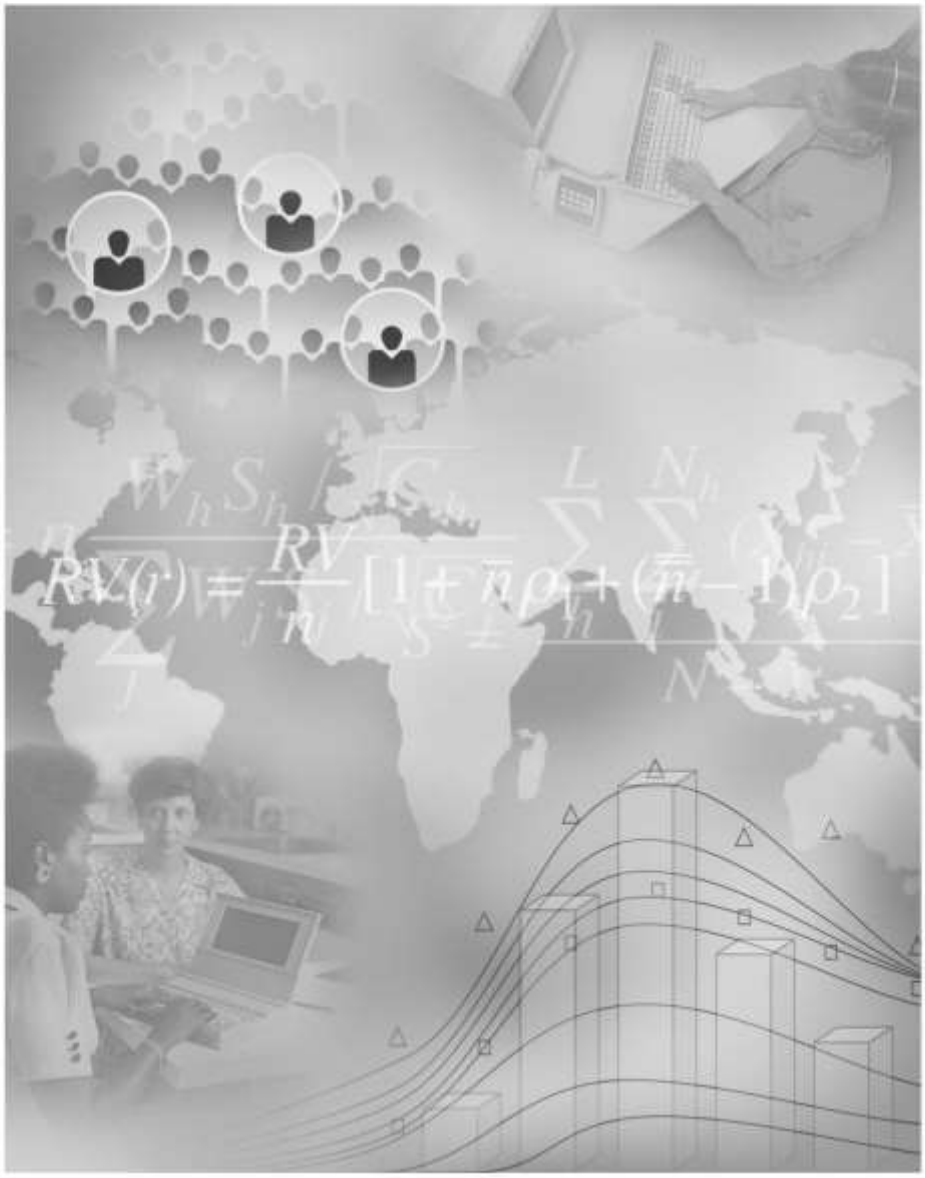
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

No. 90 **July 2024**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

The Survey Statistician No. 90, July 2024

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

Enquiries for membership in the Association or change of address for current members should be found at the section **Promoting good survey theory and practice around the world** on p.11 of this issue or addressed to: isimembership@cbs.nl

Comments on the contents or suggestions for articles in the Survey Statistician should be sent via e-mail to the editors Danutė Krapavickaitė (danute.krapavickaite@gmail.com) or Alina Matei (alina.matei@unine.ch)

ISSN 2521-991X

In this Issue

Letter from the Editors

Dear Readers,

We are happy to present you the contents of the latest issue of *The Survey Statistician!*

The IASS EC members welcome the readers. The President of IASS, Natalie Shlomo, in her *Letter from the President*, and the Scientific Secretary of IASS, Annamaria Bianchi, in her *Report of the Scientific Secretary*, overlook the activities of IASS during the first half of 2024. Although these activities are presented in a comprehensive and timely manner on the IASS homepage, the concentrated information in the TSS once again ensures that the IASS is a very active institution. Natalie Shlomo draws readers' attention to the IASS Strategic Document and invites its members to attend the IASS Annual General Meeting on 17 July at 14.00 CET.

Next, Kirk Wolter, Eric Rancourt and Wesley Yung pay tribute to two past presidents of IASS who recently passed away: Barbara Bailar and Nanjamma Chinnappa. They made significant contributions to the fields of statistics in which they worked and left good memories in the hearts of their colleagues and friends. On this occasion, we recall Nanjamma Chinnappa's congratulations to IASS members on the occasion of IASS's 50th anniversary: "May the institution continue to grow, and be effective in promoting good survey methods all over the world".

The issue covers a wide range of topics, from the analysis of historical facts to recent methods in survey statistics. The composition of the IASS Jubilee special issue (Vol. 88, July 2023) opened an interesting topic on the history of survey statistics. The historical development of survey statistics varies greatly between countries and continents, and it is worth revealing this history in TSS to make it easier to understand for future generations. Thomas Laitila gives an overview of the early Scandinavian contributions to survey sampling. Some surprising facts await the reader.

The following articles focus on recent methods in survey statistics: machine learning and the use of big data, as well as statistical inference with non-probability samples. Machine learning methods are becoming particularly popular in the era of big data and powerful computers. They can also be used in survey statistics. Camelia Goga gives an overview of some theoretical issues of machine learning in survey statistics. She and her collaborators have made an important contribution in this area.

Early career survey statisticians Camilla Salvatore and Angelo Moretti solve a modern statistical problem. They estimate the proportion of people worried about climate changes in small areas using web data as auxiliary information to improve the estimates. This work complements the topic presented in the January 2024 issue of TSS by the Italian statisticians Stefano Marchetti and Schirripa Spagnolo on the use of big data in small area estimation applications.

In the *Book & Software Review* section, Angelo Moretti reviews the book "Sampling and estimation from finite populations" by Yves Tillé, and an overview of software for statistical inference with nonprobability samples is provided by the Spanish statisticians Beatriz Cobo, Ramón Ferri-García, Jorge L. Rueda-Sánchez and María del Mar Rueda. This review therefore presents software for the application of some recent methods in survey statistics.

Finally, the current issue includes the traditional country reports which overlook the activities of survey statisticians around the world. Recent publications in other journals and planned workshops and conferences in 2024-2025 are listed as usual.

The TSS editorial team is changing. Alina Matei is the new Associate Editor of TSS and this is the first issue published with her contribution. She replaces Eric Rancourt, who worked for TSS for almost ten years - since January 2015 - and wrote the history of TSS in the July 2023 issue. We

thank Eric for his important editorial work. We are delighted that Alina has joined the editorial team. A new person usually brings some freshness to the work, and we expect to see some updates in future issues of TSS. The Technical Editor of TSS, Maciej Beręsewicz, is making his last compilation of TSS and will hand over his task to another statistician. The essence of the technical editor's work is to collect information from the content pages of other journals, to assemble the articles and all materials into a complete whole, including some editing and merging of .pdf files obtained from MS Word and Latex into a common .pdf file. Maciej has learned this from the previous Technical Editor, Martiņš Liberts, and the work is going smoothly. We thank Maciej for his important contribution and his patient work in compiling an issue again and again due to the endless editorial changes.

Most of the articles published in TSS are solicited by the editors. We encourage readers of TSS to take the initiative and submit articles themselves, outlining some topics in survey statistics, its history and choosing the appropriate section. You can contact any of the Editorial Board members listed on the 2nd page of TSS.

Danutė Krapavickaitė

TSS Editor

Letter from the President

Dear IASS Members,

The IASS Executive Committee have been working very hard to promote and support the many activities that serve our community of survey statisticians. I'm very grateful to Partha Lahiri, Eric Rancourt, Andres Gutierrez, Jiraphan Suntornchost and Annamaria Bianchi for their dedication in undertaking all of the IASS initiatives. I am also grateful to the ISI Permanent Office, particularly to Conchita Kleijweg and her team, for supporting IASS operations and administration and ensuring that everything is run smoothly.

Our IASS Strategic Document is open for consultation and available here: [Strategic Document]. Please provide comments to natalie.shlomo@manchester.ac.uk before we bring it to the vote at our 2024 IASS Annual General Assembly (AGM) to be held on July 17th at 2 pm CET. It is important that you make your voice heard on the future direction of the IASS by attending the AGM. The registration page is here: [2024 Annual General Assembly].

In the Strategic Document, we have identified new avenues to increase our visibility and support for our community. The first strategy is to actively support relevant survey statistics conferences by organizing special invited sessions. I am pleased to say that we will now have a sponsored session at the Statistics Canada International Methodology Symposium (October 29–November 1, 2024) titled: Developments in Small Area Estimation. We are also sponsoring an invited session at the upcoming 3rd Workshop on Methodologies for Official Statistics at ISTAT (December 4–5, 2024) titled: Innovations in Small Area Estimation, focusing on the use of innovative non-survey data in small area estimation.

Another important strategy is to sponsor an IASS satellite conference prior to the 65th World Statistics Congress to be held in The Hague. The IASS has a tradition of providing monetary support to various conferences in survey statistics, but does not necessarily brand an official IASS satellite conference to the WSC. This will now change. The IASS is happy to join with the biannual 2025 9th Italian Conference on Survey Methodology (ITACOSM 2025). The joint conference will be held July 1–4, 2025 in Bologna, Italy. Moreover, we are also delighted to announce that the 2025 Small Area Estimation (SAE 2025) conference will be held during the second week of July also in Italy. Therefore, both conferences will be labelled as IASS satellite conferences to the 65th WSC. Note that although the 65h WSC has had to be moved from July to October 5–9, 2025, we have decided to keep the IASS satellite conference in July as originally scheduled to allow for more flexibility during the summer months. Watch this space as more information emerges on these exciting July 2025 conferences now in planning.

This year the IASS supported two events: a Workshop on Survey Data Visualization using R, held in Ogun State, Nigeria on January 24, 2024, and the upcoming 13th edition of the International Francophone Conference on Sample Surveys to be held on November 5–8, 2024 in Luxembourg. The list of upcoming conferences and workshops are featured on our events page on the IASS website: [Events], in our monthly newsletters: [Newsletters], as well as listed in The Survey Statistician.

The Survey Statistician (past issues can be found here: [The Survey Statistician]) continues to be our flagship IASS publication. An essential and historical part of The Survey Statistician are the country reports and we are very grateful to the IASS country representatives for their dedication in submitting timely articles. We are also very grateful to our editorial team of The Survey Statistician,

in particular our Editor-in-Chiefs, Danutė Krapavickaitė and Alina Matei, for producing and publishing this outstanding publication.

The Survey Statistician serves our community scientifically with informative articles, particularly highlighting new developments and opportunities as we make the cross-over into AI and Data Science. Beyond traditional components of survey statistics covering survey design, questionnaire development, data collection and modes of response, nonresponse adjustments and weighting, small area estimation and confidentiality practices, the field of survey statistics is moving into more formal statistical and modelling theory as we deal with ongoing challenges related to the quality of probability-based surveys. New directions of research include measuring and compensating for measurement errors and informative nonresponse, nonprobability sampling, data integration and multi-source statistics, and Bayesian survey estimation. We aim to ensure that The Survey Statistician remains relevant with timely articles dedicated to these areas and to the future of survey statistics.

Our highly successful IASS webinar series continues to provide stimulating and informative talks from leading survey statisticians around the world and I thank Andres Gutierrez for organizing this successful series. We provide regular updates on our IASS website [Home] and through our increasing social media presence and I thank Annamaria Bianchi for posting and coordinating our social media accounts on Facebook, LinkedIn and X. We are also now in the process of selecting the recipient of the 2024 Hukum Chandra Prize and I thank Eric Rancourt for chairing the selection committee.

Now is the time to spread the word about IASS to your networks and to help increase our membership. The website for joining the IASS is here: [Join IASS]. Please download our brochure here: [Brochure] to help recruit new members from your networks. I am also pleased to say that we continue to expand our institutional members and have recently welcomed two new institutional members to the IASS: the Office for National Statistics in the UK and IPSOS Italy. We thank all institutional members for their continuing support to the IASS. The list of institutional members to the IASS are included in the back page of The Survey Statistician.

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, particularly related to organizing sessions at relevant conferences and workshops.

With best wishes,

Natalie Shlomo

President IASS, 2023-2025

Report from the Scientific Secretary

One year has already passed since my election and appointment as Scientific Secretary of the IASS. It has been a very intense year, with a lot of existing and new activities sponsored by the IASS.

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 41 with the speaker Andrew Mercer, Senior Research Methodologist at Pew Research Center, 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 first six months of 2024 have covered Latent Variable Models for Finite Population Inference by Maria Giovanna Ranalli (University of Perugia), New Developments in Small Area Estimation from a Practitioner's Perspective by David Newhouse (World Bank), Data Integration, Data Linkage, and Linked Data Analysis by Tiziana Tuoto (Istat, Italian National Institute of Statistics), Model-Based Optimal Designs for a Multipurpose Farm Survey by Jay Breidt (NORC at the University of Chicago), An Estimation of Variance of Random Effects to Solve Multiple Problems in Small Area Estimation by Masayo Y. Hirose (Institute of Mathematics for Indystry, Kyushu University, Japan).

One upcoming **webinar** in October will be devoted to the **winner of the second edition of the Biennial Hukum Chandra Memorial Prize.** The prize will be awarded by the end of July to a mid-career researcher, defined as someone with more than 10 years of experience after PhD or in employment, who has made an important contribution in research areas of Hukum Chandra's work, namely survey sampling, small area estimation, official statistics, spatial analysis applied to official and survey statistics and agricultural statistics. The definition of a mid-career researcher in this call aims to recognise researchers who are close to Dr. Chandra's career trajectory. The prize committee has been appointed by the IASS EC and is composed by Eric Rancourt (Statistics Canada), Nancy McBeth (New Zealand), Siu-Ming Tam (Australian Bureau of Statistics), and Sharon Lohr (Arizona State University). Please follow the updates on our social media pages and our Newsletter.

As we approach the next **ISI World Statistics Congress** in The Hague, we have been working on it. The Invited Paper Session (IPS) submissions have closed and there have been 14 submitted proposals sponsored by the IASS. Also the Call for Contributed Paper Submissions (CPS) closed on 17 June 2024. Furthermore, we are planning an **IASS satellite meeting to the WSC**. Details on it will be announced shortly. For more information, visit the WSC webpage https://www.isi-next.org/conferences/isi-wsc2025/.

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. The number of followers to the pages are increasing: since January, the followers increased from 392 to 416, LinkedIn followers from 1434 to 1801 and Facebook followers from 178 to 216. Thus, please, follow the updates on the life of the IASS via **social media** and reading our **monthly Newsletter** available at http://isi-iass.org/home/services/newsletters/. Other than webinars, information on conferences, on the recipients of awards and on call for nominations, it also 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

IASS Scientific Secretary 2023-2025

News and Announcements

Barbara A. Bailar's passing

I write to inform her many friends and colleagues that Barbara Bailar passed away June 13, 2023, at her home in Houston, Texas, where she had moved during the pandemic to be near family

Barbara grew up in upstate New York and received a bachelor's degree in mathematics from the State University of New York at Albany. She was trained in statistics at Virginia Tech in Blacksburg (MS) and American University in Washington, DC (PhD).

She spent most of her career at the US Census Bureau, where she worked from 1958 to 1987. She rose to become Associate Director for Statistical Standards and Methodology and, in that role, was instrumental in establishing a Computer Assisted Telephone Interviewing capability (then a relatively new mode of data collection) and survey methodology as a distinct discipline at the Census Bureau. Barbara sought and organized the financial resources and institutional backing necessary to develop a cognitive-laboratory organizational unit within the Bureau. She defended the 1980 Decennial Census in federal court and led efforts to develop and test new methods of census taking that would reduce differential undercount in future censuses. Barbara founded the Census Bureau's annual research conference in the mid-1980's, a forerunner of today's FCSM Research and Policy Conference. She was an executive with NORC at the University of Chicago from 1995 to 2001.

Barbara conducted research on non-sampling errors in social surveys and censuses. She designed and analysed special studies to measure the correlated component of response variance brought by interviewer effects. Her work on rotation group bias and an error profile for the Current Population Survey is well known.

Barbara was a prominent statistician. She was the 82nd president of the American Statistical Association (ASA) in 1987, president of the International Association of Survey Statistician in 1989-1991, and vice president of the International Statistical Institute in1993-1995. She served as Executive Director of the ASA for seven years in the late 1980s and early 1990's. She was active in the Washington Statistical Society throughout her career. Barbara was a Fellow of the ASA and an Elected Member of the International Statistical Institute.

Barbara was married to John C. Bailar III (deceased), a prominent bio-medical statistician who was founding chair of the Department of Public Health Sciences, University of Chicago. She is survived by two daughters, Pamela Monaco (Ocean County College) and Melissa Bailar (Rice University), and one grandchild.

I had the distinct honour and great pleasure of working with Barbara for 20 years.

Kirk Wolter

Tribute to Nanjamma Chinnappa

Dr. Boverianda Nanjamma Chinnappa, loving wife to Boverlanda Chinnappa and dear mother to Kaveri and Gouri passed away peacefully in Mysuru, India on March 31, 2024, at the age of 89.

Nanjamma held a master's degree in Statistics from Madras University and one in Advanced Statistics from the Indian Statistical Institute, Kolkata. She joined Statistics Canada as a Senior Methodologist in the Institutions and Agriculture Survey Methods Division (IASMD) in 1975. In 1987, she became the Director of the Business Survey Methods Division (BSMD) and retired in 1995. During her statistical career, she was also an international consultant and was very active in professional associations. From 1997 to 1999, she was president of the IASS. A renowned scholar, she was elected as a Fellow of the American Statistical Association in 1993 and was awarded an honorary doctorate degree (D. Litt.) by Mangalore University in 2006.

Prior to joining Statistics Canada, she was a Visiting fellow at Cambridge University. After retiring, she and her husband went back to India where they deployed enormous amounts of effort in successfully contributing to the preservation and publication of information on local sacred heritage sites. This included the publication of two books related to the Kodava community.

As soon as anyone would pronounce the name Nanjamma, anyone who had the great chance of working with her or meeting her would immediately have recollections of how an exceptionally nice and genuinely caring person she was. At Statistics Canada, her office door was always opened to listen to and advise any and all on any topic. She approached issues in ways that often left people struck by her openness and by how she made so many work problems become easier to solve when looked at more humanly.

Nanjamma was icon. She will be dearly missed by all who knew her.

Eric Rancourt and Wesley Yung

The Survey Statistician, 2024, Vol. 90, 11–17.

J. Mohn (1838–1882)

A. N. Kiaer (1838–1919)

Early Scandinavian Contributions to Survey Sampling

Thomas Laitila

Örebro University, Sweden, thomas.laitila@oru.se

Abstract

This paper is concerned with the developments of the representative method from its suggestion by Kiaer in 1895 until its acceptance at the International Statistical Institute meeting in Bern 1925. The focus is on contributions from the Scandinavian countries while not intending for a full account of all work made during the 1895–1925 period. A main aim is to provide plausible explanations to the developments of the method with respect to applications in official statistics. Relating to the non-sampling errors facing National Statistical Institutes in our time, it is relevant to ask if we would have been better off today if our methods would have stayed within the original methodology of the representative method.

Keywords: the representative method, random selection, purposive selection, official statistics.

1 Introduction

The initial idea was to write a paper covering Scandinavian contributions to survey sampling from the presentation by Anders Kiaer at the 1895 International Statistical Institute (ISI) meeting in Bern up to the 1980s. When collecting information on contributions three vital curiosities emerged. The first was the domination of contributions from Sweden from the 1950s onward. One explanation might be how the reference search was conducted. Partly this may be so, but publications from the 1950s and onward are relatively well documented in different data bases.

The second curiosity was the late adaption of the Neyman (1934) theory. Statistics Sweden started employing sample surveys in the 1950s, but probability sampling were rarely used in the beginning. The theory was generally known and also suggested by users of official statistics. The third curiosity was the seemingly inactivity in developments of the representative method up to its acceptance at the 1925 ISI meeting.

It is a cliché but it is true, understanding the present is not possible without knowing the history. This is also true with statistics. An example is the way National Statistical Institutes (NSI) are publishing survey statistics, i. e. point estimate plus/minus a margin of error. This has historical roots and users may in those days have been able to correctly interpret the statistics. With the wide use of official statistics today it is reasonable to assume that alternative ways of publications of statistics are required for correct interpretation.

The purpose of this paper is therefore to study early developments of survey sampling, with a focus on contributions from the Scandinavian countries. The aim is to obtain a picture of the developments from Kiaer's suggestion in 1895 until its acceptance at the 1925 ISI meeting.

Kiaer's contributions are given attention in the next section, followed by a section describing applications of the representative method in the period. Section 4 includes some methodological issues and a discussion is saved for the final section.

2 Kiaer and the Representative Method

The importance of the contributions made by the Norwegian statistician Anders Kiaer (1838–1919) for the development of the survey sampling methodology cannot be understated. The idea of taking a sample instead of a census was presented by Kiaer at the ISI meeting in Bern 1895; he called it the "representative method". There he suggested selection of samples giving miniature representations of the populations as a complement to censuses. The arguments were lower costs and faster dissemination of results.

His importance for the development of sample surveys is demonstrated by the many sources describing his work and efforts. One is the Encyclopedia of Mathematics (Kiaer, Undated), another is Statisticians of the Centuries (Heyde and Seneta, 2001). A third source is Kruskal and Mosteller (1980) which gives a thorough description of Kiaer's struggle with the ISI to have the representative method accepted. A story different from the one generally told, on Kiaer's role in developing the representative method idea, is given by Lie (2002).

For increased precision (reduced bias error) of sample statistics, Kiaer and his colleagues at the Norwegian statistical bureau used sampling designs involving e. g. stratified sampling, cluster sampling and its special case of systematic sampling. They used auxiliary census data in evaluation of sample representativeness. These basic methodologies are found in later applications of the representative method and they are standard tools in survey sampling design of today.

Kiaer did not use random sampling, but Kiaer (1897)[p. 39] argues the selection of units made are as if they would have been drawn by lots, i. e. the sample can be treated as a random one. Jensen (1925b)[p. 548] supports this view considering a two-stage sampling design with purposive selection in the first stage of the social survey. In the second stage, men of specific ages were considered and within those age categories men with names starting on a set of specific letters are selected to the sample. This second sampling stage is by Jensen considered as random sampling.

Kruskal and Mosteller (1980) do not classify Kiaer's sample selection methods as random sampling, in the meaning of a probability sample. The earliest known application of random sampling is, according to Kruskal and Mosteller (1980), a study of housing conditions in Gothenburg around 1910. (The same survey is in Jensen (1925b) dated to 1911.) The survey was conducted by K. A. Edin (1880–1937) and reported in Edin (1912). The sample selection procedure is described in Dalenius (1957)[p. 40].

Lie (2002) draws attention to three studies using samples instead of censuses conducted prior to Kiaer's application of the representative method. According to Lie, the first study was made at the Norwegian State's statistical office of which Kiaer was the head. The design of the survey, aiming for statistics on agricultural production, was made by Jacob Mohn (1838–1882) and conducted by Mohn and Boye Strøm (1847–1930) around 1875. Both Mohn and Strøm were colleagues with Kiaer. The second study was conducted in the 1870s by Mohn. The third, a survey on household consumption in 1888, was conducted by Strøm.

The last survey was undertaken at the Norwegian Central Bureau of Statistics (CBS), formed in 1876

by reorganization of the state statistical office. Kiaer was appointed director at the CBS in 1877. Thus, prior to Kiaer's own study in 1897, his colleagues had several times employed the basic idea of choosing a sample. This has to be kept in mind when attributing the start of survey sampling to Kiaer and his representative method. According to Lie it is reasonable to consider Mohn's work in the mid 1870s as the first step of the development of the representative method.

Another enlightening story in Lie (2002) provides an answer to a question raised by Kruskal and Mosteller (1980). They describe the 1903 ISI meeting as the last one Kiaer promoted the representative method. Thereafter it was not brought up again until the 1925 meeting, six years after the death of Kiaer. Kiaer seems not to have brought up the topic elsewhere either. So, why was not the representative method on the agenda between 1903 and 1925? Lie gives an explanation.

Kiaer was not a mathematician while probability calculations could be done by treating the sample as if it were taken at random. Mathematicians did such calculations and sometimes they implied that Kiaer's survey results were questionable. Kiaer could not respond and explain the differences. The most damaging example described by Lie, is an estimate of the number of disabled people in Norway in 1906. With probability calculations and by scrutinizing the survey design, the critics claimed Kiaer's estimate was too low. Kiaer did a complementary survey showing the critics were right! Kiaer's potentially erroneous estimate may not had been such a deal if it was not for its purpose. In this case it was to be used for decisions on a general disability insurance. The insurance plan was controversial and publicly debated.

The story on the distrust and public discussion on Kiaer's survey results may explain why Edin (1912) chose to use random sampling by lot (stratified simple random sampling). Dalenius (1957) gives the following citation of Edin:

> *The main object of the procedure here described has naturally been that no one could possibly have the slightest reason for saying that the sample was biased, or on the whole, for whatever reason, that preferably worse apartments have been included in the survey.*

3 The Representative Method in Practice

Adolph Jensen (1866–1948) from Denmark was a member (Rapporteur) of the ISI committee preparing the report on the representative method for the ISI meeting in 1925 (Jensen, 1925a). In an appendix to the report (Jensen, 1925b) he fulfilled a decision made by the ISI already in 1903; a report on the practice of the representative method. The reports give valuable insights on the state-of-the art at that time and the earlier evolution of the representative method.

Jensen (1925b) reports on 50 studies between the late 1890s up to the early 1920s. Over this period he found a pattern where the representative method where applied to some extent around the time Kiaer was promoting the method. Thereafter a decline in interest of the method up to the mid 1910s was indicated. World War I (WWI) seems to have fueled an increasing interest in the representative method followed by an even higher interest after the war. Jensen concluded the driving force behind the increasing number of applications was the growing demand of statistics paired with limited resources for statistics production. Notably, many of the earliest studies comprise of selection of samples from census data. Apart from experimentation these were motivated by timeliness and resource limitations. Instead of analyzing all data from a census, sample results could be disseminated faster to a lower cost.

Jensen's survey covers studies made in 13 different countries with a tendency of a center of gravity towards the Scandinavian countries. He found this not to be a surprising result even when taking into account he himself was Danish. One explanation given is that representative sampling is a Norwegian idea and may have had a particular influence on the methods used in the neighboring countries,

Denmark and Sweden. His explanation can further be supported by the long history shared by these three countries.

Jensen's paper is divided into sections by how the sample is selected. Two of the sections deal with random sampling of units and random sampling of groups, respectively. Around half of the studies reported on are classified into these two categories of sample selection. However, it seems the meaning of a random sample was that the sample could be treated as a random sample. Many studies reported as using random sampling involve systematic sampling but it does not seem the starting points were selected at random. The only study, where the random sampling can be interpreted in the meaning of probability sampling, is the earlier mentioned Edin (1912).

Other sections in Jensen's paper cover applications of purposive sampling. In one section, two studies are considered with purposive selection of groups and random selection of units in selected groups. These are, of course, examples of two-stage cluster sampling, although not using probability sampling in the two stages.

Stratified sampling is a frequently applied sampling design in the studies reported, a design also applied by Kiaer. This is obviously necessary in purposive sampling where the ambition is to have a sample representing a miniature of the population. One stage cluster sampling is reported in several studies where the purposes were to sample from census data or registers to reduce time and costs in calculation of statistics. Some examples are experiments to test the representative method.

Upon reading the report for the ISI 1925 meeting (Jensen, 1925a) it is striking how close the reasoning on and recommendations for designs of representative sampling is to modern theory and practice. These do not include probability sampling. However, on page 487 the following is stated regarding the desire to avoid sample data being "one-sidedly coloured":

> *The handiest method would seem to be that the inquiry is made according to some mechanical principle or other which is unconnected with the subject and purpose of the inquiry, with essential condition that every unit in the population or universe in question shall have an equal chance of inclusion in the sample.*

Later on Jensen suggests to draw the sample by lot if there is no "mechanic principle", e. g. systematic selection, deemed to yield a representative sample.

4 Methodology

A discussion of the two meanings of statistics, statistics and statistical methodology (mathematics), seems to have emerged during the 1910s when interpreting Edin (1916). Kiaer's idéa of taking a sample enabled studies of the society in new areas and in more detail. Perhaps because of the method being new, Edin (1916) suggests prioritizing, for a time ahead, a deeper understanding of the statistics rather developing new mathematical formulas for "fine tuning" of statistics.

This implied division among statisticians may have been much deeper when considering the general use of statistics. The Danish statistician Harlad Westergaard (1853–1936) has the the following paragraph in Westergaard (1916).

> *Still it is a fact, that there are at present not one, but several corps of statisticians, each trying earnestly to promote the science, but hardly able to cooperate for lack of mutual sympathy and sometimes acting in direct opposition to one another.*

In the paper and in Westergaard (1918) he promotes use of simplicity over complexity. There was a trend towards finding models or formulas giving more universal relations between variables. In many cases simple tabulation of data would do equally well or better than correlation calculations and fitted

models. Notably he wrote about the importance of keeping calculated statistics close to the original data.

The first volume of the statistical journal *Nordisk Statistisk Tidskrift* was published in 1922. In 1929 the first volume of an English edition, *Nordic Statistical Journal*, was published. The founder was Thor Andersson (1869–1935) and the last volumes of both journal editions were published in 1932. The journal included papers on a variety of aspects on statistics. Some of the contributions were concerned on problems related to the representative method.

One contribution in the journal is Jensen (1923) who proposed some labor saving methods in statistics. Statistical agencies could not with existing resources meet the increasing demand of new statistics. One argument of his was that statistical agencies took on too big tasks than necessary. An example is the use of one and the same questionnaire to all units in the population, while often a shorter version would do for most units. Another argument is time placed on calculation of too detailed statistics.

Jensen also brings up the idea of the representative method. He points to it being put on the table by Kiaer for almost 30 years ago. The initial skepticism towards the method he believes to be grounded in the fear of unprofessional applications. This skepticism can be understood, but Jensen means it cannot be an argument today. There are enough of examples of applications of the representative method yielding satisfactory results.

Another author with several contributions in the two journals is Tor Jerneman (1897–1965). His dissertation (Jerneman, 1931) includes a number of interesting statements. In the first paragraph of the paper he states he prefer to use the English word “sampling” instead of the Swedish “den representativa metoden”. He finds the Swedish name misleading and rather cumbersome.

In the second paragraph he motivates the use of sampling for reduction of response burden.

Now, one of Jerneman’s contributions is a derivation of the standard error for the sample mean under SRS:

$$\sigma_n = \frac{1}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\sigma. \tag{1}$$

Here N and n denotes population and sample sizes, respectively, and σ denotes the population standard deviation of population values.

He notes A. A. Tschuprow (1874–1926) had much earlier in Tschuprow (1918) claimed to be the first to derive this expression.

Jerneman realizes replacing the unknown σ with a corresponding sample value (σ_s) will either give a higher or a lower value. To protect against an underestimation of the standard error he suggests the alternative value

$$\sigma_n = \frac{1}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\sqrt{\sigma_s^2 + \bar{\sigma}^2} \tag{2}$$

where $\bar{\sigma}$ is an expression for the standard error of σ_s^2.

5 Discussion

It is accurate to think of this paper as merely a prestudy or a qualitative study upon which hypotheses can be formulated on the early developments of survey methods. Here focus has been on Scandinavia where, of course, developments were integrated with those elsewhere. The report to the ISI in 1925, for example, was based on the work of a six member committee of which one were from the Nordic countries. There, Jensen’s survey of applications covered 13 countries, so the topic of

the representative method was of a world wide interest. Furthermore, the names and contributions cited here cannot be claimed to cover all the important contributions from the Scandinavian countries either.

However, based on what is included, one can formulate a number of hypotheses causing the evolution of the survey methodology in the years between Kiaer and Neyman. Production of official statistics is not made in isolation. There is a production organization and there are users and stakeholders. So there is a whole environment to convince of the pros of a new methodology. Going from a census to a sample survey would also imply the need of education in interpretion of statistical results. This is likely one part of the explanation of the early skepticism at the ISI meetings. Members would have to consider their organizations and users, and they might themselves have dual interests.

WWI was a mediator forcing statistical agencies to produce new statistics in a short time with limited budgets. Several studies around 1915 reported on by Jensen (1925b) consider countries food production and consumption, which were important war time issues. The postwar time further demanded new statistics with probably less resources in most European countries.

With the representative method being accepted, by necessity, in the early 1920s, the problem of quantifying the estimation error gained interest among mathematicians. It seems plausible the statisticians interpreted the inference problem differently from the mathematicians. The latter introducing randomness saw a range of possible values on the statistic out of which one would be realized. On the other hand statisticians had gone from censuses, providing "true" population values, to selection of a sample in such a way the sample based values are close to the ones in the population. Remember that stratification was used by Kiaer to account for differences among groups of the population in order to obtain a more correct value. It was not used for a reduction in standard errors of estimates.

A similar explanation can be placed on why systematic samples without randomization of the first unit are interpreted as random samples. Of the around 25 studies using random selection reported by Jensen (1925b), only one study used proper randomization. An interpretation is that if the selection of a cluster matters, your survey design has excluded important factors. Thus, with a design taking into account of all the important factors affecting what you are studying, the selection of a cluster only marginally affects your result. Without having considered the period after Neyman (1934), one can foresee an even further distance between statisticians and mathematicians with regard to statistics.

It is interesting to note that non-sampling errors were of concern already from the beginning. Non-response was an issue in several studies reported by Jensen (1925b). This problem has drastically increased, particularly in household surveys, and today in 2024 the levels of non-response threatens validity of statistics and drains NSI's resources. The trend is to abandon traditional sample surveys and adapt new approaches to produce statistics.

After Neyman (1934) mathematical statisticians were spurred to further develop the theory and have made important contributions with applications in most research areas. These developments of the theory generally assumes full response. There is also a vast literature on estimation under survey non-response. However, as stated in the following quote by Brick (2013) on the literature on non-response,

> *the central problem, in our opinion, is that even after decades of research on nonresponse we remain woefully ignorant of the causes of nonresponse at a profound level.*

This quote begs the question: With regard to household surveys, would survey statisticians of today have been better off if the representative method had stayed within the methodology of Kiaer?

References

Brick, J. Michael (2013), Unit nonresponse and weighting adjustments: A critical review, *Journal of Official Statistics*, 29:3, pp. 329-353.

Dalenius, Tore (1957), *Sampling in Sweden*, Almqvist & Wicksell, Stockholm.

Edin, Karl Arvid (1912), De mindre bemedlades bostadsförhållanden (första bostadsbeskrifningen) i Göteborg, Statistisk undersökning utförd år 1911 på uppdrag af Kommittén för stadens kommunalstatistik. Göteborg.

Edin, Karl Arvid (1916), Statistik, In: Meijer B. and T. Westrin (eds) *Nordisk familjebok*, AB Familjeboken, Stockholm, pp. 1060-1066.

Heyde, Christopher C. and Eugene Seneta (2001), *Statisticians of the Centuries*, Springer, New York.

Jensen, Adolph (1923), Arbeidsbesparende metoder i statistikken, *Nordisk Statistisk Tidskrift*, Band 2, pp. 409-434.

Jensen, Adolph (1925a), The representative method in Statistics, *Nordisk Statistisk Tidskrift*, Band 4, pp. 481-503.

Jensen, Adolph (1925b), The representative method in practice, *Nordisk Statistisk Tidskrift*, Band 4, pp. 504-566.

Jerneman, Tor (1931), Ett bidrag till de representativa undersökningarnas metodik, *Nordisk Statistisk Tidsskrift*, Band 10, pp.

Kiaer, Anders Nicolai (1897), The representative method of statistical surveys, In: Statistics Norway, (1976), Samfunnsøkonomiske Studier 27, pp. 37-56.

Kiaer, Anders Nicolai *Encyclopedia of Mathematics*. URL: http://encyclopediaofmath.org/index.php?title=Kiaer,_Anders_Nicolai&oldid=39220. Downloaded May 20, 2024.

Kruskal, William and Frederick Mosteller (1980), Representative sampling, IV: The history of the concept in statistics, 1895-1939, *International Statistical Review*, 48, pp. 169-195.

Lie, Einar (2002), The rise and fall of sampling surveys in Norway, 1875-1906, *Science in Context*, 15:3, pp. 385-409.

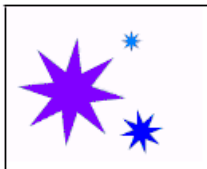
Neyman, Jerzy (1934), On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97:4, pp. 558-625.

Tschuprow, Alexander A. (1918), Zur theorie der stabilität statisticher reihen, *Skandinavisk Aktuarietidskrift*, 1, pp. 199-135.

Westergaard, Harald (1916), Scope and methods of Statistics, *Publications of the American Statistical Association*, 15:115, pp. 229-276.

Westergaard, Harald (1918), On the future of Statistics, *Journal of the Royal Statistical Society*, 81:3, pp. 499-520.

The Survey Statistician, 2024, Vol. 90, 18–29.



New and Emerging Methods

From traditional to modern machine learning estimation methods for survey sampling

Camelia Goga

Université de Franche-Comté, LMB, Besançon, France
camelia.goga@univ-fcomte.fr

Abstract

Modern parametric and nonparametric estimation methods based on machine learning are becoming increasingly popular in surveys. This paper intends presenting a synthetic review of different uses of recent modern parametric and nonparametric methods for estimating finite population totals by means of probabilistic surveys, with full data as well as with missing data.

Keywords: bagging, boosting, overfitting, penalized regression, random forests

1 Introduction

Since the first proposals for the use of probability surveys and estimation in the presence of auxiliary information in the 19th century, sample surveys have undergone significant developments. The objectives pursued, the data collected as well as their acquisition methods, and the statistical techniques used for their processing have all been deeply transformed. Over the past decade, due to the emergence of big data driven by advancements in technology and computational capabilities, we have witnessed a major transformation in this field, both in statistics in general and in sample surveys in particular, national statistical institutes being not spared from this constantly evolving reality. This paper intends giving an overview of recent machine learning methods used in survey sampling focusing on estimation and prediction issues with probability surveys.

We present in Section 2 the historical development of estimation methods in survey sampling, highlighting the major steps starting from the ratio estimator proposed during the 19th century to contemporary estimation methods, particularly non-parametric estimation methods. Section 3 focuses on modern machine learning estimation methods that have been proposed in survey sapling over the last decade. Finally, Section 4 concludes the paper and discusses new challenges associated with the use of machine learning methods in survey sampling as well as some caveats regarding the automatic use of them, including model interpretability and overfitting.

2 From parametric to traditional non-parametric estimation methods

We will consider as usual a target population U of size N. A probability sample $s \subseteq U$ is selected from U according to a sampling design $p(\cdot)$. Given $p(\cdot)$, each unit k from the population has a known inclusion probability $\pi_k = \mathbb{P}(k \in s)$ supposed to be strictly positive and a corresponding sampling design weight $d_k = 1/\pi_k$. In a survey, we are usually interested in estimating several study parameters. The simplest study parameter is the finite population total of the study variable y on U, $t_{yU} = \sum_{k \in U} y_k$. More complex study parameters such as means, ratios or quantiles as well as concentration measures (*i.e.* Gini index) may be also of interest but we devote our analysis to the finite population totals. Some discussions on more complex parameters are given in the conclusion.

With full data, the unknown total t_{yU} may be estimated by the Horvitz-Thomspon estimator (Horvitz and Thompson, 1952):

$$\hat{t}_{yd} = \sum_{k \in s} d_k y_k, \tag{1}$$

which is design unbiased, namely $\mathbb{E}_p(\hat{t}_{yd}) = t_{yU}$, where $\mathbb{E}_p(\cdot)$ is the expectation computed with respect to the sampling design $p(\cdot)$. The design-variance of $\hat{t}_{yd}$ is equal to $\mathbb{V}_p(\hat{t}_{yd}) = \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l)(y_k/\pi_k)(y_l/\pi_l)$, where $\pi_{kl} = \mathbb{P}(k, l \in s)$ is the second-order inclusion probability of units k and l in the sample. If $\pi_{kl} > 0$ for all $k, l \in U$, then the variance $\mathbb{V}_p(\hat{t}_{yd})$ may be estimated unbiasedly by $\hat{\mathbb{V}}_p(\hat{t}_{yd}) = \sum_{k \in s} \sum_{l \in s} ((\pi_{kl} - \pi_k \pi_l)/\pi_{kl})(y_k/\pi_k)(y_l/\pi_l)$. Variance and variance estimation are important issues in the analysis of survey data, as national statistical or private institutes may desire computing confidence intervals.

2.1 First use of auxiliary information

If auxiliary information is present in the sampling frame, then it can be used to construct effective sampling strategies for the estimation of t_{yU}. When the auxiliary information is available prior to sampling, we may use it to build sampling designs, such as the stratified or balanced sampling, under which the Horvitz-Thompson estimator may be highly efficient (*i.e.* with small design variance). An alternative way is to build new estimators based on such auxiliary information and exhibiting low design variance. We focus on this paper on the second approach.
One of the first estimators to use auxiliary information was the ratio estimator (Laplace, 1814) used to estimate the total number of habitants from France in 1802:

$$\hat{t}_{yrat} = t_{xU} \frac{\hat{t}_{yd}}{\hat{t}_{xd}} = \sum_{k \in s} w_{ks} y_k, \tag{2}$$

where $\hat{t}_{xd} = \sum_{k \in s} d_k x_k$ is the Horvitz-Thomspon estimator of the x-total, $t_{xU} = \sum_{k \in U} x_k$. The ratio estimator only needs the total, t_{xU}, of the univariate x-variable on U, without needing values of x for the non-sampled individuals, which is particularly interesting when the auxiliary information is accessible only in aggregate form. In addition, y_k and x_k must be available for all $k \in s$. Laplace considered as auxiliary information the number of births, with known total thanks to the national birth registers. From (2), the ratio estimator is a weighted sum of the y-values recorded for the sampled individuals with weights $w_{ks} = d_k t_{xU}/\hat{t}_{xd}$ depending only on the x-variable and independent of the study y-variable. The ratio estimator is no longer unbiased for t_{yU}, as it is a non-linear function of finite population totals, but it can be proven that it is asymptotically unbiased under mild asymptotic assumptions (Särndal et al., 1992). It is highly efficient, namely its asymptotic variance is low, if the relationship between y and x may be modeled by a straight line through the origin with the variance around the line increasing proportionally to x. Laplace had visionary ideas since the ratio estimator is

one of the most widely used estimators for a population total and many more complicated estimators are in fact based on the ratio estimator.

2.2 Traditional linear regression based estimators

Since the ratio estimator, several estimators have been suggested in order to improve the estimation of t_{yU} under a given sampling design $p(\cdot)$ by using several auxiliary variables $x_1, \ldots, x_p$. Usually, we know only the finite population total of $\mathbf{x}_k = (x_{kj})_{j=1}^p$ denoted by $t_{\mathbf{x}U}$. With multipurpose surveys, the main goal is to derive a unique system of weights w_{ks} for each unit $k \in s$ with w_{ks} independent of the study variables, making so possible the simultaneous estimation of any linear combination of totals and other finite population parameters.There are mainly two ways to incorporate auxiliary information at the estimation stage: the *model-assisted* (Särndal et al., 1992) or the *model-based* (Valliant et al., 2000) approaches if a model is considered, and the *calibration* approach (Deville and Särndal, 1992) otherwise.

We assume that the y_k values are realizations from an infinite super-population model ξ relating y_k to the vector $\mathbf{x}_k$, as follows:

$$\xi: \quad y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \quad k \in U, \tag{3}$$

where the error terms ε_k are supposed to be independent, with zero mean, $\mathbb{E}_\xi(\varepsilon_k) = 0$ and variance $\mathbb{V}_\xi(\varepsilon_k) = v_k$. The model-assisted approach is based on the generalized difference estimator (Cassel et al., 1976):

$$t_{y,\mathbf{x}}^{\text{diff}} \;=\; \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}U})^\top \boldsymbol{\beta} = \sum_{k\in s} d_k (y_k - \mathbf{x}_k^\top \boldsymbol{\beta}) + \sum_{k\in U} \mathbf{x}_k^\top \boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta}$ is the true regression coefficient. It is in fact the difference between the Horvitz-Thompson estimator $\hat{t}_{yd}$ and the bias of $\hat{t}_{yd} - t_{yU}$ under the model ξ. It can be also seen as the prediction of t_{yU} under the model ξ plus a design-bias adjustment. The unknown true $\boldsymbol{\beta}$ is estimated by design-based weighted least square criterion as follows:

$$\hat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \sum_{k\in s} d_k v_k^{-1} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta})^2. \tag{5}$$

The solution is given by $\hat{\boldsymbol{\beta}} = \left(\sum_{k\in s} d_k v_k^{-1} \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\sum_{k\in s} d_k v_k^{-1} \mathbf{x}_k y_k\right)$, assuming that the matrix $\mathbf{X}_s = (\mathbf{x}_k^\top)_{k\in s}$ is of full rank. The model-assisted estimator of t_{yU} is obtained by plugging $\hat{\boldsymbol{\beta}}$ instead of $\boldsymbol{\beta}$ in (4), we get $\hat{t}_{y,\mathbf{x}}^{\text{ma}} = \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}U})^\top \hat{\boldsymbol{\beta}} = \sum_{k\in s} d_k (y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}) + \sum_{k\in U} \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}$. For univariate x variable and variance function given by $v_k = \sigma^2 x_k, k \in U$, we get the ratio estimator (2). The widely used *poststratified* estimator of t_{yU} is obtained for $\mathbf{x}_k = (\mathbf{1}_{\{k\in U_g\}})_{g=1}^G$, where $\mathbf{1}_{\{k\in U_g\}} = 1$ if the unit k belongs to U_g and zero otherwise, $U_g, g = 1, \ldots, G$ being a partition of the population U according to some classification criterion. The variance function is supposed to be constant over the whole population, namely $v_k = v$ for all $k \in U$. The regression coefficient estimator is in this case given by $\hat{\boldsymbol{\beta}} = (\hat{\bar{y}}_g)_{g=1}^G$, where $\hat{\bar{y}}_g = \sum_{k\in s_g} d_k y_k / \hat{N}_g$, with $\hat{N}_g = \sum_{k\in s_g} d_k$ and $s_g = s \cap U_g$ for all $g = 1, \ldots, G$. The poststratified estimator reduces to the sum of the estimated predictions of y_k under the super-population model and given by $\hat{t}_{y,\mathbf{x}}^{\text{post}} = \sum_{g=1}^G N_g \hat{\bar{y}}_g$, where N_g is the size of U_g. The poststratified estimator is the sum of G ratio estimators of totals of y over the poststrata U_g, for $g = 1, \ldots, G$.
The model-based estimator of t_{yU} is built on a prediction approach,

$$t_{y,\mathbf{x}}^{\text{pred}} = \sum_{k\in s} y_k + \sum_{k\in U-s} \mathbf{x}_k^\top \boldsymbol{\beta} = \sum_{k\in s} (y_k - \mathbf{x}_k^\top \boldsymbol{\beta}) + \sum_{k\in U} \mathbf{x}_k^\top \boldsymbol{\beta}. \tag{6}$$

The unknown $\boldsymbol{\beta}$ is estimated as in (5) but without considering the sampling weights in the optimisation criterion leading to the *model-based* estimator $\hat{t}_{y,\mathbf{x}}^{\mathrm{mb}} = \sum_{k\in s}(y_k - \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}) + \sum_{k\in U} \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}$. Finally, the calibration approach consists in finding weights $(w_{ks}^{\mathrm{cal}})_{k\in s}$, such that they are as close as possible (from a pseudo distance point of view) to the sampling weights $(d_k)_{k\in s}$ while satisfying the *calibration constraints*: $\sum_{k\in s} w_{ks}^{\mathrm{cal}}\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k$. The calibration estimator is equal to the model-assisted estimator under some conditions (Deville and Särndal, 1992).

Several important properties are shared by the above three-type estimators. All estimators need only the population total of the auxiliary variables contained in the vector $t_{\mathbf{x}U}$. They may be written as weighted sums $\sum_{k\in s} w_{ks}y_k$ of sampled y-values with weights $w_{ks}, k \in s$, depending only on $\mathbf{x}_k$ recorded for sampled units and importantly, they do not depend on the y-variable. As for the ratio estimator, such weights are useful in multipurpose surveys.

2.3 Modern linear regression estimators

In the context of big data, the number p of auxiliary variables may be very large with respect to the sample size and the efficiency of estimators based on the whole set of auxiliary information may be highly deteriorated. The first issues appeared in a model-based approach since the estimators are model-dependent and many auxiliary variables may be considered to protect from model misspecification, while the model-assisted or the calibration estimators are more robust to model misspecification since they are asymptotically design unbiased and consistent for t_{yU}, whether the model is correct or not.

The weights of estimators built on the traditional linear model (3) with a large number of auxiliary variables become very instable (very large or very small) and they did not meet the predefined upper and lower range limits; they hardly satisfy a large number of calibration constraints. Finally, the design-based precision of estimators may be deteriorated when p is large with respect to the sample size, as it was noticed by Silva and Skinner (1997) by means of simulation studies and shown recently theoretically by Goga and Chauvet (2022). To correct these drawbacks, ridge-type penalized optimization criteria were suggested to relax the weight constraints (Bardsley and Chambers, 1984, Rao and Singh, 1997) leading to penalized estimators of t_{yU}. Beaumont and Bocci (2008) studied the properties of penalized calibration estimators and Guggemos and Tillé (2010) suggested a new optimisation criterion to ensure partial penalized calibration, namely a small number of important calibration equations are exactly satisfied while the other ones are approximately satisfied. As shown in Goga (2024), these penalized estimators for t_{yU} may be also obtained by considering ridge-type penalized optimization criterion to compute the regression coefficient as in classical statistics:

$$\hat{\boldsymbol{\beta}}^{\mathrm{pen}} \quad = \quad \arg\min_{\boldsymbol{\beta}\in\mathbf{R}^p} \sum_{k\in s} c_k(y_k - \mathbf{x}_k^\top\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2, \tag{7}$$

where c_k are positive constants; with $c_k = d_k$, we get the penalized calibration estimator. The solution of (7) is a ridge-type regression coefficient estimator, $\hat{\boldsymbol{\beta}}^{\mathrm{pen}} = \left(\sum_{k\in s} c_k\mathbf{x}_k\mathbf{x}_k^\top + \lambda\mathbf{I}_p\right)^{-1}\left(\sum_{k\in s} c_k\mathbf{x}_k y_k\right)$, where $\mathbf{I}_p$ is the identity matrix of size p. The ridge-type penalized estimators of t_{yU} are next obtained by plugging $\hat{\boldsymbol{\beta}}^{\mathrm{pen}}$ in (4) or (6). The resulting ridge-type estimator of t_{yU} holds the same properties as the non-penalized estimator, namely it needs only t_{xU} and it is a weighted sum of y-values, with weights not depending on the study variable. Different penalty functions in (7) lead to different penalized estimators of $\boldsymbol{\beta}$ and so, to different penalized estimators of t_{yU}. McConville et al. (2017) used the penalty $\lambda\sum_{j=1}^{p}|\beta_j|$ in (7), leading to the lasso estimator of $\boldsymbol{\beta}$ (Tibshirani, 1996) and studied the lasso-penalized estimator of t_{yU}. This penalty has the effect of shrinking the β-coefficients to zero and, unlike the ridge, it can set some of them to zero, acting as a variable selection method. Dagdoug et al. (2023b) used $\lambda[\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2]$ in (7) leading to the elastic-net estimator of $\boldsymbol{\beta}$

(Zou and Hastie, 2005), which can be viewed as a trade-off between the ridge estimator and the lasso estimator, realizing variable selection and regularization simultaneously. Alternatively, dimension reduction methods based on principal component analysis may be used to estimate $\boldsymbol{\beta}$ and to build new class of improved model-assisted or calibration estimators in presence of high-dimensional auxiliary information (Cardot et al., 2017).

The penalized estimators of t_{yU} may exhibit better efficiency than the non-penalized estimators; however, their efficiency highly depends on the tuning parameter values, which are sample data dependent. Several algorithms have been suggested to compute λ in the case of ridge regression, such as the bisection method (Beaumont and Bocci, 2008), the Fisher algorithm (Guggemos and Tillé, 2010) or simply, choosing the value for which all the weights are positive (Bardsley and Chambers, 1984, Cardot et al., 2017). More research is needed in this field.

With the emergence of smart connected objets, variables may be recorded at a very fine scale leading to another kind of high-dimensional data. The study objects are functions or curves now and called *functional data*. Cardot et al. (2013) considered the functional linear model, $y_k(t) = \mathbf{x}_k^\top \boldsymbol{\beta}(t) + \varepsilon_k(t)$, for $t \in [0, \mathcal{T}]$ and extended the model-assisted estimator to estimate the total curve of some function y over the target population. New goals and challenges appear in this new setting, such as computing global confidence bands, and Lardin-Puech et al. (2014) give a review of works related to these issues.

2.4 Traditional non-parametric model-based estimators

Estimation methods presented in sections 2.1-2.3 are all related to a linear relationship between the study variable and the auxiliary ones. Datasets are nowadays more and more complex and nonparametric models are more flexible to model the relationship between y and the x-variables:

$$\xi: \quad y_k = m(\mathbf{x}_k) + \epsilon_k, \quad k \in U,$$

where the regression function $m(\cdot)$ is unknown, but supposed to be a smooth function. Again, it was in a model-based approach that nonparametric models have been employed for the first time as a protection against model misspecification (Kuo, 1988). In model-assisted or calibration approaches, nonparametric methods have emerged later, at the beginning of the 2000's, with the seminal work of Breidt and Opsomer (2000).

With nonparametric models, we need to estimate the unknown regression function $m(\cdot)$. Traditional nonparametric methods consist in estimating $m(\cdot)$ by using kernel functions or by projecting onto a known basis function such as the truncated polynomials or the B-spline functions. Both approaches need to specify some tuning parameters, the bandwidth for kernel-based methods or the number of knots and the polynomial degree for the latter one. Once m is estimated by $\hat{m}$, the nonparametric model-assisted estimator is built from the difference estimator: $\hat{t}_{y,x}^{\text{np}} = \sum_{k\in s} d_k(y_k - \widehat{m}(\mathbf{x}_k)) + \sum_{k\in U} \widehat{m}(\mathbf{x}_k)$ and the nonparametric model-based from the prediction estimator: $\hat{t}_{y,x}^{\text{np}} = \sum_{k\in s}(y_k - \widehat{m}(\mathbf{x}_k)) + \sum_{k\in U} \widehat{m}(\mathbf{x}_k)$. As usual, the design weights are included in $\widehat{m}$ for the model-assisted case, while they are neglected for the model-based one. The nonparametric model-assisted estimators based on spline functions (Breidt et al., 2005, Goga, 2005, McConville and Breidt, 2013) inherit many desirable properties from the linear case. They may be written as traditional model-assisted estimators with explicative variables given by the basis functions, they are weighted sums of y-values with weights depending only on the x-values. However, the nonparametric estimators need x_k to be known for all the population units as we need to compute $\sum_{k\in U} \widehat{m}(\mathbf{x}_k)$. For several auxiliary variables, the additive models are the simplest way to incorporate them; for exemple, with two variables, the model is $y_k = m_1(x_{k1}) + m_1(x_{k2}) + \epsilon_k$ and $m_1(\cdot)$ and $m_2(\cdot)$ may be estimated by using one of the above method. Breidt and Opsomer (2017) give a recent review of nonparametric model-assisted

estimation techniques and Goga (2021) of B-spline nonparametric estimation methods in surveys.

In case of the traditional calibration estimator, the underlying relationship between y and $\mathbf{x}_k$ is implicitly a linear one, so it is not adapted to account for nonlinear relationships. Montanari and Ranalli (2005) suggest the *nonparametric model-calibration* estimator, which consists in finding weights satisfying $\sum_{k\in s} w_{ks}\hat{m}(x_k) = \sum_{k\in U} \hat{m}(x_k)$, while Goga (2021) suggests calibrating directly on the basis functions (B-spline functions) instead of the regression function estimator, allowing in this way to obtain weights not depending on the y-variable, property not owned by the nonparametric model-calibration.

3 Modern non-parametric estimation methods

The traditional nonparametric models based on kernel or spline smoothing are relatively easy to use and efficient if the number of auxiliary variables is low, at most three or four. As the number p of auxiliary variables is becoming large, these models tend to breakdown as they need extremely large sample sizes, phenomenon known as the curse of dimensionality. Moreover, additive models do not account for interactions between the x-variables. Semiparametric models (Breidt et al., 2007), containing linear terms as well as nonlinear terms, may be used as a tradeoff between completely parametric and nonparametric models. Alternately, the K-nearest neighbors may be used as it is a simple nonparametric method which can be used with multivariate auxiliary information (Baffeta et al., 2010).

Modern machine learning algorithms, as suggested lately in the statistical literature, are nonparametric methods which can handle easily a large number of auxiliary variables. Broadly speaking, these methods may be classified into two classes as they are based on *bagging* or on *boosting* (Hastie et al., 2011). Bagging produces a large number B of predictions and combines them to produce more accurate predictions than a single model would do:

$$\widehat{m}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{m}^{(b)}(\mathbf{x}), \tag{8}$$

where $\widehat{m}^{(b)}$ is the prediction of m obtained by some nonparametric method. Any nonparametric method may be used to obtain $\widehat{m}^{(b)}$, however bagging is particularly interesting for regression trees. To obtain B different models, initial dataset is bootstrapped (with replacement). Boosting works differently, it starts with a weak fit (or learner) and improves it at each step of the algorithm by predicting the residuals of prior models and adding them together to make the final prediction:

$$\widehat{m}^{(b)}(\mathbf{x}) = \widehat{m}^{(b-1)}(\mathbf{x}) + \widehat{m}(\mathbf{x}, \varepsilon^{(b-1)}), \quad b = 1, \ldots, B,$$

where $\widehat{m}(\mathbf{x}, \varepsilon^{(b-1)})$ is the prediction based on data $\mathbf{x}_k$ and the residuals $\varepsilon_k^{(b-1)} = y_k - \widehat{m}^{(b-1)}(\mathbf{x}_k)$ computed from the previous model. While B should be very large to improve the efficiency of bagging methods, it should be small for boosting to avoid overfitting. To cope with overfitting issues, a large value of B is considered and a penalty term is added in the boosting algorithm (Hastie et al., 2011).

3.1 Tree-based estimation methods

Regression trees based on CART algorithm as suggested by Breiman et al. (1984) are simple to use in practice and useful for interpretation. Toth and Eltinge (2011) studied the asymptotic behavior of regression trees for survey data and McConville and Toth (2019) used them in a model-assisted context. Regression tree prediction of x in some point $\mathbf{x}$ is obtained in two steps. The predictor space spanned by the x-variables measured on data is partitioned, according to some criterion, into

A_j, $j = 1, \ldots, J$ disjointed zones called *terminal nodes* and the unknown regression function in a point $\mathbf{x}$ is approximated as $m(\mathbf{x}) \simeq \beta_1 \mathbf{1}_{\{\mathbf{x} \in A_1\}} + \ldots + \beta_J \mathbf{1}_{\{\mathbf{x} \in A_J\}}$. The β-coefficients are estimated with survey data by weighted least-square criterion (McConville and Toth, 2019) leading to $\widehat{\beta}_j = \sum_{k \in A_j} d_k y_k / \hat{N}_j = \hat{\bar{y}}_j$, where $\hat{N}_j = \sum_{k \in A_j} d_k$. For every unit from A_j, tree-based predictions are the same and equal to the weighted mean of y-values of units belonging to A_j, namely $\widehat{m}(\mathbf{x}) = \widehat{\beta}_j$ for all $\mathbf{x} \in A_j$. The tree model-assisted estimator $\hat{t}_{y,\mathbf{x}}^{\text{tree}}$ of t_{yU} is obtained by plugging $\widehat{m}(\mathbf{x}_k)$ in (4). As $(A_j)_{j=1}^J$ is a partition of the predictor space, then $\sum_{k \in s} d_k (y_k - \widehat{m}(\mathbf{x}_k)) = 0$ leading to $\hat{t}_{y,\mathbf{x}}^{\text{tree}} = \sum_{k \in U} \widehat{m}(\mathbf{x}_k) = \sum_{j=1}^J N_j \hat{\bar{y}}_j$, which is a poststratified-type estimator with random poststrata A_j of size N_j built on the sample data $(\mathbf{x}_k, y_k)_{k \in s}$ (see section 2.2 for the traditional poststratified estimator).

To determine the terminal nodes, we may use the greedy CART algorithm (Breiman et al., 1984) which recursively searches for the splitting variable and the splitting position (i.e., the coordinates on the predictor space where to split) leading to the greatest possible reduction in the residual sum of squares. More exactly, let $\mathcal{C}_A$ be the set of all possible pairs (j, z) in A and $A_L(j, z) = \{\mathbf{x}_k \in A; x_{kj} < z\}$, $A_R(j, z) = \{\mathbf{x}_k \in A; x_{kj} \geq z\}$. The best split (j^*, z^*) in a region A is $(j^*, z^*) = \arg\min_{(j,z) \in \mathcal{C}_A} \{\sum_{k \in s: \mathbf{x}_k \in A_L(j,z)} (y_k - \bar{y}_{A_L})^2 + \sum_{k \in s: \mathbf{x}_k \in A_R(j,z)} (y_k - \bar{y}_{A_R})^2\}$, where $\bar{y}_{A_L}$ (respectively $\bar{y}_{A_R}$) is the average of the y-values of units belonging to the node $A_L(j, z)$ (respectively $A_R(j, z)$). The procedure continues until a stopping criterion is reached. The random non-overlapping regions obtained by the CART algorithm depend on the sample data $(\mathbf{x}_k, y_k), k \in s$. Nalenz et al. (2024) suggest a CART criterion based on design-based estimation of the residual sum of square and Beaumont et al. (2024) adapt the CART criterion to a classification problem and data integration issues by considering well-chosen stopping criteria.

Regression trees are simply to use and interpret, however they are not appropriate with high-dimensional data and deep trees are known to have large variance and to lead to overfitting. The random forest algorithm (Breiman, 2001) is an ensemble method that corrects the tree defaults by a large number of randomized deep decorrelated trees. More exactly, the random forest prediction of $m(\cdot)$ is a bagging estimator as in (8), where each $\widehat{m}^{(b)}(\cdot)$ is a regression tree prediction of $m(\cdot)$ built on a bootstrap sample data and selecting randomly at each split in the tree a new set of p_0 auxiliary variables from the p initial variables. In this way, a fresh set of variables is considered at each tree split. Random forests are very popular methods due to their predictive performances and ability to handle large data sets, however their theoretical properties have been proved only recently for particular algorithms (Scornet et al., 2015, Klusowski and Tian, 2024). Random forest algorithms have been only recently used with survey data starting with Tipton et al. (2013), Buskirk and Kolenikov (2015) for missing data, De Moliner and Goga (2018) for small area estimation. Very recently, Dagdoug et al. (2023c) suggested the random forest model-assisted estimator and studied its asymptotical properties. The random forest model-assisted estimator may be also written as a weighted sum of y-values, however the weights depend now on the study variable as the partitions are built on sample data $(\mathbf{x}_k, y_k), k \in s$. With multipurpose surveys, the user has the choice between two options: use random forest algorithms not depending on the study variable (Devroye et al., 2013) or use a model-calibration procedure as suggested in Dagdoug et al. (2023c) to determine weights for estimating simultaneously several totals. Dagdoug et al. (2023c) use a without replacement bootstrap resampling procedure and they show that the random forest model-assisted estimator can be written as the total of the estimated prediction of $m(\cdot)$ plus a correction term equal to the weighted sum of residuals computed for the non-resampled units, also called the *out-of-bag* individuals, from each of the B trees. This correction term brings additional information from the units not used in computing the prediction and preventing in this way from overfitting. Nalenz et al. (2024) suggest bootstrapping individuals with unequal probabilities.

The random forest algorithms depend on several hyper parameters: the number B of trees, the number p_0 of the selected variables and the number n_0 of individuals from the terminal nodes, the

hyper parameter values affecting the precision of the model-assisted estimators and finding the best values of such hyper parameters may be difficult with complex sampling designs (Dagdoug et al., 2023c,b). Another issue with random forests, and even with nonparametric methods in general, is the result interpretability. These methods are known to have high predictive performances, however the predictions are difficult to interpret and this may be a problem with surveys conducted by national institutes.

3.2 Missing data

The estimators presented above supposed that all the sampled individuals respond, so we have complete sample data $y_k, k \in s$. In practice however, due to various reasons, some individuals respond only partially (*item nonresponse*) or do not respond to the survey questionnaire (*unit nonresponse*). Item nonresponse is treated by imputation while unit nonresponse is treated by weighting methods.

With item nonresponse, the imputed estimator $\hat{t}_I$ of t_{yU} is obtained from the Horvitz-Thompson estimator given in (1) by replacing the missing values y_k by predicted values $\hat{y}_k$, $\hat{t}_I = \sum_{k\in s_r} d_k y_k + \sum_{k\in s_m} d_k \hat{y}_k$, where s_r is the respondent subset and s_m, the subset of s containing the nonrespondents. The imputed values are obtained by fitting an imputation model, assuming usually that the response mechanism is MAR (*missing at random*). Recently, Dagdoug et al. (2023a) conduct a large simulation study of parametric as well as nonparametric and machine-learning imputation procedures in terms of bias and efficiency in a wide variety of settings, including high-dimensional data sets. They considered methods such as B-spline additive model and K-nearest neighbor as well as regression trees, random forest and boosting with the XGBoost algorithm (Chen and Guestrin, 2016), Bayesian additive regression trees (Chipman et al., 2010), cubist algorithm (Quinlan, 1993). Their simulation results show that, in general, the non-parametric imputation models are superior to parametric models to capture the non-linear trend in the data. However, in high-dimension settings (*i.e.* a large number p of auxiliary variables), the K-nearest neighbor or the additive models are out-performed by machine learning methods which are more robust in such contexts. Dagdoug et al. (2024) study the asymptotic properties of the regression tree and the random forest imputed estimator.

With unit nonresponse, the weighted estimator of t_{yU} is $\hat{t}_w = \sum_{k\in s} d_k y_k/\hat{p}_k$, where $\hat{p}_k$ is the estimated response probability of unit k. The response probabilities p_k may be estimated by parametric logistic regression, or through nonparametric regression. Da Silva and Opsomer (2006) studied the kernel smoothing and Da Silva and Opsomer (2009) extended it to local polynomial regression. Very recently, Larbi et al. (2023) make a large simulation study of nonparametric and machine learning methods for estimating the response probabilities.

3.3 Variance estimation

Variance estimation with survey data is a very important but difficult issue. In a design-based approach, the variance of estimators derived under the sampling design is desired in order to deduce next estimated confidence intervals. All estimators presented in this paper are non linear estimators, so their variances are not computable and in the best case, only asymptotic variances may be deduced. Using linearization techniques and adapted asymptotic framework including assumptions on the sampling design, the study and the auxiliary variable, the asymptotic variance of model-assisted or calibration estimators are equal to Horvitz-Thompson variance applied to residuals $y_k - m(\mathbf{x}_k)$, $A\mathbb{V}(\hat{t}^{\mathrm{ma}}_{y\mathbf{x}}) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)((y_k - m(\mathbf{x}_k))/\pi_k)((y_l - m(\mathbf{x}_l))/\pi_l)$ and estimated by the Horvitz-Thompson variance estimator applied to estimated residuals $\widehat{\epsilon}_k = y_k - \widehat{m}(\mathbf{x}_k), k \in s$. With nonparametric methods, overfitting usually happens leading to underestimated residuals $\widehat{\epsilon}_k, k \in s$, so confidence intervals based on such variance estimator will not have the desired rate. This issue was already raised by Opsomer and Miller (2005) in the context of local polynomial regression. To

cope with this issue, Dagdoug et al. (2023c) suggested a variance estimator based on a K-fold cross-validation criterion, widely used in machine learning community for determining for example the tuning hyper parameters. More specifically, the sample s is split randomly into K groups $s_\kappa, \kappa = 1, \ldots, K$, of approximately equal size. For $k \in s_\kappa$, let $\widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ denote the prediction at the point $\mathbf{x}_k$ fitted on $s - s_\kappa$ and $\widehat{\epsilon}_k^{(-\kappa)} = y_k - \widehat{m}^{(-\kappa)}(\mathbf{x}_k)$ the associated residual. The proposed K-fold variance estimator is given by $\widehat{\mathbb{V}}^{(K)} = \sum_{\kappa_1=1}^{K} \sum_{\kappa_2=1}^{K} \sum_{k \in s_{\kappa_1}} \sum_{l \in s_{\kappa_2}} ((\pi_{kl} - \pi_k \pi_l)/\pi_{k\ell})(\widehat{\epsilon}_k^{(-\kappa_1)}/\pi_k)(\widehat{\epsilon}_l^{(-\kappa_2)}/\pi_l)$. In practice, the number of groups (or folds) is often set to $K = 5$ or $K = 10$. This variance estimation procedure allowed to greatly improve the symmetric confidence interval rates for the random forest model-assisted estimator and Dagdoug et al. (2024) adapted the method to account for the item-nonresponse.

4 Conclusion

This paper provides a synthetic presentation of estimation methods for totals in surveys, as their objectives and type of data have evolved over time. Modern machine learning methods are becoming increasingly popular in surveys. However, their automatic use in survey sampling comes with its own set of challenges and caveats, including concerns about model interpretability, overfitting, and bias amplification. The implementation of such modern estimation methods is not straightforward with complex sampling designs. Most of machine learning algorithms have been implemented for non survey data and they do not allow considering the sampling weights in the predictions $\widehat{m}(\cdot)$, leading to potentially biased estimators for unequal and complex survey designs (Dagdoug et al., 2023b). As such, it is essential for researchers and users to exercise caution and rigor in applying these techniques and to complement them with traditional estimation methods to ensure the validity and reliability of survey estimates. Many research perspectives open up: the choice of hyper-parameters to use in machine learning algorithms, the bootstrapping of individuals or the estimation of variance are really important questions that need to be explored further more deeply. The estimation of non-linear study parameters with high-dimensional auxiliary information may be also of interest, however there is little research on this field. Goga and Ruiz-Gazen (2014) used B-spline nonparametric estimation for nonlinear functions such as median, Gini index but we are not aware of use of machine learning methods for the estimation of such parameters. This paper treated the estimation issues with probabilistic samples. Non-probabilistic surveys are used more and more often nowadays and recent works started treating estimation issues with such samples by using also machine learning methods. Another item not treated in this paper is the use of machine learning methods for small-area estimation, the reader is referred to the excellent paper of Krennmair et al. (2022) for a review on this area.

References

Baffeta, F., Corona, P. and Fattorini, L. . (2010). Design-based diagnostics for k-nn estimators of forest resources. *Canadian Journal of Forest Research*, 41:59–72.

Bardsley, P. and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33:290–299.

Beaumont, J.-F. and Bocci, C. (2008). Another look at ridge calibration. *Metron-International Journal of Statistics*, LXVI:260–262.

Beaumont, J. F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *To appear in Survey Methodology*.

Breidt, F., Claeskens, G. and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92:831–846.

Breidt, F. J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, pages 190–205.

Breidt, F. J., Opsomer, J. D., Johnson, A. A. and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, 33:35–44.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Advanced Books and Software, Belmont, CA., MR0726392.

Buskirk, T. D. and Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*, pages 1–17.

Cardot, H., Goga, C. and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7:562–596.

Cardot, H., Goga, C. and Shehzad, M.-A. (2017). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *Statistica Sinica*, 27(243-260).

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

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*. ACM Press.

Chipman, H., George, E. and McCulloch, R. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Da Silva, D. N. and Opsomer, J. D. (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. *Canadian Journal of Statistics*, 4:563–579.

Da Silva, D. N. and Opsomer, J. D. (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. *Survey Methodology*, 35(2):165–176.

Dagdoug, M., Goga, C. and Haziza, D. (2023a). Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: an Empirical Comparison. *Journal of Survey Statistics and Methodology*, 11(1):141–188.

Dagdoug, M., Goga, C. and Haziza, D. (2023b). Model-assisted estimation in high-dimensional settings for survey data. *Journal of Applied Statistics*, 50(3):761–785.

Dagdoug, M., Goga, C. and Haziza, D. (2023c). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 118(542):1234–1251.

Dagdoug, M., Goga, C. and Haziza, D. (2024). Statistical inference in the presence of imputed survey data through regression trees and random forests. *Submitted*.

De Moliner, A. and Goga, C. (2018). Sample-based estimation of mean electricity consumption curves for small domains. *Survey Methodology*, 44(2):193–214.

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

Devroye, L., Györfi, L. and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *Canad. J. Statist.*, 33(2):163–180.

Goga, C. (2021). B-spline estimation in a survey sampling framework. In Daouia, A. and Ruiz-Gazen, A., editors, *Advances in Contemporary Statistics and Econometrics, Festschrift in Honor of Christine Thomas-Agnan*, pages 79–99. Springer.

Goga, C. (2024). High-dimensional estimation in a survey sampling framework, model-assisted and calibration points of view. *Accepted for publication in Metron*.

Goga, C. and Chauvet, G. (2022). Asymptotic efficiency of the calibration estimator in a high-dimensional data setting. *Journal of Statistical Planning and Inference*, 217:177–187.

Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76:113–140.

Guggemos, F. and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *J. of Statistical Planning and Inference*, 140:3199–3212.

Hastie, T., Tibshirani, R. and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

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

Klusowski, J. and Tian, P. (2024). Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119:525–537.

Krennmair, P., Wurz, N. and Schmid, T. (2022). Tree-based machine learning in small area estimation. *The Survey Statistician*, 86:22–31.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In Association, A. S., editor, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 280–285.

Laplace, P. (1814). Essai philosophique sur les probabilités. *Paris: MME VE Courcier, Imprimeur-Libraire pour les Mathématiques, quai des Augustins, no. 57*.

Larbi, K., Tsang, J., Haziza, D. and M., D. (2023). Treatment of unit nonresponse in surveys through machine learning methods: an empirical comparison. *Submitted*.

Lardin-Puech, P., Cardot, H. and Goga, C. (2014). Analysing large datasets of functional data: a survey sampling point of view. *Journal de la Société Française de Statistique*, 155(4):70–94.

McConville, K. and Breidt, F. J. (2013). Survey design asymptotics for the model-assisted penalized spline regression estimator. *Journal of Nonparametrics Statistics*, 25:745–763.

McConville, K. and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2):389–413.

McConville, K. S., Breidt, F. J., Lee, T. C. and Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2):131–158.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *J. Amer. Statist. Assoc.*, 100:1429–1442.

Nalenz, M., Rodemann, J. and Augustin, T. (2024). Learning de-biased regression trees and forests from complex samples. *Machine learning*, https://doi.org/10.1007/s10994-023-06439-1.

Opsomer, J. and Miller, C. (2005). Selecting the amount of smoothing in nonparametric regression estimation for complex surveys. *Nonparametric Statistics*, 17(5):593–611.

Quinlan, J. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning*, pages 236–243.

Rao, J. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741.

Silva, P. and Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23:23–32.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tipton, J., Opsomer, J. and Moisen, G. (2013). Properties of endogenous post-stratified estimation using remote sensing data. *Remote sensing of environment*, 139:130–137.

Toth, D. and Eltinge, J. L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496):1626–1636.

Valliant, R., Dorfman, A. and Royall, R. M. (2000). *Finite Population Sampling and Inference*. Wiley Series in Probability and Statistics.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

The Survey Statistician, 2024, Vol. 90, 30–37.

The Use of New Data Sources in Small Area Estimation of Attitudes towards Climate Change

Camilla Salvatore[1] **and Angelo Moretti**[2]

[1]Utrecht University, The Netherlands, c.salvatore@uu.nl
[2]Utrecht University, The Netherlands, a.moretti@uu.nl

Abstract

Climate change is a global problem that has a significant impact on the world's economy and society. To effectively address climate change and other societal challenges, policymakers often require reliable estimates of relevant variables at a sub-national level. Nationally representative surveys are not often designed for this purpose. In this study we propose to use small area estimation techniques to obtain reliable estimates of the proportion of people very and extremely worried about climate change at regional level. A novel aspect of our approach is that we include non-traditional auxiliary information, specifically web data, into our model. For the data used in this paper, our results show that incorporating web data yields more reliable estimates than the model without them. Finally, we also acknowledge and address certain limitations associated with the use of web data in small area estimation.

Keywords: Digital Trace Data, Data Integration, Attitudes, Fay-Herriot model

1 Introduction

Climate change is one of the greatest challenges of the present century, with consequences for ecosystems, the economy, and society [Lee et al., 2023]. This global issue has promoted collaborative cross-national efforts, such as the Paris Agreement, which provides a common framework to combat climate change and mitigate its impact. Furthermore, in 2020, the European Union (EU) approved the European Green Deal, a comprehensive strategy with the goal of achieving climate neutrality by 2050.

As governments worldwide engage in policies concerning climate change and sustainable development, the necessity for high-quality data to monitor these processes and the public opinion becomes increasingly important. Beyond environmental metrics, understanding societal attitudes and behaviours towards climate change is necessary for developing effective strategies taking into account the perspectives of the local communities [Prakash and Bernauer, 2020].

Large-scale nationally representative survey data are the base for the construction of such indicators. They are designed to produce precise and accurate estimates for large population domains, e.g., at country-level. However, policymakers and researchers are often interested in sub-national indicators, i.e., at regional or province-level. Direct estimates obtained for these areas may return large variability

to small sample sizes [Rao and Molina, 2015]. A popular approach to derive reliable estimates at sub-national level is Small Area Estimation (SAE).

The key idea of SAE models is to "*borrow strength*" from the other areas and auxiliary information from nationally representative surveys, administrative data or non-traditional data sources.

In recent years, the integration of digital trace data (e.g. from websites, social media, google trends) with survey data has gained importance [Salvatore, 2023]. In SAE the use of these non-traditional sources is very promising because this data can provide additional and relevant information that characterize the small-area of interest. Some examples in SAE include the use of Twitter (now X) data to improve the estimation of food consumption expenditure and the use of Google trends data to estimate relative changes in rates of household Spanish-speaking in the United States [Marchetti and Schirripa Spagnolo, 2024, Marchetti et al., 2015, 2016, Porter et al., 2014].

In this study, our focus is to estimate attitudes towards climate change at a regional level (NUTS2). We present some preliminary results on one country only, Spain, with plans to extend the analyses to more countries in future work. To do that, we utilise the European Social Survey (ESS) data and we apply a SAE model, named Fay-Herriot (FH). As an innovative methodological element in our analyses, we consider auxiliary information in the model, from both traditional data sources (Eurostat archive) and non-traditional ones (web data from Booking.com).

The aim of this paper is twofold. Firstly, given the importance of climate change in the current political and social debate, the aim of this paper is to show how reliable indicators of attitudes towards climate change at a local level can be obtained through SAE. Secondly, we study whether the use of new data sources can be beneficial in the estimation process.

The remainder of this article is structured as follows. In Section 2 we present the data and the modeling strategy. In Section 3 we discuss the results and we draw conclusions in Section 4.

2 Data and Methods

2.1 The European Social Survey

We employ data from ESS round 10. ESS is a nationally representative European cross-national survey that has been running every two years since 2001 [European Social Survey, 2022]. The survey collects data for a large umber of countries in Europe, however, in this article, we focus on Spain only. We consider the NUTS2 level of classification, which pertains to 17 autonomous regions and 2 autonomous cities (Melilla and Ceuta). The latter areas are from our analyses since they did not have all the auxiliary variables available. This results in the selection of 17 autonomous regions for a total sample size of 2214 units. The first row in Table 1 presents descriptive statistics for the regional sample sizes. It is evident that some regions present very small sample sizes, thus, they do not allow for reliable direct estimates. Indeed, the ESS is not designed to produce accurate and precise estimates at the sub-national level [Moretti and Whitworth, 2020, Santi and Moretti, 2021], hence SAE methods are needed.

We focus on a specific indicator, i.e., the proportion of individuals who are very and extremely worried about climate change. The survey participants were asked to answer this question: "How worried are you about climate change?" on a 5-point scale (1 not at all worried, 2 not very worried, 3 somewhat worried, 4 very worried, and 5 extremely worried). We dichotomies the variable following the median split method (median = 4). Thus, we re-code the variable as "very and extremely worried" (score equal to 4 or 5) versus "other" (score from 1 to 3). The weighted proportion at the national level of people very concerned about climate change is 56.40%. The second and third rows in Table 1 show descriptive statistics of the direct estimates and the coefficient of variation (CV) across the 17

regions. As a rule of thumb, a CV larger than 16% is considered non-reliable, hence, SAE is needed [Marchetti and Schirripa Spagnolo, 2024].

Table 1: Summary statistics of regional sample sizes, direct regional estimates and their CVs

	Min.	Median	Mean	Max.
Sample Size	21	73	130	368
Direct estimate	41.66	57.37	54.97	69.58
CV%	6.71	26.27	16.02	31.20

2.2 The auxiliary data

In the FH model, we use auxiliary data from both traditional sources, i.e., official statistics, and also from non-traditional data sources, specifically web data.

In terms of traditional variables, we consider the following ones: proportion of people with tertiary education, long term unemployment rate, robbery rate and population density. These variables show a good spatial heterogeneity at regional level in Europe, and they were used in public attitudes context before [Santi and Moretti, 2021, Moretti and Whitworth, 2020]. These variables can be obtained by Eurostat data archive[1].

As web data, we consider information about sustainable hotels available on Booking.com. Until recently, the *Travel Sustainable*[2] programme included four different labels that could be assigned to hotels that satisfied some sustainability requirements: namely, from the lowest to the highest, Level 1, Level 2, Level 3 and Level 3+. However, in March 2024 the programme has undertaken significantly changes and now only one label, named *Sustainability certification*, is available. The data we use for our analysis was scraped prior to the change, thus, in this article we use the old classification system. In the concluding remarks, we discuss this issue further. The rational for using the proportion of hotels with a specific label is the following. The number of hotels with an environmental certification can give valuable insights into the spatial socio-environmental context of different areas, in particular with respect to the environmental awareness and sustainability practices. Thus, this can help in estimating attitudes towards climate change at a local level.

In order to obtain the data we perform web-scraping using R [R Core Team, 2021] and the *rvest* package [Wickham, 2024]. Our research query relates only 1 room for 1 adult. For each region, we gather data on the total number of hotels and we calculate the proportions of hotels categorized under different sustainability levels. Due to fluctuations in hotel availability over time, we randomly select 84 dates, roughly 7 days per month, spanning form February 1st 2024 and January 31st 2025. Thus, we proceed by averaging the proportions across the 84-day period for each region. Across all regions, the proportion of hotels with label Level 1 is 17.2%, with Level 2 is 9.26%, with Level 3 is 2.74% and with Level 3+ is 2.11%.

2.3 The Fay-Herriot model

We assume that we have a finite population P with dimension N partitioned into $d = 1, ..., D$ disjoint small areas. N_d is the population dimension in area d, thus $\sum_{d=1}^{D} N_d = N$. From P a random sample s with dimension n is selected, $\sum_{d=1}^{D} n_d = n$, where n_d denotes the sample size in area d. We are interested in estimating the mean of a variable denoted by Y, denoting worry about climate change, for area d, and this is denoted by $\bar{Y}_d$. A direct estimator for this is $\hat{\bar{Y}}_d^{DIR} = \sum_{i=1}^{n_d} y_{di} w_{di} / \sum_{i=1}^{n_d} w_{di}$,

[1] https://ec.europa.eu/eurostat/web/regions/database

[2] see https://news.booking.com/en/bookingcom-celebrates-one-year-of-travel-sustainable-with-new-product-features-for-accommodations-rental-cars-and-flights/

where w_{di} denotes the survey weight for unit i in area d. However, in case of small area sample sizes this will be unreliable [Rao and Molina, 2015]. In article, we apply the Fay-Herriot model in order to provide accurate and precise estimates of our study phenomena [Fay and Herriot, 1979]. This model consists in a sampling model:

$$\hat{\bar{Y}}_d^{DIR} = \bar{Y}_d + e_d, d = 1, ..., D \tag{1}$$

where e_d is the sampling error of the direct estimator, and a linking model:

$$\bar{Y}_d = \bar{\mathbf{X}}_d^T \beta + u_d, d = 1, ..., D, \tag{2}$$

and combining the two we obtain the following:

$$\hat{\bar{Y}}_d^{DIR} = \bar{\mathbf{X}}_d^T \beta + u_d + e_d, d = 1, ..., D, \tag{3}$$

where $u_d \sim N(0, \sigma_u^2)$ and $e_d \sim N(0, \sigma_{e_d}^2)$, with $\sigma_{e_d}^2$ (variance of the direct estimates) is assumed to be known. $\bar{\mathbf{X}}_d$ are the auxiliary variables (e.g., area level means) for area d. The Empirical Best Linear Unbiased Predictor (EBLUP) of $\bar{Y}_d$ under model 3 is given by [Fay and Herriot, 1979]:

$$\hat{\bar{Y}}_d^{EBLUP,FH} = \hat{\gamma}_d \hat{\bar{Y}}_d^{DIR} + (1 - \hat{\gamma}_d) \bar{\mathbf{X}}_d^T \hat{\beta}, \tag{4}$$

where $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{e_d}^2}$ is the shrinkage factor. Thus, when the n_d is small more weight will be given to the direct estimates, since they will reliable, and vice-versa. This optimises the trade-off between large variance of the direct estimator $\hat{\bar{Y}}_d^{DIR}$ and bias of the synthetic estimator $\bar{\mathbf{X}}_d \hat{\beta}$. The Mean Squared Error (MSE) of 4 can be estimated following Prasad and Rao [1990]. In this article, we use the Maximum Likelihood (ML) estimation technique to obtain the FH model estimates [Marhuenda et al., 2014].

3 Results

3.1 Modelling strategy

In order to provide model-based estimates of the proportion of individuals who is very worried and extremely worried about climate change, based on the FH model, we consider two scenarios: 1) only traditional variables from the Eurostat archive are available, and 2) additional variables are derived from web data. We address model selection and estimation issues in both scenarios separately and then we compare the results obtained in the two scenarios. We begin with a full model that includes all available covariates and we use a stepwise algorithm[3] to identify which covariates include in the FH model. To compare models, we employ the Kullback symmetric divergence (KICb2) criterion. This information criterion, developed for FH models, is used to assess dissimilarities between two statistical models and the model with the lowest KICb2 value should be preferred [Marhuenda et al., 2014]. Thus, the model with the lowest KICb2 value is selected for analysis in both scenarios.

3.1.1 Scenario 1: only traditional data

In this scenario, only the traditional data from the Eurostat data archive are available (see Section 2.2). Table 2 shows the results of model selection. The model with the lowest KICb2 value is selected, i.e., the one with long term unemployment (LTU) and proportion of people with tertiary education (t.edu).

[3] see the step function in the emdi R package Harmening et al. [2023]

Table 2: Model selection results based on the KICb2 for Scenario 1 (traditional data only, i,e, ESS and Eurostat Archive).

Predictors	**KICb2**
t. edu., LTU, robbery, pop. density	-8.24
t. edu., LTU, pop. density	-16.46
LTU, t. edu.	**-21.63**

3.1.2 Scenario 2: traditional and web data

In Scenario 2, in addition to the traditional variables included in Scenario 1, we also consider web data. The web data refers to the proportion of hotels with a specific environmental label level from Booking.com (see Section 2.2). Table 3 shows the values of the KICb2 metrics. We select the model with the lowest KICb2 value, which is the model that includes long term unemployment (LTU), population density (pop. density), and the proportion of hotels with level 1 and level 3 labels.

Table 3: Model selection results based on the KICb2 for Scenario 2 (traditional data, i.e, ESS and Eurostat data supplemented by web data from Booking.com).

Predictors	**KICb2**
t. edu., LTU, robbery, pop. density, level 1, level 2, level 3, level 3 plus	-10.42
t. edu., LTU, pop. density, level 1, level 2, level 3, level 3 plus	-14.25
LTU, pop. density, level 1, level 2, level 3, level 3 plus	-17.81
LTU, pop. density, level 1, level 3, level 3 plus	-22.05
LTU, pop. density, level 1, level 3	**-25.26**

3.2 Diagnostics

In order to evaluate whether we introduce bias in the final regional estimates produced by the FH models, we perform some diagnostics measures in both scenarios. We implement the Brown test [Brown et al., 2001] in order to evaluate the quality of the EBLUPs. For this the Wald statistics is used, with null hypothesis being the EBLUP estimates do not differ significantly from the direct estimates:

$$W = \sum_{d=1}^{D} \frac{(\hat{\bar{Y}}_d^{DIR} - \hat{\bar{Y}}_d^{EBLUP,FH})^2}{\hat{var}(\hat{\bar{Y}}_d^{DIR}) + \hat{MSE}(\hat{\bar{Y}}_d^{EBLUP,FH})} \tag{5}$$

This is approximately distributed as a χ^2 with D (number of areas) degrees of freedom, under the null hypothesis.

In Scenario 1 the correlation between synthetic part and direct estimator is 0.19, however, the EBLUP estimates do not differ significantly from the direct estimates (W=9.38, df=17 and p-value=0.93). In Scenario 2 the correlation between synthetic part and direct estimator is larger, i.e., equal to 0.62 and the EBLUP estimates do not differ significantly from the direct estimates (W=8.36, df=17 and p-value=0.96). According to these results, Scenario 2 should be preferred given the higher correlation between the two sets of estimates.

3.3 Comparing results

Figure 1 shows the the CV% and percentage Relative Root Mean Squared Error (RRMSE) % of the direct estimates and EBLUPs, respectively, across the regions. The RRMSE is defined as the ratio between the squared root of the MSE and the EBLUP. Here, we compare the direct estimates (Direct)

to the EBLUPs with (EBLUP with web data - Scenario 2) and without the (EBLUP with Traditional Data - Scenario 1) use of web data.

We can see that the use of auxiliary information from web data helps reducing the RRMSE% of the estimates considerably, for the areas with small sample size (i.e. less than 200 units). This gain is larger compared to the use of the model with traditional data only.

It can be seen that, in case of large regional sample sizes the RRMSE% estimates of the EBLUP with traditional data only are similar to the CV% of the direct estimates. On the contrary, the RRMSE% of the EBLUP obtained using web data is higher than the CV% of the direct estimates. However, in this case, the direct estimates are reliable due to very small CV%, thus, these should be used in practice.

Given the RRMSE results discussed above and the diagnostics results presented in the previous section, we consider Scenario 2 when mapping the estimates.

In Figure 2 we map the regional estimates for NUTS2 level in Spain obtained via the EBLUP approach and the following auxiliary variables LTU, pop. density, level 1, and level 3. It can be seen that areas with a larger presence of tourists and especially coastal regions, show larger level of worries about climate change. This is also valid for Madrid region. The region with the lowest value of the indicator is Balearic islands, followed by Castile-La Mancha and Castile-León.

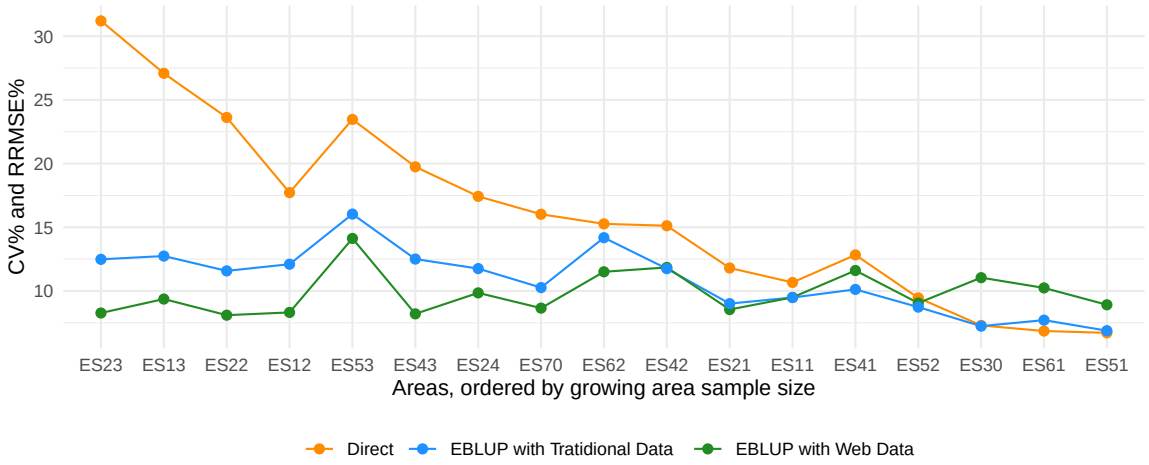


Figure 1: CV% (for the direct estimates) and RRMSE% (for the EBLUPs considering the two scenarios) of the regional estimates of worry about climate change for NUTS2 level in Spain.

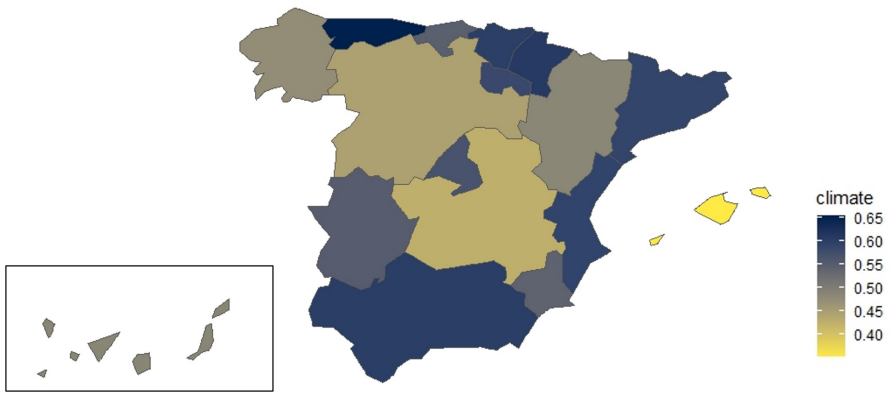


Figure 2: Regional Estimates NUTS2 level Spain of worry about climate change.

4 Concluding remarks

To effectively address climate change and other societal challenges, policymakers often require reliable estimates of relevant indicators at a sub-national level. Nationally representative surveys are not designed for this purpose. In this article, we discuss how SAE can address this issue. We construct a SAE model to estimate the proportion of people very and extremely worried about climate change at regional level in Spain. We employ both traditional and non-traditional (web data) variables as auxiliary information. Our results demonstrate that incorporating web data yields more reliable estimates than the ones produced by the model without those variables. Thus, our empirical analyses highlight the opportunity of using non-traditional data in SAE.

Future work will investigate the problem discussed in this article with a larger number of countries in the ESS, which means that the number of regions will also be larger. This will possibly show more evident gains in efficiency in the model-based small area estimates. Furthermore, users will be able to carry out comparisons between countries.

In addition, it is crucial to acknowledge and address certain drawbacks associated with the use of digital trace data for survey research. While these data sources offer valuable and innovative auxiliary information, their quality may be questionable. For instance, in our analysis, the availability of hotel data depends on scraping timing and request parameters, introducing selection biases. Additionally, web data are subject to volatility; changes in company policies, such as the *Travel Sustainable* programme in our case, can alter what we can measure with web data.

Research about the use of digital trace data in small area estimation, and more in general, in survey research is expanding. However, there remains a necessity for additional empirical evaluations and deeper exploration into the quality aspects of these data to fully understand the benefits and limitations of incorporating them in statistical models [Keusch and Kreuter, 2021].

References

Gary Brown, Ray Chambers, Patrick Heady, and Dick Heasman. Evaluation of small area estimation methods—an application to unemployment estimates from the UK LFS. In *Proceedings of statistics Canada symposium*, volume 2001, pages 1–10. Statistics Canada, 2001.

European Social Survey. *European Social Survey round 10*, 2022. URL `https://www.europeansocialsurvey.org/news/article/round-10-data-now-available`.

Robert E Fay and Roger A Herriot. 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, 1979.

Sylvia Harmening, Ann-Kristin Kreutzmann, Sören Schmidt, Nicola Salvati, and Timo Schmid. A framework for producing small area estimates based on area-level models in r. *R Journal*, 15(1), 2023.

Florian Keusch and Frauke Kreuter. Digital trace data: Modes of data collection, applications, and errors at a glance. In *Handbook of Computational Social Science, Vol 1*. Taylor & Francis, 2021.

Hoesung Lee, Katherine Calvin, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter Thorne, Christopher Trisos, José Romero, Paulina Aldunce, Ko Barret, et al. Ipcc, 2023: Climate change 2023: Synthesis report, summary for policymakers. Technical report, IPCC, Geneva, Switzerland., 2023. URL `https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_FullVolume.pdf`.

Stefano Marchetti and Francesco Schirripa Spagnolo. Social big data to enhance small area estimates. *The Survey Statistician*, 89:59–67, 2024.

Stefano Marchetti, Caterina Giusti, Monica Pratesi, Nicola Salvati, Fosca Giannotti, Dino Pedreschi, Salvatore Rinzivillo, Luca Pappalardo, and Lorenzo Gabrielli. Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2):263–281, 2015.

Stefano Marchetti, Caterina Giusti, Monica Pratesi, et al. The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. *AStA. Wirtschafts- und sozialstatistisches Archiv*, pages 1–15, 2016.

Yolanda Marhuenda, Domingo Morales, and María del Carmen Pardo. Information criteria for fay–herriot model selection. *Computational statistics & data analysis*, 70:268–280, 2014.

Angelo Moretti and Adam Whitworth. European regional welfare attitudes: a sub-national multidimensional analysis. *Applied Spatial Analysis and Policy*, 13(2):393–410, 2020.

Aaron T Porter, Scott H Holan, Christopher K Wikle, and Noel Cressie. Spatial fay–herriot models for small area estimation with functional covariates. *Spatial Statistics*, 10:27–42, 2014.

Aseem Prakash and Thomas Bernauer. Survey research in environmental politics: why it is important and what the challenges are. *Environmental Politics*, 29(7):1127–1134, 2020.

NG Narasimha Prasad and Jon NK Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association*, 85(409):163–171, 1990.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

John NK Rao and Isabel Molina. *Small area estimation*. John Wiley & Sons, 2015.

Camilla Salvatore. Inference with non-probability samples and survey data integration: a science mapping study. *Metron*, 81(1):83–107, 2023.

Caterina Santi and Angelo Moretti. Carbon risk premium and worries about climate change. *Available at SSRN 3942738*, 2021. URL `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3942738`.

Hadley Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2024. URL `https://rvest.tidyverse.org/`. R package version 1.0.4, https://github.com/tidyverse/rvest.

The Survey Statistician, 2024, Vol. 90, 38-39

Book and Software Review

Sampling and Estimation from Finite Populations

Angelo Moretti[1]

[1] Department of Methodology and Statistics, Utrecht University, The Netherlands, a.moretti@uu.nl

Abstract

The book *Sampling and Estimation from Finite Populations* by Yves Tillé was published by Wiley in 2020 in the series Probability and Statistics. The book constitutes an important reference on survey sampling and estimation for a wide range of scholars, users, and practitioners, focusing on the 'representation side' of the Total Survey Error (TSE) framework (Groves and Lyberg, 2010). The book contains 16 chapters presenting both classical and modern approaches to survey sampling and estimation. The main focus of the book is on design-based inference; however, there is also some discussion on model-based techniques and design-based estimators that use auxiliary information in survey sampling and estimation. Importantly, it covers also issues on nonresponse, which, nowadays is a great deal in data collection studies.

Keywords: sampling, complex survey design, design-based inference, variance.

The book *Sampling and Estimation from Finite Populations* by Yves Tillé, published by Wiley in 2020, provides a comprehensive overview of modern survey sampling and estimation problems, and the author importantly defines the discipline as a 'living discipline'. The book has a technical nature, however, sometimes, some demonstrations are omitted; in this case, the author provides bibliographical reference for these, which can be helpful to interested readers. The main focus of the book is on the design-based approach; given that, as Tillé stresses, that is consistent and ethically acceptable to public statistics.

After two introductory chapters, i.e., Chapters 1 and 2, where an historical overview is given and the concepts of population, sample and estimation are introduced, the book covers two areas of the TSE framework, i.e., sampling and estimation issues. Various sampling designs including simple and systematic sampling, stratified sampling, unequal probability sampling designs, balanced sampling, cluster and two-stage sampling are extensively presented and discussed. Furthermore, topics such as spatial sampling, sampling coordination, and multiple survey frames are addressed. A comprehensive discussion on some important estimators is presented, i.e., ratio, difference, regression, poststratified, and calibration estimators. Each chapter contains exercises, and their solutions are provided in Chapter 17.

Chapter 3 is about simple designs, such simple random sampling, Bernoulli sampling, and systematic sampling. It also focuses on specific problems such as sample size determination and entropy calculation. Chapter 4 introduces the problem of stratification, which allows the use of auxiliary information in the sampling design. Different types of allocation are discussed and the notion of optimality is applied to estimators other than totals. Other issues, such as taking into account for costs and power allocation are presented. Chapter 5 is about unequal probabilities

designs, i.e., systematic sampling with unequal probabilities, Poisson sampling, the Rao-Sampford and the Brewer methods. Other methods are also presented and discussed in detail. The problems of variance estimation and entropy are also presented. Chapter 6 is on balanced sampling, which draws samples whose expansion estimators are equal to or at least very close to the population totals for one or more auxiliary variables. Here, different methods are presented, putting an emphasis to the Cube method. Variance approximation and estimation procedures are presented. The chapter concludes with special cases and practical aspects of this sampling design**.** Chapter 7 is about clustering and two-stage designs, where auxiliary information may be used to improve the organisation of the survey. More advanced issues are discussed here, such as self-weighting two-stage design and multi-stage designs. Chapter 8 presents a wide range of problems, e.g., among the other problems, spatial sampling, modifications of stratified sampling, coordination in repeated surveys, multiple frames survey and adaptive sampling designs. Overall, this aspect is particularly innovative as these topics are not always covered in a basic sampling textbook.

Design-based estimators that use auxiliary information are presented in Chapter 9 (ratio estimator), Chapter 10 (post-stratification and calibration), and Chapter 11 (regression estimator). The important problem of calibration is treated in Chapter 12. A wide range of approaches are presented here, focusing also on the problem of generalised calibration and its application in practise. Furthermore, the author provides the reader with references to various software (among others R, SAS, SPSS) through which it is possible to develop calibration estimators.

Although the book focuses on design-based approaches mainly, Chapter 13 introduces the model-based inference problem. Here, some model failures issues are discussed. Interestingly, the chapter concludes with a robust approach by looking at the inference from both the design and the model. The issue of estimation of complex population parameters is presented in chapter 14. For example, the problems of estimating the covariance, and the Gini index are presented, as well as quantile estimation.

Chapter 15 is about an important challenge, i.e., variance estimation. Here, many different linearization methods are proposed, which constitutes a great review of approaches. Chapter 16 is about non-response and how to treat it. Reweighting and imputation procedures are discussed and, the link between regression imputation and reweighting is presented. The topic of nonresponse is often overlooked in similar books on survey sampling, however, it is a crucial issue in nowadays data collection studies, especially in social surveys where nonresponse rates are dramatically increasing. Importantly, in this chapter, the author considers both target parameter estimation problems but also variance estimation and provide software suggestions to deal with this usually underestimated problem.

In conclusion, the book begins with detailed introductory chapters, ensuring it can be understood by a wide range of readers, including statisticians, professionals, and researchers alike. With a collection of exercises featuring clear, synthetic solutions, it serves as a valuable resource for advanced survey sampling courses. Certain more advanced chapters delve deeper into technical aspects, demanding a solid understanding of survey sampling theory. To highlight is the extensive bibliography, often updated with recent years, presented in this book to allow the reader to delve into topics of their own interest. Undoubtedly, this book stands out as a significant contribution in the field of survey sampling theory.

References

Groves, R.M. and Lyberg, L., 2010. Total survey error: Past, present, and future. Public opinion quarterly, 74(5), pp.849-879.

The Survey Statistician, 2024, Vol. 90, 40–47.

Book and Software Review

Software review for inference with non-probability surveys

Beatriz Cobo[1], Ramón Ferri-García[2], Jorge L. Rueda-Sánchez[3], María del Mar Rueda[4],

[1,2,3,4]University of Granada, Spain
[1]beacr@ugr.es, [2]rferri@ugr.es, [3]jorgerueda@ugr.es, [4]mrueda@ugr.es

Abstract

Implementing probability sampling methods has become more challenging as there has been a noticeable decline in response rates with a consequent increase in survey costs. Furthermore, new data sources that have emerged in recent years could be considered alternatives to survey data. Examples include large data sets from sources such as registries or geolocation and web surveys that have the potential to provide estimates, as well as offer easier access to data and lower data collection costs in comparison to traditional probability sampling, leading to larger sample sizes. Given these new forms of sampling, specific software is needed to support theoretical development. We are going to carry out a review of the existing software for this purpose in the most used programming languages in this field (R and Python), indicating the strengths of each of them.

Keywords: non-probability surveys, inference, software, R, Python.

1 Introduction

Survey methodologies are currently in flux due to social and technological changes that have led to a significant increase in refusals to participate and difficulties in accessing individuals to interview. However, the development of new technologies has facilitated the emergence of new data acquisition techniques, such as web surveys, that present great advantages in terms of speed in obtaining data, reduced costs and the possibility of accessing specific population sectors.
Web surveys have replaced face-to-face and computer-assisted telephone interviews as the primary mode of data collection in most countries, and this trend was reinforced as a result of restrictions related to the COVID-19 pandemic. Although web surveys can be probabilistic, in practice many operate via self-selection, meaning that the principles of probability sampling are not applied.

Non-probability surveys have serious problems in calculating estimates because the principles of probability sampling inference are invalid. The first problem is the lack of an adequate sampling frame that allows the selection of samples from the general population in a probabilistic manner, causing selection bias if the covered population differs from the target population (Elliott and Valliant, 2017), therefore, the generalization of the results under these biases is compromised. However, these types of surveys can be useful in some situations, such as complementing a small probability

sample with a larger non-probability sample to improve the efficiency of estimates or mitigating selection biases by intentionally focusing on respondent profiles that tend to be underrepresented.

Powerful new methodologies have been developed to infer parameters using data from non-probability samples, and this research has been reviewed by Buelens et. al. (2018), Rao (2022), Valliant (2020), Yang and Kim (2020), among others. The methods considered include propensity score adjustment (Lee and Valliant, 2009), tree-based inverse propensity weighting (Chu and Beaumont, 2019), propensity-adjusted probability prediction (Elliott and Valliant, 2017), inverse sampling (Kim and Wang, 2019), mass imputation or statistical matching (Rivers, 2007), doubly robust methods (Chen et. al., 2020), kernel smoothing methods (Kern et. al., 2021), superpopulation modelling (Buelens et. al., 2018) and combinations of these techniques (Castro et. al., 2022; Liu and Valliant, 2023). In the following we will briefly describe some of the open source software that implement some of these techniques (R and Python) and their main features.

2 R Software

Over the years, software has been developed to carry out estimates in probability surveys, but it cannot be used directly with non-probability surveys. To fill this gap, researchers have developed new functions in which they consider that the samples are obtained without a probability sampling design. If we focus on the free software Rwe find some packages **NonProbEst** (Martín et. al., 2020), **nonprobsy** (Chrostowski and Beręsewicz, 2024), **nppR** (Beaumont and Dhushenthen, 2024), and **KWML** (Wang and Kern, 2023), that implement different estimation methods with non-probability surveys.

2.1 "NonProbEst" Package

The R package **NonProbEst** (Martín et. al., 2020) is stored in the official R repository `https://cran.r-project.org/web/packages/NonProbEst/index.html` and is explained in detail in Rueda et. al. (2020). This package was the first that implemented the estimation of linear parameters using Propensity Score Adjustment (PSA) (Valliant and Dever, 2011). The package computes estimates on the propensity to participate in the convenience sample based on classification models to be selected by the user. From these propensities the program calculates pseudo-weights for the non-probability sample units using 4 methods: inverting the propensity ($valliant\_weights$), inverting the propensity minus 1 as in Schonlau and Couper (2017) ($sc\_weights$), using the propensity score averaging design weights formula introduced in Lee and Valliant (2009) ($lee\_weights$), and using the propensity stratification averaging formula introduced in Valliant and Dever (2011) ($vd\_weights$). These weights can be downloaded for use in any subsequent statistical analysis that the researcher wishes to perform with his non-probability data.

The package implements functions to calculate the estimator of the total, mean, and the proportion using these four versions of PSA, as well as other techniques as PSA plus calibration (Lee and Valliant, 2009; Ferri and Rueda, 2018), mass imputation (Rivers, 2007; Beaumont, 2020), or superpopulation models (Buelens et. al., 2018), including model-based, model-calibrated, and model-assisted estimator. One of the main features of the package is that a wide range of statistical models and machine learning (ML) algorithms can be used to leverage the information provided by the auxiliary variables, because all the functions that compute the estimators have an argument in which we can include the machine learning model of our choice from those available in the *train()* function of the **caret** package (Kuhn, 2008).

NonProbEst package, in addition to offering the estimates, also allows users to calculate the variance using two alternatives of Leave-One-Out jackknife (Quenouille, 1956), with and without reweighting in each iteration, and the confidence intervals. It also provides a data set of a simulated fictitious

population of 50000 individuals, a probability sample (drawn with simple random sampling from the simulated population and sample size 500) and a non-probability sample (drawn from the simulated population and sample size 1000), to test the implemented functions.

2.2 "nonprobsvy" Package

The newly released package **nonprobsvy** (Chrostowski and Beręsewicz, 2024) is deposited in `https://github.com/ncn-foreigners/nonprobsvy` and is fully explained in Chrostowski and Beresewicz (2023). The goal of this R package is to carry out statistical inferences with non-probability survey samples (including big data) when auxiliary information from external sources like probability samples or population totals/means are available.

The R package **nonprobsvy** implement the first alternative of PSA, also called inverse probability weighting estimators (IPW) (see Chen et. al., 2020), with possible calibration constraints and using as predictive model for propensity scores (selection model) logistic regression (GLM) with logit, probit, and log-log functions; mass imputation (Yang et. al., 2021), using logistic regression (GLM), nearest neighbour (NN) and predictive mean matching (PMM) (Kim et. al., 2021); and the doubly robust estimator (Yang et. al., 2020). The package also allows variable selection in high-dimensional spaces using SCAD (Yang et. al., 2020), Lasso and MCP penalty (via the ncvreg, Rcpp, RcppArmadillo packages), estimation of variance using analytical and bootstrap approaches (Wu, 2022) and computation of the estimated covariance matrix for model coefficients. The software therefore stands out for the wide variety of current techniques implemented for parameter estimation. It also contains a simulated population, described in Chen et. al. (2020), with size desired by the user, and from this population it extracts a non-probability sample, also with size selected by the user.

2.3 "nppR" Package

The R package **nppR** (Beaumont and Dhushenthen, 2024) is stored on the Github site `https://github.com/StatCan/nppR`. This package provides us with some tools to carry out estimations with non-probability samples, using the information of a probability sample in a complementary way. In particular, this package is revolutionary because allows us to use in our experiments the tree-based inverse propensity weighting (TrIPW) estimator, which has not yet been developed in any software. The package was developed in 2022 and is based on the method explained in Chu and Beaumont (2019), which estimates the inclusion probabilities of individuals in the non-probability sample using a modification of the Classification And Regression Tree (CART) algorithm, using information from both samples to more accurately predict whether an individual belongs to the volunteer sample or not. This package also allows us to implement the doubly robust estimator, developed by Chen et. al. (2020), although we can only use linear regression as a predictive model for predicting the variable of interest and logistic regression for estimating inclusion probabilities of the non-probability sample. Finally, the package offers the possibility of creating a synthetic population and probability and non-probability samples from this population.

2.4 "KWML" Package

The R package **KWML** (Wang and Kern, 2023) is deposited in `https://rdrr.io/github/chkern/KWML/` and is explained in detail in Kern et. al. (2023). This package of functions mainly allows us to compute pseudo-weights for non-probability samples, especially using kernel weighting (KW), with the possibility of estimating propensities with logistic regression models or with ML techniques. As ML techniques, we can use conditional random forests (CRF), gradient boosting machines (GBM) and model-based recursive partitioning (MOB). This package also allows to compute the pseudo KW weights given the previously computed propensity scores, so that we can use any ML technique we want if we have previously estimated the inclusion probabilities of the non-probability sample using

the method of our choice. It also contains a dataset containing a simulated non-probability (size 2000) and probability (size 2000) sample.

3 Python

Python is a free software that is being increasingly used by researchers who have to analyze data. If we focus on the Python programming language, the package **inps** (Castro-Martín, 2024) is the first library in Python that implements the main bias adjustment techniques in non-probability surveys. The package is available from the Python Package Index (PyPI) at `https://pypi.org/project/inps/` and is fully detailed in `https://github.com/luiscastro193/inps`.

The library allows statistical inference from non-probability samples and emphasizes an innovative integration of advanced statistical models and machine learning algorithms for bias correction. More precisely, **inps** implements some of the most promising methods for selection bias mitigation such as propensity score adjustment, calibration, methods based on superpopulation modelling (mass imputation and model-based, model-adjusted and model-calibrated estimators), the doubly robust estimator and the PSA-weighted mass imputation estimator from Castro et. al. (2022). Moreover, all of the implemented methods can be easily applied with any machine learning model with a standard API. This allows us to immediately benefit from the huge pool of models already implemented (and those to be added in the future) in Python libraries such us SciPy or scikit-learn. This package also includes a function to pre-process an estimator, which is essential for those models that require numerical data with no missing values. The implementation of these methods is unprecedented in Python and has the potential to allow researchers and practitioners to leverage all the statistical functionalities that Python offers while being able to use the same methods for inference in non-probability samples that were already present in other software.

Six datasets are available in the package. The first three (*nonprobEd, nonprobCovid* and *nonprobHealth*) correspond to three real nonprobability surveys carried out in Spain that address different topics, such as eating disorders, attitudes towards the coronavirus pandemic and satisfaction with life of health professionals. Datasets *probEd* and *probCovid* contain two well-designed probability surveys conducted by official agencies in the same study populations that *nonprobEd* and *nonprobCovid* respectively attempted to cover. They are intended to be used as reference surveys in each of the two cases to adjust for selection biases. The last dataset, *censusHealth*, contains the population frame of units from which the non-probability survey *nonprobHealth* was obtained. These examples show the potential of the package in tackling selection bias and the wide usage possibilities for non-probability sampling estimation for both academics and practitioners.

4 Discussion

In the last few years, a plethora of methods have emerged to mitigate selection bias, employing techniques based on superpopulation models or based on the quasi-aleatorization of the sample, or even the combination of both, which use auxiliary information provided by alternative sources of data. In practice, however, the majority of applications where non-probability samples are considered do not apply any of these corrections, making at most only simple adjustments using basic sociodemographic variables such as gender or age. The scarce use of these methods in practice may be due to their difficult implementation for people who are not experts in sampling. In the last years, several free-use packages have appeared that can contribute to increasing its use in real surveys by researchers.

In this work we have briefly introduced the main packages available in R and Python to obtain estimates from non-probability samples. Now we show a summary table to have all the previously

mentioned information concentrated in an easy-to-understand table.

Table 1: Summary table of packages for non-probability bias reduction (see the text for the following abbreviations)

Packages	Software	Techniques	ML algorithms	Datasets	Extra Features
NonProbEst	R	Calibration	"caret" package algorithms	Simulated population with 50000 individuals	Variance and CI with re-sampling techniques
		PSA (4 alternatives)			
		Superpopulation Models			
		Mass Imputation			
		PSA + Calibration			
nonprobsvy	R	IPW	GLM	Simulated population and non-probability sample with selected sizes	Variable selection, Variance with analytical and bootstrap approach, Control parameters for outcome, selection, and inference model, Weights adjustment with probit, logit, and log-log models, Different links for outcome variables, ...
		Mass Imputation	GLM, NN, PMM		
		Doubly Robust	GLM / GLM, NN, PMM		
		IPW + Calibration	GLM		
nppR	R	Doubly Robust	GLM / GLM	Creation of a synthetic population and drawing of probability and non-probability samples	None
		TrIPW	CART		
KWML	R	KW	GLM, CRF, GBM, MOB	Simulated probability and non-probability samples	Can use any ML algorithm if you compute propensity scores separately
		IPW	GLM		
inps	Python	Calibration	"scikit-learn" API compatible packages algorithms	Six datasets (three non-probability samples, two probability samples, and one real population)	Data preprocessing
		PSA (4 alternatives)			
		Superpopulation Models			
		Mass Imputation			
		Doubly Robust			
		KW			
		PSA + Mass Imputation			

We can see in Table 1 that packages that implement the largest amount of techniques are **NonProbEst** package and **inps**, especially the Python package, as it also implements the kernel weighting (KW) estimator, the doubly robust estimator, and mass imputation with PSA weights. This package is unique because it allows such techniques to be used in Python, it is the first to implement the PSA-weighted mass imputation estimator (Castro et. al., 2022) and includes a function to pre-process data for an estimator. We highlight that both packages allow researchers to use their estimators with the vast majority of machine learning (ML) algorithms as predictive models. The **nonprobsvy** package allows estimation using Inverse Propensity Weighting (IPW) estimator, innovating since logit, log-log and probit functions can be used to predict propensity scores; mass imputation estimator, emphasising predictive mean matching (PMM) algorithm as predictive models for the variable of interest; and Doubly Robust estimator, with the aforementioned innovative predictive algorithms for the IPW and Mass Imputation estimators. This package also allows other functionalities such as variable selection before calculating the estimator (SCAD, Lasso and MCP penalty) or variance estimation

with the bootstrap technique and analytical formulas, both unique to this package. In the case of the **nppR** package, the doubly robust estimator can be calculated and it is worth noting that we can calculate the tree-based inverse propensity weighting estimator (TrIPW), something that cannot be done with any other package. If we want to compute the KW estimator in R software, we have to use the **KWML** package, which also allows us to use any ML model if the propensity scores have been computed beforehand (for example, with the **NonProbEst** package) with the desired ML algorithm.

Some things are missing in the various packages, for example, a mean square error estimation procedure that takes into account all the sources of randomization that involve the various methods and that are valid for any machine learning technique or estimators to integrate probability and non-probability data (Kim and Tam, 2021; Rueda et. al., 2023; Rueda et. al., 2024). A measure of the final bias in the estimate and its reduction compared to estimators that do not use these adjustment methods would also be useful.

In our opinion, these software described here can serve to advance inference in non-probability sampling, offering a very broad set of specific tools that can be useful both for academic research and for practical implementation.

References

Beaumont, J. F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology, Statistics Canada*, **46**1. http://www.statcan.gc.ca/pub/12-001-x/2020001/article/00001-eng.htm

Beaumont J. and Dhushenthen J. (2024). nppR: Inference on non-probability sample data via integrating probabilty sample data. R package version 1.13.003.

Castro-Martín, L. (2024). INPS: Inference from Non-Probability Samples. Python package version 1.0.

Buelens, B., Burger J. and van den Brakel J. A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review*, **86**(2), 322–343.

Castro L., Rueda M. and Ferri-García R. (2022). Combining Statistical Matching and Propensity Score Adjustment for inference from non-probability surveys. *Journal of Computational and Applied Mathematics*, **404**, 113414.

Chen, Y., Li, P. and Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, **115**(532), 2011–2021.

Chrostowski L. and Beręsewicz M. (2023). nonprobsvy: Package for Inference Based on Nonprobability Samples. R package version 0.1.0. https://cran.r-project.org/web/packages/nonprobsvy/index.html

Chrostowski L. and Beręsewicz M (2024). nonprobsvy: Inference Based on Non-Probability Samples. R package version 0.1.0, ¡https://CRAN.R-project.org/package=nonprobsvy¿.

Chu, K. C. K. and Beaumont, J. F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. In Proceedings of the Survey Methods Section: SSC Annual Meeting, Calgary, AB, Canada **26**.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Stat. Sci.* **32** , 249–264.

Ferri-García, R. and Rueda, M. d. M. (2018). Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Stat. Oper. Res. T.*, **42**(2), 159–162.

Kern, C., Li, Y. and Wang, L. (2021). Boosted Kernel Weighting – Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, **9**,(5),1088—1113. https://doi.org/10.1093/jssam/smaa028.

Kern, C., Li, Y. and Wang, L. (2023). KWML: Boosted Kernel Weighting - Using Statistical Learning to Improve Inference from Nonprobability Samples. `https://rdrr.io/github/chkern/KWML/`

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.

Kim, J. K., Park S., Chen Y. and Wu C. (2021). Combining Non-Probability and Probability Survey Samples Through Mass Imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society* **184** (3): 941—963. https://doi.org/10.1111/rssa.12696.

Kim, J. K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review,* **87**, 177–191.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26. `https://doi.org/10.18637/jss.v028.i05`

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, **37**(3), 319–343.

Liu, Z. and R. Valliant. (2023). Investigating an Alternative for Estimation from a Nonprobability Sample: Matching plus Calibration. *Journal of Official Statistics* **39** (1), 45—78.
DOI: http://dx.doi.org/10.2478/JOS-2023-0003.

Martín L. C., García, R. F. and Rueda, M. d. M. (2020). _NonProbEst: Estimation in Nonprobability Sampling_. R package version 0.2.4, ¡`https://CRAN.R-project.org/package=NonProbEst`¿.

Quenouille, M. H. (1956). Notes on bias in estimation. Biometrika, 43(3/4), 353-–360. [p408]

Rao, J. N. K. (2022) On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*, **83**, 242–272.

Rivers, D. (2007). Sampling for web surveys. *In Proceedings of the 2007 Joint Statistical Meetings, Salt Lake City, UT, USA*

Rueda, M. d. M., Cobo, B., Rueda-Sánchez, J. L., Ferri-García, R. and Castro-Martín, L. (2024). Kernel Weighting for blending probability and non-probability survey samples. *SORT,* **48**(1), 1–32.

Rueda, M. d. M., Ferri-García, R. and Castro, L. (2020). The R package Non-ProbEst for estimation in non-probability surveys. *The R Journal,* **12**(1), 406–418.

Rueda, M. d. M., Pasadas-del-Amo, S., Rodríguez, B. C., Castro-Martín, L. and Ferri-García, R. (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, **65**(2), 2200035.

Schonlau, M., Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science,* **32**(2), 279–292.

Valliant, R., and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, **40**(1), 105–137.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, **8**(2), 231–263.

Wang L., Kern, C. (2023). KWML: Boosted Kernel Weighting. R package version 1.0.1.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology, Statistics Canada*, **48**(2), 283—311.

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. https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-eng.html

Yang, S., Kim, J. K. and Song R. (2020). Doubly Robust Inference When Combining Probability and Nonprobability Samples with High Dimensional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **82**(2), 445-–65. https://doi.org/10.1111/rssb.12354

Country Reports

ARGENTINA

Reporting: **Verónica Beritich**

2022 Census: The INDEC presented the final results on fertility

The National Institute of Statistics and Censuses (INDEC) released more definitive results of the 2022 National Census of Population, Households and Dwellings. On this occasion, indicators compiled in a thematic publication on fertility are presented, with information on women aged 14 to 49 years.

This publication on fertility is added to those on Housing conditions of the population, households, and dwellings; population structure, by sex and age; collective housing and homeless population; health and social security; education; gender identity; migrations; economic characteristics of the population; indigenous population; and afro-descendant population.

- The 2022 Census counted 12,382,860 women between 14 and 49 years old: 57.5% had daughters and sons born alive. On average, each woman had 1.4 children in 2022.

- Santiago del Estero, Misiones and Formosa, with 1.7 children per woman, were the jurisdictions with the highest average number of children. Opposite, the Autonomous City of Buenos Aires had 0.9 children per woman. The provinces of Córdoba, Neuquén, Tierra del Fuego, Antarctica, and the South Atlantic Islands followed with 1.3.

- The percentage of women between 15 and 19 years old with daughters and sons born alive in 2022 had a very pronounced decrease compared to the results of the previous census: it went from 13.1% in 2010 to 6.4% in this census.

- Afro-descendant women had an average of 1.4 children; foreign women, 1.8; and indigenous women, 1.9.

- From the total number of women between 14 and 49 years old with incomplete secondary school, 84.6% had daughters or sons born alive. This percentage decreases with the educational level: From those with a complete or incomplete higher level, 64.6% had children born alive.

- Forty-two out of every 100 14-year-old women with daughters and sons responded that they were not attending an educational establishment in 2022. That number is reduced to 4 out of every 100 in women of the same age who declared that they did not have daughters and sons born alive.

- 71.8% of the women, aged between 30 and 34 years old who had daughters or sons born alive, were economically active and 88% of the women who did not have children.

General information can be found at https://www.indec.gob.ar/ftp/cuadros/poblacion/censo2022_fecundidad.pdf

For further information, please contact https://www.indec.gob.ar/indec/web/Institucional-Indec-Contacto.

CANADA

Reporting: **Steve Matthews and François Brisebois**

Fasten your seat belt for Statistics Canada's Methodological Acceleration!

A comprehensive review of the agency's activities and practices was conducted at Statistics Canada in the fall of 2022, in order to better understand and take stock of the investments made throughout its modernization journey up to now. This initiative required extensive agency-wide consultations and resulted in 12 strategic recommendations aimed at achieving efficiencies, mainly in the form of reduced costs for survey programs, for example through reducing manual processing and avoiding duplication of efforts.

Methodological Acceleration was the title given to one of these twelve recommendations centered around statistical methods. The motivation for this initiative comes from the need to take a significant step forward in the Agency's ability to effectively and quickly implement modern solutions to the data challenges of today and tomorrow. To propel the *Acceleration*, a series of six projects were launched in early 2023, each targeting different sources of potential efficiency. Three of these projects were more specifically associated with programs on social statistics and aimed to explore methodological solutions likely to improve the efficiency of data collection. The other three projects rather concerned processes common in the world of economic statistics, where here the efficiency gains would mainly come from a reduction in manual processing. These projects made it possible to concretely launch efforts aimed at efficiency and helped to design and define what is now known as the *Methodological Acceleration*. Although these proposed projects are evaluated in the context of very specific existing programs, they were each reviewed and developed from a horizontal perspective to potentially generate even greater efficiencies by applying the developed scalable solutions to multiple programs.

The first six projects were carried out within a year, from April of 2023 to March of 2024, and in that time each project first established baseline indicators linked to the targeted area of efficiency, then explored different solutions to achieve efficiencies, and finally provided recommendations on the implementation of the recommended solution, including indications on how it could be extended to other statistical programs.

At the end of this first exploration phase, four of the six projects concluded that the proposed idea could indeed generate efficiency gains and therefore recommended moving to the implementation phase. The current fiscal year will therefore include work aimed at these implementations, as well as the addition of a new series of projects to explore other new ideas that have potential for more efficiency gains across the various statistical programs of the agency. With the momentum gained in the first year, the plan is to keep our foot on the gas and continue to fuel the acceleration with new and innovative ideas.

CROATIA

Reporting: **Ksenija Dumicic**

Generations and Gender Programme (GGP) Survey launched in Croatia in 2023 based on GGP Push-to-Web Experiment pilot study experience

The Generations & Gender Programme (GGP), was launched in 2000 by the Population Unit of the UNECE and has been coordinated by the Netherlands Interdisciplinary Demographic Institute (NIDI) since 2009. It is a large international demographic survey that provides harmonized, large-scale, longitudinal, cross-national panel data on individual life courses and family dynamics. Over time, the

GGP Survey follows respondents through relationships, marriages, parenthood, divorces, deaths, and many of the opportunities and challenges that people face along the way. The European Commission gave it the status of top research infrastructure. The GGP Survey in the second round (GGS-II) research methodology is generally explained by Gauthier et al. (2023), and The Technical Guidelines (Generations and Gender Programme, 2023), which specify that all data collections have to include a web component and might be a mixed-mode survey.

In Croatia, jointly with the NIDI, the institutions that worked on the GGS-II are the Central State Office for Demography and Youth, IPSOS Croatia, the University of Zagreb Faculty of Economics, and GGP Central Hub. The target population of the GGS-II round in Croatia is the resident non-institutionalized population within a specified age range of 18 to 54. The sampling frame was the list of residents provided by the Croatian Ministry of Interior Affairs. A random stratified sample with proportionate allocation across all 21 counties was applied, enabling the calculation of estimates for four NUTS 2 regions of Croatia. The gross sample was sufficient to achieve the aimed net sample size for Croatia of 5,000 as planned in Gauthier et al. (2023). The GGS Push-to-Web Experiment in the pilot study discovered that GGS in Croatia is feasible for Computer-Assisted Web Interviews (CAWI), and also looked to advance the knowledge on the introduction of an incentives approach. A questionnaire was sent on May 9, 2023, and over the course of seven months, more than 6,000 questionnaires were collected for an effective sample size. Response rate varies across countries, but was the highest in Croatia, being around 28%. Each respondent received an unconditional incentive of 6.64 euros. Minor adjustments in the survey design were made to make GGS-II more suitable for web interviews (Emery et al., 2023). The results of GGS-II Round, Wave 1, executed in Croatia in 2023 were presented on March 19, 2024, Croatian Large Families Association “Obitelji 3plus“ (2024).

Bibliography:

Croatian Large Families Association "Obitelji 3plus" (2024). Croatia: Presentation of the Results of the International Research of the GGP-Generations & Gender Program. Held on March 19, 2024. Available at: https://www.elfac.org/croatia-presentation-of-the-results-of-the-international-research-of-the-ggp-generations-gender-program/

Emery, T.; Cabaco, S.; Fadel, L.; Lugtig, P.; Toepoel, V.; Schumann, A.; Lück, D.; Bujard, M. (2023): Breakoffs in an hour-long, online survey. Survey Practice 16, 1. https://doi.org/10.29115/SP-2023-0008.

Gauthier, A. H., Kong, S., Grünwald, O., Bujard, M., Caporali, A., Deimantas, V. J., Emery, T., Jablonski, W., Koops, J., Rijken, A., & Schumann, A. (2023). "Data Brief: The Generations and Gender Survey second round (GGS-II)." GGP Technical Paper Series. https://doi.org/10.5281/zenodo.10220746.

Generations and Gender Programme. (2023). The Generations and Gender Programme: Technical Guidelines. Available at: https://zenodo.org/records/10812889.

ESTONIA

Reporting: **Ene-Margit Tiit**

Estonian Statistical Society conference “Science of Big Data”

On 19 and 20 April 2024, the Estonian Statistical Society held a conference titled "Science of Big Data" (“Suurte andmete teadus”) in Tartu. This was the thirty-first in a series of conferences organised by the Estonian Statistical Society, founded in 1992, and was dedicated to the 90th birthday of Ene-Margit Tiit, Professor Emeritus at the University of Tartu and a founding member of

the Society. The main organisers of the conference were Krista Fischer, Professor of Mathematical Statistics at the University of Tartu, and Terje Trasberg, team lead of population and education statistics at Statistics Estonia.

On the first day of the conference, Associate Professor Liina-Mai Tooding gave a historical overview of the development of data analysis at the University of Tartu, which began in the early 1960s when the university received a Ural-type mainframe computer that became an important tool for the statistical analysis of data collected in the university's research activities and the drawing of reliable conclusions. In her presentation, Dagmar Kutsar, Associate Professor in Social Policy, highlighted the reliability and truthfulness of the information and testimonies given in surveys by children, even though they are often mistakenly underestimated by researchers. Professor of Probability Theory Kalev Pärna spoke about the work of statisticians in estimating forest resources, which has been a topic of lively and popular discussion in Estonia for a number of years since Estonians as a forest people – forests cover more than 50% of the country's territory – are concerned about the future of forests. The presentations by Andres Võrk and Kadri Rootalu focused on the use of data and innovative models in labour and wage policy.

The last two presentations of the day covered linguistic statistics. Professor Mare Koit, laureate of the prestigious Wiedemann Language Prize, spoke about modelling the Estonian language and its history. The topic is particularly relevant today, on the one hand because of the topicality of large language models, and on the other because of the active language policy of the Estonian state – the introduction of a uniform Estonian-language education system. Estonian is probably one of the languages with the smallest number of speakers (less than 1 million) for which machine translation and speech synthesis models have been developed. A distinguished colleague, Professor Emeritus Esa Läärä (University of Oulu, Finland) compared the development and situation of statistics vocabulary in the Estonian and Finnish languages, which are closely related. The most entertaining presentation was given by the legendary 95-year-old Professor of Astronomy, Academician Jaan Einasto, who, accompanied by fitting music, showed a series of photos from about 70 years ago.

The presentations on the second day of the conference focused on the latest population and housing census, in which Statistics Estonia implemented several original innovations. The residency index used to determine the total population of the census was presented by Ethel Maasing, one of the authors and implementers of the idea of this indicator which integrates information from several dozen state registers. Helle Visk talked about the idea, potential applications and practical use of another original concept – the partnership and location index. Her results make it possible to statistically correct inaccuracies in registers in order to obtain more adequate statistics.

The conference continued with a presentation of the printed word. Professor of Mathematical Statistics Tõnu Kollo gave an overview of the nearly fifty published books by Ene-Margit Tiit, which were also on display in the conference room. The most important of these were deemed to be the university's teaching materials offset printed between 1960 and 1990, which aided students in the fields of probability theory and mathematical statistics.

This was followed by the presentation of the book "Eesti rahvastik. Loendamata loendatud" ("Estonian Population. Counted without Counting"). Its author, Ene-Margit Tiit, was interviewed by professors Krista Fischer and Jaak Kikas. The book introduces the results of the 2021 population census in Estonia, offers a brief comparison with the results of the previous census, highlighting the most important changes, and based on census data, gives an overview of the population of all 15 Estonian counties.

Link to the book (available only in Estonian):

https://www.stat.ee/et/uudised/eesti-rahvastik-loendamata-loendatud.

HONG KONG S.A.R.

Reporting: **Ada Cheung**

Re-engineering of Population Census of Hong Kong

The Census and Statistics Department (C&SD) of Hong Kong, China is actively preparing for the re-engineering of the workflow of Population Census starting from the 2026 round. Since 1961, Hong Kong's Population Census was conducted every ten years in years ending "1", which comprised a simple enumeration of about nine-tenths of the entire population using a "short form" questionnaire and a detailed enquiry of the remaining one-tenth of the population using a "long form" questionnaire. A Population By-census was conducted in the middle of each intercensal period in years ending "6", with only the "long form" enumeration of one-tenth of the population.

Starting from 2026, Population Census of Hong Kong will be re-engineered into a sample survey on par with the scale of a Population By-census to be conducted every five years. With the availability of administrative data on passenger movements from Immigration Department, accurate population size and age-sex structure can be compiled without the need to conduct the "short form" enumeration in full Censuses. Simulation studies using the 2021 Population Census (21C) data have confirmed the accuracy of the new approach. Besides passenger movement records, the feasibility of matching administrative data from other Government bureaux and departments with the census sample file at record level using addresses is being explored, with a view to trimming some census questions to reduce respondent burden and enhance data quality.

The experience of conducting 21C during the COVID-19 pandemic pointed to the vast risks of conducting large-scale survey operation in a short period of time (1.5 months in past Census operations). The re-engineered Census of Hong Kong will take place over a full year, significantly reducing the operational risks brought forth by uncontrollable events. With a longer data collection period, a smaller and better trained workforce can be deployed for fieldwork. Data quality can be enhanced as a result while savings in salary and associated overheads can also be realised. Statistical methods will be deployed to adjust the data collected to the mid-year position, such that the results will continue to be generally comparable to those of the previous rounds based on data collected near mid-year.

For more information, please contact Ada Cheung (akycheung@censtatd.gov.hk).

HUNGARY

Reporting: **Gáborné Székely**

The HCSO-ingatlan.com rent index – experimental statistics

The vast majority of Hungarian households live in their own property. The proportion of rental dwellings in the Hungarian housing market is extremely low, even in European comparison. 6.7% of occupied dwellings, about 270,000 dwellings, were used by residents who rented it from a private owner, according to the 2022 census. In Hungary, 4.8% of the population lived in market rental housing in 2018, while on average in the EU Member States this proportion was close to 21%.

Home renting belongs to the 'gray zone' of the Hungarian housing market. Many of its actors are hiding, their relationships are changing rapidly, in many respects they do not fit into the conceptual framework of traditional statistical data collection, and its observation is also difficult to solve relying on traditional statistical techniques. Therefore, it is extremely difficult to reliably measure the size, occupants and, last but not least, prices of the rental sector.

At the same time, renting out private dwellings has spread appreciably in recent years and has now become a key factor in the metropolitan housing market. Today – in addition to policymakers, market participants, and researchers – lay people also demand statistical information on rental trends. In the absence of reliable data, partial information, data of uncertain origin and validity are circulating in the market.

A significant step forward in this situation was that ingatlan.com, a large online real estate advertisement company offered its advertising database to the Hungarian Central Statistical Office (HCSO) for statistical utilization, to monitor changes in rents. Ingatlan.com has an extremely extensive database of hundreds of thousands of unique advertisements, which is also structured for statistical processing. Linking this database with relevant data available from the HCSO system makes it possible to prepare analyses that none of the contributors could solve individually. Another advantage of the cooperation is that it provides much faster access to data than official channels, so that information on this particularly sensitive segment of the housing market can be available within a few days after the reference period.

In contrast to official statistics, these results do not cover the entire examined population, i.e. the rental market, as they are based only on home rental data that appear among the advertisements of ingatlan.com. Nearly half of Budapest landlords and more than a third of those living in county seats contacted their tenants via internet real estate portals, and this ratio is likely to have risen further in the time since, according to the results of the 2018 HCSO rent survey preceding the present research (Private apartment rental, rental prices – main results of the 2018 rent survey).

Our results provide information on changes in supply prices of rentals, as they reflect the rental price level of dwellings available to those entering the market, which – as we have also shown before – differs significantly from the rental prices of dwellings actually rented in the given period. We emphasize that the results are experimental statistics, which are based on some novel, innovative solution. Their novelty also lies in the use of new data sources and new methods, but unlike official statistics, they are less robust, they may not cover all aspects of the given phenomenon, but the results still meet most quality expectations.

HCSO-ingatlan.com rent index was first published in August 2020, very soon after the collapse of rental housing market due to the Covid-19 pandemic. That situation offered a unique opportunity to present a housing market indicator which (despite its experimental character) could contribute to better understand housing market changes.
https://www.ksh.hu/s/en/experimental-statistics/publications/hcsoingatlancom-rent-index-april-2024/

THE NETHERLANDS

Reporting: **Deirdre Giesen**

CBS launches new statistic: goods transport by pipeline

At the request of the Ministry of Infrastructure and Water Management, CBS has recently developed a new statistic: goods transport by pipeline. This statistic is an addition to the mandatory European statistics on transport for sea, air and inland waterway transport and road and rail traffic. Examples of goods that can be transported by pipeline are CO2, natural gas, petroleum (products), chemicals, (industrial) water and heat.

The new statistic was developed in close collaboration with the association of pipeline owners in the Netherlands (VELIN). The first publication for this statistic was in April 2024, about 2022 data. The data show that transport by pipeline accounts for 16 percent of the total Dutch transport. That is slightly more than inland shipping (15 per cent) and much more than transport by rail (2 per cent). The national government and regional authorities can use the new statistics on transport by pipeline

for policy decisions. For example, can a branch be made from an existing pipeline to a new project? The statistic will be updated annually. For more information: Goods transport; modes and flows of transport to and from the Netherlands | CBS (https://www.cbs.nl/en-gb/figures/detail/83101ENG?q=transport) or contact Mathijs Jacobs, Program Manager, mj.jacobs@cbs.nl.

NEW ZEALAND

Reporting: **Rosalia Rohwer**

Gender and sexual identity variables in the 2023 Census

For the first time, New Zealand's 2023 Census asked questions about gender, sexual identity, and variations of sex characteristics (generally known as intersex). We have developed a data standard for gender, sex, and variations of sex characteristics (https://www.stats.govt.nz/methods/data-standard-for-gender-sex-and-variations-of-sex-characteristics/). This includes the "Gender by default principle" that defaults to the collection and output of gender data as opposed to sex. In the 2023 Census we collected both sex at birth and gender, allowing us to derive information about transgender populations.

Methodologies for filling gaps in the gender and sex at birth variables have been updated for New Zealand's 2023 Census. (https://www.stats.govt.nz/methods/methodologies-for-filling-gaps-in-gender-and-sex-at-birth-concepts-for-the-2023-census/) These methods have been informed by engagement with key stakeholders, a public consultation and discussion with our Data Ethics Advisory Group.

For more information, please contact Micah.Davison@stats.govt.nz or Digby.Carter@stats.govt.nz.

The 2023 Post-enumeration Survey: A targeted approach to improve survey response

The post-enumeration survey (PES) is a household survey undertaken in New Zealand shortly after the 2023 Census to evaluate the completeness of census coverage. For the 2023 PES, we developed a targeted collection approach in areas where we expected it to be harder to obtain a response. We recruited a diverse field-force who were representative of the communities we were surveying. A survey incentive scheme was implemented in parts of Auckland with relatively high socio-economic deprivation – this was a supermarket voucher included with the PES information sent out to households in these areas prior to interviewing.

Initial analyses suggest the combined impact of the targeted initiatives was significant. PES results will be released on the Stats NZ website (stats.govt.nz) on 9 December 2024.

For more information, please contact Ben.Faulks@stats.govt.nz or Joel.Watkins@stats.govt.nz.

The 2023 Household Disability Survey: An innovative community approach for Deaf respondents

The established collection method for the Household Disability Survey is telephone calls. Historically, Deaf people selected for Stats NZ telephone surveys have either relied on external New Zealand Sign Language (NZSL) interpreters or been unable to participate.

We contracted a qualified NZSL interpreter to directly engage with the Deaf community and design a tailored Deaf-for-Deaf collection approach. The outcome was that Stats NZ hired Deaf interviewers to collect data from Deaf survey participants using NZSL via Microsoft Teams video calls. We also provided extensive information about the survey in alternate formats, including in NZSL. 2023 HDS results will be released on the Stats NZ website (www.stats.govt.nz) later in 2024.

For more information, please contact Chris.Pooch@stats.govt.nz.

NIGERIA

Reporting: **Olaniyi Mathew Olayiwola**

IASS WORKSHOP ON JANUARY 24, 2024.

On January 24th, 2024, a full-day workshop titled "Visualization of Survey Data using R" was inaugurated at 1:00 pm, generously funded by the International Association of Survey Statisticians (IASS). This workshop was held in partnership with the Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria. It was conducted at Pisad Auditorium at the Federal University of Agriculture.

Dr. (Mrs.) Wale-Orojo, from the Department of Statistics at the Federal University of Agriculture Abeokuta, Nigeria, officially opened the workshop and welcomed participants. During her address, she briefed attendees on the process of becoming a member of IASS, providing a link to the association's webpage for membership details.

The second session of the workshop, which focused on big data, was led by Henry Ekong from the Department of Statistics at the College of Physical Sciences, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.

Throughout the workshop, Ekong delved into the significance of Nominal and Ordinal scaled variables, elucidating on the impact of Survey weights on the analysis of survey data. He provided valuable insights into the composition of survey data, covering various aspects such as Demographic Data, Dietary Data, Medical Examination Data, Laboratory Data, Questionnaire Data, and Geographic Data.

Leveraging the R software extensively, Ekong demonstrated the visualization of survey data with categorical variables, including Nominal and Ordinal responses. The showcased visualizations included Correlation Matrix Plots, Upper Triangular Correlation Matrix Plots, Cumulative Density Plots, Pie Chart Plots, Grouped Pie Chart Plots, Bar Chart Plots, Grouped Bar Charts, Stacked Bar Charts, and Item-person maps.

Ekong also explored Survey Data with weights, addressing Sampling weights and finite population corrections. The presentation featured Bar Plots of Proportions and Spatial visualization of Survey data.

The workshop garnered participants up to One Hundred and Thirty attendees, representing academic institutions, ministries, and various agencies. The participants were actively engaged throughout the lectures and practical sessions, demonstrating enthusiasm in acquiring knowledge in the Sample Survey.

The feedback from participants universally reflected satisfaction with the workshop's content and delivery. The workshop concluded at 4:00 pm, and the organizing team expressed gratitude to IASS for their financial support, extending prayers for future collaboration and support.

POLAND

Reporting: **Tomasz Żądło**

9th Conference on Survey Sampling in Economic and Social Research

The 9th Conference on Survey Sampling in Economic and Social Research is scheduled for **December 3-4, 2024** in Katowice. The conference is organized by the Department of Statistics, Econometrics, and Mathematics at the University of Economics in Katowice, along with the Polish Statistical Association – Katowice branch. This event will be held entirely **online**. This conference aims to continue the tradition of fostering collaboration and knowledge exchange among experts in the field of survey methodology.

The conference website is: https://web.ue.katowice.pl/metoda

UKRAINE

Reporting: **Volodymyr Sarioglo**

Current situation in the State Statistics Service of Ukraine

The humanitarian crisis brought on by the February 2022 Russian full-scale invasion of Ukraine has led to large-scale displacement, both internally and across borders, of the Ukrainian population. This has necessitated the development of updated approaches for obtaining relevant and actual statistics to inform stakeholders, first of all international organizations, and internal users for social, economic, and community reconstruction efforts [1].

The situation was significantly complicated by the fact that after the beginning of the war, the State Statistics Service of Ukraine (SSSU) stopped conducting any surveys of the population and households, as well as the formation of population statistics. This is due to the provisions of the Law of Ukraine "On Protection of the Interests of the Subjects of Submission of Reports and Other Documents During the Period of Martial Law or State of War".

Under such conditions, there was an urgent need to get estimates of the most important indicators on the basis of available data from other sources, namely data from administrative registers, in particular data of the Pension Fund of Ukraine regarding the number and characteristics of pensioners and contributors, of the Ministry of Education regarding the number, distribution and gender age structure of students, data of Mobile Operators regarding the number and placement of subscribers, results of sample surveys conducted without the participation of the SSSU, etc. Together with UNFPA and other international organizations an approach has been developed to estimate the number, structure and location of the population as of mid-2023. In particular, it was established that there were about 31.7 million people in the Government Controlled Area of Ukraine. Currently, studies are being conducted to estimate the characteristics of the population as of mid-

2024. These estimates are essential both for humanitarian purposes and for the activities of authorities at all levels.

With the technical support of UNICEF, at the end of 2023, a household sample survey on the living conditions and economic activity of the population was conducted with a probability sample of 8,000 households and according to a methodology close to the methodology of state sample surveys. This made it possible to assess important indicators of the composition of households, poverty by various criteria, the effectiveness of social programs, employment and unemployment, etc. Preparations are now being made to disseminate the results of the survey to stakeholders. The possibility of conducting at least one more round of this survey in 2024 is being explored.

The SSSU, taking advantage of the forced pause, is piloting a number of surveys that were planned for implementation in 2023-2025. We are primarily talking about the EU-SILC survey and the Time use survey, which users were counting on before the war. At the same time, the efforts of some government officials under the guise of digitization and digital transformation and saving public resources to make the transition from official statistics to the collection of data from the population and households exclusively through online or telephone surveys are cause for concern, which, given the situation in Ukraine, in particular the level of Internet use, may actually destroy the system of state sample household surveys, which has been in development for 25 years [2].

References

[1] Silva, R., Snyder, M., Han, M. D., Sariogolo, V., & Libanova, E. (2023). Subnational population projections for humanitarian response in Ukraine: integration and cross-validation of traditional and non-traditional sources: Quetelet 2023 Seminar, (November 9–10, 2023). Leuven. Retrieved from https://uclouvain.odoo.com/en_US/event/quetelet-2023-seminar-403/page/introduction-quetelet-2023-seminar

[2] Sarioglo, V. G. (2023). 25 Years of Experience in Household Sample Surveys in the Official Statistics of Ukraine: Main Achievements, Problems, Prospects. Statistics of Ukraine, 1, 27–39. Doi: 10.31767/su.1(100)2023.01.03 [in Ukrainian].

UNITED STATES

Reporting: **Andreea Erciulescu, Linda Young, Jenny Thompson and Blynda Metcalf**

Updated Race and Ethnicity Standards

Over the past two years, dozens of federal agencies, with input from the public, have been working on updating the standards for maintaining, collecting, and presenting federal data on race and ethnicity. The changes include collecting race and ethnicity information using one combined question, adding Middle Eastern or North African as a new minimum category, collecting detailed race and ethnicity categories as default, and updating terminology such as discontinuing the use of the terms ‘majority’ and ‘minority,’ with some exceptions. The new standards became effective on March 28, 2024, and are to be implemented in censuses and surveys. They help improve consistency, accuracy, and utility of race and ethnicity data across federal programs. For additional details, see the following notice published in the Federal Register: Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity.

(https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and)

Webinar Series on Artificial Intelligence

The U.S. Federal Committee on Statistical Methodology (FCSM) and the National Institute of Statistical Sciences (NISS) collaborated to produce a series of five webinars, beginning in October 2023, aimed to equip U.S. federal practitioners and managers with insights into the applications and challenges of Artificial Intelligence (AI) adoption within their respective agencies. Methods and use cases for text and image analyses as well as organizational, managerial, and ethical concerns were discussed. The series of webinars provided a solid foundation for an in-person event at the historic National Academies of Sciences building in collaboration with the Committee on National Statistics in May. This event brought together experts and practitioners in the field of AI to explore its integration into federal statistical practices and was a culmination of collaborative efforts aimed at advancing AI integration in federal statistical practices. One of the highlights of the event was the opportunity for participants to engage in discussions about the latest developments in AI technology and its applications in federal statistics. From machine learning algorithms to data analytics techniques, attendees gained valuable insights into how AI is reshaping the landscape of federal data collection and analysis.

Annual Integrated Economic Survey

The U.S. Census Bureau launched its newest survey, the Annual Integrated Economic Survey (AIES), on March 15, 2024. (https://www.census.gov/programs-surveys/aies.html) This ambitious initiative consolidates seven separate annual business surveys into one comprehensive program encompassing all sectors of the U.S. economy. The AIES marks a significant departure in the Economic Directorate's practices, prompting the reimagination of data collection, processing, analysis, and dissemination methods. It incorporates many of the key recommendations from a comprehensive review of our economic programs by a panel from the National Academy of Sciences (NAS) issued in the 2018 report. (https://nap.nationalacademies.org/catalog/25098/reengineering-the-census-bureaus-annual-economic-surveys) As a result, the Census Bureau consolidated the Annual Retail Trade Survey (ARTS), the Annual Wholesale Trade Survey (AWTS), the Service Annual Survey (SAS), the Annual Survey of Manufactures (ASM), the Annual Capital Expenditures Survey (ACES), the Manufacturers' Unfilled Orders Survey (M3UFO), and the Report of Organization (COS) into a single survey.

The AIES leverages key Census Bureau infrastructure investments to improve data collection, processing, and dissemination. This allows the AIES significantly greater adaptability in adjusting survey content, sampling methods, and integrating non-survey source data compared to what was achievable with various specialized collection platforms employed in previous annual surveys. The AIES sample design requirements are informed by the user community's longstanding data needs (e.g., national and sub-national tabulations), as well as by extensive respondent research on collection. The AIES sample unit is the company (firm), which can operate in one or more industries and can own one or more establishments. This reduces the reporting burden for medium- and large-scale companies over the previous practice of requesting separate reports from the same company by sector activity (i.e., use a "company-centric, rather than industry-centric, collection strategy").

The AIES sampling units are clusters of unequal size, with the sampling procedure selecting all establishments within the sampled company. Since economic activity is measured within industry using the North American Industry Classification System (NAICS), the AIES sample design provides a "crosswalk" between company and industry. Inclusion probabilities account for company contributions to industries at national and subnational levels, and the allocation procedure likewise for unit size. The sample design accounts for the inherent skewness in the economic population(s) under consideration, designating many of the largest or most organizationally complex companies on the frame as self-representing (included with probability one). The remaining companies are stratified into disjoint "noncertainty" strata, from which unequal probability samples are selected

using sequential random (Chromy) sampling. Ongoing research activities include post-data collection edit and imputation methods, replication variance estimation, and quality metrics (e.g., response rate, imputation rate computation).

AIES data will release estimates from the 2023 data collection in July 2025.

Conferences on survey statistics and related areas

Workshop on Modern Methods in Survey Sampling

July 8- to 10, 2024
University of Ottawa, Canada
https://www.eventbrite.ca/e/canssi-crt-workshop-on-modern-methods-in-survey-sampling-tickets-852432878687?aff=oddtdtcreator

Complex surveys play an important role in providing information for policy makers and the general public as well as many scientific areas, such as public health and social science research. The objective of this workshop is to take stock of new developments in the field of survey data, to bring together some of the most active researchers in the field, and to identify the current challenges. The workshop is the final activity of a three-year Collaborative Research Team project funded by the Canadian Statistical Sciences Institute. The project is called “Modern Methods in Survey Sampling”. The workshop will cover a range of topics, including:

- Machine learning methods
- Data integration techniques
- High-dimensional data
- Small area estimation

The BNU Workshop on Survey Statistics

August 26 to 30 2024
"**Data Integration and Population Size Estimation**"
Poznań University of Economics and Business, Poland

https://wiki.helsinki.fi/xwiki/bin/view/BNU/Events/Workshop%20on%20Survey%20Statistics%202024/

Australasian Applied Statistics Conference

September 2 to 4, 2024
Rottnest Island, Western Australia
https://aasc2024conf.netlify.app/

This conference is in the series which originated with the Australasian Genstat Conference held in Canberra in 1979. It became the Australasian Applied Statistics Conference in 2011 to encompass the wider community of applied statisticians. AASC conferences have been held in Palm (Aus), Bermagui (Aus), Rotorua (NZ) and Inverloch (Aus).

The organizers passionately welcome anyone studying or working in statistics, data science, and related fields to join us for an enlightening journey into how statistical methodologies can be leveraged across various biological disciplines.

ITSEW2024 - International Total Survey Error Workshop (ITSEW)

September 18 to 24, 2024

National Institute of Statistical sciences (NISS), United States

Location: George Washington University, Washington

https://www.niss.org/events/international-total-survey-error-workshop-itsew

The goal of ITSEW is to promote discussion of questions of research, methodology and practice relating to Total Survey Error (TSE) and Total Survey Quality, and Blended Data.

Conference PSD2024: Privacy in Statistical Databases 2024

September 25 to 27, 2024.

Antibes Juan-les-Pins, France

https://crises-deim.urv.cat/psd2024/

The purpose of PSD 2024 is to attract worldwide, high-level research in statistical database privacy.

The conference is organized by the CRISES research group with proceedings published by Springer-Verlag in Lecture Notes in Computer Science.

2024 International Methodology Symposium

"Shaping the future of official statistics"
October 29 – November 1, 2024
Ottawa, Ontario, Canada
Offices of Statistics Canada (with a virtual option)

Email: statcan.symposium2024-symposium2024.statcan@statcan.gc.ca
http://www.statcan.gc.ca/eng/conferences/symposium2024/index.

The conference proceedings will be published online.

13th International Francophone Conference on Sample Surveys

November **5** to **8, 2024**
University of Luxembourg in Esch-Belval
https://sondages2024.sciencesconf.org/

STATEC , the University of Luxembourg and the Société Luxembourgeoise de Statistique will host on the Campus of the University of Luxembourg in Esch-Belval, the *13th international French-speaking conference on surveys*. Tuesday, November 5, 2024, will be devoted to training workshops that will be offered to event participants.

This international scientific meeting has been organized since 1997, every two to three years, under the aegis of the French Statistical Society (SFdS) . It brings together researchers and practitioners, from public institutes or the private sector, who carry out or use sample surveys.

The purpose of the 2024 conference is to take stock of the state of practices and research in the various areas of survey and polling methodology. It is also about bringing together all the people working in the field of surveys, whether study designers, collection managers or data users.

The R Project - The Use of R in Official Statistics – uRos2024

November 27 to 29, 2024

Hellenic Statistical Authority (ELSTAT), Pireos, Greece

https://r-project.ro/conference2024.html

Over the last two decades R has become the *lingua franca* for statisticians, methodologists and data scientists worldwide. The reasons why the official statistics community is rapidly embracing R are clear: it has an active worldwide community of users, there is wide support from the industry and it combines a vast amount of functionalities for data preparation, methodology, visualisation and application building. Moreover, R-based software is exchanged through strictly enforced technical standards: *"R is probably the most thoroughly validated statistics software on Earth." – Uwe Ligges, CRAN maintainer (useR!2017).*

The 9th Conference Survey Sampling in Economic and Social Research

December 3 to 4, 2024
University of Economics in Katowice, Poland
Online
https://sites.google.com/uekat.pl/survey-sampling

NTTS2025: New Techniques and Technologies for Statistics

March 11 to 13, 2025
Bruxelles, Belgium.

https://cros.ec.europa.eu/dashboard/ntts-2025

The NTTS conference is organized by Eurostat and devoted to statistical innovation and how to best use Data and Statistics for taking informed decisions. Topics include data collection and integration, innovation in statistics, Artificial Intelligence, data analytics, estimation and analysis, data reuse and sharing, outreach to users, communications and dissemination.
NTTS 2025 is financed by the European Commission and there are no attendance fees.

The 9th Italian Conference on Survey Methodology (ITACOSM 2025)

July 1 to 4, 2025
The University of Bologna, Italy

ITACOSM is a bi-annual international conference organized by the Survey Sampling Group (S2G) of the Italian Statistical Society (SIS) whose aim is promoting the scientific discussion on the developments of theory and application of survey sampling methodologies in the fields of economics, social and demographic sciences, of official statistics and in the studies on biological and environmental phenomena.

IASS satellite meeting SAE2025

July 8 to 12, 2025, Turin, Italy
Venue: Castello del Valentino
https://castellodelvalentino.polito.it/?page_id=47 [castellodelvalentino.polito.it

11th European Survey Research Association Conference ESRA 2025

July 14 to 18, 2025

Venue: Utrecht University in Utrecht, The Netherlands

The conference theme "**Promises and problems of new and alternative data sources and data formats for survey research. Methodological challenges and substantive conclusions**"

https://www.europeansurveyresearch.org/conference/utrecht-2025/call-sessions/

Call for Session Proposals to ESRA 2025 is now open!

65th ISI World Statistics Congress 2025

October 5 to 9, 2025

The Hague, The Netherlands

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

The ISI World Statistics Congress (WSC) is the leading congress on Statistics & Data Science worldwide. It is held every two years, since 1887 by the International Statistical Institute (ISI). The organizers welcome you at the 65th edition: ISI WSC 2025!

In Other Journals

Journal of Survey Statistics and Methodology

Volume 12, Issue 1, February 2024

https://academic.oup.com/jssam/issue/12/1

Survey Methodology

Interviewer Involvement in Respondent Selection Moderates the Relationship between Response Rates and Sample Bias in Cross-National Survey Projects in Europe
Marta Kołczynska, Piotr Jabkowski, and Stephanie Eckman

Detecting Interviewer Fraud Using Multilevel Models
Lukas Olbrich, Yuliya Kosyakova, Joseph W. Sakshaug, and Silvia Schwanha

Panel Conditioning in a Probability-Based Longitudinal Study: A Comparison of Respondents with Different Levels of Survey Experience
Fabienne Kraemer, Henning Silber, Bella Struminskaya, Matthias Sand, Michael Bosnjak, Joanna Koßmann, and Bernd Weiß

Reducing Burden in a Web Survey through Dependent Interviewing
Curtiss Engstrom and Jennifer Sinibaldi

Equipping the Offline Population with Internet Access in an Online Panel: Does It Make a Difference?
Ruben L. Bach, Carina Cornesse, and Jessica Daikeler

Is there a Day of the Week Effect on Panel Response Rate to an Online Questionnaire Email Invitation?
Chloe Howard, Lara M. Greaves, Danny Osborne, and Chris G. Sibley

Survey Statistics

Handling Missing Values in Surveys With Complex Study Design: A Simulation Study
Natasa Kalpourtzi, James R. Carpenter, and Giota Touloumi

Modeling Public Opinion Over Time: a Simulation Study of Latent Trend Models
Marta Kołczynska and Paul-Christian Burkner

Using Auxiliary Marginal Distributions in Imputations for Nonresponse while Accounting for Survey Weights, with Application to Estimating Voter Turnout
Jiurui Tang, D. Sunshine Hillygus, and Jerome P. Reiter

Joint Imputation of General Data
Michael W. Robbins

Jackknife Bias-Corrected Generalized Regression Estimator in Survey Sampling
Marius Stefan and Michael A. Hidiroglou

Maximum Entropy Design by a Markov Chain Process
Yves Tillé and Bardia Panahbehagh

Applications

Modeling Group-Specific Interviewer Effects on Survey Participation Using Separate Coding for Random Slopes in Multilevel Models
Jessica M. E. Herzing, Annelies G. Blom, and Bart Meuleman

Volume 12, Issue 2, April 2024

https://academic.oup.com/jssam/issue/12/2

2022 Morris Hansen Lecture

Hansen Lecture 2022: The Evolution of the Use of Models in Survey Sampling
Richard Valliant

Discussion of the 2022 Hansen Lecture: "The Evolution of the Use of Models in Survey Sampling"
F. Jay Breidt

Modeling in Sample Surveys: Discussion of Professor Valliant's Hansen Lecture 2022
Trivellore Raghunathan

Survey Statistics

Multivariate Small-Area Estimation for Mixed-type Response Variables with Item Nonresponse
Hao Sun, Emily Berg, and Zhengyuan Zhu

Pseudo-Bayesian Small-Area Estimation
Gauri Datta, Juhyung Lee, and Jiacheng Li

Small Area Poverty Estimation Under Heteroskedasticity
Sumonkanti Das and Ray Chambers

Poverty Mapping Under Area-Level Random Regression Coefficient Poisson Models
Naomi Diz-Rosales, Maria Jose Lombardia, and Domingo Morales

Survey Methodology

Leveraging Predictive Modelling from Multiple Sources of Big Data to Improve Sample Efficiency and Reduce Survey Nonresponse Error
David Dutwin, Patrick Coyle, Joshua Lerner, Ipek Bilgen, and Ned English

Bayesian Integration of Probability and Nonprobability Samples for Logistic Regression
Camilla Salvatore, Silvia Biffignandi, Joseph W. Sakshaug, Arkadiusz Wisniowski, and Bella Struminskaya

Automated Classification for Open-Ended Questions with BERT
Hyukjun Gweon and Matthias Schonlau

Correction to: Leveraging Predictive Modelling from Multiple Sources of Big Data to Improve Sample Efficiency and Reduce Survey Nonresponse Error
David Dutwin, Patrick Coyle, Joshua Lerner, Ipek Bilgen, and Ned English

Journal of Official Statistics

Volume 40 (2024): Issue 1 (March 2024)

https://journals.sagepub.com/toc/jofa/40/1

Steps Toward a Shared Infrastructure for Multi-Party Secure Private Computing in Official Statistics
Fabio Ricciato

Some Open Questions on Multiple-Source Extensions of Adaptive-Survey Design Concepts and Methods
Stephanie M. Coffey, Jaya Damineni, John Eltinge, Anup Mathur, Kayla Varela and Allison Zotti

Visualizing the Shelf Life of Population Forecasts: A Simple Approach to Communicating Forecast Uncertainty
Tom Wilson

Nonresponse Bias of Japanese Wage Statistics
Daiji Kawaguchi and Takahiro Toriyabe

Structural Break in the Norwegian Labor Force Survey Due to a Redesign During a Pandemic
Håvard Hungnes, Terje Skjerpen, Jørn Ivar Hamre, Xiaoming Chen Jansen, Dinh Quang Pham and Ole Sandvik

Bayesian Inference for Repeated Measures Under Informative Sampling
Terrance D. Savitsky, Luis G. León-Novelo and Helen Engle

On the Validity of Using Webpage Texts to Identify the Target Population of a Survey: An Application to Detect Online Platforms
Piet Daas, Wolter Hassink and Bart Klijs

Survey Research Methods

Journal of the European Survey Research Association

Vol 18 No 1 (2023)

https://ojs.ub.uni-konstanz.de/srm/issue/view/238

Modeling Public Opinion Over Time and Space: Trust in State Institutions in Europe, 1989-2019
Marta Kołczyńska, Paul-Christian Bürkner, Lauren Kennedy, Aki Vehtari

The Poisson Extension of the Unrelated Question Model: Improving Surveys With Time-Constrained Questions on Sensitive Topics
Benedikt Iberl, Anesa Aljovic, Rolf Ulrich, Fabiola Reiber

Question Wording Matters in Measuring Frequency of Fear of Crime: A Survey Experiment of the Anchoring Effect
Aubrey L. Etopio, Emily R. Berthelot

Evaluating the Effect of Monetary Incentives on Web Survey Response Rates in the UK Millennium Cohort Study
Charlotte Booth, Erica Wong, Matt Brown, Emla Fitzsimons

We Have Come a Long Way and We Have a Long Way to Go: A Cross-Survey Comparison of Data Quality in 16 Arab Countries in the Arab Barometer vs the World Values Survey
Saskia Glas, Veronica Kostenko

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.	Christopher	Antoun	United States
MR.	Kwamena Leo	Arkafra	Ghana
PROF. DR.	Jonathan	Auerbach	United States
DR.	Veronica	Ballerini	Italy
MRS	Rosalyn	Berry	United States
DRS	Gaia	Bertarelli	Italy
DR.	Haoyi	Chen	United States
DR.	Mamadou S	Diallo	United States
PROF	Maria	Ferrante	Italy
DR.	Pramod Kumar	Gupta	India
MR.	Karol	Krotki	United States
DR.	Yukiko	Kurihara	Japan
PROF	Thomas	Lumley	New Zealand
MR.	Kouadio J. Stephane	N'zi	Côte d'Ivoire
DR.	Jennifer	Park	United States
MR.	Francisco	Quiroa	Guatemala
MR.	Benjamin	Schneider	United States
DR.	Christine	Wells	United States

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 Cimmittee 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 & Inter., Luxembourg

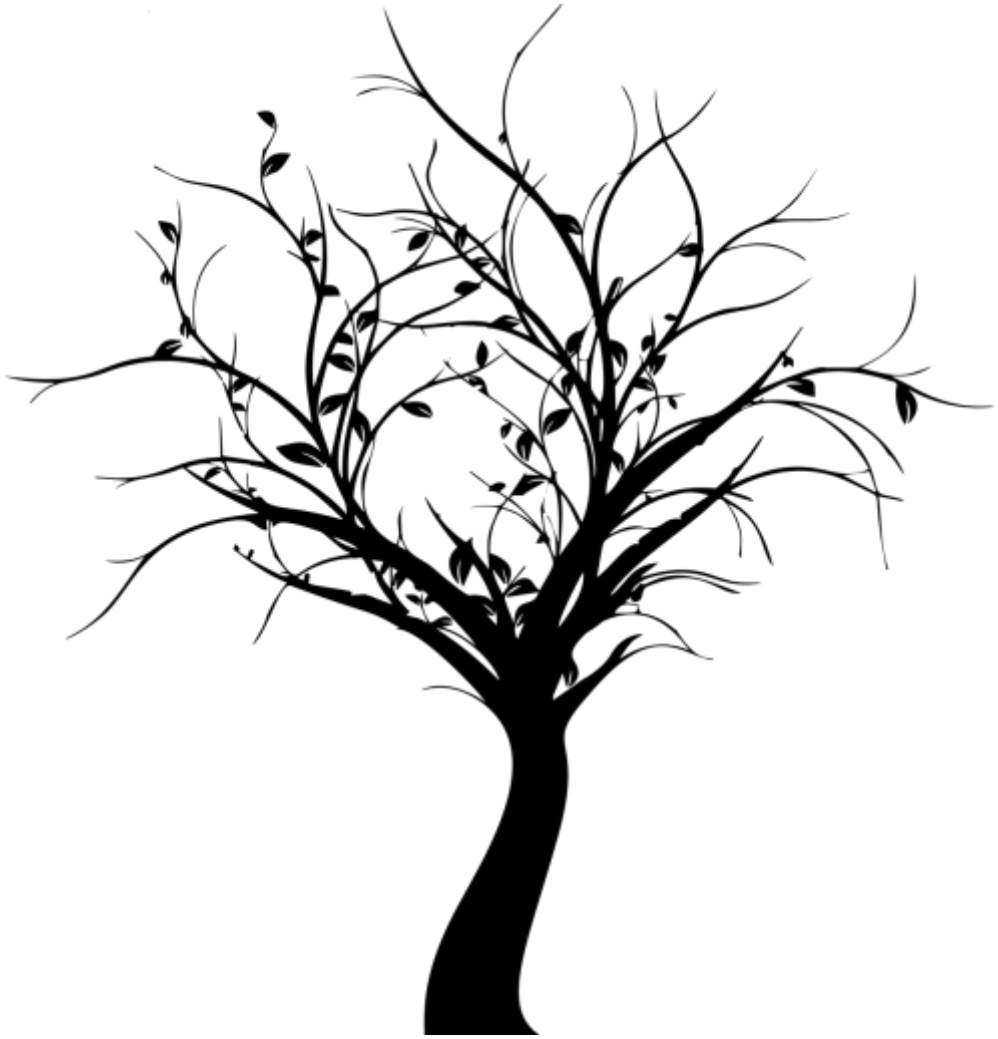
National statistical offices:

- Australian Bureau of Statistics, Australia
- Instituto Brasileno de Geografia y Estatística (IBGE), Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistisches Bundesamt (Destatis), Germany
- International Rel. & Statistical Coordination, Israel
- Istituto nazionale di statistica (ISTAT), Italy
- Univ. of Tuscany, Dept. Economics & Management, Italy
- Statistics Korea (KOSTAT), Republic of Korea
- Direcção dos Serviços de Estatística e Census (DSEC), Macao, SAR China
- Statistics Mauritius, Mauritius
- Statistics New Zealand, New Zealand
- Statistics Norway, Norway
- Instituto Nacional de Estatística (INE), Portugal
- Statistics Sweden, Sweden
- Office for National Statistics Service (ONS), United Kingdom
- National Agricultural Statistics Service (NASS), United States

Other statistical organizations:

- Institut Public de Sondage d'Opinion Secteur (Ipsos), Italy
- WESTAT Inc., United States

Read *the Survey Statistician* online!



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