, domains or areas of study are often different from the ones originally planned in survey designs. This leads to small sample sizes and consequent unreliable estimates. In this context, the Small Area Estimation (SAE) framework provides a set of indirect estimation techniques that relies on external information, borrowing strength across areas and producing estimates of interest with an acceptable level of uncertainty.

The SAE estimators that are based on regression models are named model-based estimators. The small area models are divided into two strands: the models defined at the area level and the ones defined at the unit level. The first strategy relates area-specific survey estimators to areal covariates, while the second one ties individual observations of the underlying variable to individual covariates. We focus on area-level models as they prove to be highly convenient in practice. Indeed, they only require covariates defined at the area level and account for design-based properties; whereas unit-level models generally need auxiliary information available for the entire population and do not consider survey weights.

We focus on indicators defined on the unit interval, being common in SAE modelling because of the high presence of rates and proportions releases in official statistics. Two different bodies of literature relate with this field in the area-level context, revolving around linear mixed models with suitable transformations (Rao and Molina, 2015) and Beta regression models (Janicki, 2020). An additional strand, common in disease mapping, explicitly models counts via Binomial or Poisson models (Hobza et al., 2018; Wakefield, 2007; MacNab, 2003).

Several R packages implement SAE tools. By focusing on model-based methods, the most complete packages are **sae** (Molina and Marhuenda, 2015), **emdi** (Kreutzmann et al., 2019) and **mcmcsae** (Boonstra, 2021). The **sae** and **emdi** packages implements both area-level and unit-level models from a frequentist perspective. The latter one provides suitable model diagnostics, plots, and exporting tools. Conversely, the **mcmcsae** package implements SAE models in a Bayesian framework via Markov chain Monte Carlo (MCMC) simulation. Such package considers both area and unit level models, allowing also for spatial and temporal dependencies. It includes different prior settings, model diagnostics, and posterior predictive checks functions. Lastly, count models are implemented in the **zipsae** package (Utomo and Wulansari, 2021) with zero-inflated poisson models and **saeeb** package (Fauziah and Wulansari, 2020) with poisson-gamma models.

Among those listed, only the **emdi** package directly accounts for unit interval responses by implementing the Fay-Herriot model with arc-sin transformation (Schmid et al., 2017). In this context, small area models based on the Beta regression lack of a proper implementation and the **tipsae** package fills this gap (De Nicolò and Gardini, 2022). The package, available on CRAN, implements area-level models comprising the standard Beta regression model, Zero and/or One Inflated Beta (Wieczorek et al., 2012) and Flexible Beta (De Nicolò et al., 2023) models. Moreover, spatial, temporal and spatio-temporal dependence structures can be included. Note that Beta mixed regression models are already implemented in R through specific packages that do not accommodate peculiarities of small area models, such as the assumption of known dispersion parameter and popular structured random effects.

The Bayesian inferential framework, implemented through the Stan routine (Carpenter et al., 2017), allows to manage non-conventional likelihood assumptions and capture the uncertainty on target parameters through posterior inference. It assists the user in carrying out a complete SAE analysis, from data loading step, the implemented functions contemplate the entire process of data exploration, model estimation and validation, presentation and exportation of results. This facilitate the use Bayesian models and complex SAE methods for practitioners, building bridges between methodological and applied fields. The package comes equipped with a Shiny application to further facilitate the workflow for non-expert users.

The paper is organized as follows. Section 2 briefly illustrates the **tipsae** package and its main features, describing the various inferential settings that can be implemented. Section 3 outlines the workflow of a SAE procedure using the R Shiny application and then presents the corresponding R code. Concluding remarks are drawn in Section 4.

2 tipsae in a nutshell

The main feature of the **tipsae** package is that it includes a variety of area-level models based on the Beta likelihood with different prior settings to be defined by the users. The inclusion of several methodological tools facilitates the customization and suitability of the model given different practical situations. The statistical methods implemented in the **tipsae** package can be estimated using the function `fit_sae()` and are briefly summarized in the following.

The standard Beta-regression model is useful to handle responses with double-bounded support. It

easily manages different distributional shapes while having a simple location-dispersion parametrization. However, its support does not include 0 and 1 values which may be observed in some applications. A model able to encompass them is required, thus, the package includes also the Zero and/or One Inflated Beta (ZOIB) model (Wieczorek et al., 2012). Lastly, when the distribution of the response is characterized by heavy tails and/or high skewness, the standard Beta regression could fail (Bayes et al., 2012; Migliorati et al., 2018). In this case, the package includes the more suitable Flexible Beta model. It is defined as a mixture of two Beta random variables and is proposed in small area estimation by De Nicolò et al. (2023).

To facilitate practitioners, standard wide-range prior distributions are assumed for model parameters. The priors for the random effects include the case of unstructured random effects, spatially structured random effects, and temporal random effects. The unstructured random effect accounts for area-specific deviations from the synthetic predictor. The package includes three different strategies to specify its prior distribution, that can be chosen through the `prior_reff` argument of `fit_sae()` function. As a first option, a zero-mean normal prior is considered (the default option, namely `prior_reff ="normal"`). When covariates have poor explanatory power in some domains, a more flexible handling of random effect is required. As a second option, the package implements a robust Student's t prior with exponential hyperprior for degrees of freedom (`prior_reff ="t"` option, Fabrizi and Trivisano, 2016). The third option contemplates the presence of very informative covariates, in that case the variability of the small area parameters may not require the inclusion of a random effect term (Datta et al., 2011). Therefore, the variance gamma shrinkage prior is included as prior choice (`prior_reff ="VG"` option, Fabrizi et al., 2018). In all the cases, the scale (or global scale) parameter has an half-normal hyperprior whose scale can be set using the `scale_prior` argument.

By setting the argument `spatial_error = TRUE`, the user can specify a spatially structured effect to the linear predictor, in addition to the unstructured one. The vector of spatial random effect has an intrinsic conditional autoregressive (ICAR) prior (Besag et al., 1991). The spatial structure of the areas, in the form of an object of class `'SpatialPolygonsDataFrame'` (from the **sp** package, Bivand et al., 2013) or class `'sf'` (from the **sf** package, Pebesma, 2018), is required as input of the `spatial_df` argument. When multiple observations of the target indicator are observed for different time periods, a temporal model can be specified in order to borrow strength from time repetitions. The user may choose to add a temporal random effect in the linear predictor in addition to the unstructured one by setting `temporal_error = TRUE` and declare the name of the temporal variable in `temporal_variable`. Its prior is a random walk prior of order 1, assuming independence among the areas. If both temporal and spatial error argument are set to `TRUE`, a spatio-temporal model is fitted.

Posterior inference is carried out through an efficient Hamiltonian Monte Carlo (HMC) fitting algorithm and parallel computing imported from **rstan** (Stan Development Team, 2020). The package provides out-of-sample predictions available through the `export()` function. In practice, when an area is out-of-sample but its auxiliary information is observed, its direct estimate can be included in the dataset labeled as `NA`. In such a way, the function automatically draws a sample from the posterior distribution of the predictor by combining the samples drawn from the posterior of all the involved parameters.

One of the output of the `fit_sae` function is the `'stanfit' S4` object produced by the **rstan** package. This can be exploited to check convergence, monitor sampler diagnostics, and, lastly, perform an exhaustive posterior analysis by relying on existing tools, e.g. **loo** and **bayesplot** packages (Gabry et al., 2019). Moreover, specific diagnostics for small area models are produced by ad-hoc functions, facing the most relevant aspects to deepen within a SAE framework. The package supplies both visualization tool for graphical assessments and functions that easily export the final results. Lastly, variance smoothing routines for pre-processing and benchmarking procedures for post-processing are provided as complementary tools. A Shiny application (Chang et al., 2021) with an intuitive

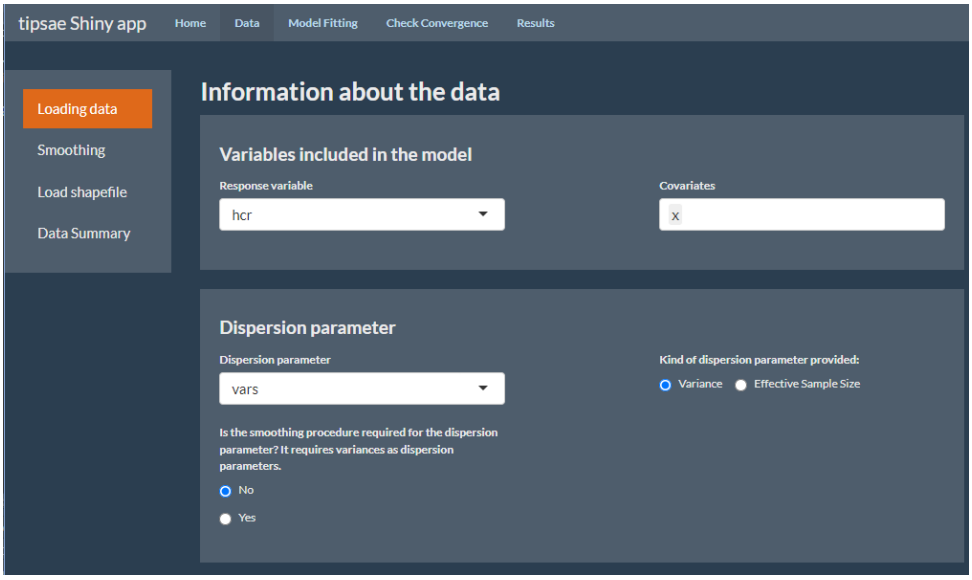


Figure 1: Focus on the Data tab: overview on the section where it is possible to load data and specify the interesting variables.

graphical user interface can be launched through `runShiny_tipsae()`.

3 Workflow of a SAE procedure: R code and R Shiny

The objective of this section is to illustrate the value of the Shiny application as a powerful tool for encouraging practitioners to embrace advanced statistical techniques. Specifically, we compare the input and output features of the Shiny app with those obtained using the R language. The application can be easily accessed by executing the following commands[1].

```
R> library(tipsae)
R> runShiny_tipsae()
```

When the application is launched, a browser window appears, providing users with easy navigation through the interface. The interface is thoughtfully organized into five main sections. The **Home** page features a schematic description of the application along with pertinent references. Within this page, the 'Load dataset' button facilitates testing the application by executing an analysis using the toy dataset `emilia_cs` provided within the package. Clicking this button is equivalent to running the R command:

```
R> data(emilia_cs)
```

3.1 Data input and exploratory anlysis

The **Data** page is organized into four subsections, with the initial three tabs dedicated to data entry steps and pre-treatment, while the last tab offers graphical exploratory tools. In the *Loading Data* tab, users can upload a CSV file and visualize the loaded data by clicking the "View loaded dataset" button. Essential details about the data must be provided, such as the nature of inserted variables, labeling the response, covariates, and specifying dispersion values, indicating whether it is variance or effective sample size (refer to Figure 1). This tab also allows users to configure smoothing procedures if necessary. Additional information, including a possible time variable, domain ids, and sample sizes, can be specified for subsequent visual diagnostics. When a smoothing procedure is

[1]The following R packages should be installed before running the commands: tipsae, shinythemes, shinyFeedback, shinybusy, shinyWidgets, leaflet, shinyjs, sp, bayesplot.

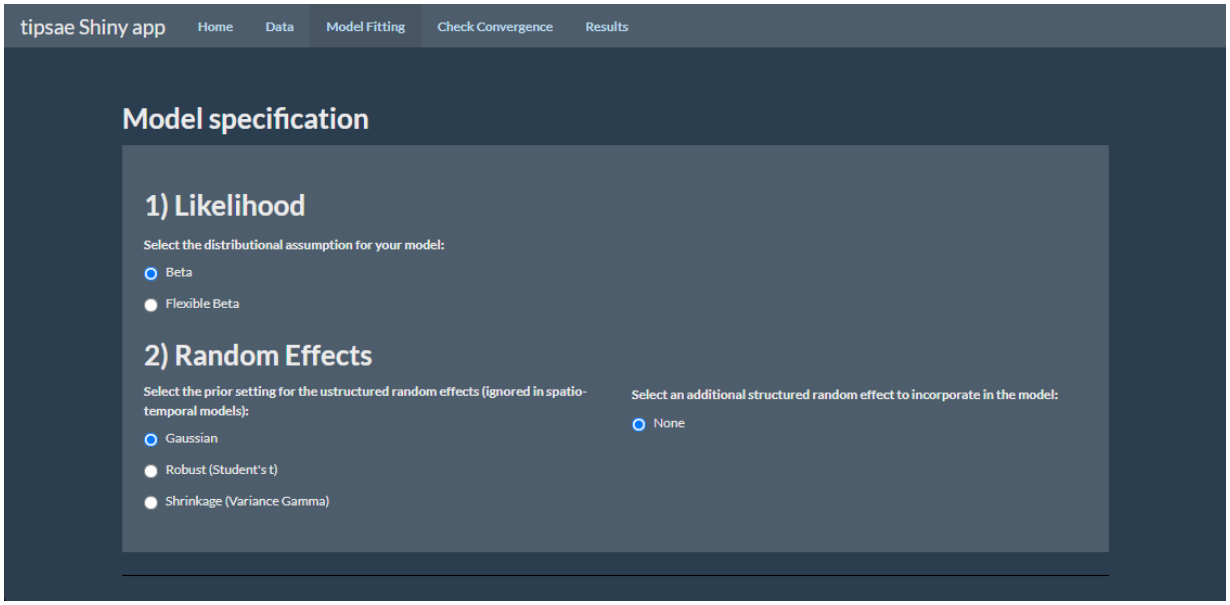


Figure 2: Model Fitting Section of **tipsae** Shiny App

configured, the *Smoothing* subsection permits users to adjust settings and visually inspect the procedure output. The implemented procedure is a Generalized Variance Function smoothing technique (Wolter, 2007, Ch.7); for a detailed methodological explanation, please consult the package vignette. In subsection *Load Shapefile*, a spatial structure can be incorporated, loading either a SHP file either an RDS file containing a `'SpatialPolygonsDataFrame'` or `'sf'` object. Such object would enable to account for spatial dependencies in the model and/or plot maps with relevant quantities. The last subsection, called *Data Summary*, provides an accurate data exploration before moving to the modelling step, depicting the distribution of the response variable and its relationship with the covariates and the dispersion measure.

3.2 Model fitting

The **Model Fitting** section employs the Stan routine for the estimation of the Bayesian models, as detailed in Carpenter et al. (2017). The *Model Specification* tab allows users to define the model likelihood, while the application automatically constraints the model choice depending on the input data. Specifically, when all response observations $y_i \in (0, 1)$, $\forall i$, the allowed options are Beta and Flexible Beta models. If some observations are recorded as zeros and/or ones, the application automatically limits the choice to Zero and/or One Inflated Beta models. Furthermore, users have the flexibility to choose their preferred prior distribution for random effects from the options listed in Section 2. If a shape file is loaded, users can also decide whether to include a spatial random effect in the predictor. It is noteworthy that when analyzing a panel dataset with an available temporal variable, the application automatically incorporates a temporal random effect.

Users can customize certain algorithm settings in the *Settings about the MCMC Algorithm* tab. This includes parameters such as the number of iterations per chain, including warm-up (set as half of the total iterations), enabling parallel computation with the appropriate tick, determining the number of chains, and specifying the number of cores. Additional HMC options include the maximum allowed tree depth (Maximum treedepth) and the target average proposal acceptance probability (adapt_delta). For a more in-depth understanding, users can refer to the Stan documentation. The ”Fit Model” button initiates the estimation process, whose progress is displayed through an iterative printed output.

Note that the model, as configured in Figure 2 and guided by the data input decisions in Figure 1,

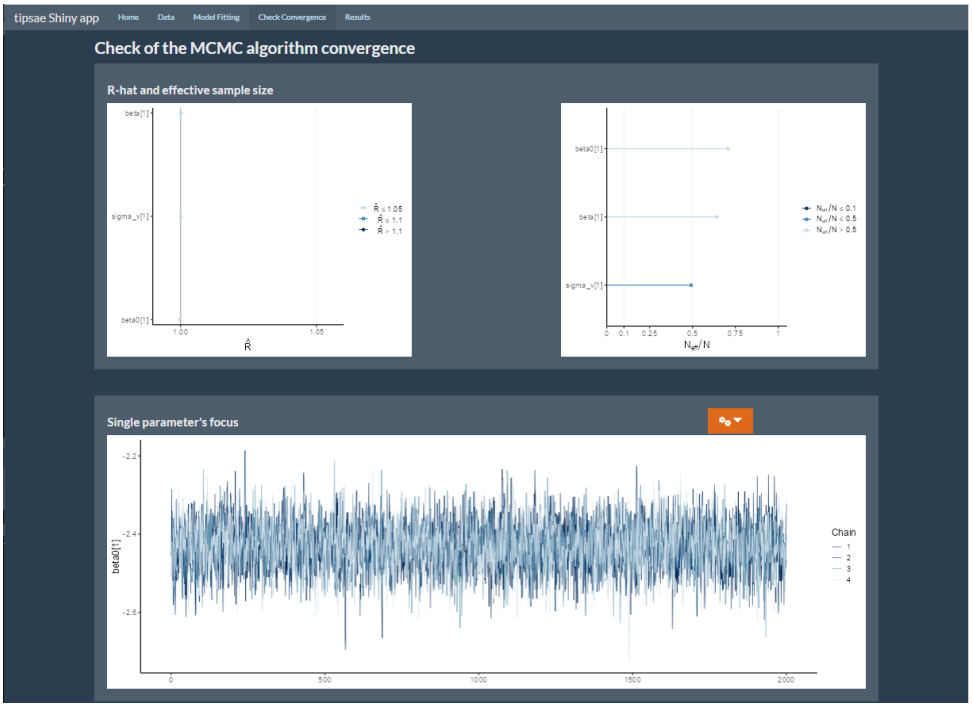


Figure 3: Check convergence section of the Shiny app.

can also be implemented using the following lines of R code. This code produces an object with the class `'fitsae'`.

```
R> fit_beta <- fit_sae(formula_fixed = hcr ~ x, data = emilia_cs,
+                      domains = "id", type_disp = "var",
+                      disp_direct = "vars", domain_size = "n")
```

3.3 Analysing posterior results

After fitting a Bayesian model through a MCMC methods, the proper convergence of the algorithm needs to be assessed before moving to the analysis of the results. This can be done by relying on the useful tool provided by the `bayesplot` package (Gabry and Mahr, 2021). Such functionalities are exploited in the **Check Convergence** section of the Shiny application, where posterior densities, chains trace-plots, autocorrelation functions and rank plots are displayed (Figure 3).

The same plots can be obtained through R code, recalling that the `'stanfit'` objects is contained by the result of the `fit_sae` function.

```
R> library(bayesplot)
R> mcmc_rhat(rhat(fit_beta$stanfit, pars = c("beta0", "beta", "sigma_v")))
R> mcmc_neff(neff_ratio(fit_beta$stanfit, pars = c("beta0", "beta", "sigma_v")))
R> mcmc_trace(as.array(fit_beta$stanfit, pars = c("beta0")))
```

When the convergence of the MCMC algorithm is achieved, the user can move to the evaluation of the model results, accessing the **Results** section of the application. The *Model Summaries* subsection provides posterior syntheses of regression coefficients and random effects. Furthermore, a summary of residuals is also reported, including an histogram, and the LOO Information Criterion can be computed through the button "Click to compute LOOIC", enabling for model selection (Vehtari et al., 2017). Such summary measures can be explored also in R by printing the result of the `summary` method applied to a `'fitsae'` object:

```
R> summ_model<-summary(fit_beta)
```

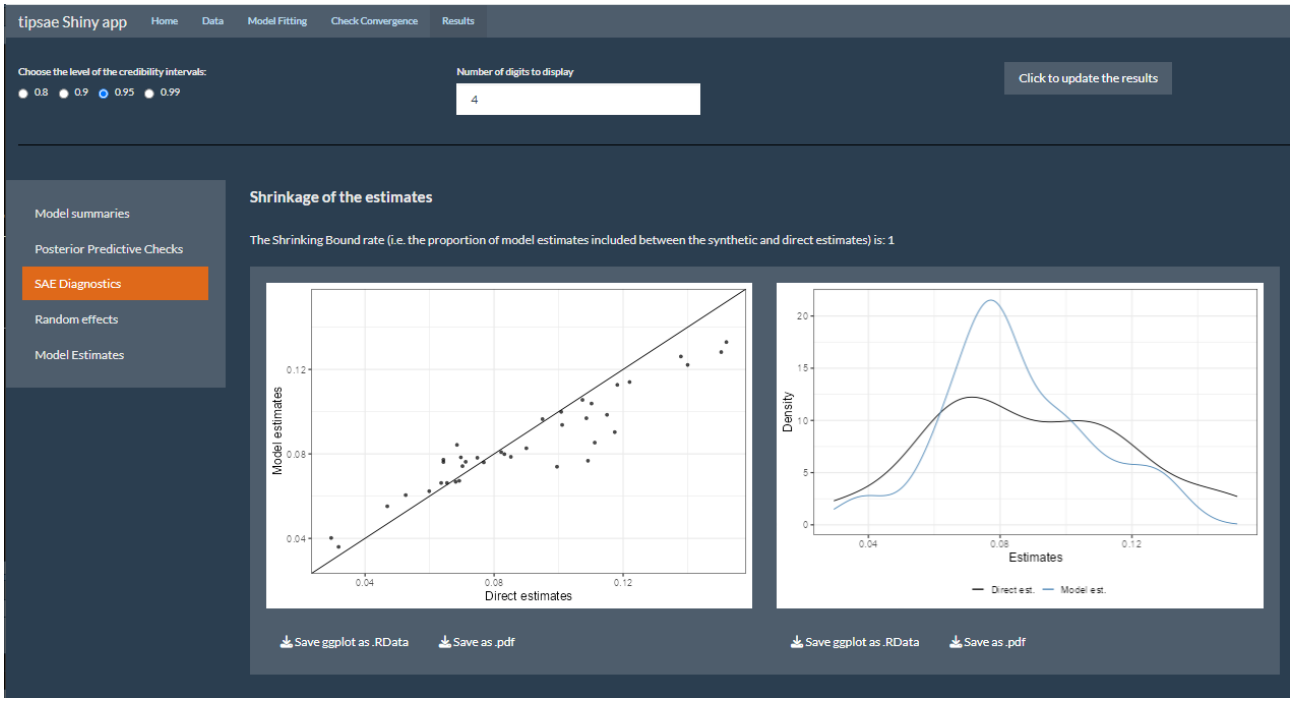


Figure 4: Results section of the Shiny application. A focus on the available SAE diagnostics.

```
R> print(summ_model)
```

The *Posterior Predictive Check* subsection displays the sample data kernel density versus those of the datasets generated from the posterior predictive distribution, denoted with $Y_d^{\bullet}|\boldsymbol{y},\ d = 1, \ldots, D$, in order to assess the goodness of fit. Here, a specific tab focuses on area-specific Bayesian p-values, defined as $BP_d = \mathbb{P}[Y_d^{\bullet} > y_d|\boldsymbol{y}]$ (Fabrizi et al., 2011). In absence of systematic deviations, the expected Bayesian p-value is 0.5, whereas values near 0 or 1 highlight issues of over-estimation and under-estimation, respectively. Summary statistics of Bayesian p-values are printed also in the output of the `summary` method applied to `'fitsae'` objects. In this section, also visual inspections of posterior predictive checks are shown, being implemented through the functions available in the `bayesplot` package.

Small area specific diagnostics have a proper subsection (*SAE diagnostics*), visually illustrating the shrinking process induced on model-based estimates and comparing direct and model-based estimates (Table 4). The standard deviations of both type of estimates are compared and summaries of a measure of standard deviation reduction are provided. The *Random Effect* subsection compares the standardized effects density versus the one of a standard normal and a caterpillar plot, comparing their posterior distributions for each area. Many basic plots reported in this part of the Shiny application can be obtained through the following R code:

```
R> plot(summ_model)
R> density(summ_model)
```

Lastly, the *Model Estimates* subsection displays a table with direct and model-based estimates, including relevant posterior summaries of target parameters. Such object can be downloaded in CSV format via a proper button (Figure 5). A caterpillar plot of the target parameter posteriors is also provided. To extract the whole set of estimated from the model fitted in R, the following command can be used:

```
R> extract(summ_model)
```

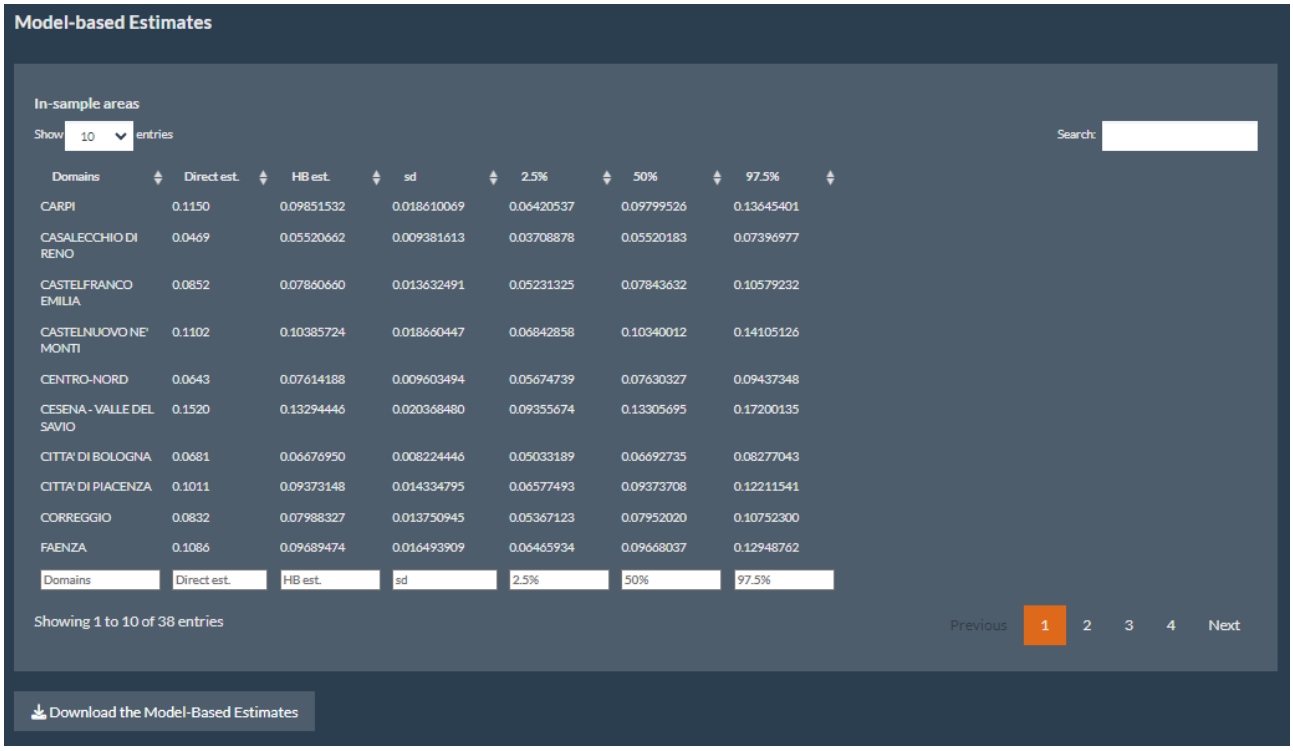


Figure 5: Results section of the Shiny application. A focus on the table with model-based estimates.

4 Concluding remarks

The **tipsae** Shiny application serves as a simple and convenient tool to map indicators on the unit interval in a SAE framework, offering a user-friendly interface for data visualization and analysis. However, it is essential to acknowledge the inherent limitations in terms of flexibility that come with its ease of use. We refer to the use of the **tipsae** R package for a higher level of customization and for additional tools, such as the benchmarking procedure and the estimation of area-specific design effects, not implemented in the Shiny app. For a comprehensive understanding of the package itself and its features, it is recommended to refer to the CRAN vignette [2], which provides detailed information on what sets it apart. While recognizing its constraints, we encourage the survey statistics community to actively engage with Shiny applications. Its potential for fostering collaboration and enhancing data communication is noteworthy.

References

C. L. Bayes, J. L. Bazán, and C. García. A new robust regression model for proportions. *Bayesian Analysis*, 7(4):841–866, 2012.

J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, 1991.

R. S. Bivand, E. Pebesma, and V. Gomez-Rubio. *Applied Spatial Data Analysis with R, Second edition*. Springer-Verlag, 2013. URL `https://asdar-book.org`.

H. J. Boonstra. *mcmcsae: Markov Chain Monte Carlo Small Area Estimation*, 2021. URL `https://CRAN.R-project.org/package=mcmcsae`. R package version 0.7.0.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo,

[2] see https://cran.r-project.org/web/packages/tipsae/index.html

P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

W. Chang, J. Cheng, J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. shiny: Web application framework for r, 2021. URL `https://CRAN.R-project.org/package=shiny`. R package version 1.6.0.

G. S. Datta, P. Hall, and A. Mandal. Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106(493):362–374, 2011.

S. De Nicolò and A. Gardini. *tipsae: Tools for Handling Indices and Proportions in Small Area Estimation*, 2022. URL `https://CRAN.R-project.org/package=tipsae`. R package version 0.0.4.

S. De Nicolò, M. R. Ferrante, and S. Pacei. Small area estimation of inequality measures using mixtures of Beta. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page NA, 2023. ISSN 0964-1998. URL `https://doi.org/10.1093/jrsssa/qnad083`.

E. Fabrizi and C. Trivisano. Small area estimation of the gini concentration coefficient. *Computational Statistics & Data Analysis*, 99:223–234, 2016.

E. Fabrizi, M. R. Ferrante, S. Pacei, and C. Trivisano. Hierarchical bayes multivariate estimation of poverty rates based on increasing thresholds for small domains. *Computational Statistics & Data Analysis*, 55(4):1736–1747, 2011.

E. Fabrizi, M. R. Ferrante, and C. Trivisano. Bayesian small area estimation for skewed business survey variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4): 861–879, 2018.

R. A. Fauziah and I. Y. Wulansari. Package 'saeeb'. 2020.

J. Gabry and T. Mahr. *bayesplot: Plotting for Bayesian Models*, 2021. URL `https://mc-stan.org/bayesplot`. R package version 1.8.1.

J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society A (Statistics in Society)*, 182(2):389–402, 2019.

T. Hobza, D. Morales, and L. Santamaría. Small area estimation of poverty proportions under unit-level temporal binomial-logit mixed models. *Test*, 27:270–294, 2018.

R. Janicki. Properties of the beta regression model for small area estimation of proportions and application to estimation of poverty rates. *Communications in Statistics-Theory and Methods*, 49 (9):2264–2284, 2020.

A.-K. Kreutzmann, S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis. The R package emdi for estimating and mapping regionally disaggregated indicators. *Journal of Statistical Software*, 91(7):1–33, 2019. doi: $10.18637/\text{jss.v091.i07}$.

Y. C. MacNab. Hierarchical bayesian spatial modelling of small-area rates of non-rare disease. *Statistics in Medicine*, 22(10):1761–1773, 2003.

S. Migliorati, A. M. Di Brisco, and A. Ongaro. A New Regression Model for Bounded Responses. *Bayesian Analysis*, 13(3):845–872, 2018.

I. Molina and Y. Marhuenda. sae: An R package for small area estimation. *The R Journal*, 7(1):81, 2015.

E. Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10 (1):439–446, 2018. doi: 10.32614/RJ-2018-009. URL https://doi.org/10.32614/RJ-2018-009.

J. N. Rao and I. Molina. *Small-Area Estimation*. Wiley Series in Survey Methodology, 2015.

T. Schmid, F. Bruckschen, N. Salvati, and T. Zbiranski. Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in senegal. *Journal of the Royal Statistical Society A (Statistics in Society)*, 180(4):1163–1190, 2017.

Stan Development Team. *RStan: The R Interface to Stan*, 2020. URL http://mc-stan.org. R package version 2.21.2.

F. W. Utomo and I. Y. Wulansari. Package 'zipsae', 2021.

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, 2017.

J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

J. Wieczorek, C. Nugent, and S. Hawala. A Bayesian zero-one inflated beta model for small area shrinkage estimation. In *Proceedings of the 2012 Joint Statistical Meetings, American Statistical Association, Alexandria, VA*, 2012.

K. M. Wolter. *Introduction to variance estimation*, volume 53. Springer, 2007.

ARGENTINA

Reporting: **Verónica Beritich**

System of Environmental and Economic Accounting of Argentina: INDEC began its development and implementation

The National Institute of Statistics and Censuses (INDEC) presented the working document "Towards the construction of a system of environmental and economic accounts", which describes for the first time the roadmap for the development and implementation of the environmental domain in the official production of statistics of Argentina, within the framework of the 2021-2026 Strategic Plan. At the same time, a new web section is inaugurated (indec.gob.ar > Statistics > Territory > Environment > Environmental Statistics) that contains the compilation of the 78 environmental indicators that are already regularly produced by the organizations of the National Statistical System. This preliminary inventory not only provides a more complete and multipurpose view of the national environmental landscape, but also contributes to the monitoring of the United Nations Sustainable Development Goals (SDGs).

With this approach, the Institute began to walk along the path for the development and implementation of the System of Environmental and Economic Accounting (SEEA), which integrates the concepts and principles of the System of National Accounts (SNA). The construction of the SEEA constitutes a work proposal in different stages, which must be deepened and operationalized. In this first phase of this initiative, the objective of INDEC was to compile, reorder and systematize statistics that were already available in various publications, systems and portals, following the structure proposed by the Framework for the Development of Environmental Statistics (FDES), which establishes the Basic Set of Environmental Statistics organized into three levels.

The work, which had technical support from the World Bank, was prepared to begin the joint tasks that will be carried out by the organizations that make up the National Statistical System (NSS). The publication contains three large sections in which the global status of environmental statistics is described, the evolution of the international methodological frameworks that will guide the work, and the INDEC´s roadmap to incorporate the environmental domain to official statistics in different time phases. The document also provides details on the result of the initial process of collecting environmental statistics, which led to the publication of the first preliminary inventory on the INDEC website. From now on, it will be possible to consult and download from one site the statistical tables produced by all NSS organizations in .xlsx format, organized according to each component of the FDES.

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

CANADA

Reporting: **Shirin Roshanafshar** and **Eric Rancourt**

Implementation of Generative AI at Statistics Canada

In the last five years, Statistics Canada has progressively integrated Artificial Intelligence (AI) into its operations. The advent of advanced Generative AI (GenAI) technologies has recently ignited increased interest and innovative applications within the agency and other Government of Canada departments. Various experimentations are underway to evaluate the use of Generative AI in the Agency.

More recently, Statistics Canada has made strides in the practical application of GenAI in a responsible manner. A prime example is our development of advanced chatbots using Retrieval Augmented Generation (RAG) technology. We have successfully created an experimental chatbot by providing the model with many long, thoroughly prepared, relevant PDF documents as the knowledge base. This approach not only enhances the accuracy of responses but also minimizes the risk of erroneous information generation. Our chatbot is adept at summarizing and responding to complex queries, always referencing the specific document from the knowledge base from which the information is sourced.

Another area of our experimentation lies in the domain of automated draft generation. We have pioneered the use of GenAI for creating textual summaries from data presented in tables, images, and infographics. Furthermore, we are finetuning models using relevant documents previously translated by our translation experts. With further development, evaluation, and experimentation, such a tool can be used to create first drafts of short summary documents based on data tables, charts and infographics, in two official languages, significantly accelerating the data interpretation and reporting process in both of Canada's official languages.

Concurrently, Statistics Canada is rigorously evaluating and addressing the legal, ethical, and responsible usage of GenAI. This involves creating Generative AI guidelines for use by all employees, specialized GenAI training programs, and fostering communities of practice to ensure a responsible and effective implementation of GenAI technologies. Furthermore, a lot of effort is underway to create an enabling, robust IT infrastructure for rapid prototyping and production of GenAI use cases.

A highly impactful event in Canada: The World Statistics Congress 2023

From July 15th to 20th, 2023 Ottawa, Canada hosted the 64th World Statistics Congress (WSC). It had been 60 years since Canada had first hosted the event and this time again it had marked impact. Hundreds of Canadians participated in the Congress in one way or another. Some were involved in Invited Paper Sessions, some through contributed sessions, others in meetings, events, but all were involved in developing and/or nurturing their international statistical network. Statistics Canada, the Statistical Society of Canada as well as multiple Federal departments, provincial ones, municipal and also private sector organizations participated.

In parallel to the congress, Statistics Canada organized 24 side events covering a wide variety of topics such as Indigenous data, Community Safety, Gender and Diversity, Data Literacy, Consumer Prices, Linkable Data and Agriculture. Further, a good number of in-person meetings with experts and groups (for example the Chief Methodologists Network) met at Statistics Canada during or after the WSC.

Many young statisticians in Canada had been working on research and/or applications that are using methods developed by people who attended the WSC and so it was a unique occasion for them to speak with them in person. The WSC also gave a chance to many university students in statistics

as well as employees from Statistics Canada to contribute their time as a volunteer as well as to have an opportunity to attend sessions.

Many months after the WSC2023, we can still hear people in associations, in committees and other statistical activities referring to the discussion, the decision, the connexion made in Ottawa as the WSC left a clear mark on them. The World Statistics Congress 2023 in Ottawa did have a significant impact!

FIJI

Reporting: **M.G.M. Khan**

Fiji Standard Classification of Occupations (FISCO) Upgrade

The department is working to concord Fiji Classification of Individual Consumption according to Purpose (FCOICOP 2010) to Pacific Classification of Individual Consumption According to Purpose 2020 (PACCOICOP 2020) for the compilation of Prices and Consumption Statistics.

Contact persons: Mr. Tawaketini Autiko tautiko@statsfiji.gov.fj , Ms. Shaista Bi shaistab@statsfiji.gov.fj and Mr. Viliame Raduva vraduva@statsfiji.gov.fj

Trade Database

The International Trade unit is collaborating with the Information Dissemination Division to develop an internal database for International Merchandise Trade Statistics. This database will provide comprehensive time series data on International Merchandise Trade.

The system is expected to improve the storage, management and extraction of Trade statistics. With this database, the Fiji Bureau of Statistics will be able to respond to data requests more efficiently and in a timely manner. Discussions are underway with Asian Development Bank for their collaboration on the project.

Contact persons: Mr. Shonal Deo shonal.deo@statsfiji.gov.fj and Mr. Abdul Sahib asahib@statsfiji.gov.fj

2019 GDP Rebase

Preparatory work is in progress for rebasing the Gross Domestic Product to the year 2019. The 2019 GDP rebased numbers will be published in 2025 after the successful completion of 3-year survey results (2019-2021), 2019 supply and use table; and the rebase of the price indices. All rebases relating to volume indexes and price indices are expected to be finalized by end of 2023.

Contact persons: Mr. Bimlesh Krishna bkrishna@statsfiji.gov.fj and Ms. Artika Devi artikad@statsfiji.gov.fj

Rebase of Indicators and Deflators

The rebasing of indicators and deflators used for Gross Domestic Product estimation mainly the Industrial Production Index, Consumer Price Index, Import & Export Price Index, Producer Price Index and Building Material Price Index are in progress. The department is closely working with the Pacific Community SPC for rebasing the Consumer Price Index to 2019.

Contact persons: Ms. Radhika Kumar radhikak@statsfiji.gov.fj and Mr. Sitiveni Sikivou ssikivou@statsfiji.gov.fj

FINLAND

Reporting: **Maria Valaste** and **Risto Lehtonen**

Report on the BaNoCoSS 2023 Conference in Helsinki

The Baltic-Nordic Conference on Survey Statistics – BaNoCoSS 2023 was organized in August 2023 in Helsinki. The event was of a hybrid type, which in addition to the on-site participation in Helsinki enabled online participation from all over the world. The event was the 6th in the series of scientific conference events of the Baltic-Nordic-Ukrainian BNU Network on Survey Statistics. Previous conferences were organized in 2002 in Ammarnäs (Sweden), 2007 in Kuusamo (Finland), 2011 in Norrfällsviken (Sweden), 2015 in Helsinki (Finland) and 2019 in Örebro (Sweden).

BaNoCoSS-2023 was a scientific conference that presented recent theoretical and practical developments in survey statistics and data science, as well as the interaction of disciplines. As keynote speakers we had Jan van den Brakel of Statistics Netherlands and Maastricht University, Andrew Gelman of Columbia University, Camelia Goga of University of Franche-Comté, Jae Kwang Kim of Iowa State University, and Li-Chun Zhang of University of Southampton; Statistics Norway; University of Oslo. The scientific program featured 16 other invited speakers on interesting topics and more than 30 other great presentations by speakers from 10 different countries. A workshop on Multilevel Regression and Poststratification (MRP) was also held, led by Philipp Christian Broniecki of University of Oslo, Norway.

Abstracts of the presentations are available in "Proceedings of the 6th Baltic-Nordic conference on survey statistics 2023 of the BNU Network on Survey Statistics", published by National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (2023). The publication is freely available via the web site of the BNU network https://wiki.helsinki.fi/display/bnu/events.

The conference was partially funded by the International Association on Survey Statistics (IASS), a section of the International Statistical Institute (ISI). The support was crucial for supporting Ukrainian teacher and students to take part in the event. The event was sponsored by the Statistics Finland and University of Helsinki, via the organizing the onsite event and hosting of the virtual Zoom sessions. Keynote talks and invite sessions were recorded and made available to the participants.

Cooperation in education and research in the field of survey statistics between universities in the Baltic and Nordic countries began in 1992 via the initiative of Professor Gunnar Kulldorff (University of Umeå, Sweden) and has been developed since 1996 as the Baltic-Nordic-Ukrainian Network on Survey Statistics. Today, the network includes partner organizations (universities and national statistical agencies) of seven countries: Estonia, Finland, Latvia, Lithuania, Poland, Sweden, and Ukraine. The network's next event, a workshop, will take place in August 2024 in Poznan, Poland. More information about the BNU network and the other activities of the network can be found on the BNU website at https://wiki.helsinki.fi/display/BNU/.

FRANCE

Reporting: **Danièle Guillemot**

A methodological household survey for studying the mixed-mode collection effects

Household surveys (conducted by official statistical offices or research institutes) do use more and more mixed-mode collection devices, with for example self-administered internet questionnaires along with traditional interviews, by telephone or face-to-face. This evolution, that among other things enables to maintain or improve the participation rates, raises the question of potential biases linked

to the collection mode, particularly for the variables of the common household set (CHS) of household surveys. This CHS contains questions intended for identifying the composition of the households and is used for all INSEE (the French National Statistical Institute) household surveys. When developing mixed-mode collection, it is important for INSEE to ensure that the way of asking questions is robust in regard to the choice of the mode: data provided by the respondents have to be the same, whatever the collection mode.

Between October 2023 and January 2024, INSEE is conducting a methodological survey dedicated to that question. More precisely, this survey concerns the CHS that has to be adapted for all collection modes, as well as some thematic questions on living conditions that have been considered as sensitive to collection modes.

Six sub-samples are being used, five with a "one-mode" collection (self-administered on internet, with responses on smartphone, with responses on a computer, telephone interview conducted by INSEE, telephone interview conducted by an external provider, face-to-face interview), the sixth sub-sample using a sequential mode internet – telephone – face-to-face. The sub-samples have been randomly drawn in the household sampling frame, and the questionnaire is identical regardless of the collection mode; its filling should take around twenty minutes.

If significant differences are noticed in the responses of the sub-samples, it will be possible to identify, with this protocol, what is due to a selection effect (the composition of the respondents may differ between the sub-samples) or to a measure effect (the same person may answer differently when using one mode or another one).

This methodological survey should provide different opportunities for the French statistical authorities:

- Improve and validate the new CHS, so that this CHS is used for all mixed-mode surveys;
- Give a robust measure of potential measure biases coming from the collection modes, on a set of variables used for official statistics or research surveys;
- Depending on observed measure biases, propose action plans (tests, statistical estimates, etc.).

The first challenge, for this survey whose data collection is in progress, is to obtain a response rate sufficient to be able to draw the lessons expected form this methodological operation.

Contact : daniele.guillemot@insee.fr

HONG KONG S.A.R.

Reporting: **Ronald Chan**

Strategies for Application of Data Analytics in Official Statistics

The Census and Statistics Department (C&SD) of Hong Kong, China has been exploring the application of data analytics in official statistics to improve the relevance, cost-effectiveness, quality and timeliness of the statistics. Recognising the demand for better coordination and collaboration within the department in developing and implementing such initiatives, a new Data Analytics Branch (DAB) has been established since April 2023. The key mandate of DAB is to promote and facilitate the use of data analytics in official statistics work by adopting a Hub-and-Spoke Model within the C&SD.

DAB, acting as the central point of coordination (the hub), fosters collaboration and knowledge sharing among different branches (the spokes) in the C&SD on data analytics projects involving professional statisticians and provides support and technical assistance to the spokes. DAB

identifies, promotes and facilitates applications of data analytics in different areas of official statistics work, allowing statistical staff to gain practical experience and enhance their techniques and knowledge. Project owners are invited to participate in different Departmental Theme Groups on Data Analytics (Theme Groups) coordinated by DAB, which facilitate knowledge inheritance and experience transfer among relevant branches. Examples of such Theme Groups include Text Analytics and Web-scraped data. Theme Groups regularly publish reports and methodology papers to promote the literacy of applying data analytics in official statistics within the department. These initiatives are considered conducive to the sustainable development of data analytics in the department.

Furthermore, DAB is responsible for formulating a Departmental Guideline on Data Analytics, which aims to standardise the terminologies and approaches of data analytics within the C&SD. It also ensures the appropriate, ethical and effective use of such techniques in the course of official statistics work.

For more information, please contact Ronald Chan (rchchan@censtatd.gov.hk).

THE NETHERLANDS

Reporting: **I.D.N. (Deirdre) Giesen**

CBS Academy

In a world where developments follow each other in rapid succession, it is important for organizations and their staff to grow with them. To facilitate this, in 2019 Statistics Netherlands (CBS) set up the CBS Academy, and it has rapidly become an integral part of the organization. The Academy aims to be a single point of entry for the learning needs of management and professionals of all levels.

In addition to setting out the HR strategy in which the development of employees was given an important place, an internal sounding board group was populated with members from various divisions of CBS. Its purpose is to engage with employees about new programmes and initiatives, or advise them about all kinds of learning questions from their division. The Academy also consults the divisional directors regularly about the strategic learning objectives and what the focus should be on. An inventory of these things are made and programmes are developed based on that. Furthermore, 80% of decentralized training budget has been transferred to the Academy to make the programmes possible.

For professional skills such as training on Data Science, Statistical Methods and IT, the Academy leans on more than 100 internal colleagues who train as part of their regular work. Its curriculum is geared to support the modernization programs occurring within the statistical divisions. The Academy values modern training methods and offers internal trainers masterclasses in didactic skills.

CBS as a learning organisation

CBS, with its wide variety of staff, has a strong focus on rejuvenation: last year, for example, CBS recruited 200 new employees. The CBS Academy is tasked with supporting them and the other employees. It is important to keep the different target groups connected and to ensure that the youngest generation also ends up in a good place. The young generation no longer chooses the same job for life. In addition, the statistical field is subject to major changes. Data Science is playing an increasingly important role both within the government and in society as a whole. And don't forget the digital developments. To attract and retain people in this difficult market, you want to challenge them and let them grow. The key here is to have a good conversation about their development. CBS focuses on its management to be talent-driven and development-oriented whereby employees are

motivated in their work, receive constructive feedback, learn to reflect with each other and receive coaching when necessary. The organization, supported by the CBS Academy, continuously trains and develops all layers of management. For personal soft skills and management development, the CBS Academy works with a varied range of external training agencies.

Develhub award

In December 2021, the CBS Academy was surprised with the nomination for the Develhub Learning Organization Award. Develhub is a professional organisation for people involved in learning and development within companies, and it presents three awards every year. In the nomination, setting up an Academy from scratch in three years was hugely praised. There was also great appreciation for the development of the learning portal and the fact that everything is evaluated in terms of content, presentation, impact and application.

The CBS Academy will continue this development and support its employees with formal and informal learning interventions in order to maintain professionalism, as well as to continue developing personal skills. CBS will focus more than ever before on its role and task of helping the government to increase the accessibility of data and will develop training and education in the use of data and statistics for our government partners. The CBS Academy therefore aims to collaborate with other educational institutions of government partners, such as the National Academy for Finance and Economics.

Please contact Sandra Duijndam, manager CBS Academy, at sam.duijndam@cbs.nl for more information on the CBS Academy.

NEW ZEALAND

Reporting: **Christine Bycroft**

Harmonising ethnicity from multiple administrative data sources using latent class modelling

Stats NZ has recently published Harmonising ethnicity from multiple administrative data sources using latent class modelling. The paper describes a statistical model for resolving observed differences in ethnicity for individuals when ethnicity is reported across multiple data sources.

Ethnicity is a measure of cultural identity in New Zealand and is a key social factor used to describe the population. Ethnicity is self-perceived and people can belong to more than one ethnic group. Because of its importance, ethnicity is collected by Stats NZ and by several government agencies. The nature of ethnicity means there will be legitimate differences in recorded ethnicity between collections, as well as some level of introduced error.

Stats NZ's Integrated Data Infrastructure (IDI) brings together data from different sectors which reveals differences in recorded ethnicity for individuals. To produce consistent results from the IDI data, a harmonised view of ethnicity is needed. Previous research has looked at rules-based methods for determining the 'best' values of ethnicity when there are conflicting values (Reid et al, 2016; Stats NZ, 2018). Currently the IDI uses a method based on ranking sources by quality. The most recent census is assumed to provide the most reliable measure of ethnicity. Other data sources are ranked on the basis of their agreement rate with the census, and each individual is assigned the ethnic profile from the highest-ranked source available. While this source ranking method generally performs well, it does not consider all the observations available for an individual, whereas statistical modelling has the potential to make use of all available data.

In this paper we develop latent class models to identify the most probable ethnic groups an individual belongs to, based on observations reported across multiple data sources. A key assumption normally applied in latent class analysis is that the modelled latent variable can be interpreted as an underlying

true value, in this case a ‘true’ ethnic status. However, it is a matter of debate among sociologists whether a true underlying ethnic status in fact exists as a concept, given that ethnicity is self-perceived, and the fluidity of reporting that occurs across different contexts and over time. However, despite these limitations, ethnicity is widely used in social science, and there is a need for researchers to be able to apply the same ethnicity indicator across different studies. The ethnicity predicted by the latent class model can be interpreted pragmatically as an ‘ethnicity indicator’ based on finding the most consistent values observed from reporting across sources.

We assess the performance of the modelled values by comparison with responses to the New Zealand 2018 Census of Population and Dwellings, which we assume to be the highest-quality available source. By including 2018 Census in the model as a data source, we were able to confirm our assumption that 2018 Census is the highest quality source of those available, while data from birth registrations performs very well. When comparing aggregate measures for the total population, latent class models and source ranking both compare well with official estimates of ethnic proportions, with some variation by ethnic group. The main benefit of the modelled latent class predictions over the source ranking method is individual-level predicted values which are closer to those reported in the census.

We plan to develop the model further, including investigating the use of covariates such as age and country of birth, and applying it at a more detailed level of the ethnicity classification.

For more information, please contact censustransformation@stats.govt.nz.

References

Bycroft, C, Elleouet, J, & Tran, H (2023). Harmonising ethnicity from multiple administrative data sources using latent class modelling. Retrieved from www.stats.govt.nz.

Reid, G, Bycroft, C & Gleisner, F (2016). Comparison of ethnicity information in administrative data and the census. Retrieved from www.stats.govt.nz.

Stats NZ (2018). Experimental ethnic population estimates from linked administrative data. Retrieved from www.stats.govt.nz.

PERU

Reporting: **Leonor Laguna**

Update from the INEI

(1) The INEI (Instituto Nacional de Estadística e Informática) is using tablets to collect the information for its surveys. The information is sent in real time to the INEI and used to enrich the data base corresponding to that survey. This data base and the corresponding documentation of each survey are accessible to the public through the link:

https://proyectos.inei.gob.pe/microdatos/

(2) The INEI is preparing for the upcoming census programme, CENSOS NACIONALES 2025: XIII DE POBLACIÓN, VIII DE VIVIENDA Y IV DE COMUNIDADES INDÍGENAS, by thoroughly updating all necessary cartographic material.

SWITZERLAND

Reporting: **Michèle Gillard** and **Alina Matei**

The use of artificial intelligence for land use statistics: the ADELE tool

The Federal Statistical Office (FSO) is Switzerland's national centre of competence for official statistics. Among many other statistics, the land use statistics produced by the FSO provide information on land use and land cover in Switzerland and how they change over time, with the aim of ensuring long-term spatial monitoring. Land use statistics are based on a number of sample points and on the interpretation of aerial photographs, traditionally carried out by human experts.

Since 2022, artificial intelligence has been contributing for the first time "in real use" to the production of an official statistic in Switzerland. The deep learning tool ADELE (“Arealstatistik Deep Learning”) has enabled partial automation of the labour-intensive task of interpreting the aerial photographs. While its use offers major advantages in terms of saving human time, there are conflicts between the way artificial intelligence works and the need to produce reliable series of statistics for long-term use. ADELE determines the land use and land cover by means of a statistical model consisting of a large number of parameters set by the tool itself during the training phase of the used deep learning algorithm. The tool fails to achieve the precision required to ensure the consistency of past and future survey results. Despite ongoing improvements, this is unlikely to change in the medium to long term. Some categories simply look too similar, while in others – such as parks – there is a diversity of shapes and colors that the tool struggles to handle. It is therefore necessary to test which categories of land use and cover ADELE can identify with the required accuracy. The published land use statistics can only draw on these parts of ADELE’s output. In addition, the published results incorporate those sample points where the probability (calculated by ADELE itself) of correct identification exceeds a certain threshold. All other sampled points still require human interpretation.

The degree of automation will increase over time as the volume of training data continues to grow with each new survey. As the level of automation increases, further increases in automation become more and more difficult – and therefore more and more expensive. At a certain point, it becomes more efficient to allocate the resources required for further automation directly to human interpretation of the sample points. It is also important to bear in mind that the human handling of the basic data – mass viewing of aerial photos on the screen, discussions with colleagues and the field visits – goes beyond mere classification. Human interpretation also leads to a deeper understanding of the data and the categories used in Switzerland’s land use statistics, and thus to the development of the expertise absolutely needed to interpret the statistical results correctly.

Despite all these shortcomings, it is noted that the ADELE tool is now able to determine land cover and land use in about 27% of the sample points. This means that about 1.1 million points do not require human interpretation, which translates into considerable time savings. ADELE is thus a good example of the new possibilities offered by artificial intelligence in the production of official statistics.

General information about ADELE can be found at

The use of artificial intelligence in the FSO's land use statistics - The ADELE deep learning tool: how it works, capabilities and limitations | Publication | Federal Statistical Office (admin.ch)

Contact: arealstatistik@bfs.admin.ch

UNITED STATES

Reporting: **Andreea L. Erciulescu**

The Joint Program in Survey Methodology Celebrated its 30th Anniversary

The Joint Program in Survey Methodology was founded in 1993 with the goal to educate the next generation of survey statisticians and survey methodologists. Ever since its inception, the program has been a collaboration between two academic institutions (University of Maryland, College Park, and University of Michigan) and one research firm (Westat). On November 2nd, 2023, it celebrated its 30th anniversary, acknowledging more than 270 program graduates working in academia, government, and private sector. In particular, the capabilities of the U.S. federal statistical workforce have been substantially enhanced due to this program.

.

Conferences on survey statistics and related areas

SAE 2023-2024 Conference – Small Area Estimation Survey & Data Science. A conference celebrating the 65th birthday of Prof. Partha Lahiri

SAE 2023-2024 will be held on June 3-7, 2024 at Pontificia Universidad Católica del Perú. The deadline for invited paper session is March 15, 2024. The deadline of abstract submission is March 15, 2024.

More information: https://sae2023.pucp.edu.pe/ .

Sampling Program for Survey Statisticians (SPSS)

Survey Michigan | Summer Institute in Survey Research Techniques

Date: 02 June 2024 - 26 July 2024

Location: Michigan

More information: https://si.isr.umich.edu/course-offerings/sampling-program-for-survey-statisticians/

15th International Congress on Mathematical Education
7-14 July 2024 • ICC Sydney, Australia

The 15th International Congress on Mathematics Education (ICME-15) will be held on **7-14 July 2024** at **International Convention Centre in Sydney, Australia**.

Please visit Congress website for more information: https://icme15.com

Workshop on Survey Statistics

The Workshop on Survey Statistics organized by the Baltic-Nordic-Ukrainan Network on Survey Statistics under the title **Data Integration and Population Size Estimation** will be held in Poznań, Poland, in **August, 26-30, 2024**.

More information: https://wiki-emerita.it.helsinki.fi/display/BNU/Events

IAOS-ISI 2024, Mexico City
Improving Decision-Making for All

May 15, 2024 - May 17, 2024

Mexico's National Institute of Statistics and Geography (INEGI), the International Association of Official Statistics (IAOS) and the International Statistical Institute (ISI) welcome you to the website for the 19th IAOS Conference: https://www.isi-next.org/conferences/iaos-isi-2024/

ISI WSC 2025

65th ISI World Statistics Congress 2025 will be held in The Hague, The Netherlands

July 13, 2025 – July 17, 2025

More information: https://www.isi-next.org/conferences/wsc2025/

Special Call: *Invited Paper Sessions* (IPS) is Open!

In Other Journals

Journal of Survey Statistics and Methodology

Volume 11, Issue 3, June 2023

Special Issue: Recent Advances in Data Integration

https://academic.oup.com/jssam/issue/11/3

Introduction

Recent Advances in Data Integration
Joseph W. Sakshaug and Rebecca C. Steorts

Survey Methodology

Experiments on Multiple Requests for Consent to Data Linkage in Surveys
Sandra Walzenbach, Jonathan Burton, Mick P. Couper, Thomas F. Crossley, and Annette Jackle

Augmenting Survey Data with Digital Trace Data: Is There a Threat to Panel Retention?
Mark Trappmann, Georg-Christoph Haas, Sonja Malich, Florian Keusch, Sebastian Bahr, Frauke Kreuter, and Stefan Schwarz

Survey Statistics

A Primer on the Data Cleaning Pipeline
Rebecca C. Steorts

Bayesian Graphical Entity Resolution using Exchangeable Random Partition Priors
Neil G. Marchant, Benjamin I. P. Rubinstein, and Rebecca C. Steorts

Implicates as Instrumental Variables: An Approach for Estimation and Inference with Probabilistically Matched Data
Dhiren Patki and Matthew D. Shapiro

Improving Statistical Matching when Auxiliary Information is Available
Angelo Moretti and Natalie Shlomo

Evaluating Data Fusion Methods to Improve Income Modeling
Jana Emmenegger, Ralf Munnich, and Jannik Schaller

Applications

Integrating Administrative and Survey Data to Estimate WIC Eligibility and Access
Linden McBride, Thomas B. Foster, Renuka Bhaskar, Mark Prell, Maria Perez-Patron, Erik Vickstrom, Brian Knop, and Michaela Dillon

Constructing State and National Estimates of Vaccination Rates from Immunization Information Systems
Trivellore Raghunathan, Karen Kirtland, Ji Li, Kevin White, Bhavini Murthy, Xia Michelle Lin, Latreace Harris, Lynn Gibbs-Scharf, and Elizabeth Zell

Combining National Surveys with Composite Calibration to Improve the Precision of Estimates from the United Kingdom's Living Costs and Food Survey
Takis Merkouris, Paul A. Smith, and Andy Fallows

Volume 11, Issue 4, September 2023

https://academic.oup.com/jssam/issue/11/4

Survey Methodology

Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA)
Andrew B. Seidenberg, Richard P. Moser, and Brady T. West

The Precision of Estimates of Nonresponse Bias in Means
Stephanie Eckman, Jennifer Unangst, Jill A. Dever, and Christopher Antoun

An Experimental Evaluation of Alternative Methods for Case Prioritization in Responsive Survey Design
Brady T. West, Wen Chang, and Alexa Zmich

Simulating the Consequences of Adaptive Survey Design in Two Household Panel Studies
Nicole Watson and Alexandru Cernat

Survey Statistics

A Comprehensive Overview of Unit-Level Modeling of Survey Data for Small Area Estimation Under Informative Sampling
Paul A. Parker, Ryan Janicki, and Scott H. Holan

Comparison of Unit-Level Small Area Estimation Modeling Approaches for Survey Data Under Informative Sampling
Paul A. Parker, Ryan Janicki, and Scott H. Holan

Empirical Best Prediction of Small Area Means Based on a Unit-Level Gamma-Poisson Model
Emily Berg

Bayesian Nonparametric Joint Model for Domain Point Estimates and Variances Under Biased Observed Variances
Terrance Dean Savitsky and Julie Gershunskaya

Item Count Technique with a Continuous or Count Control Variable for Analyzing Sensitive Questions in Surveys
Barbara Kowalczyk, Wojciech Niemiro, and Robert Wieczorkowski

Applications

Accuracy of Estimated Ratios as Affected by Dynamic Classification Errors
Arnout van Delden, Sander Scholtus, Joep Burger, and Quinten Meertens

Joinpoint Regression Methods of Aggregate Outcomes for Complex Survey Data
Benmei Liu, Hyune-Ju Kim, Eric J. Feuer, and Barry I. Graubard

Volume 11, Issue 5, September 2023

https://academic.oup.com/jssam/issue/11/5

Survey Methodology

Statistics Canada Statistique Canada

Survey Methodology, December 2023, vol. 49, no.1

Special paper in memory of Professor Chris Skinner – Winner of the 2019 Waksberg Award

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2023001-eng.htm

Survey Methodology, December 2023, vol. 49, no.2

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2023002-eng.htm

Small area prediction of general small area parameters for unit-level count data
by Emily Berg

A method for estimating the effect of classification errors on statistics for two domains
by Yanzhe Li, Sander Scholtus and Arnout van Delden

Model-based stratification of payment populations in Medicare integrity investigations
by Don Edwards, Piaomu Liu and Alexandria Delage

Journal of Official Statistics

Volume 39 (2023): Issue 3 (September 2023)

https://sciendo.com/issue/jos/39/3

Letter to Editor Quality of 2017 Population Census of Pakistan by Age and Sex
Asif Wazir and Anne Goujon

Looking for a New Approach to Measuring the Spatial Concentration of the Human Population
Federico Benassi, Massimo Mucciardi and Giovanni Pirrotta

Predicting Days to Respondent Contact in Cross-Sectional Surveys Using a Bayesian Approach
Stephanie Coffey and Michael R. Elliott

Towards Demand-Driven On-The-Fly Statistics
Tjalling Gelsema and Guido van den Heuvel

Database Reconstruction Is Not So Easy and Is Different from Reidentification
Krishnamurty Muralidhar and Josep Domingo-Ferrer

Comment to Muralidhar and Domingo-Ferrer (2023) – Legacy Statistical Disclosure Limitation Techniques Were Not An Option for the 2020 US Census of Population And Housing
Simson Garfinkel

A Rejoinder to Garfinkel (2023) – Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial US Census: Still a Viable Option
Krishnamurty Muralidhar and Josep Domingo-Ferrer

A Note on the Optimum Allocation of Resources to Follow up Unit Nonrespondents in Probability Surveys
Siu-Ming Tam, Anders Holmberg and Summer Wang

Volume 39 (2023): Issue 4 (December 2023)

https://sciendo.com/issue/jos/39/4

Small Area Estimates of Poverty Incidence in Costa Rica under a Structure Preserving Estimation (SPREE) Approach
Alejandra Arias-Salazar

Block Weighted Least Squares Estimation for Nonlinear Cost-based Split Questionnaire Design
Yang Li, Le Qi, Yichen Qin, Cunjie Lin and Yuhong Yang

Answering Current Challenges of and Changes in Producing Official Time Use Statistics Using the Data Collection Platform MOTUS
Joeri Minnen, Sven Rymenants, Ignace Glorieux and Theun Pieter van Tienoven

Small Area with Multiply Imputed Survey Data
Marina Runge and Timo Schmid

Temporally Consistent Present Population from Mobile Network Signaling Data for Official Statistics
Milena Suarez Castillo, Francois Sémécurbe, Cezary Ziemlicki, Haixuan Xavier Tao and Tom Seimandi

Application of Sampling Variance Smoothing Methods for Small Area Proportion Estimation
Yong You and Mike Hidiroglou

Book Review: Silvia Biffignandi and Jelke Bethlehem. Handbook of Web Surveys, 2nd edition. 2021 Wiley, ISBN: 978-1-119-37168-7, 624 pps
Maria del Mar Rueda Garcia

Survey Research Methods

Journal of the European Survey Research Association

Vol 17 No 2 (2023)

https://ojs.ub.uni-konstanz.de/srm/issue/view/234

Articles

What Parcel Tax Records Tell Us About Homeownership Measurement in Surveys
Shiyu Zhang, James Wagner, Elisabeth R. Gerber, Jeffrey D. Morenoff

Observing Interviewer Performance in Slices or by Traces: A Comparison of Methods to Predict Interviewers' Individual Contributions to Interviewer Variance
Celine Wuyts, Geert Loosveldt

Boosting Survey Response Rates by Announcing Undefined Lottery Prizes in Invitation Email Subject Lines Evidence from a Global Randomized Controlled Trial
Syedah Ahmad, Robert Lensink, Annika Mueller

The Role of the Interviewer in Producing Mode Effects: Results From a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration
Steven Hope, Pamela Campanelli, Gerry Nicolaas, Peter Lynn, Annette Jäckle

Answer Refused: Exploring How Item Non-response on Domestic Abuse Questions in a Social Survey Affects Analysis
Valeria Skafida, Fiona Morrison, John Devaney

Replication Studies

Comparing Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia: A Total Survey Error Perspective
Paul John Lavrakas, Darren Pennay, Dina Neiger, Ben Phillips

Vol 17 No 3 (2023)

Recent Methodological Advances in Panel Data Collection, Analysis, and Application

https://ojs.ub.uni-konstanz.de/srm/issue/view/235

Editorial

Recent Methodological Advances in Panel Data Collection, Analysis, and Application
Sabine Zinn, Tobias Wolbring

Articles

The Researcher, the Incentive, the Panelists and Their Response: The Role of Strong Reciprocity for the Panelists' Survey Participation
Rolf Becker

Case Prioritization in a Panel Survey Based on Predicting Hard to Survey Households by Machine Learning Algorithms: An Experimental Study
Jonas Beste, Corinna Frodermann, Mark Trappmann, Stefanie Unger

Satisficing Response Behavior Across Time: Assessing Negative Panel Conditioning Using an Experimental Design with Six Repetitions
Fabienne Kraemer, Henning Silber, Bella Struminskaya, Bernd Weiß, Michael Bosnjak, Joanna Koßmann, Matthias Sand

Memory Effects in Online Panel Surveys: Investigating Respondents' Ability to Recall Responses from a Previous Panel Wave
Tobias Rettig, Bella Struminskaya

Experimental Evidence on Panel Conditioning Effects when Increasing the Surveying Frequency in a Probability-Based Online Panel
Carina Cornesse, Annelies Blom, Marie-Lou Sohnius, Marisabel González Ocanto, Tobias Rettig, Marina Ungefucht

Question order and panel conditioning analysing self-reported data
Omar Paccagnella, Mariangela Guidolin

Use of Panel Surveys to Measure Employment Precarity in a Cross-National Framework: An Integrated Approach to Harmonize Research Concepts and Longitudinal Data
Katarzyna Kopycka, Anna Kiersztyn, Zbigniew Sawiński, Stefan Bieńkowski, Viktoriia Sovpenchuk

Assessing Rental Price Dynamics in Two Gentrified Neighbourhoods in Cologne by Means of a Dwelling Panel
Alice Barth, Jörg Blasius

Vol 17 No 4 (2023)

https://ojs.ub.uni-konstanz.de/srm/issue/view/237

Articles

Transitioning to a Mixed-Mode Study Design in a National Household Panel Study: Effects on Fieldwork Outcomes, Sample Composition and Costs
Katherine A. McGonagle, Narayan Sastry

The Feasibility of Using Consumer-Level Activity Trackers in Population Monitoring of Physical Activity: Comparing Representativeness and Measurement Quality With Self-Report and a Professional Research-Grade Accelerometer
Rianne Kraakman, Annemieke Luiten, Vera Toepoel, Maaike Kompier

Testing Schwartz's Model of Cultural Value Orientations in Europe with the European Social Survey: An Empirical Comparison of Additive Indexes with Factor Scores
Hermann Duelmer, Shalom H. Schwartz, Jan Cieciuch, Eldad Davidov, Peter Schmidt

Late Responding in Web and Mail Surveys: A Systematic Review and Meta-Analysis
Ellen Laupper, Esther Kaufmann, Ulf-Dietrich Reips

Detecting and Explaining Missing Comparability in Cross-National Studies: The Case of Citizen Evaluation of Patriotism
Katharina Meitinger, Peter Schmidt, Michael Braun

Impact of Mode Switching on Nonresponse and Bias in a Multimode Longitudinal Study of Young Adults
Ting Yan, Jonathan Wivagg, William Young, Cristine Delnevo, Daniel Gundersen

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.	Mingmeng	Geng	Italy
PROF. DR.	Montserrat	Guillen	Spain
Mr.	Yonghyun	Kwon	Republic of Korea
MS	Ivana	Levacic	Croatia
DR.	Andrew	Padovani	United States
MR.	Mallo	Paul Lokiru	Uganda
DR.	Andrea	Troisi	Italy

IASS Executive Committee Members

Executive officers (2023 – 2025)

President:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
President-elect:	Partha Lahiri (US)	plahiri@umd.edu
Vice-Presidents:		
Scientific Secretary and Social Media Coordinator	Annamaria Bianchi (Italy)	annamaria.bianchi@unibg.it
Monthly newsletter	Jiraphan Suntornchost (Thailand)	jiraphan.s@chula.ac.th
VP Finance and IASS conferences support 2024, 2025	Natalie Shlomo (UK) Partha Lahiri (US)	natalie.shlomo@manchester.ac.uk plahiri@umd.edu
Liaising with ISI EC and ISI PO plus administrative matters	Partha Lahiri (US)	plahiri@umd.edu
Chair of the 2025 Cochran-Hansen Prize Committee, Chair of the 2024 Hukum Chandra Prize Committee and IASS representative on the ISI Awards Committee	Eric Rancourt (Canada)	eric.rancourt@statcan.gc.ca
IASS representatives on the World Statistics Congress Scientific Programme Committee	Partha Lahiri (US)	plahiri@umd.edu
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	Partha Lahiri (US)	plahiri@umd.edu
IASS Webinars 2023-2025	Andres Gutierrez (Chile)	andres.gutierrez@cepal.org
Ex Officio Member:	Conchita Kleijweg (The Netherlands)	c.kleijweg@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

IASS Facebook Account: https://www.facebook.com/iass.isi/

Institutional Members

International organisations:

- Eurostat (European Statistical Office) – Unit 01: External & Interinst.

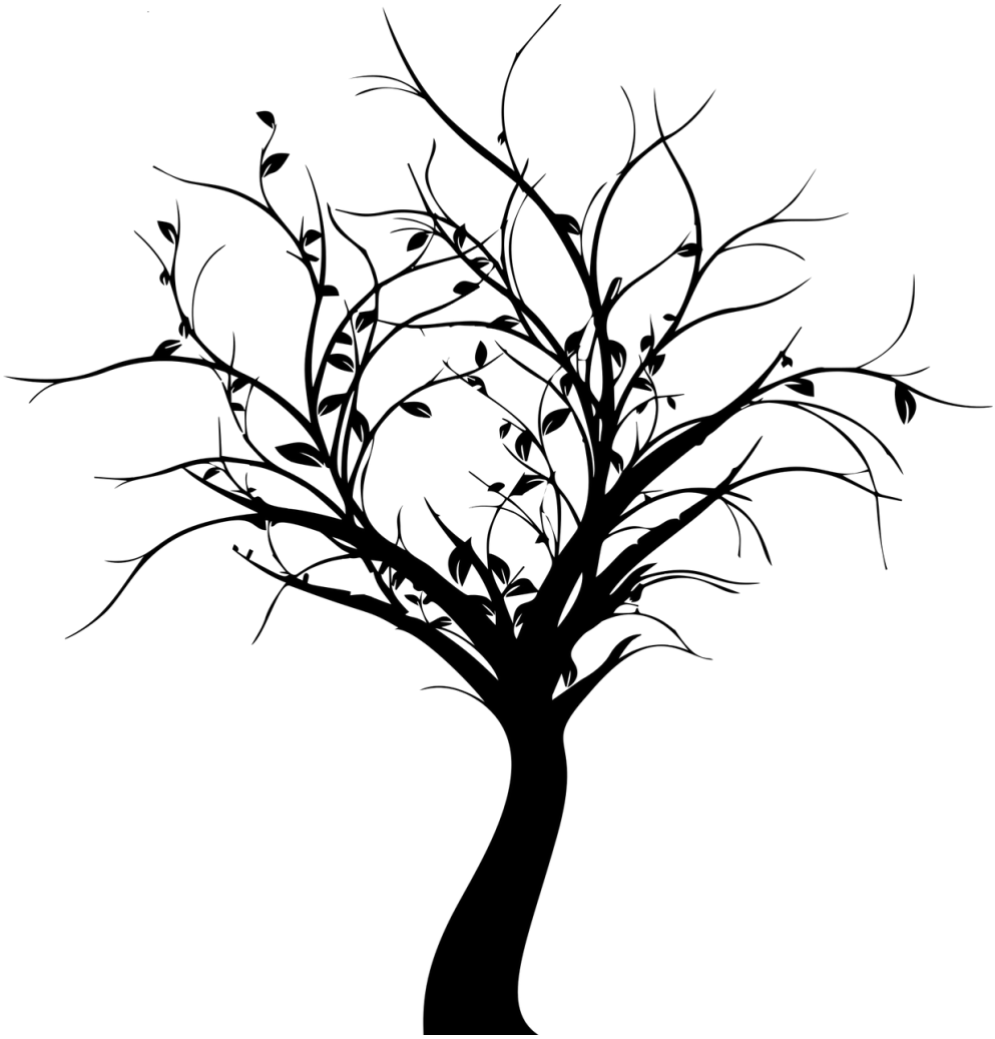
National statistical offices:

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

Private companies:

- Westat, United States

Read *the Survey Statistician* online!



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