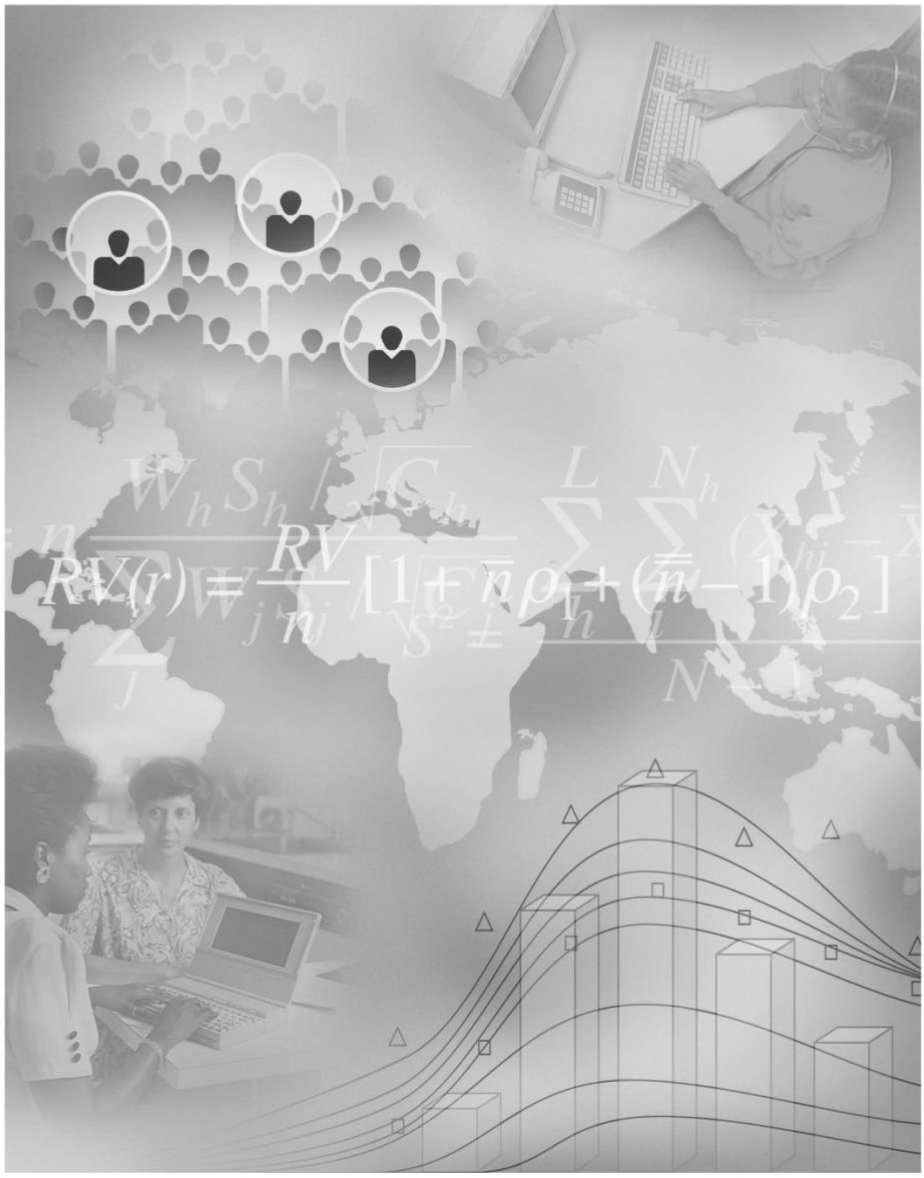
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

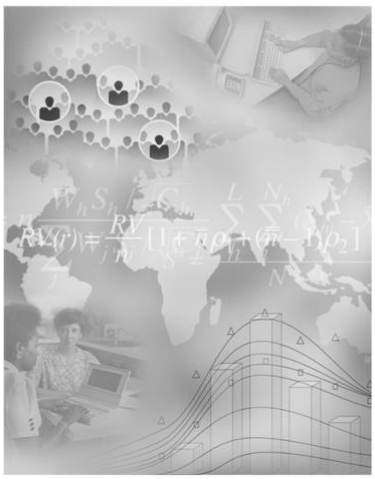
No. 86 **July 2022**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

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



The Survey Statistician No. 86, July 2022

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

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

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

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

ISSN 2521-991X

In this Issue

Letter from the Editors

Dear readers,

We are happy to present the July 2022 issue of The Survey Statistician.

Before going any further in this letter, we cannot simply walk by the fact that the work of fellow survey statisticians has been interrupted in Ukraine since February 24, 2022, when Russia started the war against Ukraine. Not only is survey work compromised, but the population itself has been profoundly and cruelly affected. Population size is decreasing and unknown due to the fact that women and children have been leaving Ukraine, that people have been killed by bombs and rockets and that people have been lifted from Ukraine to Russia. Internet data collection is not adequate because a lot of infrastructure and residential buildings are destroyed and citizens are left without roof, computers, internet connection and some of them even without life. Face to face interviewing is not possible in large parts of the country because of the risks to both interviewers and respondents. Many enterprises are destroyed, agriculture fields are mined and so traditional ways to conducting surveys are no longer feasible. Some of the University teachers and students cannot be engaged into survey statistics studies and research because they are defending their country, and not all of them are still alive.

So, President Putin has to stop this war in Ukraine! Russian soldiers should be going home away from Ukraine. Destruction needs to cease for statisticians to carry on their work, for people to continue / go back to their normal life. In the meantime, let us all continue to support them to the best that we can.

This issue starts with the *Letter from the President* by Monica Pratesi, where she explains the importance of the IASS strategy and highlights the vision for our important association. It is followed by the *Report from the Scientific Secretary*, Giovanna Ranalli.

In the News and Announcement section, after the announcement of an award and of the World Statistics Congress 2023, a tribute is paid to two giants of the survey world: Fred Smith and Chris Skinner who passed away no so long ago and for whom a conference took place in their honour.

In the Ask-the-Experts section, Natalie Shlomo from the Social Statistics Department, School of Social Sciences at the University of Manchester, UK explains how to measure disclosure risk in microdata. She describes the two types of disclosure risks (identity and attribute disclosure) and goes on to state how this applies also to synthetic data and that there still development needed in this area. In the New and Emerging section, Patrick Krennmair, Nora Würz from Freie Universitat Berlin and Timo Schmid from Otto-Friedrich-Universitat Bamberg, Germany, demonstrate how tree-based Machine Learning techniques can be applied to small area estimation.

In the Book and Software Review section, Mamadou S. Diallo from the Saudi Center for opinion Polling (SCP), Saudi Arabia, presents SAMPLICS, a library for survey sampling in Python. It includes techniques of sampling, weighting and estimation, including small area estimation among other methods. It is followed by a review of the Handbook of Web Surveys, a book by Silvia Biffignardi and Jelke Bethlehem presented by Gaia Bertarelli from the EMbeDS Department, Institute of Management, Sant'Anna School of Advanced Studies, Pisa (Italy).

Then follows the country reports, the list of upcoming conferences and recently published articles in various journals. We would like to express our many thanks to the section editors for their attentive and timely work. In particular, thank you to Natalie Shlomo for having updated the list of the country

representatives. This work resulted in 18 interesting country reports to the current issue of TSS. Thank you also to Peter Wright for editing these 18 country reports.

The health of *The Survey Statistician* depends on participation by many members. If you have any information about conferences, events or just ideas you would like to share with other statisticians – please do go ahead and contact any member of the editorial board of the newsletter.

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

Danutė Krapavickaitė (danute.krapavickaite@gmail.com)

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

Letter from the President

Dear IASS Members,

In this letter I explain what the IASS strategy is and what it means for the members. The draft strategy was made available for comments of the IASS members with a deadline of 15 June 2022. Comments received have been incorporated and the revised version will be ratified by the IASS Assembly on July, 14th 2022 (link will be sent in due time).

The vision of the IASS is a world where good survey theory and practice provide governments, businesses and civil society with the information they need to make good decisions. The IASS motto "Promoting good survey theory and practice around the world" is a fundamental task to promote progress in our society.

In fact, the data production process is at the hearth of understanding phenomena and of decision making. Survey methods, sample designs, register based statistics, integration of new data sources and the resulting design-based, model-assisted and model-based statistics play a strategic role in it.

Our strategy to move towards this vision consider the survey statistician:

- as a researcher in Universities or in other Research Institutions and companies, teaching the discipline of statistics and survey methods
- as a professional, involved in the data collection process and in the analysis of survey data and statistics
- as a scholar/ young researcher, seeking for mentoring and advising from senior experts

We consider also that all of us are promoting the use of statistics in the public interest; and are interested in improving statistical literacy and understanding of how data collection and survey methods are pivotal elements. Indeed, they are of crucial importance in relation to the development or success of understanding phenomena and of decision making.

During our current strategic period of 2022-2023 we have two particular themes we are focused upon across all of our strategic objectives. The first is engaging our members to help us deliver these goals. We do this through several actions: increasing the contact and interaction with our renovated country representatives, supporting local conferences, as well as regular communication with members, also via Twitter and LinkedIn. The second cross-cutting theme we are working on is the rise of new survey methods and integration of data sources and what it means for our work and that of survey methodologist. For this we are promoting and organizing the IASS Webinar series. They are open to the wide public and treat the emerging issues in survey sampling and survey data analysis.

Obviously, there are a wide range of views about what new survey methods and data integration is and its implications - the EC IASS took a clear view that we should be as inclusive as possible. In particular there are new challenges for survey statisticians in analysing large unstructured datasets via machine learning and data analytics, but equally there are specific and analytical skills which survey statisticians bring to the discussion which not all those coming to 'data science' may be familiar with. Uncertainty in data collection and sources of errors, measuring errors and selection bias, profiling the quality of collected data, sampling designs applied to reduce datasets dimensionality …just to cite some of them.

In spite of being a relatively small organisation, we hope to have a big impact. The only way we can achieve this is by having you, our members, involved. That's why we've decided to disseminate this

letter and the report from the Scientific Secretary- to showcase all of the work we have done in the first months of 2022 and what we are doing together through the Association.

You can view and download our draft strategy to finalize during the July IASS Assembly here (http://isi-iass.org/home/wp-content/uploads/Draft-IASS-Strategy-2022.pdf)

With my best wishes,

Monica Pratesi

IASS President

Report from the Scientific Secretary

I have been appointed Scientific Secretary of IASS during the first meeting of the newly elected IASS EC in September. I am very grateful to the members of the EC for their trust, and I am indebted to James Chipperfield for his legacy on this role. As my first duty, I had to choose a topic for the

The past months have been busy with drafting the strategic plan of the IASS as illustrated in the letter from the President. This activity has involved all members of the EC and has helped us to have a clearer insight on the goals of our Association and to better tailor our activities in order to achieve them. On a personal side, the past months have also been busy with the organization of the IASS supported conference **ITACOSM-2022** held in Perugia (Italy) on June 8-10. The 7th biannual ITAlian Conference on Survey Methodology has had a focus on Survey Methods for Statistical Data Integration and New Data Sources and has involved 101 registered participants from all over the world, among which 72 IASS individual members or affiliates of an institutional IASS member. The program has been very rich and featured three keynote presentations by Prof. Jae-Kwang Kim (Iowa State University), Anne Ruiz-Gazen (Toulouse School of Economics) and Peter Van Der Heijden (Utrecht University and University of Southampton), 36 invited papers and 30 contributed papers. Prof. Kim also delivered a truly insightful one-day short course on data integration methods on June 7. The conference has been the first occasion after the pandemic for many survey statisticians to gather again in person and to look more deeply into all those research topics that we are so passionate about.

I noticed that a good number of talks at the conference had a focus on the use of machine learning methods for survey estimation, and these came both from researchers in the Academia and at National Statistical Institutes. This has reassured me in the choice I had made for the topic for the **New and emerging methods** section of this issue of The Survey Statistician. I am very grateful to Prof. Patrick Krennmair and Nora Würz (Freie Universität Berlin) and to Prof. Timo Schmid (Otto-Friedrich-Universität Bamberg) for having accepted my invitation to write a paper on **Tree-Based Machine Learning in Small Area Estimation.**

Small area estimation methods are fundamental to obtain estimates of the spatial distribution of socio-economic indicators when direct estimates from survey data are not reliable because of a small domain sample size. The contribution relaxes the assumption of a linear relationship between the covariates and the variable of interest in unit level small area models by means of random forests. In addition, it accounts for hierarchically dependent data extending random forests to include area random effects. The method is illustrated using open-source income data from Austria. Please, contact me if you are interested in writing an article for the “New and emerging methods” of future editions of The Survey Statistician.

This issue of The Survey Statistician also features a contribution of our President Elect, Prof. Natalie Shlomo (Univ. of Manchester), in the **Ask the Expert** section on “**How to Measure Disclosure Risk in Microdata?”**. In particular, Prof. Shlomo distinguishes between microdata released from social surveys that have undergone statistical disclosure control methods and synthetic microdata generated from statistical modelling.

The organization of the monthly **Webinar series** has continued, and we are particularly thankful to Andrea Diniz da Silva for her engagement. We have now reached Webinar number 18 and we are happy to have made it a monthly appointment that has attracted an audience of up to three hundred registered participants. Please, visit the webinar section of our website http://isi-iass.org/home/webinars/ for slides, that of ISI https://www.isi-web.org/events/webinars for upcoming

and recorded webinars, and contact Andrea andrea.silva@ibge.gov.br if you have suggestions for topics and/or speakers for the upcoming Webinars. Those held in the first six months of 2022 have covered inference from non-probability samples through data integration and using model-based methods, approaches for combining data from multiple probability samples, three-form split questionnaire designs for panel surveys, and spatial sampling and geospatial information for monitoring agriculture. In addition, a special IASS Webinar was held on May 25th in **memory of Prof. Jean-Claude Deville**: Prof. Alina Matei (Univ. de Neuchâtel) gave an insightful introduction to the life and research of Prof. Deville, while Prof. Camelia Goga (Univ. de Bourgogne Franche-Comté) gave a talk on recent advances of calibration estimation in a high dimensional setting and Prof. Yves Tillé (Univ. de Neuchâtel) discussed a new method for statistical matching that uses calibration and highly stratified balanced sampling. This initiative paired that organized at the ITACOSM conference where one invited session has been in memory of Prof. Jean-Claude Deville with talks by Prof. Changbao Wu (University of Waterloo) on new developments of calibration, by Dr. Francesco Pantalone (University of Southampton) on recent advances of balanced sampling, and by Prof. Guillaume Chauvet (ENSAI) on bootstrap methods for variance estimation. Prof. Yves Tillé (Univ. de Neuchâtel) has chaired and discussed the session that aimed to be a homage to a milestone researcher in survey sampling and estimation methods as well as a great man.

One upcoming Webinar in the Fall will be devoted to the first winner of the recently established **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 recognize researchers who are close to Dr Chandra's career trajectory. The IASS prize committee has been appointed by the IASS EC and is composed by Nikos Tzavidis (Univ. of Southampton and member of the IASS EC), Alina Matei (Univ. de Neuchâtel), David Haziza (University of Ottawa), and Aberash Tariku (Ethiopian Statistics Service). Please follow the updates on this and on the life of IASS reading our **monthly Newsletter**. Other than webinars, information on conferences, on the recipients of awards and on call for nominations, it now features a new **book of the month** section. Please, feel free to contact me for news and info to be added in the Newsletter by the 15th of each month.

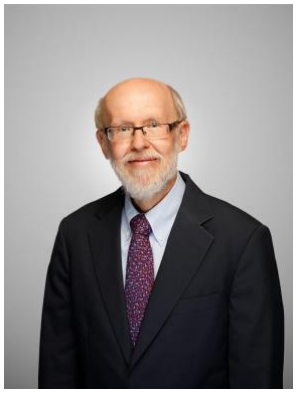
Maria Giovanna Ranalli

maria.ranalli@unipg.it

IASS Scientific Secretary

News and Announcements

Award



Small Area Estimation Outstanding Contribution Robert E. Fay of Westat has been awarded the 2022 Award for Outstanding Contribution to Small Area Estimation (SAE), which was presented at the 2022 SAE conference in May. The award recognises contribution to the research, application, and education of SAE. Previous awardees are J.N.K. Rao, Danny Pfeffermann, Malay Ghosh, Partha Lahiri, and Wayne Fuller.

The 64th ISI World Statistics Congress

The International Statistical Institute will be holding the 64th World Statistics Congress in Ottawa, Canada from July 16th to July 20th, 2023. The World Statistics Congress is a unique opportunity to share information, meet friends and colleagues and increase networks. The conference hopes to see many IASS members participating in the event. More information can be found at: 64th ISI World Statistics Congress - Ottawa, Canada | ISI (isi2023.org).

Conference report - Conference in Honour of Fred Smith & Chris Skinner, 7-9 July 2021

The IASS satellite meeting of the 2021 World Statistics Congress was hosted by the University of Southampton in association with the London School of Economics and Political Science (LSE) for a Conference in honour of Fred Smith and Chris Skinner, two giants of the survey world who passed away close together in winter 2019/20 (you can find their obituaries at https://rss.onlinelibrary.wiley.com/doi/full/10.1111/rssa.12580 and https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12600 respectively), and who both spent large portions of their careers at Southampton. The conference was connected with submissions for a special issue of the Journal of the Royal Statistical Society, Series A, which will appear in due course once the review processes have concluded.

The conference opened with a series of short reminiscences of Fred and Chris, led by Tim Holt, and containing many moving memories of the kindness, patience and questioning nature which characterised the contributions of Fred and Chris to academic life in the widest sense. This session was recorded, and is available on-line at <to follow>. Many other contributors to the conference provided stories in their own sessions, too, but these were not recorded.

There was a series of invited talks from many of those who had worked with Fred and Chris, led off by Danny Pfeffermann, who reviewed time series modelling for longitudinal data, starting from the work of Fred Smith. Chris Skinner made it onto the programme of the conference thanks to some joint work with Natalie Shlomo on measuring re-identification risk in microdata, and there was a contributed session devoted to papers on statistical disclosure control, a research area where Chris was an international leader. Jae Kwang Kim and Jon Rao finished the invited sessions on day 1 with methods for analysis of clustered data obtained by two-stage informative sampling.

At the beginning of the second day, David Steel and Ray Chambers, both of whom spent time in Southampton, presented papers in an invited session from Australia, on sample design for analysis using high influence probability sampling; and weighting, informativeness and causal inference respectively. Graham Kalton reviewed the history of probability and nonprobability sampling, and Wayne Fuller (who had just turned 90) presented a paper on post strata based on sample quantiles.

Chris Skinner was awarded the Waksberg prize in 2019, and his family graciously donated the prize money to support a student competition. A good selection of entries was received, and congratulations to the winners Loveness Dzitiki, Caio Goncalves, Dehua Tao, Estelle Medous, Fernanda Lang Schumacher and Luiz Eduardo da Silva Gomes who all made excellent presentations at the conference (two in contributed sessions and the others in a special student session).

The final day began with a discussion session on the need for a system for dealing with statistical information requirements in the time of a pandemic, and how this could be set up and joined up. Dennis Trewin led the discussion, with contributions from Pedro Silva, David Steel and Len Cook. Two contributed sessions covered a wide range of topics, from cross-classified sampling (another

topic where Chris Skinner had undertaken some research) to the use of neural networks and random forests in survey sampling and estimation.

The final session consisted of papers by Denise and Pedro Silva, who were PhD students under Fred and Chris respectively. Denise talked about compositional analysis of labour force status in the Brazilian LFS, and Pedro about fitting multilevel models under informative sampling.

The online format meant that many people joined the conference at awkward times of their day, and the organisers really appreciated the efforts of those who got up early and stayed up late to attend a wide range of sessions. It was great to have an opportunity to meet up with old and new friends, and even to have some discussions in the breaks in the programme. We would like to thank everyone who contributed to making the conference a success, and look forward to a time when we can all gather again face-to-face.

Paul Smith & Peter Smith
University of Southampton

Conference report - Day-long virtual conference on latest developments in the theory and practice of sample surveys and censuses organized by Pak Institute of Statistical Training and Research (PISTAR) on March 12, 2022

The Pak Institute of Statistical Training and Research (PISTAR) organized a day-long virtual conference on “Latest Developments in the Theory and Practice of Sample Surveys and Censuses”. on Saturday, March 12, 2022 followed by a post-conference workshop on Utilization of Remote Sensing in Sample Surveys and Censuses on Sunday, March 13, 2022. Both the conference and workshop were sponsored by the International Association of Survey Statisticians (IASS).

The first session comprised an invited talk by Mr. Isaac Shahzad on behalf of the Director General, Bureau of Statistics, Punjab, Pakistan. He described basics of sampling terminologies and various methods of random and non-random sampling, household sampling techniques. There were then many sessions on topics such as neural network calibration in surveys as well as other calibration methods, measuring adoption rates, measuring crime rates and use of GIS in agriculture.

There was a presentation by Ms. Rabia Awan on behalf of Mr. Muhammad Sarwar Gondal, Member, Pakistan Bureau of Statistics (PBS) on the 7th Population & Housing Census, the first-ever digital census of Pakistan.

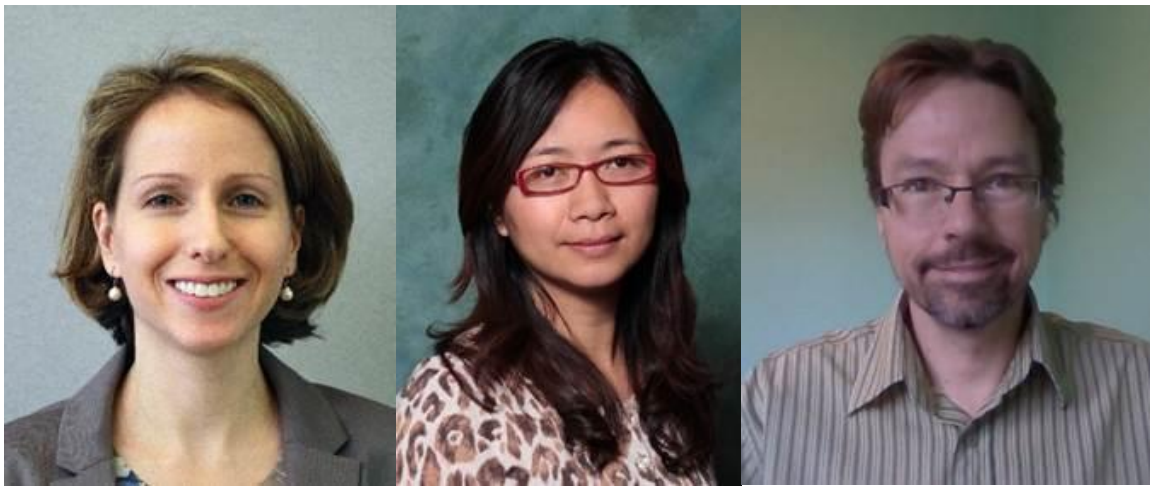
One of the highlight sessions featured Prof. Jae-Kwang Kim from Iowa State University, USA, Prof. Li-Chun Zhang from the University of Southampton in UK / Statistics Norway / University of Oslo, Norway as well as Dr. Zhonglei Wang from the Wang Yanan Institute for Studies in Economics and the School of Economics, Xiamen University, China. They presented on graph sampling, calibration for non-probability surveys, and correction of selection bias.

Prof. Monica Pratesi, President, International Association of Survey Statisticians, presented her keynote address. She discussed in detail the scope, the working, the activities and the progress of IASS.

The event also included the announcement of winners for the Best Research Paper Competition as well as winners of Best Poster Competition.

Morris-Hansen Lecture

On March 1, the Washington Statistical Society held its 29th annual Morris-Hansen Lecture Series. The theme of the event was: Working with Non-Probability Samples: Assessing and Remediating Bias. The speakers were Courtney Kennedy, Director of Survey Research, Pew Research Centre; Yan Li, Joint Program in Survey Methodology, University of Maryland; Jean-François Beaumont, Senior Statistical Advisor, Statistics Canada. Courtney Kennedy presented on: Exploring the Assumption That Online Opt-in Respondents Are Answering in Good Faith. Yan Li presented on: Exchangeability Assumption in Propensity-Score Based Adjustment Methods for Population Mean Estimation Using Non-Probability Samples. Jean-François Beaumont presented on: Reducing the bias of non-probability sample estimators through inverse probability weighting with an application to Statistics Canada's crowdsourcing data.



C. Kennedy Y. Li J.F. Beaumont

The Survey Statistician, 2022, Vol. 86, 13-21

Ask the Experts

How to Measure Disclosure Risk in Microdata?

Natalie Shlomo[1]

[1] Social Statistics Department, School of Social Sciences, University of Manchester, United Kingdom, natalie.shlomo@manchester.ac.uk

Abstract

In this article we answer the question on how to measure disclosure risk in microdata. We distinguish between two types of microdata: (1) microdata released from social surveys that have undergone statistical disclosure control methods; (2) synthetic microdata generated from statistical modelling. We define the types of disclosure risks and disclosure risk measures for each type of microdata.

Keywords: survey microdata; risk of re-identification; synthetic data; inferential disclosure; privacy models; disclosure risk measures

1 Introduction

Statistical data that are traditionally released by government agencies include microdata from social surveys and tabular data. For each of these traditional outputs, there have been decades of research on how to quantify disclosure risk, statistical disclosure control (SDC) methods and their impact on data utility. However, with increasing demands for new forms of data at higher resolution, in particular linked hierarchical data and 'open' data initiatives, there are even more pressures on government agencies to broaden access and to provide better solutions for the release of statistical data. Examples of solutions are to generate synthetic data based on models built from the original data or to provide access to data through flexible table builders and remote analysis servers. This has led to intensive research and collaboration between the computer science and statistical communities to develop more formal privacy guarantees under SDC and to adapt more perturbative techniques into the SDC tool-kit.

Synthetic data generation has been proposed as an alternative to standard SDC methods for the release of microdata. Traditional SDC methods aim to suppress and perturb existing datasets and often lead to a large loss in utility and analytical power. Synthetic data takes a different approach as it creates a new dataset having the same statistical properties as the original data but containing no data that directly corresponds to real population units. The idea of synthetic data was first introduced by Rubin (1993), who proposed treating each observed data point as if it were missing and imputing it conditional on the other observed data points using a posterior predictive distribution. The data elements are replaced with synthetic values generated from an appropriate probability model.

Several samples are drawn from the population to take into account the uncertainty of the model and to obtain proper variance estimates. See also Raghunathan, Reiter and Rubin (2003), Reiter (2005a and 2005b), Drechsler (2011) and references therein for more details on generating synthetic data. The synthetic data can also be implemented on parts of data so that a mixture of real and synthetic data is released (Little and Liu 2003).

Here we focus on calculating disclosure risk measures after the application of statistical disclosure control methods or the generation of synthetic data. This is in contrast to disclosure risk assessment in the Computer Science Literature where privacy guarantees are embedded in the perturbation method via a privacy model. These privacy models assume 'attack' scenarios which informs the parameterization of the privacy models according to thresholds. Examples of privacy models in the computer science literature are: k-anonymity, l-diversity and t-closeness where the parameters are k, l and t:

k-anonymity: The key identifying variables are coarsened within equivalence classes such that there are at least $k-1$ individuals in the equivalence class (Sweeney 2002). Equivalence classes are typically defined from quasi-identifying variables such as sex, age group, place of residence.

l-diversity: Determines how well-represented the values of a sensitive variable are within equivalence classes and that there are at least l well-represented values of the variable.

Entropy l-diversity (Machanavajjhala et al. 2006) is defined as follows: Let $p(EC, c)$ be the probability that a record has a value c for a categorical variable C in equivalence class EC. The entropy is: $H(EC) = -\sum_{c \epsilon C} p(EC, c) \log [p(EC, c)]$

A dataset possesses entropy l-diversity if for each EC the entropy $H(EC) \geq \log(l)$.

t-closeness (Li, et al. 2007): Requires the distribution of values of a sensitive variable within equivalence classes to be close (up to t) compared to the univariate distribution of the sensitive variable in the whole dataset.

More on these privacy models can also be found in Domingo-Ferrer, et al. (2008) and Xiao et al. (2010).

Another privacy model gaining much traction in the statistical community is differential privacy (Dwork et al. 2006). A 'worst case' scenario is allowed for, in which the potential intruder has complete information about all the units in the database except for one unit of interest. The definition of a perturbation mechanism M satisfies ε-differential privacy if for all queries on neighbouring databases a and a' differing by one individual and for all possible outcomes defined as subsets $S \in Range(M)$ we have: $p(M(a) \in S) \leq e^{\varepsilon} p(M(a') \in S)$.

This means that observing a perturbed output S, little can be learnt (up to a degree of e^{ε}) and the intruder is unable to decipher whether the output was generated from database a or a'. In other words, the ratio $p(M(a) \in S) / p(M(a') \in S)$ is very small (at most e^{ε}). The solution to guarantee differential privacy in the computer science literature is by adding noise/perturbation to the outputs of the queries under specific parameterizations based on the privacy budget ε and the sensitivity of the query, which is the maximum difference in the possible output of a query with and without the presence of a single individual.

In Section 2, I discuss the types of disclosure risks for microdata. In Section 3, I describe how to estimate a disclosure risk measure to assess the risk of re-identification in disclosure-controlled survey microdata. In section 4, I describe disclosure risk measures that can be used after the generation of synthetic data to assess attribute and inferential disclosure. I close with a conclusion in section 5.

2 Types of Disclosure Risks for Microdata

In the SDC literature, we define the notion of an 'intruder' as someone who wants to attack statistical data for malicious intent and cause a breach in confidentiality. Two main disclosure risks are: (1)

identity disclosure where a statistical unit can be identified based on a set of cross-classified quasi-identifying variables that are typically categorical, such as age, gender, occupation and place of residence; (2) attribute disclosure where new information can be learnt about an individual or a group of individuals. Disclosure risk scenarios form the basis of possible means of disclosure, for example, the ability of an intruder to match a dataset to an external public file based on a common set of quasi-identifying variables; the ability of an intruder to identify unique individuals through visible and rare attributes; the ability of an intruder to difference nested tables and obtain small counts; and the ability of an intruder to form coalitions with other intruders.

For the release of survey microdata that are disseminated from social surveys, the main concern is the risk of re-identification since this is a prerequisite for individual attribute disclosure where many sensitive variables such as income or health outcomes, can be revealed following an identification. Naturally, sampling from the population provides a priori protection since an intruder cannot be certain whether a sample unique, i.e. a unit that is unique in the sample with respect to some quasi-identifying variables, is a population unique. Note there is an implicit assumption of no 'response knowledge' meaning that the intruder does not know who was drawn into the sample of the survey.

Inferential disclosure is another type of disclosure risk that is becoming more prominent with the ongoing research and development into web-based interactive data dissemination. Inferential disclosure risk is the ability to learn new attributes with high probability and thus is a more general form of individual and group attribute disclosure and the terms are often used interchangeably. For example, datasets can be manipulated and combined in such a way that there is a high prediction power between variables in the dataset or combinations of data releases that can be differenced to reveal individual data points. Attribute disclosure and the more general inferential disclosure are particularly relevant for assessing disclosure risks in fully synthetic data. This is because there is a break in the link between quasi-identifying and sensitive variables in a fully synthetic dataset, but it may still be possible to disclose sensitive information about groups of individuals.

3 Quantifying the Risk of Re-identification in Survey Microdata

The basic definition of the risk of re-identification is the probability of correctly being able to match the survey microdata with a unit in the population. If the characteristics of the population are known, such as measured in a population register or census, this probability would be relatively straightforward to calculate. However, this is rarely the case since within government agencies, samples are often drawn from area or address-based sample frames. A statistical modelling framework is then needed to estimate the probability of re-identification. This probability is conditional on the released data and information available to the intruder and defined with respect to a probabilistic model and assumptions about how the data is generated (knowledge of the sampling process). The model is based on the set of quasi-identifiers available to the intruder and available in released data which, when cross-classified for the released data, form a contingency table that can be used to identify cells with small sample sizes, and we particularly focus on the sample uniques. The risk of re-identification is based on the notion of population uniqueness in the contingency table: given an observed sample unique, what is the probability that the cell is also a population unique?

The probabilistic modelling to estimate population uniqueness from the observed survey microdata was developed under two approaches: a fully model-based framework taking into account all of the information available to intruders and modelling their behaviour (Duncan and Lambert 1989, Lambert 1993 and later Reiter 2005c) and a more simplified approach that restricts the information that would be known to intruders (Bethlehem, et al. 1990, Benedetti, et al. 1998, Skinner and Holmes 1998, Elamir and Skinner 2006).

Individual per-record risk measures in the form of a probability of re-identification are estimated. These per-record risk measures are then aggregated to obtain global risk measures for the entire file. Denote by F_k the population size in cell k of a table spanned by quasi-identifying variables having K cells and by f_k the sample size. We have $\sum_k F_k = N$ and $\sum_k f_k = n$ with N the total population size and n the size of the released sample. The set of sample uniques is defined as: $SU = \{k: f_k = 1\}$

since these are the potential high-risk records with the potential to be population uniques. Two global disclosure risk measures (where I is the indicator function) are the following:

Number of sample uniques that are population uniques: $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$

Expected number of correct matches for sample uniques assuming a random assignment within cell k (the match probability) $\tau_2 = \sum_k I(f_k = 1)\ \frac{1}{F_k}$

We assume that the population frequencies F_k are unknown and need to be estimated from a probabilistic model where the risk measures are then:

$$\hat{\tau}_1 = \sum_k I(f_k = 1)\hat{P}(\ F_k = 1|f_k = 1) \text{ and } \hat{\tau}_2 = \sum_k I(f_k = 1)\hat{E}(\frac{1}{F_k}\ |f_k = 1) \quad (1)$$

Skinner and Holmes (1998) and Elamir and Skinner (2006) propose a Poisson distribution and a log-linear model to estimate disclosure risk measures in (1). In this model, they assume that $F_k \sim Pois(\lambda_k)$ for each cell k. A sample is drawn by Poisson or Bernoulli sampling with a sampling fraction π_k in cell k: $f_k|F_k \sim Bin(F_k, \pi_k)$. It follows that:

$$f_k \sim Pois(\pi_k\lambda_k) \text{ and } F_k|f_k \sim Pois(\lambda_k(1 - \pi_k)) \quad (2)$$

where the population cell counts F_k are assumed independent given the sample cell counts f_k.

The parameters λ_k are estimated using log-linear modeling. The sample frequencies f_k are independent Poisson distributed with a mean of $\mu_k = \pi_k\lambda_k$. A log-linear model for the μ_k is expressed as: $\log(\mu_k) = x_k{}'\boldsymbol{\beta}$ where x_k is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is obtained by solving the score equations:

$$\sum_k(\ f_k - \pi_k \exp(x_k'\boldsymbol{\beta}))x_k = 0 \quad (3)$$

The fitted values are then calculated by: $\hat{\mu}_k = \exp(x_k'\widehat{\boldsymbol{\beta}})$ and $\hat{\lambda}_k = \hat{\mu}_k/\pi_k$. Individual disclosure risk measures for cell k are:

$$P(\ F_k = 1|f_k = 1) = \exp\ (\lambda_k(1 - \pi_k)) \text{ and } E(\frac{1}{F_k}\ |f_k = 1) = (1 - \exp\ (\lambda_k(1 - \pi_k)))/\ (\lambda_k(1 - \pi_k)) \quad (4)$$

Plugging $\hat{\lambda}_k$ for λ_k in (4) leads to the estimates $\hat{P}(\ F_k = 1|f_k = 1)$ and $\hat{E}(\frac{1}{F_k}\ |f_k = 1)$ and then to $\hat{\tau}_1$ and $\hat{\tau}_2$ of (1). Rinott and Shlomo (2007b) consider confidence intervals for these global risk measures.

Skinner and Shlomo (2008) develop goodness-of-fit criteria for selecting the main effects and interactions of the quasi-identifying variables for the log-linear model based on estimating and (approximately) minimizing the bias of the risk estimates $\hat{\tau}_1$ and $\hat{\tau}_2$. In addition, they address the estimation of disclosure risk measures under complex survey designs with stratification, clustering and survey weights. While the method described assumes that all individuals within cell k are selected independently using Bernoulli sampling, i.e. $P(f_k = 1|F_k) = F_k\pi_k(1 - \pi_k)^{F_k - 1}$, this may not be the case when sampling clusters (e.g. households). In practice, key variables typically include variables such as age, sex and occupation that tend to cut across clusters. Therefore, the above assumption holds in practice in most household surveys and does not cause bias in the estimation of the risk measures. Inclusion probabilities may vary across strata, the most common stratification is on geography. Strata indicators should always be included in the quasi-identifying variables to take into account differential inclusion probabilities in the log-linear model. Under complex sampling, the λ_k can be estimated consistently using pseudo-maximum likelihood estimation (Rao and Thomas 2003), where the estimating equation in (3) is modified as:

$$\sum_k(\ \hat{F}_k - \exp(x_k'\boldsymbol{\beta}))x_k = 0 \quad (5)$$

and $\hat{F}_k$ is obtained by summing the survey weights in cell k: $\hat{F}_k = \sum_{i \in k} w_i$. The resulting estimates λ_k are plugged into expressions in (4) and π_k is replaced by the estimate $\hat{\pi}_k = f_k / \hat{F}_k$. The goodness-of-fit criteria are also adapted to the pseudo-maximum likelihood approach.

The probabilistic modelling presented here and in other related work in the literature assumes that there is no measurement error in the way the data is recorded. Besides typical errors in data capture, key variables can also purposely be perturbed as a means of masking the data, for example through record swapping or the post randomization method (PRAM) (Gouweleeuw, et al. 1998). Shlomo and Skinner (2010) adapt the estimation of the risk of re-identification to take into account measurement (perturbation) errors. We denote the cross-classified quasi-identifying variables in the population and the microdata as X and assume that X in the microdata have undergone some perturbation error denoted by the value $\tilde{X}$ and determined independently by a misclassification matrix M:

$$M_{kj} = P(\tilde{X} = k | X = j) \tag{6}$$

Under small sampling fractions and small rates of perturbation as reflected in the misclassification matrix in (6), we can assume that only the diagonal of the misclassification matrix is needed, i.e. the probabilities of not being perturbed. The estimate of $\hat{\tau}_2$ in (1) can be obtained by the probabilistic modelling framework described above on the misclassified sample:

$$\hat{\tau}_2 = \sum_k I\left(\tilde{f}_k = 1\right) M_{kk} \, \hat{E}\left(\frac{1}{\tilde{F}_k} \middle| \tilde{f}_k\right) \tag{7}$$

where $\tilde{f}_k$ are the observed perturbed counts and $\tilde{F}_k$ represent population counts.

There have been many other contributions extending the Poisson-log linear modelling framework for estimating the risk of re-identification in survey microdata. Ichim (2008) considers extensions by introducing the survey weights in the analysis of the contingency tables and also proposes a maximum penalized-likelihood approach to obtain smoother estimates of the risk of re-identification. Forster and Webb (2007) extend the log-linear modelling framework to a model averaging approach rather than requiring to choose a model a priori. They use a Bayesian model averaging technique according to several possible log-linear models but limit the models to decomposable geographical models. The posterior distribution under model uncertainty is hence obtained as a weighted average of the posterior distribution under the various models. Rinott and Shlomo (2006 and 2007a) generalize the probabilistic modelling using the Negative Binomial distribution rather than the Poisson distribution and implement the probabilistic modelling framework on local 'neighbourhoods' of the sample uniques. Manrique-Vallier and Reiter (2012) propose an alternative to log-linear models for datasets with sparse contingency tables according to the quasi-identifying variables using a Bayesian version of grade of membership models and they use a Markov Chain Monte Carlo algorithm for fitting the model. Carota, et al. (2015) applied a Bayesian semi-parametric version of log-linear models, specifically a mixed effects log-linear model with a Dirichlet process prior.

A new direction is currently under development to measure the risk of re–identification for non-probability data sources. More specifically, there are registers in the public domain where the membership of the register is not known and is sensitive. Examples of registers are of individuals with a medical condition, such as Cancer or HIV, or registers that include membership to a loyalty card scheme. Shlomo and Skinner (forthcoming) focus on this new setting by extending the framework of probabilistic modelling. The microdata from a random sample can still be used to estimate population parameters under the probabilistic modelling framework for estimating the risk of re-identification, however the complication is that another set of parameters needs to be estimated: the propensities of membership for the individuals in the register. This accounts for the selection bias in the register and the deviation from the general population.

For partially synthetic data, assessing disclosure risk where some values of variables are not changed has been further shown in Reiter and Mitra (2009) and Drechsler and Reiter (2011). There, the authors assume that an intruder knows the values of a single target record and then searches the released data to identify the record. Other work on identity disclosure for fully synthetic data has

been shown in Reiter et al. (2014). The authors assume that an intruder has prior knowledge of the entire dataset except for one record and then attempts to quantify the risk of re-identification using Bayesian estimation to obtain the posterior distributions of confidential data given the released data. The intruder then evaluates the posterior distribution of possible original values for the one unknown record, given the released synthetic data and information about the data generation mechanism and uses values with high probability as reasonable guesses for the unknown true values.

4 Quantifying the Risk of Attribute Disclosure for Synthetic Data

Fully synthetic data should lead to a break between the identifying variables and the sensitive target variables, and hence the main focus for quantifying disclosure risk in fully synthetic data is to measure attribute disclosure (and more generally, inferential disclosure). This disclosure risk is based on being able to infer characteristics of individuals in the datasets, particularly groups of individuals.

With respect to developing disclosure risk measures after the generation of the data, one measure that can be used to identify skewness in the distribution of categories c of a variable C in equivalence class EC is the entropy. The entropy of the distribution obtains a maximum value if the distribution of the categories is uniform and a minimum value if the distribution is degenerate (there is only one category represented). We can transform the entropy defined in Section 1 to the E measure so that we obtain a value between 0 and 1 as follows: $E = 1 - H(EC)/\log(K)$ where K is the number of categories of the variable (Antal et al. 2014). We also define the L measure which measures the percentage of the number of categories of the sensitive variable similar to the principle of l-diversity.

We can develop distance metrics that compare the overall distributions in the original data versus synthetic data for a particular variable and more specifically within equivalence classes EC. Distance metrics include Kullback-Leibler distance, the Total Variation (TV) and Hellinger's Distance (HD). For a categorical variable C in equivalence class EC, the Hellinger's Distance is equal to:

$HD_{EC}(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{c\epsilon C}(\sqrt{p(EC,c)} - \sqrt{q(EC,c)})^2}$ where $p(EC,c)$ is the distribution of C in the original data and $q(EC,c)$ is the distribution of C in the synthetic data. The Total Variation is equal to: $TV_{EC}(P,Q) = \frac{1}{2}\sum_{c\epsilon C}|p(EC,c) - q(EC,c)|$.

Note that these distance metrics can also be used for utility measures, i.e. measures that express the usefulness of the data for statistical analysis, and hence we blur the lines about what constitutes measures of disclosure risk and what measures utility.

Similar to the privacy model of t-closeness, we can use distance metrics comparing the distribution in the synthetic data for variable C in an equivalence class EC with the overall univariate distribution in the original data, denoted $Q(c)$. In this case, the Total Variation is $TV(P,Q) = \frac{1}{2}\sum_{c\epsilon C}|p(EC,c) - Q(c)|$.

Elliot (2014) and Taub et al. (2018) defined the Differential Correct Attribution Probability ($DCAP$) framework. It assumes that the intruder has access to the synthetic data s and has knowledge of an equivalence class denoted $EC_{o,i}$ for individual i in the original dataset o and wants to learn the value of a sensitive variable $T_{s,i}$. The intruder then identifies all the records that match on EC_o in the synthetic data s. If the proportion of records in the equivalence class on $\{EC_s, T_s\}$ is high then the intruder can infer the value $T_{s,i}$ for $T_{o,i}$. In summary, $DCAP$ measures the proportion of records for equivalence class EC_o that have the same target value in the synthetic data as the original value. More formally, define D_o the original data composed of equivalence classes EC_o and sensitive variables T_o: $D_o = \{EC_o, T_o\}$ and similarly, the synthetic data is defined as: $D_s = \{EC_s, T_s\}$. For each individual i we define: $DCAP_{o,i} = \sum_{i=1}^{N} I(T_{oi} = T_{si} \text{ and } EC_{oi} = EC_{si})/\sum_{i=1}^{N} I(EC_{oi} = EC_{si})$ where N is the size of the dataset (assumes the same N in the synthetic and original data) and I is the indicator function taking a value of 1 if the condition is satisfied, otherwise 0. Similarly calculate $DCAP_{s,i}$ in the

synthetic data. Note that it is possible that the denominator in $DCAP_{s,i}$ can be 0 and may be undefined. In that case, we can define the measure as 0. The baseline is: $DCAP_{b,i} = \frac{1}{N}\sum_{i=1}^{N} I(T_{oi} = T_{si})$. The original and baseline measures serve as bounds for comparing the $DCAP_{s,i}$ and ensuring that it is sufficiently reduced.

Chen et al. (2019) noted that this original measure of $DCAP$ is similar to the distance-based utility measures and proposed to adapt the $DCAP$ framework to only those records that are unique in the synthetic data in the EC. The risk measure is defined as Targeted Correct Attribution Probability ($TCAP$).

We can see that there is a clear connection between $DCAP$ and the l-diversity privacy model as the less diverse the sensitive variables in the synthetic data, the higher risk of discovering a sensitive attribute.

5 Conclusion

The framework for measuring the risk of re-identification as discussed in Section 3 based on estimating the probability of population uniqueness is well established although many different approaches have been proposed in the SDC literature to estimate these disclosure risk measures. However, as can be seen in Section 4, disclosure risk measures for synthetic data after its generation are still ad-hoc and a more formal framework is needed for measuring the risk of attribute disclosure. In addition, appropriate software needs to be developed which will enable the framework to be embedded in the SDC tool-kit at government agencies.

References

Antal, L., Shlomo, N. and Elliot, M. (2014) Measuring Disclosure Risk with Entropy in Population Based Frequency Tables. In Privacy in Statistical Databases 2014, (Ed. J. Domingo-Ferrer), Springer LNCS 8744, 62-78.

Bethlehem, J., Keller, W., and Pannekoek, J. (1990) Disclosure Control of Microdata. Journal of the American Statistical Association, **85**, 38-45.

Benedetti, R., Capobianchi, A., and Franconi, L. (1998) Individual Risk of Disclosure Using Sampling Design. Contributi Istat.

Carota, C., Filippone, M. Leombruni, R. and Polettini, S. (2015) Bayesian Nonparametric Disclosure Risk Estimation via Mixed Effects Log-linear Models. Annals of Applied Statistics, **9(1)**, 525 – 546.

Chen, Y., Taub, J. and Elliot, M. (2019) Trade-off Between Information Utility and Disclosure Risk in GA Synthetic Data Generator. Conference of European Statisticians, Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 29-31 October 2019, The Hague, the Netherlands. https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2019/mtg1/SDC2019_S3_UK_Chen_Taub_Elliot_AD.pdf

Domingo-Ferrer, J. and Torra, V. (2008) A Critique of k-anonymity and Some of its Enhancements. In 2008 Third International Conference on Availability, Reliability and Security. IEEE, 990-993.

Drechsler, J. (2011) Synthetic Datasets for Statistical Disclosure Control. Lecture Notes in Statistics (LNS) 201, NY: Springer.

Drechsler, J. and Reiter, J. (2011) An Empirical Evaluation of Easily Implemented, Non-parametric Methods for Generating Synthetic Data. Computational Statistics and Data Analysis, **55**, 3232-3243.

Duncan, G. and Lambert, D. (1989) The Risk of Disclosure for Microdata. Journal of Business and Economic Statistics, **7**, 207-217.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.

Elamir, E. and Skinner, C.J. (2006) Record-Level Measures of Disclosure Risk for Survey Micro-data. Journal of Official Statistics, **22**, 525-539.

Elliot, M. (2014) Final Report on the Disclosure Risk Associated with the Synthetic Data Produced by the SYLLS Team. Available at: https://tinyurl.com/syllsDR

Forster, J.J. and Webb, E.L. (2007) Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables. Journal of Royal Statistical Society Series C, **56**, 551–570.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, **14**, 463-478.

Ichim D. (2008) Extensions of the Re-identification Risk Measures based on Log-linear Models. In Privacy in Statistical Databases (eds. J. Domingo-Ferrer and Y. Saygin), Lecture Notes in Computer Science 5262. Springer, Berlin, 203-212.

Lambert, D. (1993) Measures of Disclosure Risk and Harm. Journal of Official Statistics, **9**, 313-331.

Li, N., Li, T. and Venkatasubramanian, S. (2007) t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering.

Little, R.J.A., and Liu, F. (2003) Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6. http://www.bepress.com/umichbiostat/paper6

Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M. (2006) l-diversity: Privacy Beyond k-anonymity. In 22nd International Conference on Data Engineering (ICDE'06), IEEE, 24.

Manrique-Vallier, D. and Reiter, J. P. (2012) Estimating Identification Disclosure Risk Using Mixed Membership Models. Journal of the American Statistical Association, **107**, 1385–1394.

Raghunathan, T.E., Reiter, J. and Rubin, D. (2003) Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, **19(1),** 1-16.

Rao, J. N. K., and Thomas, D. R. (2003) Analysis of Categorical Response Data from Complex Surveys: An Appraisal and Update. In Analysis of Survey Data (eds. R. L. Chambers and C. J. Skinner), Wiley, Chichester, UK, 85–108.

Reiter, J.P. (2005a). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. Journal of the Royal Statistical Society A, **168(1),** 185–205.

Reiter, J. (2005b) Using CART to Generate Partially Synthetic, Public Use Microdata. Journal of Official Statistics, **21,** 441–462.

Reiter, J.P. (2005c) Estimating Risks of Identification Disclosure in Microdata. Journal of the American Statistical Association, **100**, 1103-1112.

Reiter, J. and Mitra, R. (2009) Estimating Risks of Identification Disclosure in Partially Synthetic Data. Journal of Privacy and Confidentiality, **1(1)**, 99–110.

Reiter, J., Wang, Q. and Zhang, B. (2014) Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. Journal of Privacy and Confidentiality, **6(1)**, 17–33.

Rinott, Y. and Shlomo, N. (2006) A Generalized Negative Binomial Smoothing Model for Sample Disclosure Risk Estimation. In Privacy in Statistical Databases (eds. J. Domingo-Ferrer and L. Franconi), Lecture Notes in Computer Science 4302 Springer, Berlin, 82–93.

Rinott, Y. and Shlomo, N. (2007a) A Smoothing Model for Sample Disclosure Risk Estimation. In Complex Datasets and Inverse Problems (eds. R. Liu, W. Strawderman and C.-H. Zhang), Institute of Mathematical Statistics Lecture Notes, Monograph Series 54, Ohio, 161-171.

Rinott, Y. and Shlomo, N. (2007b) Variances and Confidence Intervals for Sample Disclosure Risk Measures. In Bulletin of the International Statistical Institute: Proceedings of the 56th Session of the International Statistical Institute, ISI'07, Lisbon, 1090–1096.

Shlomo, N. and Skinner, C.J. (2010) Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. Annals of Applied Statistics, **4(3)**, 1291-1310.

Shlomo, N. and Skinner, C.J. (forthcoming) Measuring Risk of Re-identification in Microdata: State-of-the Art and New Directions, Journal of the Royal Statistical Society, Series A.

Skinner, C.J. and Holmes, D. (1998) Estimating the Re-identification Risk per Record in Microdata. Journal of Official Statistics, **14**, 361-372.

Skinner, C. J. and Shlomo, N. (2008) Assessing Identification Risk in Survey Micro-data Using Log Linear Models. Journal of American Statistical Association, **103(483),** 989-1001.

Sweeney, L. (2002) k-anonymity: A Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10(5)**, 557-570.

Taub J., Elliot M., Pampaka M. and Smith D. (2018) Differential Correct Attribution Probability for Synthetic Data: An Exploration. In: (eds. Domingo-Ferrer J., Montes F.) Privacy in Statistical Databases. PSD 2018. Lecture Notes in Computer Science, Vol 11126. Springer.

Xiao, X., Yi, K. and Tao, Y. (2010) The Hardness and Approximation Algorithms for l-diversity. In Proceedings of the 13th International Conference on Extending Database Technology, ACM, 135-146.

The Survey Statistician, **2022, Vol. 86, 22–31.**

Tree-Based Machine Learning in Small Area Estimation

Patrick Krennmair [1], Nora Würz [2] and Timo Schmid[3]

[1]Freie Universität Berlin, Germany, patrick.krennmair@fu-berlin.de
[2]Freie Universität Berlin, Germany, nora.wuerz@fu-berlin.de
[3]Otto-Friedrich-Universität Bamberg, Germany, timo.schmid@uni-bamberg.de

Abstract

Reliable estimators of the spatial distribution of socio-economic indicators are essential for evidence-based policy-making. As the accuracy of direct estimates from survey data decrease with spatially finer target levels, small area estimation approaches are promising. In this article, we outline new approaches that combine small area methodology with machine learning methods. The presented semi-parametric approach is promising as it avoids the assumptions of linear mixed models in contrast to classical small area models and builds on random forests. These tree-based machine learning predictors have the advantage of robustness against outliers and implicit model-selection. As for classical small area models, we account for hierarchically dependent data. We present point estimators applicable to full as well as aggregated auxiliary data access and outline their uncertainty measure. We compare methods based on a reproducible and illustrative example using open-source income data from Austria.

Keywords: Official statistics; Mean squared error; Tree-based methods; Prediction

1 Introduction

Evidenced-based policy decisions require a solid and transparent empirical basis. An effective way to produce empirical findings is the construction of the target indicator using sampled information from individual and household surveys. Typically, we can partition a population into geographic, social, or political sub-units that are referred to as 'domains' or 'areas', which allows for the additional perspective of the spatial distribution of targeted indicators. Due to cost and efficiency constraints, the survey sample size is limited and at high spatial resolution the sample size within a domain might become small or even zero. Direct indicator estimates only use existing domain-level survey information. The implicit reduction of area-specific sample sizes as the level of required detail increases, leads to unreliable and imprecise direct estimates. A methodology that provides reliable and detailed estimates for this particular challenge is referred to as Small Area Estimation (SAE) (Pfeffermann, 2013; Rao & Molina, 2015; Tzavidis et al., 2018).

Model-based SAE methods improve estimates by linking survey data and available secondary auxiliary information (e.g. census or administrative data) via predictive models. This combination of information increases the effective sample sizes and subsequently the precision of domain-specific estimates. We broadly divide, SAE models in two classes: Area-level models - e.g Fay-Herriot models (Fay & Herriot, 1979) assuming aggregated data for survey and auxiliary information - and unit-level models - e.g. the nested error regression model by Battese et al. (1988) requiring access to a micro-level survey (Pfeffermann, 2013). Unit and area-level models alike are regression based and the hierarchical structure of observations is modelled by random effects. As a result, most of the SAE models are rooted within the methodological paradigm of Linear Mixed Models (LMM). Under the parametric framework, optimality estimators (under the assumed model) can be obtained. Jiang & Rao (2020) remind that model-based estimates follow the implied distribution of the model and predictive performance and inferences become erroneous and biased in cases of severe violations of model assumptions.

Working with social and economic datasets, we face heavily skewed and unbalanced target variables and models have to identify complex and indistinct relations between covariates. One strategy to prevent model-failure, is the assurance of normality by transforming the dependent variable improving the performance of unit-level models using a fixed logarithmic (Berg & Chandra, 2014; Molina & Martín, 2018) or data-driven (Sugasawa & Kubokawa, 2019; Rojas-Perilla et al., 2020) transformations. In cases of limited access to auxilliary information (i.e area-level aggregates of covariates from population data), small area means can be determined using robust methods like robustified linear mixed models (Sinha & Rao, 2009) or M-quantile based methods (Chambers & Tzavidis, 2006; Marchetti & Tzavidis, 2021). Another alternative is the use of models with less restrictive (parametric) assumptions to avoid model-failure. For instance, Diallo & Rao (2018) and Graf et al. (2019) formulate unit-level models under more flexible distributional assumptions. Semi- or non-parametric approaches for the estimations of area-level means were investigated among others by Opsomer et al. (2008). They use penalized splines regression, treating the coefficients of spline components as additional random effects within the LMM setting.

Machine Learning methods represent a further methodological option to avoid parametric assumptions of LMMs. These methods are not limited to parametric models and ‘learn’ predictive relations from data, including higher order interactions between covariates, without explicit model assumptions (Hastie et al., 2009; Varian, 2014; Gelman & Vehtari, 2021). Despite existing conceptual differences between machine learning and ’traditional’ statistical methods (e.g. best possible predictions vs. parametric representation and interpretation), machine learning methods became a substantial element in statistical methodology research (Efron, 2020). For instance, the training/test-set paradigm is central to machine learning and conceptually transfers to the methodology of unit-level SAE-models: the survey data serves as a training-set to construct a proper model, while supplementary data (usually census, register or administrative data) of auxiliary information is used to predict final indicators over sampled and non-sampled areas. Jiang & Rao (2020) observe that SAE research is susceptible to novel approaches from various fields of statistics, however, results from machine learning are still harder to be interpreted and justified by SAE-practitioners compared to LMM-alternatives. Especially for SAE, new methods must meet the premise of basic principles of survey and inference theory. In this sense, the objectives of SAE coincide with the general perspective of Efron (2020), maintaining that an opportunity for modern statistics lies in the critical analysis and assessment of properties of predictive algorithms to make them ’scientifically applicable’. With this paper and our research, we aim to contribute to this purpose for SAE.

Among the broad class of machine learning methods, we focus on random forests (RFs) (Breiman, 2001) because they exhibit excellent predictive performance in the presence of complex and non-linear interactions and implicitly solve problems of model-selection (Biau & Scornet, 2016). The

general idea of applying tree-based methods in SAE is not entirely new (Anderson et al., 2014; Bilton et al., 2017; De Moliner & Goga, 2018; Mendez, 2008). Recently, Dagdoug et al. (2021) analyse theoretical properties of RF in the context of complex survey data for model-assisted estimation. Krennmair & Schmid (2022) provide a consistent framework enabling a coherent use of tree-based machine learning methods in SAE and propose a non-linear, data-driven, and semi-parametric alternative for the estimation of area-level means using RFs in the methodological tradition of SAE. We will refer to this methodology combining the mixed effect model with RFs in the following as Mixed Effects Random Forest (MERFs). Section 2 introduces a general mixed effects model for SAE and its combination with RFs. Accordingly, the estimation of corresponding model-coefficients is explained and the MERF methodology to obtain domain-specific mean-estimates under unit-level and aggregated census information is elaborated in more depth. In addition, we outline the possibility of estimating the uncertainty of domain-specific indicators measured by corresponding mean squared errors (MSEs) in Section 2.3. An illustrative example on Austrian income data in Section 3 demonstrates both estimators from the theory part. Section 4 concludes and provides an outlook on further perspectives of research regarding the diversification of the model-toolbox for SAE-practitioners and researchers.

2 Using mixed effects random forests in SAE

RFs captivate with a lack of assumptions such as linearity or the distributional specification of model errors. Major benefits are the detection of higher order interactions between covariates, implicit model-selection, and the proper handling of outliers and high-dimensional covariate data without model assumptions (Hastie et al., 2009; Biau & Scornet, 2016). However, observations are assumed to be independent. Applications of SAE are characterized by the use of hierarchical data. Ignoring the correlation between observations, generally results in inferior point-predictions and inferences. Krennmair & Schmid (2022) introduce a general mixed model framework enabling the estimation of data-driven RFs, while simultaneously accounting for structural dependencies of survey data. This general formulation treats traditional LMM-based models in SAE as special cases and thus allows for a simultaneous discussion of existing SAE methods.

2.1 A general mixed effects model for SAE and MERFs

We assume a finite population U of size N consisting of D separate domains $U_1, U_2, ..., U_D$ with $N_1, N_2, ..., N_D$ units, where index $i = 1, ..., D$ indicates respective areas. The continuous target variable y_{ij} for individual observation j in area i is available for every unit within the sample. Sample s is drawn from U and consists of n units partitioned into sample sizes $n_1, n_2, ..., n_D$ for all D areas. We denote by s_i the sub-sample from area i. The vector $\mathbf{x} = (x_1, x_2, ..., x_p)^\intercal$ includes p explanatory variables and is available for every unit j within the sample s. The relationship between $\mathbf{x}_{ij}$ and y_{ij} is assumed to follow a general mixed effects regression model:

$$y_{ij} = f(\mathbf{x}_{ij}) + u_i + e_{ij} \quad \text{with} \quad u_i \sim N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \sim N(0, \sigma_e^2). \tag{1}$$

Function $f(\mathbf{x}_{ij})$ models the conditional mean of y_{ij} given $\mathbf{x}_{ij}$. Area-specific random intercepts u_i account for the hierarchical dependency structure of observations and we subsequently assume unit-level errors e_{ij} and random effects u_i to be independent.

For instance, defining $f(\mathbf{x}_{ij}) = \mathbf{x}_{ij}^\intercal \beta$, where $\beta = (\beta_1, ..., \beta_p)^\intercal$, coincides with the well known nested error regression model proposed by Battese et al. (1988). This widely used LMM with area-specific random effects forms the basis for further unit-level SAE-models, such as the EBP (Molina & Rao, 2010) or the EBP under data-driven transformation by Rojas-Perilla et al. (2020). If the assumptions of the LMMs are met, optimal estimates of fixed effects $\hat{\beta}$ and variance components $\hat{\sigma}_u^2, \hat{\sigma}_e^2$ are obtained by maximum likelihood (ML) or restricted maximum likelihood (REML) (Rao & Molina, 2015).

If we assume f in Model (1) to be a RF (Breiman, 2001), we result in a semi-parametric framework, combining predictive advantages of RFs with the ability to model hierarchical structures of survey data using random effects. The method obtains optimal estimates of model components $\hat{f}$, $\hat{u}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ based on a procedure which is reminiscent of the EM-algorithm (Hajjem et al., 2014). The proposed MERF algorithm fits optimal parameters for Model (1) (where f is a RF) by iteratively estimating a) the forest function, assuming the random effects term to be correct and b) the random effects part, assuming the Out-of-Bag-predictions (OOB-predictions) from the forest to be correct. OOB-predictions utilize the unused observations from the construction of each forest's sub-tree (Breiman, 2001; Biau & Scornet, 2016). The estimation of variance components σ_ϵ^2 and σ_u^2 is obtained implicitly by taking the expectation of ML estimators given the data. The marginal change of a generalized log-likelihood criterion of the composite model monitors the convergence of the estimation algorithm. For further methodological details, we refer to Krennmair & Schmid (2022). The resulting estimator for model-based predictions from the MERF is summarized as follows:

$$\hat{\mu}_{ij}^{\text{MERF}} = \hat{f}(\mathbf{x}_{ij}) + \hat{u}_i = \hat{f}(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{\text{OOB}}(\mathbf{x}_{ij})) \right). \tag{2}$$

2.2 Flexible domain prediction of means under unit-level and aggregated covariates

Under the assumed existence of unit-level (i.e. $\mathbf{x}_{ij}$) population data (usually census or administrative data), $\hat{\mu}_{ij}^{\text{MERF}}$ in Equation (2) can predict conditional means of a metric dependent variable. As typical for SAE, our major interest is in estimating area-level means. The domain-level mean estimator for each area i is given by:

$$\hat{\mu}_i^{\text{MERF}} = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \hat{u}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij}) + \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i} \left(\frac{1}{n_i} \sum_{j \in s_i} (y_{ij} - \hat{f}^{\text{OOB}}(\mathbf{x}_{ij})) \right), \tag{3}$$

$$\text{where} \quad \bar{\hat{f}}_i(\mathbf{x}_{ij}) = \frac{1}{N_i} \sum_{j \in U_i} \hat{f}(\mathbf{x}_{ij}).$$

While the RF part $\hat{f}()$ express the conditional mean of fixed effects, we maintain in Krennmair & Schmid (2022) that $\hat{u}_i$ is the BLUP for the linear part of Model (1). For non-sampled areas, the proposed estimator for the area-level mean reduces to the fixed part from the RF: $\hat{\mu}_i = \bar{\hat{f}}_i(\mathbf{x}_{ij})$.

The access to auxiliary population micro-data is challenging for practitioners, researchers, and even within gatekeeper organizations. The direct incorporation of aggregated auxiliary information in Equation (2) is not possible without misspecification, as for RFs $f(\bar{\mathbf{x}}_i) \neq \bar{f}_i(\mathbf{x}_{ij})$. Notably, not many methods in SAE cope with the dual problem of providing robustness against model-failure, while simultaneously working under limited auxiliary data (Jiang & Rao, 2020). Recently, Krennmair et al. (2022) solved this issue by incorporating aggregate census-level covariate information through calibration weights w_{ij}, which balance unit-level predictions from MERFs in Equation (2) achieving coherence with the area-wise covariate means from census data. In short, this estimator under reduced information for the area-level means can be written as:

$$\hat{\mu}_i^{\text{MERFagg}} = \sum_{j=1}^{n_i} \hat{w}_{ij} \left[\hat{f}(\mathbf{x}_{ij}) + \hat{u}_i \right]. \tag{4}$$

However, optimal estimated model-components ($\hat{f}$ and $\hat{u}_i$) are obtained similar to Equation (2) from survey data using the MERF algorithm as described by Krennmair & Schmid (2022), note that $\mathbf{x}_{ij}$ are unit-level covariates from the survey. Aggregated auxiliary population information ($\bar{\mathbf{x}}_{\text{pop},i}$) is incorporated through optimal weights $\hat{w}_{ij}$ inspired by Li et al. (2019) maximizing the profile empirical

likelihood function $\prod_{j=1}^{n_i} w_{ij}$ under the following three constraints:

- $\sum_{j=1}^{n_i} w_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{\text{pop},i}) = 0$, monitoring the area-wise sum of distances between survey data and the population-level mean ($\bar{\mathbf{x}}_{\text{pop},i}$) for auxiliary covariates;
- $w_{ij} \geq 0$, ensuring the non-negativity of weights;
- $\sum_{j=1}^{n_i} w_{ij} = 1$, to normalize weights.

Optimal weights are the solution to the system of equations and obtainable using the Lagrange-multiplier method (Owen, 1990, 2001; Emerson & Owen, 2009). Krennmair et al. (2022) discuss technical conditions for the feasibility of solutions in the context of SAE and propose a best practice strategy, which is compared to predominate methods in model-based SAE as well as the MERF-based estimator under unit-level data from Equation (3).

2.3 Estimation of uncertainty

A discussion on the quality of domain-specific indicators necessitates a scrutiny of inference and uncertainty. For SAE, it is convenient to use the estimated MSE of the indicators. However, even in the supposedly simple case of LMMs with block diagonal covariance matrices and estimated variances, analytical forms of the MSE can only be approximated (Prasad & Rao, 1990; Datta & Lahiri, 2000; González-Manteiga et al., 2008; Rao & Molina, 2015). The deficiency of general statistical results concerning inferences of RFs adds additional complexity. Although, from a survey perspective, Dagdoug et al. (2021) recently analyse theoretical properties of RFs in the context of model-assisted estimation methods, we propose the use of elaborate bootstrap-schemes for the assessment of uncertainty under the previously discussed methods above.

In Krennmair & Schmid (2022), we propose a non-parametric random effect block bootstrap framework for estimating the MSE for area-level means from sampled and unsampled domains as discussed given by Model (3). In short, the bootstrap-schemes builds on non-parametric generation and resampling of random components originally introduced by Chambers & Chandra (2013). Important for handling and resampling the empirical error components is to centre and scale them by a bias-adjusted residual variance proposed by Mendez & Lohr (2011). In short, the estimator of the residual variance under the MERF from Equation (2), $\hat{\sigma}_\epsilon^2$ is positively biased capturing excess uncertainty concerning the estimation of function $\hat{f}$. We argue a necessity to extrapolate this excess uncertainty before a full bootstrap pseudo-population is simulated. In the presence of aggregated census-level data, Model (4), we base the general procedure on the methodological principles of the bootstrap for finite populations introduced by González-Manteiga et al. (2008). This allows us to construct (pseudo-)true values by generating only error components instead of simulating full bootstrap populations. Details on the methodologies and the performance of proposed uncertainty estimates can found in Krennmair & Schmid (2022) and Krennmair et al. (2022).

3 Illustrative example

This section outlines the advantages of MERFs by estimating domain-level average equivalized household income for Austrian districts. Especially highly skewed distributed variables, like the household income in Austria, often lead to model violations for the classical nested error regression model from Battese et al. (1988). Therefore, semi-parametric methods for SAE, like MERFs, are very promising and needed for these kinds of empirical questions.

The used dataset consists of synthetic Austrian European Union Statistics on Income and Living Conditions (EU-SILC) from 2006 on household-level. Note that this data is exemplary data made publicly available as part of the *R*-package *emdi* (Kreutzmann et al., 2019), which contains detailed information on the data generation process. The major advantage of this illustrative example using

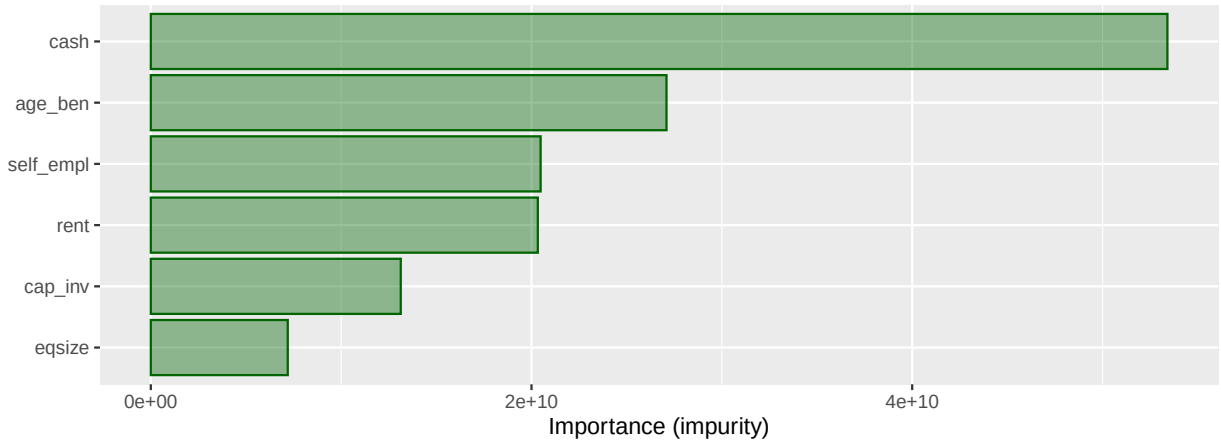


Figure 1: Variable importance for the six most influential variables

open-source data is that we provide reproducible research. The target variable is the equivalized household income (*eqIncome*), which is available in the survey but not in the census and is defined by the ratio of total household disposable income and the equivalized household size (Hagenaars et al., 1994). The illustrative Austrian population data consists of 25 000 households spread over all 94 district and 1945 households within the exemplary sample data. 70 districts are included in the sample, with sample sizes varying between 14 and 200 households (median 22.5 households). Therefore, direct area-level mean-estimates are not feasible for 24 out-of-sample districts. For this reason, and to obtain more precise estimates, SAE methods are needed.

This example displays in addition to the direct estimation, the two MERFs, Model (3) and (4), and the established EBP method (Molina & Rao, 2010) with data-driven Box-Cox transformation (Rojas-Perilla et al., 2020) as competitor. We refer to this method as EBP-BC. Please note that the MERF from Model (3), labelled as MERF_ind, as well as the EBP-BC method require micro-level population auxiliary data. Due to data security constraints, especially in developed countries, alternative estimators relying only on area-level aggregated auxiliary data are highly needed and therefore MERF_agg (from Model (4)) is also included into this example. We aim to show that mean-estimates of MERF_ind are close to the estimates from the established EBP-BC. In addition, the MERF_agg using less data is intended to be similar to both estimators using unit-level auxiliary data.

Regarding variable selection, there is a distinct difference between the EBP method and the MERFs: For EBP-BC, 13 auxiliary variables on socio-economic characteristics and income situation were selected using Bayesian Information Criterion as valid predictors for the target variable *eqIncome*. In contrast, MERFs perform an implicit variable selection. An importance plot gives the reader an impression on most influential variables for the prediction of *eqIncome*: among others, this plot highlights variables describing cash assets (cash), the receiving of age benefits (age_ben), a given situation of self-employment (self_empl) as well as income from rental of a property or land (rent) as particularly influential (cf. Figure 1). Figure 2 shows a line plot on point estimates for all four methods. The direct estimator as well as the EBP-BC are produced using the *R*-package *emdi* (Kreutzmann et al., 2019) and the code for the two MERF estimators is available from authors upon request. The two MERF estimators perform similarly to the established EBP-BC, which confirms their validity. Even under limited population data (MERF_agg), similar results are obtained as with the two methods using micro-level population data. The assessment of uncertainty of point estimates is an important step for an analysis of reliability of estimates. Thus, Figure 3 reports corresponding bootstrap MSE-estimates for point estimates of area-level means. As anticipated, the three model-based estimators are characterized by lower MSE-values in mean and median terms. For MERF_agg, however, this reduction is less pronounced than for the other two estimators, which assume access to comprehensive

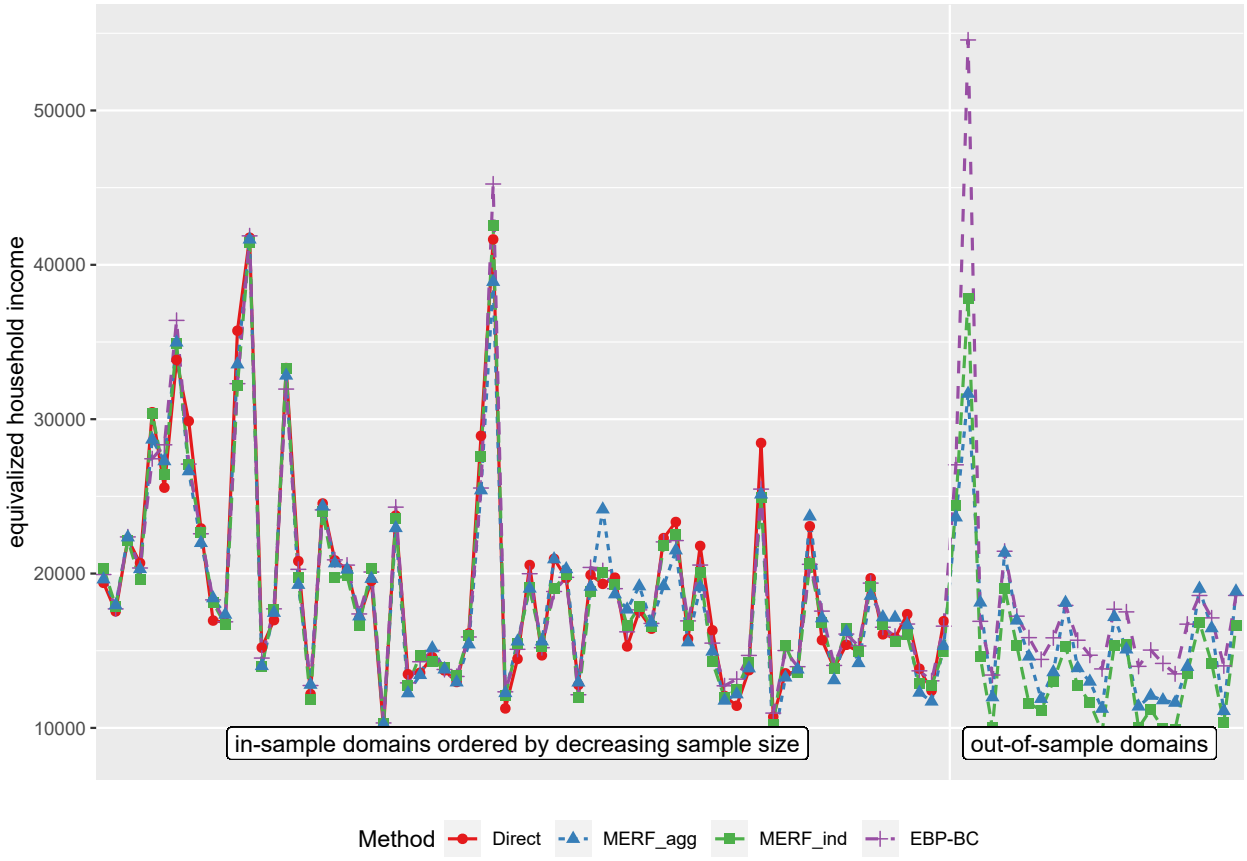


Figure 2: Point estimates for the domain-level average equivalized household income for Austrian districts.

micro-level population data. In median terms, the MSE-values of MERF_ind are the lowest among all competing methods. For detailed model-based simulations and extensive discussions, we refer to Krennmair & Schmid (2022) and Krennmair et al. (2022). The extensive analysis of properties of MERF_ind and MERF_agg reveals that, especially in the presence of complex and unknown relations between covariates, these semi-parametric methods offer substantial advantages.

4 Conclusion and Outlook

Machine learning methods became popular alternatives for predictive models in various scientific fields outside the statistical spheres of SAE. This article serves as a first step, of bridging concepts and highlighting opportunities such as the similarity of the predictive character of model-based SAE and the training/test-set paradigm in machine learning. We introduce RFs for SAE and account for dependency structures of observations using a semi-parametric framework of MERFs for the estimation of point and uncertainty estimates for domain-level indicators under unit- and aggregated auxiliary information. A reproducible example on open-source income data shows estimates for MERFs using unit-level and aggregated auxiliary data and compares them to direct estimates and the well known EBP method (Molina & Rao, 2010) with Box-Cox transformation (Rojas-Perilla et al., 2020). Benefits of RFs are the implicit model-selection and lack of specification under simultaneously high predictive power even in the presence of complex and potentially non-linear interactions between covariates. Moreover, RFs also deal with high-dimensional ($p > n$) datasets. We acknowledge that compared to predominant LMMs, the benefits of prediction serve at cost of explainability and attribution and although this 'black-box'-argument is mitigated by diagnostic tools and plots, discrepancies regarding perspectives of predictive algorithms and explanatory models remain (Efron, 2020).

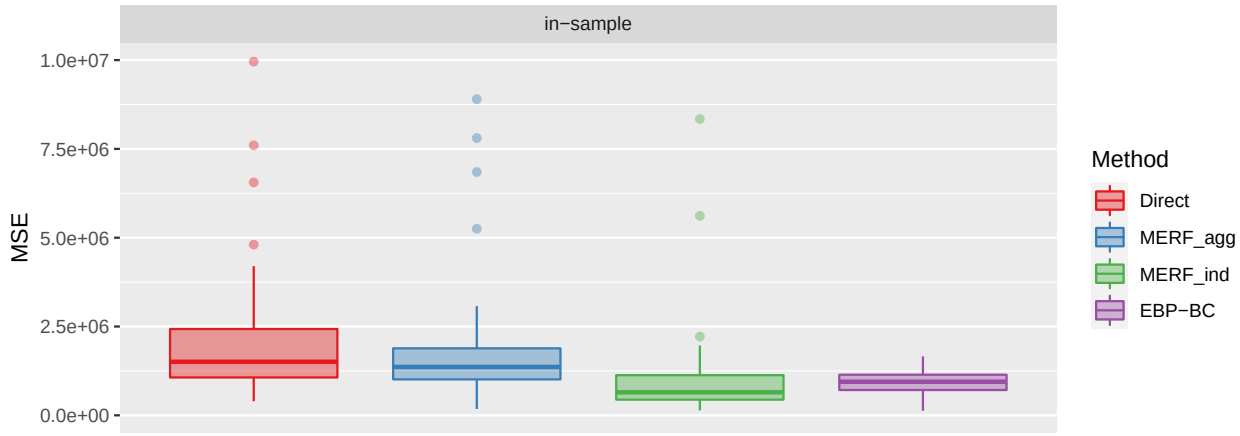


Figure 3: MSE estimates for the point estimates on average equivalized household income for Austrian districts for in-sample areas.

Overall, we conclude that machine learning methods add valuable insights and advantages to the existing repertoire of SAE methods. From our perspective, tree-based predictors perfectly align with the required emphasis on robustification of models against model-failure (e.g. providing insurances against model-misspecification, valid variable selection and the effective handling of outliers) (Jiang & Rao, 2020). The broadening of our statistical methodological toolbox must not only lie in the plain application of existing machine learning algorithms, but rather in the question how they can be made 'scientifically applicable' (Efron, 2020). For SAE, emerging methods need a clear commitment to the methodological tradition of SAE, meaning to find solutions within the context of domain-level indicators, dependent data structures, and in the broader context of survey methodology.

Our presented framework for MERFs, Model (1), is at a starting point and opens up many further research directions. Future applications might use MERFs in the presence of more complex dependency and correlations structures and increasingly compare them to existing LMM-based alternatives. The use of complex and high-dimensional covariate data is another interesting topic. Generally, there is a need for a substantial theoretical discussion on non- or semi-parametric models handling dependency structures. Concretely, our framework can be generalized to binary and count data, but also towards other model-classes, such as Support Vector Machines, Gradient Boosting, Bayesian Additive Regression Trees and many more. We firmly believe that methodological developments in SAE should be complemented by the development of suitable open-source software packages and we are currently working on an *R*-package. Facilitated access to SAE-methods promotes further development and facilitates the comparison between existing methods in model- and design-based evaluations and will result in a toolbox of tailored approaches for researchers and practitioners.

References

Anderson, W., Guikema, S., Zaitchik, B., & Pan, W. (2014). Methods for estimating population density in data-limited areas: Evaluating regression and tree-based models in Peru. *PLoS One*, *9*(7).

Battese, G. E., Harter, R. M., & Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, *83*(401), 28–36.

Berg, E., & Chandra, H. (2014). Small area prediction for a unit-level lognormal model. *Computational Statistics & Data Analysis*, *78*, 159–175.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Bilton, P., Jones, G., Ganesh, S., & Haslett, S. (2017). Classification trees for poverty mapping. *Computational Statistics & Data Analysis*, *115*, 53–66.

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

Chambers, R., & Chandra, H. (2013). A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, *22*(2), 452–470.

Chambers, R., & Tzavidis, N. (2006, 06). M-quantile models for small area estimation. *Biometrika*, *93*(2), 255-268.

Dagdoug, M., Goga, C., & Haziza, D. (2021). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 1–18.

Datta, G. S., & Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, *10*(2), 613–627.

De Moliner, A., & Goga, C. (2018, December). Sample-based setimation of mean electricity consumption curves for small domains. *Survey Methodology*, *44*(2), 193–215.

Diallo, M. S., & Rao, J. N. K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, *45*(4), 1092–1116.

Efron, B. (2020). Prediction, estimation, and attribution. *Journal of the American Statistical Association*, *115*(530), 636-655.

Emerson, S., & Owen, A. (2009). Calibration of the empirical likelihood method for a vector mean. *Electron. J. Statist*, *3*, 1161–1192.

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

Gelman, A., & Vehtari, A. (2021). What are the most important statistical ideas of the past 50 years? *Journal of the American Statistical Association*, *116*(536), 2087-2097.

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

Graf, M., Marín, J. M., & Molina, I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *Test*, *28*(2), 565–597.

Hagenaars, A. J., De Vos, K., Asghar Zaidi, M., et al. (1994). Poverty statistics in the late 1980s: Research based on micro-data.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-Effects Random Forest for Clustered Data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Jiang, J., & Rao, J. S. (2020). Robust small area estimation: An overview. *Annual Review of Statistics and its Application*, *7*(1), 337–360.

Krennmair, P., & Schmid, T. (2022). *Flexible domain prediction using mixed effects random forests.* (Working Paper)

Krennmair, P., Würz, N., & Schmid, T. (2022). *Analysing opportunity cost of care work using mixed effects random forests under aggregated census data.* (Working Paper)

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

Li, H., Liu, Y., & Zhang, R. (2019). Small area estimation under transformed nested-error regression models. *Statistical Papers*, *60*(4), 1397–1418.

Marchetti, S., & Tzavidis, N. (2021). Robust estimation of the Theil index and the Gini coeffient for small areas. *Journal of Official Statistics*, *37*(4), 955–979.

Mendez, G. (2008). *Tree-based mehtods to model dependent data* (Unpublished doctoral dissertation). Arizona State University.

Mendez, G., & Lohr, S. (2011). Estimating residual variance in random forest regression. *Computational Statistics & Data Analysis*, *55*(11), 2937–2950.

Molina, I., & Martín, N. (2018). Empirical best prediction under a nested error model with log transformation. *The Annals of Statistics*, *46*(5), 1961–1993.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369–385.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 265–286.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, *18*(1), 90–120.

Owen, A. (2001). *Empirical likelihood.* New York: Chapman and Hall.

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

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*(409), 163–171.

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2.nd ed.). New Jersey: John Wiley & Sons.

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148.

Sinha, S. K., & Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, *37*(3), 381–399.

Sugasawa, S., & Kubokawa, T. (2019). Adaptively transformed mixed-model prediction of general finite-population parameters. *Scandinavian Journal of Statistics*, *46*(4), 1025–1046.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., & Rojas-Perilla, N. (2018). From start to finish: A framework for the production of small area official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(4), 927–979.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28.

The Survey Statistician, 2022, Vol. 86, 32–43.

Book and Software Review

M. S. Diallo. Samplics: A comprehensive library for survey sampling in Python

Mamadou S. Diallo[1]

[1]The Saudi Center for Opinion Polling (SCOP), Saudi Arabia, msdiallo@samplics.org

Abstract

Survey sampling is one of the main tools used by public and private organizations of all sizes to produce statistics to guide decision-making. For example, governments regularly use large national household and non-household surveys to inform policy in numerous sectors. Similarly, opinion polling and market research surveys inform corporations and other entities on populations' views and opinions on issues and products.

Python has become a leading tool for data science and machine learning projects. Yet, survey statisticians did not have any comprehensive library in Python for designing or analyzing complex survey data. With the development of *samplics*, Python users no longer must learn another software to design or analyze complex survey samples. The library *samplics* classes and functions provide a large coverage of survey sampling topics from sample size calculation, sample selection, weight adjustments, estimation, tabulation, t-test to small area estimation. This paper discusses some of the APIs of *samplics*.

Keywords: survey, sampling, sample size, small area estimation, Python.

1 Introduction

Python is a free and open-source software; it has become one of the most popular software during the last decade. Most of its popularity is due to the explosion of data science and its applications. Python is currently one of the software of choice for machine learning and data science due to the availability of comprehensive and user friendly libraries such as scipy, numpy, matplotlib, pandas, scikit-learn, statsmodels, keras, tensorflow, and pytorch. However, until the development of *samplics*, there was no library for survey sampling techniques, see Lohr (2022).

samplics is a Python library for survey sampling techniques. The package is comprehensive and is designed to assist the survey statistician from the conception with sample size calculation to the estimation of population parameters. The main modules of the *samplics* library are sample size calculation, sample selection, weighting, population parameters estimation, tabulation and hypothesis testing, and small area estimation. A Python user no longer needs to move to the R Software or other solutions to design or analyse complex survey samples.

In this paper, we present the main APIs of the *samplics* library. It uses Python versions 3.7.x or newer and the following packages: numpy, pandas, scpy, and statsmodels. To install it, use: pip install samplics. The current version of the library is 0.3.35 (May, 2022), and its manual can be downloaded at https://samplics.readthedocs.io/en/latest/.

2 Sample size calculation

During the design of the survey, investigators collaborate with statisticians to calculate the minimum required sample sizes, see Chow et al. (2018) and Ryan (2013) for a comprehensive review of sample size calculation methods. Often at the design phase, many variables are of interest. For the sample size calculation, the investigators must reduce the number to ideally a single variable or a handful. *samplics* provides the class *SampleSize* for calculating sample size to estimate proportions, means, and totals. Its argument *parameter* can take the values "proportion", "mean" or "total", while the argument *method* takes the values "wald" or "fleiss"; for example:

SampleSize(parameter = "proportion", method = "wald", stratification = False)

To calculate the sample size, we need to provide the expected value through the argument *target* and the desired precision *half_ci* in *SampleSize.calculate*. If we are estimating a mean or a total then the standard deviation, *sigma*, is required. We have:

SampleSize.calculate(target, half_ci, sigma = None, deff = 1.0, resp_rate = 1.0,
number_strata = None, pop_size = None, alpha = 0.05)

We can use this class to calculate the sample size for simple random sampling with replacement when we estimate a proportion:

$$n_0 = \left(\frac{z_{\alpha/2}}{e}\right)^2 p(1-p),$$

where $z_{\alpha/2}$ is the quantile of order $1-\alpha/2$ of the N(0,1) distribution, p is the expected proportion, and e is the margin of error. For example, let's say we want to calculate the sample size to estimate a proportion $p = 0.5$ with a margin of error $e = 0.03$, and for $\alpha = 0.05$. We could use the following code snippet:

```
import samplics
from samplics.sampling import SampleSize
size_prop = SampleSize(parameter="proportion")
size_prop.calculate(target=0.5, half_ci=0.03)
size_samp_size
1068
```

The second line of the code above creates the object *size_prop*. In the third line, we call the method *calculate()* to compute the sample size and update the object *size_prop*. To show the content of the object *size_prop*, we can print the members using *size_prop.__dict__*.

samplics can also calculate the required sample size to conduct hypothesis testing. There are several *samplics* classes for calculating sample size for testing proportions or means in the situation of one or two samples.

The class *SampleSizeMeanOneSample* calculates the minimum required sample size for testing mean with one sample. Let's assume we have one sample and we are interested in the following hypotheses $H_0 : \mu = \mu_0 \quad versus \quad H_a : \mu \neq \mu_0$. The equation for the sample size needed to achieve power $1-\beta$ is $n = \frac{(z_{\alpha/2}+z_\beta)^2\sigma^2}{\epsilon^2}$, where ϵ is the difference $\mu - \mu_0$. For example, let's assume that we want to calculate sample size required to test the difference before and after a treatment. We have that the average before treatment is 1.5 and the same average after treatment is 2. The standard

deviation is 1, and we assume $1 - \beta = 0.8$. We could calculate the sample size by using the following code snippet for simple random sampling with replacement:

```
from samplics.sampling import SampleSizeMeanOneSample
mean_equality = SampleSizeMeanOneSample()
mean_equality.calculate(mean_0=1.5, mean_1=2, sigma=1)
mean_equality.samp_size
32
```

In a stratified design, the population is divided into H partitions or strata. The sample is selected independently from each stratum. The above *samplics* APIs integrates the notion of stratification. When instantiating the objects, we can indicate that it's a stratified design using *stratification = True*. The parameters should then be provided by stratum using Python dictionaries. For example, *mean_0 = {"North": 1.50, "South": 1.65, "West": 1.55, "East": 1.45}* where North, South, West, and East are the strata.

If more convenient, we can use the method *to_dataframe()* to convert the output data dictionary to a Pandas DataFrame, see McKinney (2010) to learn more about Pandas. Similarly, we can use *SampleSizePropOneSample* for testing proportions with one sample. In the context of two samples, we can use *SampleSizeMeanTwoSample* and *SampleSizePropTwoSample*.

3 Sample selection

The class *SampleSelection* implements several popular random selection methods such as simple random sampling (SRS), systematic (SYS), several probability proportional to size (PPS) methods. The available PPS algorithms for selecting samples with unequal probabilities of selection are systematic (*method="pps-sys"*), Brewer (*method="pps-brewer"*), Hanurav-Vijayan (*method="pps-hv"*), Murphy (*method="pps-murphy"*), and Rao-Sampford (*method="pps-rs"*) methods. Let's assume that we have a population of 100 units and we want to select 10 using the SRS method:

```
from samplics.sampling import SampleSelection
srs_sampling = SampleSelection(method="srs")
srs_sample, srs_hits, srs_probs = srs_sampling.select(samp_unit=range(1, 101),
samp_size=10)
```

As shown in the above code snippet, the method *select()* returns a tuple of three numpy arrays, see Harris et al. (2020). The first array provides the selection status of each unit in the population, the second array provides the probabilities of selection, and the third array gives the number of hits/times a unit was selected. If needed, the user may set *to_dataframe=True* to convert the output data to a pandas DataFrame from the tuple of three arrays. The resulting sample is now a Pandas DataFrame with its first 5 observations shown below.

	_samp_unit	_mos	_sample	_hits	_probs
0	1	1.0	False	0	0.1
1	2	1.0	False	0	0.1
2	3	1.0	False	0	0.1
3	4	1.0	False	0	0.1
4	5	1.0	False	0	0.1

The code above returns the entire population with the variable *_sample* indicating the selected units. We can use *sample_only=True* to subset the returned data to only contain the sample. To illustrate the PPS sample selection, we use the code below to randomly generate sizes (using the Unif(0,1) distribution) associated with each of the 100 units in our population:

```
import random
mos = []
for _ in range(100):
mos.append(round(100 * random.random(), 0))
```

Now let's select a sample of 2 units using the PPS Brewer method without replacement:

```
ss_brewer = SampleSelection(method="pps-brewer", with_replacement=False)
ss_brewer_sample = ss_brewer.select( samp_unit=range(1,101), samp_size=2, mos=mos,
to_dataframe=True, sample_only=True)
ss_brewer_sample
```

	_samp_unit	_mos	_sample	_hits	_probs
0	46	67.0	1	1	0.026677
1	50	93.0	1	1	0.037030

As discussed previously, for stratified designs, we provide the information using Python dictionaries where the keys are the strata names and the values are the sample sizes. Then, we provide the stratification variable to the method *select()* using the argument *stratum*.

4 Weighting

The *samplics* module *weighting* provides the algorithms for adjusting the sample weight for non-response, post-stratification, and calibration. The main class is *SampleWeight* and the different type of adjustments are conducted using its methods *adjust()*, *poststratify()*, and *calibrate()*.

4.1 Design weight

The design weight calculation and subsequent weight adjustments are key steps to ensuring the generalization of the sample results to the target population. The initial design weight, w_i, is obtained as the reciprocal of the probability of inclusion π_i, for unit i in the population, $w_i = \frac{1}{\pi_i}$.

samplics has a module *dataset* which provides curated datasets for running the examples. With the code below, we use the dataset module to load two datasets representing primary sampling units (PSUs) and Secondary Sampling Units (SSUs) samples:

```
from samplics.datasets import load_psu_sample, load_ssu_sample
psu_sample_dict = load_psu_sample()
psu_sample = psu_sample_dict["data"]
ssu_sample_dict = load_ssu_sample()
ssu_sample = ssu_sample_dict["data"]
```

We combine the two datasets to form the final sample data and we calculate the inclusion probability as the product of the two stage probabilities:

```
full_sample = pd.merge(
left=psu_sample[["cluster", "region", "psu_prob"]],
right=ssu_sample[["cluster", "household", "ssu_prob"]],
on="cluster")
```

Hence, the design weight follows as the reciprocal of the inclusion probability.

```
full_sample["inclusion_prob"] = full_sample["psu_prob"] * full_sample["ssu_prob"]
full_sample["design_weight"] = 1 / full_sample["inclusion_prob"]
full_sample.head()
```

	cluster	region	psu_prob	household	ssu_prob	inclusion_prob	design_weight
0	7	North	0.187726	72	0.115385	0.021661	46.166667
1	7	North	0.187726	73	0.115385	0.021661	46.166667
2	7	North	0.187726	75	0.115385	0.021661	46.166667
3	7	North	0.187726	715	0.115385	0.021661	46.166667
4	7	North	0.187726	722	0.115385	0.021661	46.166667

4.2 Non-response adjustment

For the purpose of illustrating non-response adjustments, we add non-respondent households into our example. That is, we simulate the non-response status and store it in the variable *response_status* which has four possible values: *ineligible* which indicates that the sampling unit is not eligible for the survey, *respondent* which indicates that the sampling unit responded to the survey, *non-respondent* which indicates that the sampling unit did not respond to the survey, and *unknown* means that we are not able to infer the status of the sampling unit i.e. we do not know whether the sampling unit is eligible or not for the survey.

```
np.random.seed(7)
full_sample["response_status"] = np.random.choice( ["ineligible", "respondent",
"non-respondent", "unknown"], size=full_sample.shape[0], p=(0.10, 0.70, 0.15, 0.05) )
full_sample[["cluster", "region", "design_weight", "response_status"]].head(5)
```

	cluster	region	design_weight	response_status
0	7	North	46.166667	ineligible.
1	7	North	46.166667	respondent.
2	7	North	46.166667	respondent.
3	7	North	46.166667	respondent.
4	7	North	46.166667	unknown.

In general, the sample weights are adjusted by redistributing the sample weights of all eligible units for which there is no sufficient response (nonrespondents) to the sampling units that sufficiently responded to the survey (respondents). This adjustment is done within adjustment/response classes or domains. Note that the determination of the response classes is outside of the scope of this module.

The method *adjust()* has a boolean argument *unknown_to_inelig* which controls how the sample weights of the unknown are redistributed. By default, *adjust()* redistributes the sample weights of the units with unknown eligibility to the ineligibles (*unknown_to_inelig=True*). If we do not wish to redistribute the sample weights of the unknowns to the ineligibles, we set the flag to *False*.

```
status_mapping = {"in":  "ineligible", "rr":  "respondent", "nr":  "non-respondent",
"uk":  "unknown" }
```

```
from samplics.weighting import SampleWeight
full_sample["nr_weight"] = SampleWeight().adjust(
samp_weight=full_sample["design_weight"],
adjust_class=full_sample[["region", "cluster"]],
resp_status=full_sample["response_status"],
resp_dict=status_mapping)
full_sample[["cluster", "region", "design_weight", "response_status",
"nr_weight"]].drop_duplicates().head(10)
```

	cluster	region	design_weight	response_status	nr_weight
0	7	North	46.166667	ineligible	49.464286
1	7	North	46.166667	respondent	54.410714
4	7	North	46.166667	unknown	0.000000
11	7	North	46.166667	non-respondent	0.000000
15	10	North	50.783333	non-respondent	0.000000
16	10	North	50.783333	respondent	70.733929
19	10	North	50.783333	ineligible	54.410714
21	10	North	50.783333	unknown	0.000000
30	16	South	62.149123	respondent	66.588346
35	16	South	62.149123	non-respondent	0.000000

Important. The default call of *adjust()* expects the response status variable to have values of "in", "rr", "nr", or "uk" where "in" means ineligible, "rr" means respondent, "nr" means non-respondent, and "uk" means unknown eligibility.

In the call above, if we omit the argument *resp_dict*, then the code would fail with an assertion error message. The current error message is the following: "The response status must only contains values in ('in', 'rr', 'nr', 'uk') or the mapping should be provided using *response_dict* parameter". For the call to run without specifying *resp_dict*, it is necessary that the response status takes only values in the standard codes i.e. ("in", "rr", "nr", "uk").

4.3 Post-stratification

Post-stratification is useful to compensate for under-representation of the sample or to correct for nonsampling error. Post-stratification classes can be formed using variables beyond the ones involved in the sampling design. For example, socio-economic variables such as age group, gender, race and education are often used to form post-stratification classes/cells.

Let's assume that we have a reliable external source e.g. a recent census that provides the number of households by region. The external source has the following control data: 3700 households for East, 1500 for North, 2800 for South and 6500 for West. We use the method *poststratify()* to ensure that the post-stratified sample weights (*ps_weight*) sum to the know control totals by region. Note that the control totals are provided using the Python dictionary *census_households*.

```
census_households = {"East": 3700, "North": 1500, "South": 2800, "West": 6500}
full_sample["ps_weight"] = SampleWeight().poststratify(samp_weight
=full_sample["nr_weight"], control=census_households, domain=full_sample["region"])
full_sample.head(7)
```

	cluster	region	household	design_weight	response_status	nr_weight	ps_weight
0	7	North	72	46.166667	ineligible	49.464286	51.020408
1	7	North	73	46.166667	respondent	54.410714	56.122449
2	7	North	75	46.166667	respondent	54.410714	56.122449
3	7	North	715	46.166667	respondent	54.410714	56.122449
4	7	North	722	46.166667	unknown	0.000000	0.000000
5	7	North	724	46.166667	respondent	54.410714	56.122449
6	7	North	755	46.166667	respondent	54.410714	56.122449

In some surveys, there is interest in keeping relative distribution of strata to some known distribution. For example, WHO EPI vaccination surveys, World Health Organization (2018), often poststratify sample weights to ensure that relative sizes of strata reflect official statistics e.g. census data. Assume that according to census data that East contains 25% of the households, North contains 10%, South contains 20% and West contains 45%. We can post-stratify using the snippet of code below.

```
known_ratios = {"East": 0.25, "North": 0.10, "South": 0.20, "West": 0.45}
full_sample["ps_weight2"] = SampleWeight().poststratify(samp_weight
=full_sample["nr_weight"], factor=known_ratios, domain=full_sample["region"])
```

4.4 Calibration weight

Calibration is a more general concept for adjusting sample weights to sum to known constants; see Deville & Särndal (1992). In *samplics*, we implemented the generalized regression (GREG) class of calibration. Assume that we have $\hat{\mathbf{Y}} = \sum_{i \in s} w_i y_i$ and population totals $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_p)^T$ are available. Working under the model $Y_i|\mathbf{x}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$, where $\boldsymbol{\beta}$ is the vector of parameters, and ε_i are independent error terms, for any unit i in the population, the GREG estimator of the population total is $\hat{\mathbf{Y}}_{GR} = \hat{\mathbf{Y}} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}$ where $\hat{\mathbf{B}}$ is the weighted least squares estimate of $\boldsymbol{\beta}$ and $\hat{\mathbf{X}}$ is the Horvitz-Thompson estimate of $\mathbf{X}$. The essence of the GREG approach consists of, under the regression model, finding the adjusted weights w_i^* that are the closest to w_i, by minimizing the chi-square distance between them.

Let us simulate three auxiliary variables that are *education*, *poverty* and *under_five* (number of children under five in the household) and assume that we have a total number of under five children of 6300 in the East, 4000 in the North, 6500 in the South and 14000 in the West. Similarly, we have the following number of households per poverty status (*Yes*: in poverty / *No*: not in poverty) and education level (*Low*: less than secondary, *medium*: secondary completed, and *high*: more than secondary):

```
np.random.seed(150)
full_sample["education"] = np.random.choice(
("Low", "Medium", "High"), size=150, p=(0.40, 0.50, 0.10))
full_sample["poverty"] = np.random.choice((0, 1), size=150, p=(0.70, 0.30))
full_sample["under_five"] = np.random.choice((0, 1, 2, 3, 4, 5), size=150,
p=(0.05, 0.35, 0.25, 0.20, 0.10, 0.05))
full_sample[["cluster", "region", "household", "nr_weight", "education", "poverty",
"under_five"]].head()
```

	cluster	region	household	nr_weight	education	poverty	under_five
0	7	North	72	49.464286	High	1	1
1	7	North	73	54.410714	Low	0	3
2	7	North	75	54.410714	Medium	0	2
3	7	North	715	54.410714	Medium	1	2
4	7	North	722	0.000000	Medium	0	2

We now will calibrate the nonresponse weight (*nr_weight*) to ensure that the estimated number of households in poverty is equal to 4,700 and the estimated total number of children under five is 30,800.

The class *SampleWeight* uses the method *calibrate()* to adjust the weight using the GREG approach. The control values must be stored in a Python dictionary i.e. *totals = {"poverty": 4700, "under_five": 30800}*. In this case, we have two numerical variables: *poverty* with values in 0, 1 and *under_five* with values in 0, 1, 2, 3, 4, 5. The argument *aux_vars* represents the matrix of covariates.

```
totals = {"poverty":  4700, "under_five":  30800}
full_sample["calib_weight"] = SampleWeight().calibrate(samp_Weight =
full_sample["nr_weight"],
aux_vars = full_sample[["poverty", "under_five"]], control = totals)
full_sample[["cluster", "region", "household", "nr_weight", "calib_weight"]].head()
```

	cluster	region	household	nr_weight	calib_weight
0	7	North	72	49.464286	50.432441
1	7	North	73	54.410714	57.233887
2	7	North	75	54.410714	56.292829
3	7	North	715	54.410714	56.416743
4	7	North	722	0.000000	0.000000

If we want to control by domain then we can do so using the argument *domain* from *calibrate()*. First we update the Python dictionary holding the control values for each domain. Note that the dictionary is now a nested dictionary where the higher level keys hold the domain values i.e. East, North, South and West. Then the higher level values of the dictionary are the dictionaries providing mapping for the auxiliary variables and the corresponding control values.

```
totals_by_domain = {
"East":  "poverty":  1200, "under_five":  6300,
"North":  "poverty":  200, "under_five":  4000,
"South":  "poverty":  1100, "under_five":  6500,
"West":  "poverty":  2200, "under_five":  14000,
}
```

```
full_sample["calib_weight_d"] = SampleWeight().calibrate(
samp_weight = full_sample["nr_weight"],
aux_vars = full_sample[["poverty", "under_five"]],
control = totals_by_domain,
domain = full_sample["region"])
```

```
full_sample[["cluster", "region", "household", "nr_weight", "calib_weight",
"calib_weight_d"]].head()
```

	cluster	region	household	nr_weight	calib_weight	calib_weight_d
0	7	North	72	49.464286	50.432441	40.892864
1	7	North	73	54.410714	57.233887	61.852139
2	7	North	75	54.410714	56.292829	59.371664
3	7	North	715	54.410714	56.416743	47.462625
4	7	North	722	0.000000	0.000000	0.000000

Note that the GREG domain estimates above do not have the additive property. That is the GREG domain estimates do not sum to the overall GREG estimate. To enforce the additive property of the GREG estimates, we must use *additive=True* when calling *calibrate()*.

4.5 Replicate weights

We can use *samplics* to create replicate weights. It is best to create the replicate weights from the design weights and apply the weight adjustments to each replicate. The API for creating the replicate weights is:

```
ReplicateWeight(method, stratification=True, number_reps = 500, fay_coef = 0.0,
random_seed = None)
```

```
ReplicateWeight.replicate(samp_weight, psu, stratum = None, rep_coefs = False,
rep_prefix = None, psu_varname = "_psu", str_varname)
```

The user provides the sample weight to replicate with the sampling design information, namely the PSU and stratification as applicable. In the case of the *Fay*'s method, Dippo et al. (1984) and Fay (1989) , the user may provide *fay_coef*, the coefficient to adjust the original weights.

5 Population parameters estimation

The *samplics* estimation module has two main parts: linearization (Taylor series) and replication based estimations.

5.1 Linearization (Taylor series)

The API for the Taylor-based estimation is as shown below and the argument *parameter* may take the value *mean*, *total*, *proportion*, or *ratio*. The main method of this class is *estimate()* which calculates the point estimates, the uncertainty measures, and other related statistics:

```
TaylorEstimator(parameter, alpha = 0.05, random_seed = None, ciprop_method = "logit")
```

```
TaylorEstimator.estimate(y, samp_weight = None, x = None, stratum = None,
psu = None, ssu = None, domain = None, by = None, fpc = 1.0, deff = False,
coef_variation = False, as_factor = False, remove_nan = False)
```

Some of the parameters of the method *estimate()* are : y is the variable of interest, $samp\_weight$ is the final sampling weight, x is the auxiliary variable in the case of the ratio estimation, $domain$ is the variable for the domain estimation (domain variable), and by is the variable to split the data; split the data then produce estimates for each partition (not domain estimation).

We are going to download the NHANES dataset and use it to estimate the average level of zinc:

```
from samplics.datasets import load_nhanes2
nhanes2_dict = load_nhanes2()
nhanes2 = nhanes2_dict["data"]
```

Now we estimate the average level of zinc:

```
zinc_mean_str = TaylorEstimator("mean")
zinc_mean_str.estimate(y=nhanes2["zinc"], samp_weight=nhanes2["finalwgt"],
stratum=nhanes2["stratid"], psu=nhanes2["psuid"], remove_nan=True)
print(zinc_mean_str)
```

SAMPLICS - Estimation of Mean
Number of strata: 31
Number of psus: 62
Degree of freedom: 31

MEAN	SE	LCI	UCI	CV
87.182067	0.494483	86.173563	88.190571	0.005672

The results of the estimation are stored in the dictionary *zinc_mean_str*. The users can covert the main estimation information into a *pandas DataFrame* by using the method *to_dataframe()*. The method *to_dataframe()* is more useful for domain estimation by producing a table where which row is a level of the domain of interest.

5.2 Replication

The class *replicateEstimator* provides the algorithms for the replication-based estimation. The argument *method* takes the value *brr* for the Balanced Random Replication (BRR) approach, McCarthy (1966); *bootstrap*, Rao & Wu (1992); and *jackknife*, Krewski & Rao (1981). Note that the *Fay*'s method is a generalization of the BBR method. Instead of simply taking half-size samples, we use the full sample every time but with unequal weighting: *fay_coef* for units outside the half-sample and 2 - *fay_coef* for units inside it (BRR is the case *fay_coef=0 or* None). *parameter* takes the same values as in the linearization case.

Let's load the NHANES again and use it estimate the ratio of weight over height. We want to use the BRR replicates weights.

```
from samplics.datasets import load_nhanes2brr
nhanes2brr_dict = load_nhanes2brr()
nhanes2brr = nhanes2brr_dict["data"]
```

Now we are going to estimate the ratio using the BRR replicates weights from the dataset.

```
from samplics.estimation import ReplicateEstimator
brr = ReplicateEstimator(method="brr", parameter="ratio")
ratio_wgt_hgt = brr.estimate(y=nhanes2brr["weight"],
samp_weight=nhanes2brr["finalwgt"],
x=nhanes2brr["height"], rep_weights=nhanes2brr.loc[:,
"brr_1":"brr_32"], remove_nan=True)
print(ratio_wgt_hgt)
```

SAMPLICS - Estimation of Ratio

Number of strata: None
Number of psus: None
Degree of freedom: 16

RATIO	SE	LCI	UCI	CV
0.426082	0.00273	0.420295	0.43187	0.006407

6 Categorical data

With *samplics*, users can analyze categorical data by producing tabulations and conducting t-tests.

There are two main *samplics* classes for tabulation i.e. *Tabulation* for one-way tables and *crossTabulation* for two-way tables. From the NHANES dataset downloaeded using load_nhanes2(), let's tabulate the variables *race* and *diabetes*, we can use the *tabulation* class as follows:

```
diabetes_nhanes = Tabulation("proportion")
diabetes_nhanes.tabulate(vars=nhanes2[["race", "diabetes"]], samp_weight=weight,
stratum=stratum, psu=psu, remove_nan=True)
print(diabetes_nhanes)
```

Tabulation of race
Number of strata: 31
Number of PSUs: 62
Number of observations: 10335
Degrees of freedom: 31.00

variable	category	proportion	stderror	lower_ci	upper_ci
race	1.0	0.879016	0.016722	0.840568	0.909194
race	2.0	0.095615	0.012778	0.072541	0.125039
race	3.0	0.025369	0.010554	0.010781	0.058528
diabetes	0.0	0.965715	0.001820	0.961803	0.969238
diabetes	1.0	0.034285	0.001820	0.030762	0.038197

In the case of two-way tabulation, we use the *crossTabulation* class. The APIs for *crossTabulation* is very similar to *tabulation*. Let's crosstabulate *race* by *diabetes*. We can use the *crossTabulation* class as follows:

```
crosstab_nhanes = CrossTabulation("proportion")
crosstab_nhanes.tabulate(vars=nhanes2[["race", "diabetes"]], samp_weight=weight,
stratum=stratum, psu=psu, remove_nan=True)
print(crosstab_nhanes)
```

Cross-tabulation of race and diabetes
Number of strata: 31
Number of PSUs: 62
Number of observations: 10335
Degrees of freedom: 31.00

race	diabetes	proportion	stderror	lower_ci	upper_ci
1	0.0	0.850866	0.015850	0.815577	0.880392
1	1.0	0.028123	0.001938	0.024430	0.032357
2	0.0	0.089991	0.012171	0.068062	0.118090
2	1.0	0.005646	0.000847	0.004157	0.007663
3	0.0	0.024858	0.010188	0.010702	0.056669
3	1.0	0.000516	0.000387	0.000112	0.002383

1

Pearson (with Rao-Scott adjustment):
Unadjusted - chi2(2): 21.2661 with p-value of 0.0000
Adjusted - F(1.52, 47.26): 14.9435 with p-value of 0.0000

Likelihood ratio (with Rao-Scott adjustment):
Unadjusted - chi2(2): 18.3925 with p-value of 0.0001
Adjusted - F(1.52, 47.26): 12.9242 with p-value of 0.0001.

With categorical data, users may want to compare groups. The class Ttest offers algorithms for testing means and proportions from one or two samples.

When sample sizes are too small for areas to produce reliable and stable domain estimates, the small area estimation (sae) techniques can help improve the precision of the estimates. The module *sae* of *samplics* provides routines for producing small area estimates.

7 Conclusion

As with R, Python now provides an open-source library for the design and analysis of survey sampling. The Python library *samplics* allows Python users to remain in the Python ecosystem when designing and analyzing complex samples. Furthermore, we expect that *samplics* will help bring more survey statisticians and official statistics producers to Python. Our ambition with *samplics* is to create a robust, comprehensive, and easy to use ecosystem for survey sampling and the production of official statistics.

References

Chow, S., Shao, J., Wang, H., Y, & Lokhnygina, Y. (2018). *Sample Size Calculations in Clinical Research, Third Edition*. CRC Press, Taylor & Francis Group.

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

Dippo, C. S., Fay, R. E., & Morganstein, D. H. (1984). Computing variances from complex samples with replicate weights. In *Proceedings of the Survey Research Methods Section, ASA* (pp. 489–494).

Fay, R. E. (1989). Theory and application of replicate weighting for variance calculations. In *Proceedings of the Survey Research Methods Section, ASA* (pp. 212–217).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernandez del Rio, J., Wiebe, M., Peterson, P., Gerard-Marchant, P., Sheppard, K.,

Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, **585**, 357–362.

Krewski, D. & Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, **9**, 1010–1019.

Lohr, S. L. (2022). *Sampling: Design and Analysis, Third Edition*. CRC Press, Taylor & Francis Group.

McCarthy, P. J. (1966). *Replication: An Approach to the Analysis of Data from Complex Surveys*.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (pp. 56–61).

Rao, J. N. K. & Wu, C. F. J. (1992). Resampling inference with complex survey data. *Journal of the American Statistical Association*, **83**, 231–241.

Ryan, T. P. (2013). *Sample Size Determination and Power*. Jonh Wiley & Sons, Inc.

World Health Organization (2018). *World Health Organization vaccination coverage cluster surveys: reference manual*. https://apps.who.int/iris/handle/10665/272820.

The Survey Statistician, 2022, Vol. 86, 44-46

Book and Software Review

Silvia Biffignardi & Jelke Bethlehem (2021). Handbook of Web Surveys, Wiley

Related ISBN: 9781119371687. 9781119371694. 9781119764496

Gaia Bertarelli, EMbeDS Department, Institute of Management,
Sant'Anna School of Advanced Studies, Pisa (Italy)
gaia.bertarelli@santannapisa.it

Abstract

The new-book *Handbook of Web Surveys*, 2nd edition, was released on June 2021 by Wiley. Revised and thoroughly updated, this handbook by *Silvia Biffignardi* and *Jelke Bethlehem* offers a practical and comprehensive guide for creating and conducting effective web surveys. The authors provide information on the most recent developments and techniques in the field. The book illustrates the steps needed to develop effective web surveys and explains how the survey process should be carried out. It also examines the aspects of sampling and presents several sampling designs. The book includes ideas for overcoming possible errors in measurement and nonresponse. The authors also compare the various methods of data collection. Critical information for designing questionnaires for mobile devices is also provided. Filled with real-world examples, Handbook of Web Surveys discuss the key concepts, methods, and techniques of effective web surveys. Suitable for a wide audience, the book is a useful manual for all those who wish to approach web surveys both from a theoretical and practical point of view, in the academic, official statistics or in business world.

Keywords: coverage error, adaptive design, self-selection bias, weighting adjustment techniques.

Modern society can be defined as a web society, in which technology assumes an ever greater and predominant importance, especially in the life of young adults who have always grown up with a strong technological support.

Surveys are part of the constantly evolving cultural and technological context of society and for this reason survey methodology is subject to change over time. Despite this, there are cornerstones of good data quality that must always be maintained, such as (i) good coverage of the target population, (ii) probabilistic sampling, (iii) low no-response error, (iv) accurate measurements, and (v) cost efficiency.

Web and mobile surveys allow respondents to complete questionnaires that are delivered to them and administered over the World Wide Web. Internet as data collection method offers more advantages, such as the potential for using complex questionnaires and visual and auditory incentives, the quick turnaround, and lower costs compared with other survey methods. Nevertheless, other problems arise, especially regarding the quality of data collection and its cornerstones recalled in the previous paragraph. In particular, coverage error and nonresponse error are the biggest threats to inference from Internet surveys.

The second edition of "Handbook of web surveys" by Silvia Biffignardi and Jelke Bethlehem aims to present a theoretical and practical approach to conducting and creating web surveys, combining

design and sampling issues. This can be considered both as a reference book for those who are starting to implement web surveys, and a book suitable for those who already work in the field of online surveys but want to explore newer aspects. It is suitable for students, academics and professionals in government, business, economic and social sciences organisations, as it best reflects an intersection of theoretical and practical approaches. It mainly helps treat problems on web surveys in contrast with traditional methods of data collection.

From the history of web surveys to the various ways of collecting data, to tips for detecting errors, this book introduces readers in depth to this ever-growing methodology and offers practical tips for creating successful web surveys.

The second edition of the book involves a revision of each chapter of the first edition considering the following criteria: (i) introduction of new literature and the most relevant results of recent years, and (ii) introductions of numerous examples and case studies to allow a practical study of the phenomenon, and revision of the examples present in the first edition. Updates have also been included to highlight new trends in mobile and web surveys and emerging solutions. A specific focus on mobile web surveys characterizes this edition. Two new chapters have also been introduced, one presenting a flowchart to show the steps needed to run a survey via web, and the other studying adaptive design.

The content of the book is smartly organized into twelve distinct chapters. Chapters can be addressed one after the other for beginners, while they can be seen as independent readings for the more experienced.

Chapter 1 “The road to web surveys” and Chapter 2 “About web surveys” provide an introduction into web surveys. Specifically, Chapter 1 faces the development of web surveys from a historical point of view, and examines the Blaise system and its development, going to fill a gap in the literature which was missing so far. Chapter 2 offers a basic overview of web surveys and their possible fields of use.

Chapter 3 titled “A framework for steps and errors in web surveys” reports one of the main differences with respect of the first edition. It presents a flowchart illustrating the steps and the sub-steps needed in the construction of web-surveys, explaining these steps in detail. The chapter discusses and analyses all the errors that can occur in a web survey, also placing them in the steps of the framework.

Chapter 4 “Sampling for web surveys” focuses on sampling. It underlines the need for valid probabilistic sampling to make inference and introduces the sampling frames necessary for this purpose. A number of sampling designs and estimation procedures which could be used in web surveys are discussed in order to guide the reader to the right choice considering its case study.

Chapter 5 titled “Errors in web surveys” provides a deep overview of possible errors, with great attention to errors in measurement, and their possible relations with questionnaire design. Moreover, it focuses on nonresponse errors, that can affect all the types of surveys but need particular attention in the online ones.

Chapter 6 “Web surveys and other modes of data collections” introduces other possible data collection methods, such as CAPI, CAWI, CATI, and their combination. After that, this chapter compares these methods with online data collection methods, considering advantages and disadvantages of each one.

Chapter 7 “Designing a Web Survey Questionnaire” discusses questionnaire design issues. Adaptations needed when a questionnaire is to be administered via web or mobile are taken into account.

Chapter 8 “Adaptive and Responsive Design” is written by Annamaria Bianchi and Barry Schouten. It studies methods for data collection with adaptive design when strategies are not defined in advance but must be adapted during fieldwork.

Society and technology change constantly. Efficient data collection methods must adapt to these changes, and it is not always certain that web surveys are the most appropriate solution. Mixed mode surveys with an online component included offer many advantages, but also challenges. One of the challenges is the use of mixed devices (smartphone, tablet) to complete an online questionnaire. Chapter 9 “Mixed-mode Surveys” aims to address these issues.

Chapter 10 focusses on “The Problem of Under-coverage”. Under-coverage is a problem of primary concern as it is closely linked to inequalities. In fact, in many countries the internet connection is not uniformly distributed over the territory, or accessible to the entire population in relation to income and their residence. Furthermore, the age factor must be considered, since there are still many elderly people who, not having familiarity with technology, risk not being represented by the web surveys. The chapter shows how under-coverage can lead to bias estimates and discuss several bias adjustment techniques.

“The Problem of Self-Selection” is addressed in Chapter 11. Many web surveys implemented to date do not use probabilistic sampling but are based on a self-selection mechanism. The problems that this mechanism can introduce in estimation are addressed in this chapter, which also focuses on showing how corrective methods are not always effective and how web surveys often turn out to be biased.

Chapter 12 “Weighting Adjustment Techniques” addresses several weighting techniques, such as post-stratification, ranking ratio estimation and generalised regression estimation. In addition, these techniques are explored based on their abilities to reduce under-coverage or self-selection bias.

Chapter 13 “Use of Response Propensities” introduces the idea of response probabilities, with particular attention to the response propensity weighting approach and the response propensity stratification method. The first attempts to adjust the original selection probabilities, while the second recalls the post-stratification methods.

The last Chapter, Chapter 14, explores the concept of “web panel”. Web panel is a survey system in which the same individuals are interviewed via web at different time points. Data are so collected in a longitudinal way on the same individuals, in a sort of panel design. There are some methodological challenges. This chapters gives an overview of several aspects of web panels. It describes its advantages and disadvantages, and examples of existing web panels are given.

Each chapter is structured in a first theoretical part, followed by a section on applications. At the end of each chapter there is a useful summary section and a section about the key terms related to the topic that has been addressed. Exercises and references conclude each chapter of the book.

The website https://www.web-survey-handbook.com/ is the companion of the book. It provides the survey data set which is used in the book for many applications and examples. Dataset is available in SPSS (SPSS Corporation, Chicago, IL) format. A section about the simulation of opinion pools is also available on the website.

Country Reports

ALBANIA

Reporting: **Prof. Dr. Besa Shahini**

Recent initiatives at the Algeria Institute of Statistics (INSTAT)

Official statistics are offered in Albania by INSTAT, the Bank of Albania and the Respective Ministries. Starting in 2020 INSTAT created a separate section on the website, dedicated to Covid-19, where it published the most frequent data, which helped policy makers, analysts and researchers seeking further economic-social analysis regarding the situation created in the country.

http://www.instat.gov.al/en/covid-19-statistics/

INSTAT has the mandate of collecting data for the production of official statistics from institutions, businesses, families. Pursuant to this principle, during 2020 INSTAT conducted 15 surveys of enterprises, surveying about 23,053 enterprises and 8 surveys of households, surveying about 51,302 households, and cooperated with 30 institutions to provide statistical information.

During 2020-2021 and in cooperation with UNICEF, 7 new indicators were produced, which brought the total number to 49 indicators. The new indicators are part of a special publication dedicated to children, adolescents, and young people in Albania http://instat.gov.al/al/sdgs/

A special system, called "Simona" has been made available to researchers to access statistical information remotely, not only because of the Covid-19 situation, but also to respond to those interested as quickly and in the form right.

The National Statistical System has carried out objectively and independently 125 statistical activities out of 131 planned activities, resulting in 6 failures. INSTAT has realized 113 activities: Bank of Albania 10 activities and the Ministry of Finance and Economy 2 activities. The data is made public at the same time to all users, according to the Publications Calendar. The methodology of production of statistical activities and definitions are in accordance with European and international standards, according to the generated model for statistical process management (GSBPM). In 2020, 27 statistical activities were documented / improved according to this model. For the implementation of this standard, training sessions have been conducted at two other statistical agencies.

In the year 2020, through the Time Use System (TUS) and the evaluation of financial resources used, 65 statistical activities were costed. The production of statistics aims to increase trust and satisfaction among users. In the period 2020-2021, user satisfaction increased by 0.5 percentage points, compared to a year ago, while the level of trust in statistics increased by 0.9 percentage points and the level of satisfaction with the INSTAT website, where users can easily consult the statistics, increased by 0.7 percentage points.

INSTAT in the process of statistical production follows scientific criteria for the selection of sources, methods, and procedures. The correct application of these elements has meant that the Error Handling Policy of published statistics is not used in any statistical activity. The number of INSTAT statistical publications went to 162 publications in 2021 from 160 realized in 2019, including statistical publications and books.

INSTAT during 2021 has trained the staff of INSTAT and statistical agencies, regarding the adaptation of the general model of statistical process management otherwise known as GSBPM (Generic Statistics Business Process Model) version 5.1, implemented in many European countries and further. In order to increase public confidence in official statistics, in the framework of statistical quality, for 2021, 40 quality reports have been published, to conduct 2 Self-Assessments and 1 statistical Audit.

New publications

- Accommodation structures (Quarterly publication)

- Gender Equality Index for the Republic of Albania 2020 (Book)

For more information: besashahini@feut.edu.al

ARGENTINA

Reporting: **Verónica Beritich**

The 2022 Census will take place between March 16 and May 18. The novelty of this edition is that it will be possible to autocomplete digitally.

On January 25th, through Decree 42/2022 published in the Official Gazette, it is officially announced that the National Census of Population, Households and Housing of the Argentine Republic will be held on Wednesday, May 18th. For the first time, an online questionnaire in the census web page will enable people to complete it from their homes, if they prefer. It will be accessible from March16th to May18th at 8:00. The Census Day, census takers will visit all the homes in the country on to request census receipts from those who have chosen the digital modality, or to carry out the traditional personal interview for those who have not completed it yet.

People who wish to complete the digital questionnaire will only need to have a computer, tablet or cell phone with internet access. Being an optional tool in the context of the pandemic, the objective of this development is that people, who live in private homes in rural and urban areas, can choose the moment to self-census, optimizing the completion and subsequent processing times of the information.

For the INDEC, private dwellings are those dedicated to the accommodation of one or more households where people live under a family-type regime, whether or not they are relatives.

On the Census Day, more than 600 thousand people will participate, including urban and rural census takers, national and provincial coordinators and other positions that make up the census structure.

Accredited census takers will visit all private homes in the country to carry out face-to-face interviews or, in the case of households that have completed the digital census, request proof of completion. A 6-character alphanumeric code is going to be generated automatically once the census questionnaire is completed. It can be downloaded from any device and, in addition, it is going to be sent to the email declared at the start of the digital census. If there is more than one household in the dwelling, they will all use the same proof of completion.

Between 00:00 and 20:00, the Census Day will be a national holiday. There will be no theatrical performances, film screenings, sports competitions, shows or public gatherings. Nor may clubs and shops selling food items remain open (Law 24,254).

All people who live in the national territory have to answer the questions included in the census questionnaire. It is mandatory. This information is going to be used only for statistical purposes, in accordance with the provisions of article 18 of the aforementioned Decree.

For more information, you can access the website www.censo.gob.ar or follow the official accounts of the 2022 Census on Instagram, Facebook, Twitter and YouTube.

BURKINA FASO

Reporting: **Baguinébié Bazongo**

Covid-19 monthly survey

The national statistical office (INSD), in collaboration with the World Bank, conducted covid-19 monthly survey to assess the impact of covid-19 on households living conditions. The sampling frame was a list of 7010 households surveyed during the 2018 Living Standards Measurement Survey (LSMS). Data about phone numbers were recorded during the LSMS survey to permit the selection of households and to enable phone interviews. The telephone mode was used to collect data because of covid-19 restrictions to conduct face-to-face interviews. A total of 2500 households were selected from the frame, stratified by urban and rural areas. New sampling weights were calculated by multiplying LSMS sampling weights and covid-19 sampling weights to generalize estimations to the target population. The CAPI application was designed using Survey Solutions and installed on enumerators' tablets. Mobile phones were used by enumerators to conduct interviews at home and to record responses in the CAPI application. A total of 7 rounds have been conducted from July 2020 to January 2021 from the same households sampled in the first round. To encourage the participation of households and thereby increase the response rate, the INSD provided a monthly phone credit to each sampled household after each interview. During the first interview, enumerators requested an updated phone number to households to ensure that they could be reached for the next interviews. The challenges during this survey were the displacements of some households from one stratum to another stratum due to insecurity in their region. Some of the households were out of reach because they lost their phone.

The innovation of this survey was the use of an existing frame that contained household phone numbers to conduct the survey by telephone and leading to generalizing the estimations at the country level. This approach solved the incompleteness of a sampling frame that does not have a phone number.

For more information please contact Zakaria Koncobo, Head of household surveys Unit, zakoncobo@gmail.com

CAMEROON

Reporting: **Symplice Ngah Ngah**

New methodology to estimate household consumption in Cameroon

To monitor the living conditions of households in general, and analyze issues relating to poverty in particular, the National Institute of Statistics carries out on average every 5 years a survey called the CAMEROON HOUSEHOLD SURVEY (ECAM). Unlike the four previous editions of the ECAM, the methodology of ECAM5 is modeled on that of the Harmonized Household Survey on Living Conditions carried out in many West African countries and in Chad.

Among the limits of the previous approach, we note:

i. the indicator of well-being was consumption expenditure; however, this does not contribute directly to meeting the needs of the household; a household can acquire food and give it as a gift to another household, or store it for consumption much late;

ii. the failure to take seasonality into account; and consequently
iii. the difficult comparability with other country's indicators because of methodological differences.

In the new methodology, household consumption is the main ingredient for constructing a welfare indicator, rather than income. This choice is justified by two main reasons: (i) consumption is less subject to collection errors than income; (ii) consumption is less sensitive to exogenous shocks than income and therefore better reflects the real standard of living of the household over the long term [*A. Deaton (2002), Guidelines for constructing consumption aggregate, LSMS working paper 135. The World Bank, Washington, D.C*.]. Given on the one hand seasonal variations in consumption, and on the other hand the fact that a large number of goods and services are consumed on an annual basis, the practice is to collect consumption data on an annual basis. Consumption variables are classified into two main categories: non-food goods and services, and food products. Data collection of non-food goods and services is generally done retrospectively over 7 days, 1, 3, 6 or 12 months depending on the assumed frequency of use of these goods and services.

The valuation of consumption (especially self-consumption and donations) requires the conversion of non-standard units (heap, bowl, basket, etc.) into standard units on the one hand and monetary values on the other. To this end, a survey of non-standard units was carried out prior to the main survey.

To take seasonality into account, this survey will be conducted in three waves according to the four agro-ecological zones of Cameroon.

• The first wave will cover the period from the end of September to mid-December 2021;

• The second wave will run from the end of January to mid-April 2022 (the cropping season);

• The third wave will cover the period from late May to mid-August 2022 (the harvest period).

For more information, contact Mrs. Rosalie Niekou (rosalie.niekou@ins-cameroun.cm), ECAM technical supervisor, National Institute of Statistics of Cameroon.

CANADA

Reporting: **Michelle Simard and Christos Sarakinos**

Two successful machine learning applications in the 2021 Canadian Census

In Canada, the census is conducted every five years, the latest cycle being in 2021. In every cycle, some aspects are improved and modernised from previous cycles. In 2021, one of the modernisation activities was the implementation of machine learning algorithms. These efforts were completed to reduce costs and significantly decrease the processing time while keeping the same level of high-quality data. The first activity was to code and classify open-ended questions and the second was to code and classify the comments left by the respondents.

The 2021 Canadian Census long-form questionnaire, sent to about 3.7 million dwellings, contains more than 30 questions that have the possibility of a write-in response that does not correspond to one of a few “check-box” options provided to the respondent to select. Respondents are not bound in how they may respond to any such question. Their responses are likely to include any number of different spellings or even responses completely unrelated to the question at hand. Adding to this, the number of responses requiring coding ranges from approximately 2,000 to 23,000,000 depending on the variable being coded. While Statistics Canada has experienced coders working for various programs, due to the sheer volume of census data, and the length of time between cycles, each cycle Statistics Canada is required to hire hundreds of temporary employees who spend

approximately 10 months completing this step. These factors make the coding process an extremely large undertaking which takes approximately 10 months to complete. Due to this complexity, Statistics Canada made the decision to augment its coding process with machine learning applications for the 2021 census cycle.

After an initial exploration period it was determined that the “fastText” algorithm would be the algorithm of choice for the 2021 Census. FastText is a natural language processing algorithm developed by Facebook within the past ten years. It uses a neural network to transform an input string into a “word embedding”, that is, a numerical vector representation of the string that can then be transformed into class probabilities. The algorithm is embedded in Statistics Canada’s generalized coding system, G-CODE.

In the end about 40 variable fields were processed containing more than 85 million write-ins. There were many challenges in integrating this new method in the complex census processing system, but machine learning algorithms were successfully integrated within the coding steps for almost all variables; the Place of Work variable was one of the most difficult one to code. This innovation let to reducing significantly the number of coders needed to be hired. Further improvements are planned for use with the 2026 Census of Population.

In addition, in an effort to improve the analysis of the respondent comments received on the 2021 Census of Population, Statistics Canada used machine learning techniques to quickly and objectively classify census respondent comments. As part of the project, analysts identified seventeen possible comment classes and provided previous census comments labelled with one or more of these classes. These seventeen classes included the census subject areas, such as: demography, labour, education, sex and gender, etc., as well as other general census themes, such as "experience with the electronic questionnaire", "burden of response", "positive census experience" and comments "unrelated to the census". Four different text classification algorithms were compared: SVM, CNN, semi-supervised temperature-scaled BiLSTM and transformers. Following the evaluation, a bilingual multi-label transformers model was successfully implemented in production. Incoming comments from Canada’s 2021 Census of Population were objectively categorized, achieving a high accuracy of 90%. In addition to the ML model, a simple mapping technique was also used to assign classes based on respondents’ explicit references to specific question or page numbers. As a result of this successful project, feedback from respondents was quickly directed to the appropriate subject matter analysts during collection for their information.

Statistics Canada’s largest and most visible statistical program has been modernising its methods for many cycles and will continue the automation of its processes and usage of leading-edge technologies and techniques for future cycles.

CROATIA

Reporting: **Lidija Gligorova**

Using administrative data in the Croatian CBS

In the Central Bureau of Statistics of the Republic of Croatia, usage of administrative data sources is increasing progressively. For purposes of this report, Mr. Hrvoje Žagmeštar (zagmestarh@dzs.hr), Head of the Living Conditions Statistics Unit, described actions that have already been taken as well as future actions planned in regards to usage of administrative data in the Survey on Income and Living Conditions (EU-SILC). Ms. Josipa Kalčić Ivanić (kalcicj@dzs.hr), Head of Service Statistics Department, described actions planned in regards to the usage of administrative data in the Monthly Retail Report.

Administrative Data used with the Survey on Income and Living Conditions

In the next period, the Survey on Income and Living Conditions will be prepared by obtaining data from administrative sources in order to significantly reduce the burden on interviewers themselves and on respondents as well as to shorten time needed for data processing after the fieldwork in the Living Conditions Statistics Unit. The Unit will need to provide microdata in the same year in which the fieldwork was conducted, in line with the new Regulation (EU) 2019/1700 of the European Parliament and of the Council.

In the next phases of introducing the mentioned data into the Survey, the Living Conditions Statistics Unit will use the following administrative sources:

- 'JOPPD' administrative base of the Tax Administration – it will be used to obtain data on gross and net income, obligatory contributions from the income as well as on income tax and surtax of natural persons included in the Survey on Income and Living Conditions;
- Administrative base of the Ministry of Labour, Pension System, Family and Social Policy – it will be used to obtain data on social benefits of natural persons included in the Survey on Income and Living Conditions, which are under the competence of the Ministry of Labour, Pension System, Family and Social Policy;
- Administrative base of the Ministry of the Interior – it will be used to obtain personal identification numbers (OIBs) of natural persons included in the Survey on Income and Living Conditions;
- Administrative base of the Croatian Pension Insurance Institute (HZMO) – it will be used to obtain codes of occupations and activities of employed natural persons included in the Survey (ISCO-08 and NACE Rev. 2).

Administrative data used with the Monthly Retail Report

The CBS carries out monthly calculations and publication of turnover index in retail trade. Retail trade index is calculated based on the data collected by means of regular statistical survey Monthly retail report.

Development and quality improvement by using fiscalization data began in 2020 in the scope of one of the Eurostat projects. Specific objectives for work area included: general quality improvements and in particular reductions of the revisions of the first releases of the national indicators, improved information on the retail trade via Internet and burden reductions for reporting units.

Two data sources were analysed at the first stage of the project. The first source consists of the data from the VAT database from Tax Authorities. The second source consists of the data that are obtained from the Tax Office and are primarily intended for the Fiscalization declarations (hereinafter Fiscalization data). After extensive analysis of all possible sources, the CBS staff decided that data from the fiscalization process is the best source because of timely availability.

From the beginning of 2021 fiscalization data are used in the production process as a supplement to existing business survey data for micro and medium size units (business entities employing fewer than 10 persons selected by using the random sample method).

DENMARK

Reporting: **Joakim Schollert Larsen**

Collecting data with a paper-based questionnaire – The European Social Survey

Introduction

The following is a presentation of the method used for the data collection to the European Social Survey round 10 (ESSr10). The subject for the survey is the living conditions in Denmark. Statistics Denmark had the main responsibility for the data collection in close co-operation with VIVE – The Danish Center for Social Science Research. The data collection strictly followed a data protocol given by ESS.

Since the previous round in 2017, the data collection has changed from a physical face-to-face interview to a web-based and a paper-based questionnaire. This is due to covid-19, during which personal interviews have not been possible. The paper-based questionnaire in itself is not a new way of collecting data at Statistics Denmark, but a questionnaire of this magnitude has only been used to a small extend: Each of the paper-based and the web-based questionnaire takes about 45 minutes to answer and are about 30 pages long.

Collecting the data

The data collection lasted from November 2021 to April 2022. 8000 people from the age of 15 and up comprised the sample. Of these 6000 people, who did not respond on the first invitation, received a paper-based questionnaire. After receiving and answering the questionnaire, the respondents returned it to Statistics Denmark. Answers from every single returned paper-based questionnaire were typed in manually in the web-based questionnaire. More than 800 questionnaires were returned which made this process quite extensive and also very educative. It surprised us that so many chose to use this mode, which we thought was outdated.

Things to consider

So, what is the broader potential application and interest of this? Though the use of paper-based questionnaires is not revolutionary, it has led to a broader discussion of its more frequent use in future data collections; considerations include the do's and don'ts when dealing with an international study and paper-based questionnaire in this particular context. Regarding the application and interest, the use of a paper-based questionnaire potentially opens the door to a specific segment in the sample who would otherwise not have answered, given that they may not have direct access to the web-based questionnaire via the internet. Analysis shows, not surprisingly, that a majority of elderly among the respondents chose the paper-based questionnaire. Lastly, an important point in this matter is the trade-off between collecting answers from this segment by means of a paper-based questionnaire and the amount of resources needed setting up the scheme (hence it is relatively expensive compared to a web-based questionnaire). Given this fact, reflecting upon this trade-off is of high importance when considering the application of paper-based questionnaires in a study.

ETHIOPIA

Reporting: **Aberash Tariku**

1 The 30th Annual Conference of the Ethiopian Statistical Association (ESA)

The 30th Annual Conference of the Ethiopian Statistical Association (ESA) was held on May 21 – 22, 2022 with a theme of "The Role of Statistics for National Development in the Past, Present and in

the Future Perspective of Ethiopia" in Addis Ababa, Ethiopia. For information please contact the ESA at ethstat@gmail.com.

2 The first Gender Asset Gap survey report is finalized

The main purpose of the survey was to estimate the gender gap in asset ownership, the wealth gap and to analyse intra-household dynamics of asset ownership and wealth in Ethiopia.

2.1 Estimating Gender Asset Gap

One of the objectives of the survey is to explore gender parity in asset ownership among households with couples. The United Nations guideline recommends to measure the gender asset gap primarily using two indicators, namely, the prevalence of asset ownership among women and men, and the share/ratio of women and men owning assets. While the prevalence indicator measures the percentage of women and men who own a given type of asset from the total population of each respective gender, the ratio indicator measures whether women and men are equally represented among the owners of a given asset type.

The survey also explores the different modes of asset acquisition, forms of ownership, and alienation rights by different forms of assets and socioeconomic characteristics, sex being one of the primary dimensions of interest.

2.2 Estimating Gender Wealth Gap

The gender wealth gap shows the disparity between the value of assets owned by women and men. While the gender asset gap tells us whether women and men have equal rights to own assets, the gender wealth gap provides further information about the composition, quantities and the relative values of women's and men's assets.

2.3 Intra-Household Analysis and Decision Making

The third main objective of the survey is uncovering the intra-household dynamics of asset ownership and wealth within couples or between spouses. The survey also looks into the dynamics of intra-household decision making and its association with asset ownership and wealth.

2.4 Association between asset ownership and gender-based violence

The survey provides analysis of the relationship between asset ownership and wealth on the one hand and experience of and attitude towards spousal physical violence against women and men. Asset ownership might affect the [in]dependence, self-esteem and bargaining power of women, thereby their experience of and attitude towards violence.

2.5 Asset ownership and Covid-19 Pandemic

The sale of assets to cope with the adverse effects of the Covid-19 pandemic is also covered in the survey. Assets may serve as an insurance against shocks, such as the Covid-19 pandemic.

For further information, please contact Mrs. Sorsie Gutema at sorsieg@yahoo.com

FIJI

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

Recent developments at the Fiji Bureau of Statistics

Seasonal adjustment of time-series data

The Fiji Bureau of Statistics (FBoS) is producing seasonally adjusted series for high seasonal series since 2016. The "*Introductory Guide on Seasonal Adjustment of Time Series Data"* was published

by the department in March 2022 to educate and guide compilers internally on how a Seasonally Adjusted Series is compiled. Formulation of the *JDemetra Guide* is also in progress. The guide will provide detailed step by step instructions on using JDemetra software [developed by National Bank of Belgium and provided by Eurostat] and Fiji-based series for seasonal adjustment. The Australian Bureau of Statistics (ABS) provides technical support and expert advice on all seasonal adjustment works.

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

Fiji Standard Classification of Occupations (FISCO) Upgrade

The department is working to concord the Fiji Standard Classification of Occupations (FISCO 2007) to the Pacific Standard Classification of Occupations (PACSCO 2016) for the compilation of Employment Statistics.

Contact persons: Ms. Amelia Tungi ameliat@statsfiji.gov.fj and Mr. Tawaketini Autiko tautiko@statsfiji.gov.fj

Gross Domestic Product (GDP) Rebase

Fiji's current GDP base year is 2014. With the increase in demand for a recent base year, the Fiji Bureau of Statistics is working on a GDP rebase for the year 2019. Though rebasing will take place in 2024, preparations such as rebasing the indicators and deflators used for GDP estimation mainly of the Industrial Production Index, Consumer Price Index, Import & Export Price Index, Producer Price Index and Building Material Price Index are in progress. The Supply and Use table is also in the finalizing stage. These are important development works to update Fiji's GDP by production, expenditure and income approach.

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

High Frequency Phone Survey – World Bank

FBoS is currently preparing for High Frequency Phone Survey. The World Bank is monitoring the crisis and the socioeconomic impacts of COVID-19 through a series of high-frequency phone surveys, as countries move through the pandemic and into economic recovery. In-person surveys are often impossible due to social distancing, making phone surveys an attractive option given its track record for successfully collecting timely data to inform evidence during crisis. The survey will be conducted using random digit dialing with a target sample size of 2,500 respondents. The survey will collect data on the following:

1. Behavioral change in response to COVID-19.
2. Vaccine reluctance
3. Unemployment
4. Income
5. Food Security
6. Coping Strategies
7. A general snapshot of Fiji's condition

Contact person: Avineshwar Prasad avineshwarp@statsfiji.gov.fj

Vital Statistics, Demography & GIS

The Fiji Bureau of Statistics is working with Vital Strategies, an International Vital Statistics expert organization based at the United States of America to address issues concerning backlog of Birth,

Death and Marriage Data Collecting Activities. With the support of Vital Strategies, FBoS is currently conducting the project "Developing Vital Statistics Indicators and Assessing Completeness and Inequalities in the Registration of Births and Deaths." The project is expected to complete by end of the year – 2022.

Subsequently, the VDG Unit is also running Fiji Civil Registration & Vital Statistics (CRVS) Inequality Assessment Project with the UNESCAP. The project will build on an initiative implemented by ESCAP at the beginning of 2021 "Inequalities in CRVS: Let's really get everyone in the picture!" where experts from national governments, academia and development partners come together to develop guidelines and technical support for Fiji to assess inequalities by evaluating and using secondary data sources and indirect demographic methods for estimating vital events.

In addition, the department in collaboration with the Environmental System Research Institute (ESRI), is now utilizing the demographic tool (license) which gives a more visual representation of information.

Contact persons: Ms. Amelia Tungi ameliat@statsfiji.gov.fj and Mr. Meli Nadakuca mnadakuca@statsfiji.gov.fj

FRANCE

Reporting: **Philippe Brion**

The development of mixed-mode collections for the production of the French Official Statistics

With the introduction of the Internet as a new data collection mode and the increasing difficulties in contacting households (and also enterprises), the evolution of surveys towards mixed-mode protocols has become a strong strategic orientation for official statistical offices. The recent Covid crisis has been an additional reason to accelerate this evolution.

Many mixed-mode protocols can be used, making it possible to take advantage of the benefits of each collection mode, depending on the constraints, the survey topic and the target populations. However, such protocols lead to a more complex survey process. Adaptations are necessary to guarantee the quality of the results: firstly, the questionnaire and its duration, then the definition of the collection protocol and finally the statistical processing of the data after collection.

The French Statistical Authorities, under the umbrella of INSEE, have implemented an overall approach to this issue during the last ten years: first for business surveys, then for household surveys.

Common tools have been developed, by first splitting the production process in phases. This work needed efforts of standardisation and simplification: in particular with the introduction of self-administered sequences, this evolution towards mixed-mode surveys constitutes a real paradigm shift for household surveys.

More details (in French) can be found in the two first articles of N°7 of French « Courrier des Statistiques »: https://www.insee.fr/fr/information/6035950.

KENYA

Reporting: **David I. Ojakaa**

Improvements to the sample design of the Demographic and Health Survey

On-going between February and July of 2022, the Kenya Demographic and Health Survey (KDHS) counts as the most significant of national sample surveys currently being undertaken in Kenya by the Government's Kenya National Bureau of Statistics (KNBS)[1]. Part of the Demographic and Health Surveys (DHS) programme conducted recurringly five years in at least 90 developing countries, the KDHS collects and shares accurate data on fertility, family planning, maternal and child health, gender, malaria, and nutrition.

It is in three areas that the 2022 KDHS is innovative compared to previous survey rounds. First, with the promulgation of the new Constitution in Kenya in 2010 that mandated the creation of the second and lower tier of Government after the national Government – the 48 Counties – and therefore the increased demand for data in these counties, the sample size for the 2022 KDHS is significantly augmented. Thus, compared to the 2008 KDHS whose sample of women of reproductive age (WRA) comprised only 8,444 cases and the 2014 KDHS sample size of 31,079, the 2014 KDHS sample is 397.7% higher than the 2008 survey and 35% more than that of 2014, at 42,025 WRA. The second innovation is the cause of the first above. To alleviate the burden of respondent fatigue and survey management arising from the significantly increased sample sizes, the short questionnaire which collects priority information is administered in half of all the households selected; these data can be used for county-level estimates. Data collected in the full questionnaire are however used for national estimates. Lastly, a number of new themes have been added to the 2022 survey in addition to those prior. These include questions on mental health, and COVID-19 coverage.

More information on the survey can be obtained from: directorgeneral@knbs.or.ke; archive@dhsprogram.com

MALAYSIA

Reporting: **Mahmod Othman**

Flood Disaster Impact Assessment Survey

Malaysia has been affected by the worst flood in the country history resulting from a tropical depression made landfall on the eastern coat of Peninsular Malaysia which brought torrential downpours throughout the peninsula for three days. Eight out of 16 states and federal territories was affected by flood, causing more than 71,000 residents to be displaced from their homes and affected more than 125,000 people as a whole.

A survey has been carried out to have an overall assessment on the impact of the flood on the affected states. In early assessment, the loss of houses and vehicles damaged due to flood is predicted to be RM1.4 billion and RM1.33 billion respectively. The loss faced by the manufacturing sector is predicted amounted to RM1 billion, RM600 million damages to business premises and RM49.9 million loss in agricultural sector. Further assessment will be carried out to help the authority to have a better understanding of the aftereffect of this flood; and to have a better contingency plan to face this kind of disaster on national level.

National Covid19 Statistic

[1] The views expressed here are those of the author solely and not of KNBS nor those of the DHS programme.

Up until 24th January 2022, Malaysia's National Covid-19 Immunisation Programme (PICK) has administered a total of more than 62 million doses of Covid-19 vaccine, with at least 26 million people has received at least one dose of the vaccine. In addition, the third dose, known as the booster dose, has been administered to at least 10 million people when it first started on September 2021. Statistic shows that 97.9% of the adult population and 88.3% of the adolescents aged 12 – 17 years old has been vaccinated with two doses.

In the meantime, the vaccination programme is expected to initiate the vaccine administration to kids aged 5 – 11 years old. This is after the Drug Control Authority (DCA) of Malaysia has approved the use of vaccine on this category of population. Ministry of Health Malaysia targeted to administer first vaccine dose to at least 70% of the kids in the first two months of the programme, and to have at least 80% of them with complete vaccination in 6 months.

63rd ISI WSC 2021

The International Statistical Institute's 'World Statistics Congress 2021 - The Hague' was held virtually on 11-16 July 2021 due to COVID-19. Previously, Malaysia hosted the 62nd ISI WSC 2019 and for this latest edition, about 40 delegates from the Department of Statistics Malaysia (DOSM) led by the Chief Statistician Malaysia supported the event as the participants and presenters to experience the inspiring lessons and the culture of the other countries.

NETHERLANDS

Reporting: **Deirdre Giesen**

Improving data collection and redesigning the Labour Force Survey at Statistics Netherlands

Statistics Netherlands successfully updates application landscape for data collection with Phoenix Program

Until recently Statistics Netherlands used numerous, interwoven systems for data collection that did not optimally facilitate the ever-increasing need for flexibility (e.g., switching modes). In 2015, the Phoenix program started to work step-by-step towards a completely new application landscape for data collection. The main aims of Phoenix were to be able to perform all surveys more efficiently and to ensure business continuity.

The architecture is set up in such a way that potential new forms of data collection, such as apps, can be easily implemented in a plug-and-play manner. In addition, parts of the IT landscape can be renewed over time without too much effect on other parts. This sizeable IT project was characterized by incremental delivery of production capabilities, with more and more statistics being transferred to the new applications as the project progressed. The transfer of the last survey was completed on 31 December 2021.

For more information see CBS successfully updates application landscape with Phoenix or contact Joost Huurman jwf.huurman@cbs.nl (Director Research & Development, former program manager Phoenix).

Redesign of the Dutch Labour Force Survey

Based on a Eurostat regulation, a redesign of the Dutch Labour Force Survey (LFS) is implemented in 2021 with the purpose to increase comparability of labour force data between European Member States. Statistics Netherlands has seized this opportunity to introduce additional changes in the data-gathering and derivations of the LFS to further improve data quality. A redesign of a survey process

generally changes non-sampling errors that occur in the various steps of a survey, in particular during the data collection phase. This results in systematic differences in the outcomes of a survey, which are often referred to as discontinuities. To avoid that systematic differences in measurement errors and selection bias are incorrectly interpreted as period-to-period changes of the parameters of interest, it is important to quantify discontinuities that are the result of a survey process redesign.

In the case of the Dutch LFS a method to quantify and correct for discontinuities is developed as a part of the transition to the new survey design. The method is based on a time series model that is implemented in 2010 for the production of monthly labour force figures. Discontinuities in the first wave of the Dutch LFS are quantified by collecting data under the old and new design in parallel for a period of nine months where both sample sizes are equal to the sample size of the regular survey. Reliable direct estimates for discontinuities for the first wave are obtained from the data collected during the parallel run. Initially a parallel run with a length of three months was planned for the last quarter of 2020 but due to the COVID crisis, the changeover to the new design was postponed from January 2021 to July 2021. Discontinuities in the four follow-up waves are quantified by extending the time series model with level intervention components that model the moment that the data collection changes from the old to the new design. The information obtained with the parallel run for the first wave is integrated in the time series model and can be used to produce uninterrupted time series. More details about the redesign and transition process can be obtained from the contact persons.

Contact persons : Martijn Souren (mhj.souren@cbs.nl) & Jan van den Brakel (ja.vandenbrakel@cbs.nl)

NEW ZEALAND

Reporting: **Penny Barber, Jasmine Ludwig and Keith Lyons**

New longitudinal survey to measure poverty persistence

The challenge of how to measure the persistence of poverty is being addressed by a new survey recently started by Stats NZ. 'Living in Aotearoa' will be the largest longitudinal survey to be held in New Zealand, with a panel of 7,200 households increasing up to 25,000 households by 2025 – representing around 1 in every 75 of the nation's households. Covering income, housing costs and material well-being, 'Living in Aotearoa' will survey participants once a year for six years in a row.

Stats NZ is required to report on 10 measures of child poverty under the Child Poverty Reduction Act (2018). While nine of 10 measures can be collected through the Household Economic Survey (HES), the final measure, that of persistent poverty, requires data that follows members of households over an extended period of time. The HES will eventually be replaced by the new longitudinal household survey.

Work and planning for the 'Living in Aotearoa' survey began in mid-2020, with emphasis on improved design, systems and approach to minimize the main challenge of longitudinal surveys: attrition.

Milestones so far include:

- The design of a new longitudinal sample;
- The design of a new questionnaire;
- The introduction of a new survey interviewer user interface system;
- The development of a Relationship Approach for data collection, informed by the Māori worldview (Te Ao Māori).

In addition to the launch of the survey itself, the work that has gone into its development has applications that will benefit the wider organisation. 'Living in Aotearoa' marks a move to a survey user case management application (CMA) in Blaise 5. This was presented at a recent International Blaise User Demonstration meeting, focusing on setup of the CMA, Personal and Demographic questionnaires, and the interactions between these questionnaires and the CMA system.

The Relationship Approach is an important development in the way Stats NZ conduct its social surveys. Research shows some priority groups are at higher risk of attrition than others, leading to a less-representative sample over time, which may jeopardise the validity of the findings generated. Evidence from longitudinal studies shows that establishing and maintaining a meaningful relationship with survey participants is fundamental to achieving high retention rates.

In the Māori worldview, Te Ao Māori, investing in relationships is effective in ensuring Māori have trust and confidence in survey processes and outputs. The relationship approach is designed to develop a sense of collective responsibility, known as a shared *kaupapa*, so that people are more inclined to participate and contribute their information for the 'greater good'.

Stats NZ will produce a newsletter covering the technical aspects of the survey later this year. The participant-facing webpage for 'Living in Aotearoa' is stats.govt.nz/about-the-living-in-aotearoa-survey.

For further information, please contact LivinginAotearoaTeam@stats.govt.nz

POLAND

Reporting: **Tomasz Żądło**

2021 Census

The National Population and Housing Census was performed in Poland from April to September 2021. Due to the COVID-19 pandemic and restrictions the initially planned duration of the census was extended by three months. In the 2021 Census, Poland has continued, as it did in the 2011 Census, using administrative registers and non-public data as data sources together with different methods of data collection including CAWI (online self-enumeration), CATI and CAPI. The Census made use of almost 35 registers and information systems including among others the Universal Electronic System for Registration of the Population, the Social Insurance Institution data, the Central Register of Insured Persons, the State Fund for Rehabilitation of Disabled Persons and the Agricultural Social Insurance Fund. What is more, non-administrative data sources were used as well, including data collected by operators of telecommunication networks, electric energy, water, gas and heating energy suppliers. These data sources were used as a direct source of information, a list of population units, to increase the accuracy and for imputation purposes.

The above description is based on

- https://stat.gov.pl/en/national-census/national-population-and-housing-census-2021/national-population-and-housing-census-2021/preliminary-results-of-the-national-population-and-housing-census-2021,1,1.html
- https://stat.gov.pl/en/national-census/national-population-and-housing-census-2021/national-population-and-housing-census-2021/national-population-and-housing-census-2021-research-methodology-and-organization,3,1.html

where more details can be found including the list of questions. Data based on 2021 census will be available in English from September 2022 in the Local Data Bank (https://bdl.stat.gov.pl/BDL/start) and the Geostatistics Portal (https://portal.geo.stat.gov.pl/en/home/).

SPAIN

Reporting: **Belen Gonzalez Olmos and Maria Velasco Gimeno**

Integration of data sources in Tourism statistics in the Spanish Statistical Office (INE)

In recent years the Spanish NSI has been looking for new sources and new procedures to meet the needs of the users of statistics.

The new strategy involves significant efforts to reach agreements with data owners to work together and to develop methodologies that would allow us to combine the information from these new sources with the traditional ones, while maintaining their quality.

Hereunder, three new experimental statistics are noted. They have been made possible thanks to new data sources and the development of new methodologies and processes.

1. Mobile positioning data for tourism (mobile phone)

The purpose of the study is to obtain aggregate information, through cell phone signalling, by means of active and passive events captured by telephone antennas, on the movements of resident and foreign tourists and excursionists.

The Spanish NSI has signed an agreement with the 3 most important mobile operators to carry out this project, extracting information from their databases and implementing the definitions and methodology design by the NSI.

This new source of information provides much more detailed, disaggregated and timely indicators.

2. Distribution of the expenditure made by foreign visitors on visits to Spain with credit and debit card

In this experimental statistic, information from a traditional survey (inbound tourism expenditure-EGATUR) and data from an auxiliary source (bank transactions by credit and debit cards of foreigners in Spain) are integrated.

These bank transactions include transactions made through a card in person (payments made through the Point of Sale or POS Terminal), as well as cash withdrawals at ATMs. Combining both sources of information, this statistic provides data on tourist spending by visitors in the destination where the spending was actually made.

The use of bank card data makes it possible to offer a more detailed breakdown by the traveler's country of residence, as well as to identify with greater precision the place where said expenses have been made. This information complements the information currently published in EGATUR, in which the expenditure made by travelers is shown, taking into account the main destination of trips and excursions.

In addition, this experimental statistic provides information on traveler expenditure in autonomous communities that are generally not the main destination of trips or excursions by non-residents and that therefore do not have sufficient sample coverage in EGATUR. These are stopover destinations where tourists have layovers or go to for an excursion.

3. Measurement of the number of tourist dwellings in Spain and their capacity

Traditional accommodation registers normally do not include private dwellings. The approach that the NSI has used for extracting this information is through web scraping techniques, based on computer programs, which go through the webs collecting listings and their features. The biggest challenge to deal with, when web scraping techniques are used, is to unduplicate listings being in several platforms.

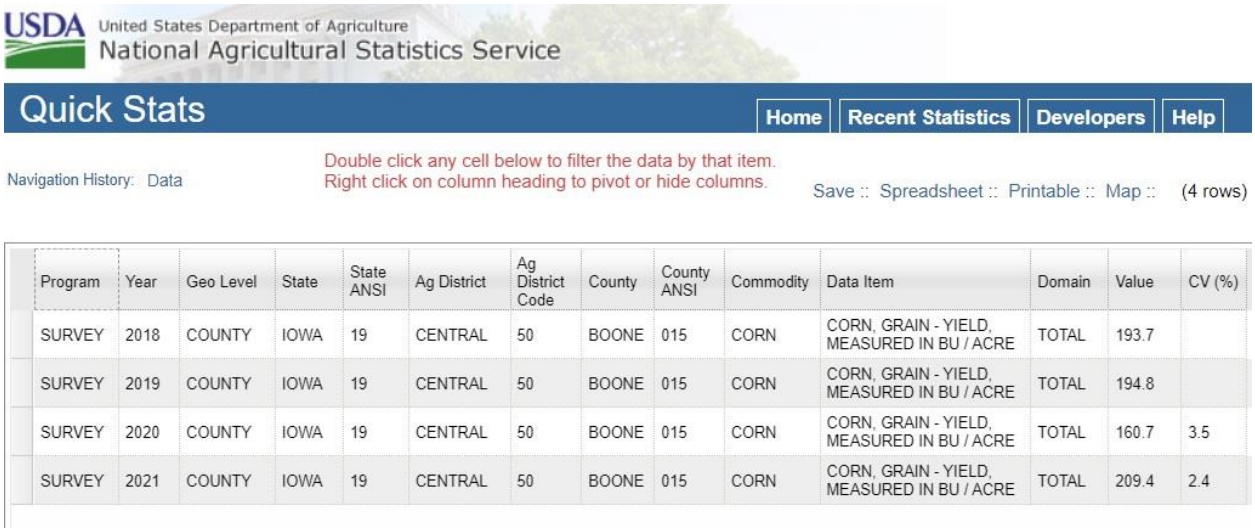
With this technique we have managed to draw a map of tourist housing throughout Spain at the census section without disturbing any informant.

UNITED STATES

Reporting: **Andreea L. Erciulescu** and **Nicole Dangermond**

U.S. Agricultural Official Statistics: Measures of uncertainty for end-of-season crop yield estimates

The county-level end-of-season crop estimates of acreage, production, and yield have been used by the U.S. Department of Agriculture (USDA) for programme administration, by a number of federal and state agencies for research and decision making, and by farmers and ranchers for planning and market assessment. Users may access these estimates, as well as many other U.S. agricultural estimates based on data from hundreds of sample surveys and the Census of Agriculture, using the USDA National Agricultural Statistics Service's Quick Stats tool. Recent statistical modeling developments have made it possible for the construction of measures of uncertainty for end-of-season crop yield estimates and these quantities started being released in the Quick Stats database in 2020. The table below provides an example of end-of-season corn-for-grain yield estimates, measured in bushels per acre, for Boone county in Iowa, for years 2018-2021.



Program	Year	Geo Level	State	State ANSI	Ag District	Ag District Code	County	County ANSI	Commodity	Data Item	Domain	Value	CV (%)
SURVEY	2018	COUNTY	IOWA	19	CENTRAL	50	BOONE	015	CORN	CORN, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	193.7	
SURVEY	2019	COUNTY	IOWA	19	CENTRAL	50	BOONE	015	CORN	CORN, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	194.8	
SURVEY	2020	COUNTY	IOWA	19	CENTRAL	50	BOONE	015	CORN	CORN, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	160.7	3.5
SURVEY	2021	COUNTY	IOWA	19	CENTRAL	50	BOONE	015	CORN	CORN, GRAIN - YIELD, MEASURED IN BU / ACRE	TOTAL	209.4	2.4

United States data on COVID-19 infections in the workplace

The county-level end-of-season crop estimates of acreage, production, and yield have been used by the U.S. Department of Agriculture (USDA) for programme administration, by a number of

Nearly 400,000 U.S. private industry workers were out of work for one or more days in 2020 due to a COVID-19 infection contracted as a result of performing their work-related duties, according to results of the Survey of Occupational Injuries and Illnesses (SOII) conducted by the U.S. Bureau of Labor Statistics (BLS). These data, released in late 2021, are the first and most comprehensive look at the effect of COVID-19 in the workplace. Three out of four of these cases occurred among workers in health care and social assistance industries, such as hospitals and nursing homes.

The SOII publishes estimates of incidence rates and counts of workplace injuries and illnesses, and provides details on the injured or ill worker and the circumstances surrounding the event or exposure for cases that involve one or more days away from work and for cases that require days of job transfer and restriction. The program relies on Occupational Safety and Health Administration (OSHA) recordkeeping requirements, which mandate employers record certain work-related injuries and illnesses.

Occupational injuries and illnesses collected in the 2020 SOII include cases of COVID-19 when a worker was infected as a result of performing their work-related duties and met other recordkeeping

criteria. In November 2021, the SOII reported that the rate of illness cases increased from 12.4 cases per 10,000 full-time equivalent workers (FTE) in 2019 to 55.9 cases in 2020. This change was driven by a dramatic increase in the respiratory illness case rate. COVID-19 is considered a respiratory illness under criteria established by OSHA.

The impact of COVID-19 is elsewhere reflected in the SOII through detailed case information for incidences requiring at least one day away from work. While the current version of the Occupational Injury and Illness Classification System (OIICS) does not have a unique code for COVID-19, these cases were classified as "other diseases due to viruses, not elsewhere classified" a rarely-used category prior to the pandemic. (For context, the last time this category was publishable was in 2015 when there were 20 cases reported.)

In 2020, private industry employers reported an estimated 390,020 cases of "other diseases due to viruses, not elsewhere classified," with an illness rate of 40.0 cases per 10,000 FTE. These cases made up about one-third of total injuries and illnesses requiring time away from work. Of these, private industry healthcare and social assistance establishments reported 288,890 cases, with a rate of 196.3 cases.

Median days away from work is an indicator of the severity of injuries and illnesses. In 2020, the median number of days away from work for all injury and illness cases was 12 days in the private sector, up from 2019 when median days away from work was 8 days. For "other diseases due to viruses, not elsewhere classified", the median days away from work was 13 days. Manufacturing, accommodation and food services, professional and technical services, and transportation and warehousing each had 14 median days away from work for this category.

Days away from work cases for Other diseases due to viruses, not elsewhere classified (n.e.c.), for selected private industries, 2020

Private Industry	Number	Rate[1]	Median Days
All private industries	390,020	40.0	13
Health care and social assistance	288,890	196.3	13
Manufacturing	30,490	25.4	14
Retail trade	19,090	17.5	13
Accommodation and food services	8,640	11.7	14
Wholesale trade	8,600	15.6	12
Administrative and waste services	7,000	13.6	13
Construction	4,690	6.8	12
Professional and technical services	4,370	5.0	14
Transportation and warehousing	3,930	7.6	14

Footnotes

(1) Rates are per 10,000 full-time equivalent workers

Source: U.S. Bureau of Labor Statistics, U.S. Department of Labor

For more tables reflecting COVID-19 in U.S. workplaces see, Nonfatal illnesses due to novel viruses by industry, Nonfatal illnesses due to novel viruses by occupation and How COVID-19 is reflected in the SOII data.

URUGUAY

Reporting: **Miguel Galmés & Juan Pablo Ferreira**

Uruguay adopts new methodology for its Continuous Household Survey (CHS)

Since 1968, the National Statistics Institute (NSI) has been conducting a monthly household survey (CHS) to obtain information on a set of socio-economic variables. During the pandemic (March 2020 - June 2021) the field survey was performed by telephone applying a reduced questionnaire to get the necessary information for continuing the labour market and income indicators monthly. During the health emergency, the CHS became a survey of rotating panels, where households were chosen at random using respondent cases of the CHS 2019 until February 2020; that is, the non-face-to-face CHS used a design in two phases: each rotating panel was a subsample of the households that had responded during in 2019 until February 2020.

Once face-to-face interviews were restarted, in July 2021, the CHS introduced a methodological change: the design used since 2006 (cross-section with monthly and independent random samples) was substituted for a design of rotating panels also with monthly periodicity, but where the one-month sample is composed of six panels or rotation groups (RG) being each RG a representative sample of the population. This implies that a household stays/participates in the CHS during six months. In the first month (implementation) it is visited in person using a form like the CHS 2019 one and using the same sample design as in previous years (random, clustered, stratified and in two stages of selection). Once the home is established, during the remaining 5 months the home is interviewed by telephone to collect labour market information for all the members that make up the working-age population only.

Each RG has an expected sample size of 2,000 households when initiated. This implies that once the rotating panel of the CHS is operational, that is, once the six-month rotation period has elapsed, the sample to estimate parameters of the monthly labour market will be composed (considering the expected attrition) of around 10,500 households. This increase in the monthly sample size[2] with respect to the previous design; the overlap of approximately 5/6 between the sample of one month and the previous one; and a new estimation method that uses composite regression/calibration estimators (using information from the labour market of the previous month, including the RG leaving the sample) allows an important reduction of sampling errors on level and net change of labour market indicators.

Because of its characteristics, the CHS with its new methodology, CHS can be seen as two different surveys: i) a cross-section multipurpose survey of living conditions and ii) a labour market survey. Because these surveys traditionally in other National Statistical Offices (NSOs) are carried out independently, as a result of the periodicity of the indicators (e.g. labour market monthly and poverty on an annual basis) the NSI of Uruguay, with its new methodology, tries to align with the rest of the NSOs but with a single survey.

For more information: https://www.ine.gub.uy/web/guest/encuesta-continua-de-hogares3

jferreir@ine.gub.uy; mgalmes@hotmail.com;

[2]Other indicators (e.g. income, poverty, living conditions) are computed using only households/individuals at the time of implantation

Conferences on survey statistics and related areas

Workshop on Survey Statistics 2022

of the Baltic-Nordic-Ukrainian Network on Survey Statistics will be held in Tartu, Estonia on 23 to 26 August, 2022. https://wiki.helsinki.fi/display/BNU/Events

ITSEW2022

The International Total Survey Error Workshop 2022 will be held in Manchester, United Kingdom from 31 August 2022 to 2 September 2022. Information is available at:

https://www.manchester.ac.uk/itsew2022

CESS2022 – The Conference of European Statistics Stakeholders 2022

will be held at the University of Rome "La Sapienza" on Oct 20 – 21. Scope of the Conference is to enhance the dialogue between European methodologists, producers, and users of European Statistics identifying the requirements of the users (ESAC), the best practices of the production (EUROSTAT, ECB, ISTAT, Banca d'Italia), with innovative ways of official statistics production based on Statistics, Data Science and Artificial Intelligence, and based on new methodological ideas for collecting and analysing data (Accademia via FENStatS).https://cess2022.dss.uniroma1.it/event/3/

2022 International Methodology Symposium

Data Disaggregation: Building a more-representative data portrait of society. Statistics Canada's 2022 International Methodology Symposium "Data Disaggregation: Building a more-representative data portrait of society" will take place virtually from November 2 to November 4, 2022, inclusively. 2022 International Methodology Symposium (statcan.gc.ca)

In Other Journals

Journal of Survey Statistics and Methodology

Volume 10, Issue 1, February 2022

https://academic.oup.com/jssam/issue/10/1

Survey Statistics

Nonparametric Mass Imputation for Data Integration
Sixia Chen, Shu Yang, Jae Kwang Kim

Targeting Key Survey Variables at the Unit Nonresponse Treatment Stage
David Haziza, Sixia Chen, Yimeng Gao

Design- and Model-Based Approaches to Small-Area Estimation in a Low- and Middle-Income Country Context: Comparisons and Recommendations
John Paige, Geir-Arne Fuglstad, Andrea Riebler, Jon Wakefield

Match Bias or Nonignorable Nonresponse? Improved Imputation and Administrative Data In the CPS Asec
Charles Hokayem, Trivellore Raghunathan, Jonathan Rothbaum

Parametric Bootstrap Confidence Intervals for the Multivariate Fay–Herriot Model
Takumi Saegusa, Shonosuke Sugasawa, Partha Lahiri

Survey methodology

Responsive and Adaptive Survey Design: Use of Bias Propensity During Data Collection to Reduce Nonresponse Bias
Andy Peytchev, Daniel Pratt, Michael Duprey

Building on a Sequential Mixed-Mode Research Design in the Monitoring the Future Study
Megan E Patrick, Mick P Couper, Bohyun Joy Jang, Virginia Laetz, John E Schulenberg ...

Risk of Nonresponse Bias and the Length of the Field Period in a Mixed-Mode General Population Panel
Bella Struminskaya, Tobias Gummer

Determined by Mode? Representation and Measurement Effects in a Dual-Mode Statewide Survey
Enrijeta Shino, Michael D Martinez, Michael Binder

Applications

Projecting Local Survey Response in a Changing Demographic Landscape: A Case Study of the Census in New York City
Annette Jacoby, Arun Peter Lobo, Joseph J Salvo

Using American Community Survey Data to Improve Estimates from Smaller U.S. Surveys Through Bivariate Small Area Estimation Models
Carolina Franco, William R Bell

CORRIGENDUM

Corrigendum to: Using American Community Survey Data to Improve Estimates from Smaller U.S. Surveys Through Bivariate Small Area Estimation Models
Carolina Franco, William R Bell

Volume 10, Issue 2, April 2022

https://academic.oup.com/jssam/issue/10/2

Survey Methodology

Positive Learning or Deviant Interviewing? Mechanisms of Experience on Interviewer Behavior
Yuliya Kosyakova, Lukas Olbrich, Joseph W Sakshaug, Silvia Schwanhäuser

Examining Interviewers' Ratings of Respondents' Health: Does Location in the Survey Matter for Interviewers' Evaluations of Respondents?
Dana Garbarski, Nora Cate Schaeffer, Jennifer Dykema

The Carryover Effects of Preceding Interviewer–Respondent Interaction on Responses in Audio Computer-Assisted Self-Interviewing (ACASI)
Hanyu Sun, Frederick G Conrad, Frauke Kreuter

Interviewer Effects in Live Video and Prerecorded Video Interviewing
Brady T West, Ai Rene Ong, Frederick G Conrad, Michael F Schober, Kallan M Larsen …

Measuring Skin Color: Consistency, Comparability, and Meaningfulness of Rating Scale Scores and Handheld Device Readings
Rachel A Gordon, Amelia R Branigan, Mariya Adnan Khan, Johanna G Nunez

A Model-Assisted Approach for Finding Coding Errors in Manual Coding of Open-Ended Questions
Zhoushanyue He, Matthias Schonlau

Survey Statistics

On the Robustness of Respondent-Driven Sampling Estimators to Measurement Error
Ian E Fellows

Estimating the Size and Distribution of Networked Populations with Snowball Sampling
Kyle Vincent, Steve Thompson

Neighborhood Bootstrap for Respondent-Driven Sampling
Mamadou Yauck, Erica E M Moodie, Herak Apelian, Alain Fourmigue, Daniel Grace ...

Model-Based Inference for Rare and Clustered Populations from Adaptive Cluster Sampling Using Auxiliary Variables
Izabel Nolau, Kelly C M Gonçalves, João B M Pereira

Applications

Challenges of Virtual RDS for Recruitment of Sexual Minority Women for a Behavioral Health Study
Deirdre Middleton, Laurie A Drabble, Deborah Krug, Katherine J Karriker-Jaffe, Amy A Mericle ...

ERRATUM

Erratum to: On the Robustness of Respondent-Driven Sampling Estimators to Measurement Error

Ian E Fellows

Volume 10, Issue 3, June 2022

https://academic.oup.com/jssam/issue/10/3

Special Issue: Privacy, Confidentiality, and Disclosure Protection

Preface

Preface to JSSAM Privacy, Confidentiality, and Disclosure Protection Special Issue

Natalie Shlomo, Anne-Sophie Charest

Survey Methodology

In an Era of Enhanced Cybersecurity: The Effect of Disclosing a Third Party's Role in Confidentiality Pledges on Response Propensity
Cleo Redline, Alfred D Tuttle

Data Privacy Concerns as a Source of Resistance to Complete Mobile Data Collection Tasks Via a Smartphone App
Caroline Roberts, Jessica M E Herzing, Jimena Sobrino Piazza, Philip Abbet, Daniel Gatica-Perez

Protecting the Identity of Participants in Qualitative Research
Joanne Pascale, Joanna Fane Lineback, Nancy Bates, Paul Beatty

Survey Statistics

A Hybrid Covariate Microaggregation Approach for Privacy-Preserving Logistic Regression
Lamin Juwara, Paramita Saha-Chaudhuri

Improving the Utility of Poisson-Distributed, Differentially Private Synthetic Data Via Prior Predictive Truncation with an Application to CDC WONDER
Harrison Quick

A Semiparametric Multiple Imputation Approach to Fully Synthetic Data for Complex Surveys
Mandi Yu, Yulei He, Trivellore E Raghunathan

Differential Privacy and Noisy Confidentiality Concepts for European Population Statistics Fabian Bach

Accuracy Gains from Privacy Amplification Through Sampling for Differential Privacy
Jingchen Hu, Jörg Drechsler, Hang J Kim

Private Tabular Survey Data Products through Synthetic Microdata Generation
Jingchen Hu, Terrance D Savitsky, Matthew R Williams

Differentially Private Synthesis and Sharing of Network Data Via Bayesian Exponential Random Graph Models
Fang Liu, Evercita C Eugenio, Ick Hoon Jin, Claire Mckay Bowen

Bayesian Inference for Estimating Subset Proportions using Differentially Private Counts Linlin Li, Jerome P Reiter

Nonparametric Differentially Private Confidence Intervals for the Median
Jörg Drechsler, Ira Globus-Harris, Audra Mcmillan, Jayshree Sarathy, Adam Smith

Applications

Incorporating Economic Conditions in Synthetic Microdata for Business Programs
Katherine Thompson, Hang Joon Kim

Survey Methodology

Statistics Canada Statistique Canada Canada

Survey Methodology, June 2022, vol. 48, no.1

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2022001-eng.htm

Maximum entropy classification for record linkage
Danhyang Lee, Li-Chun Zhang and Jae Kwang Kim

The anchoring method: Estimation of interviewer effects in the absence of interpenetrated sample assignment
Michael R. Elliott, Brady T. West, Xinyu Zhang and Stephanie Coffey

Relative performance of methods based on model-assisted survey regression estimation: A simulation study
Erin R. Lundy and J.N.K. Rao

Bayesian inference for a variance component model using pairwise composite likelihood with survey data
Mary E. Thompson, Joseph Sedransk, Junhan Fang and Grace Y. Yi

Non-response follow-up for business surveys
Elisabeth Neusy, Jean-François Beaumont, Wesley Yung, Mike Hidiroglou and David Haziza

Using Multiple Imputation of Latent Classes to construct population census tables with data from multiple sources
Laura Boeschoten, Sander Scholtus, Jacco Daalmans, Jeroen K. Vermunt and Ton de Waal

Bayesian inference for multinomial data from small areas incorporating uncertainty about order restriction
Xinyu Chen and Balgobin Nandram

A generalization of inverse probability weighting
Alain Théberge

Is undesirable answer behaviour consistent across surveys? An investigation into respondent characteristics
Frank Bais, Barry Schouten and Vera Toepoel

A simulated annealing algorithm for joint stratification and sample allocation
Mervyn O'Luing, Steven Prestwich and S. Armagan Tarim

Journal of Official Statistics

Volume 38 (2021): Issue 1 (March 2022)

Special Issue on Price Indices in Official Statistics

https://sciendo.com/issue/jos/38/1

Preface
Jörgen Dalén, Jens Mehrhoff, Olivia Ståhl and Li-Chun Zhang

Estimating Weights for Web-Scraped Data in Consumer Price Indices
Daniel Ayoubkhani and Heledd Thomas

Using Scanner Data for Computing Consumer Spatial Price Indexes at Regional Level: An Empirical Application for Grocery Products in Italy
Tiziana Laureti and Federico Polidoro

Sub-National Spatial Price Indexes for Housing: Methodological Issues and Computation for Italy
Ilaria Benedetti, Luigi Biggeri and Tiziana Laureti

Unit Value Indexes for Exports – New Developments Using Administrative Trade Data
Don Fast, Susan E. Fleck and Dominic A. Smith

Substitution Bias in the Measurement of Import and Export Price Indices: Causes and Correction
Ludwig von Auer and Alena Shumskikh

Rolling-Time-Dummy House Price Indexes: Window Length, Linking and Options for Dealing with Low Transaction Volume
Robert J. Hill, Michael Scholz, Chihiro Shimizu and Miriam Steurer

Econometric Issues in Hedonic Property Price Indices: Some Practical Help
Mick Silver

Rentals for Housing: A Property Fixed-Effects Estimator of Inflation from Administrative Data
Alan Bentley

Experimental UK Regional Consumer Price Inflation with Model-Based Expenditure Weights
James Dawber, Nora Würz, Paul A. Smith, Tanya Flower, Heledd Thomas, Timo Schmid and Nikos Tzavidis

The Geometric Young Formula for Elementary Aggregate Producer Price Indexes
Robert S. Martin, Andy Sadler, Sara Stanley, William Thompson and Jonathan Weinhagen

Measuring Inflation under Pandemic Conditions
W. Erwin Diewert and Kevin J. Fox

A Comment on the Article by W. Erwin Diewert and Kevin J. Fox
Carsten Boldsen

Creative and Exhaustive, but Less Practical – a Comment on the Article by Diewert and Fox
Bernhard Goldhammer

"Measuring Inflation under Pandemic Conditions": A Comment
Naohito Abe

Price Index Numbers under Large-Scale Demand Shocks–The Japanese Experience of the COVID-19 Pandemic
Naohito Abe, Toshikatsu Inoue and Hideyasu Sato

Early Real Estate Indicators during the COVID-19 Crisis
Norbert Pfeifer and Miriam Steurer

Volume 38 (2021): Issue 2 (June 2022)

https://sciendo.com/issue/JOS/37/3

In Memory of Dr. Lars Lyberg Remembering a Giant in Survey Research 1944–2021

Spatial Sampling Design to Improve the Efficiency of the Estimation of the Critical Parameters of the SARS-CoV-2 Epidemic
Giorgio Alleva, Giuseppe Arbia, Piero Demetrio Falorsi, Vincenzo Nardelli and Alberto Zuliani

Assessing Residual Seasonality in the U.S. National Income and Product Accounts Aggregates
Baoline Chen, Tucker S. McElroy and Osbert C. Pang

Improved Assessment of the Accuracy of Record Linkage via an Extended MaCSim Approach
Shovanur Haque and Kerrie Mengersen

If They Don't Understand the Question, They Don't answer. Language Mismatch in Face-to-Face Interviews
Jannes Jacobsen

Improving the Output Quality of Official Statistics Based on Machine Learning Algorithms
Q.A. Meertens, C.G.H. Diks, H.J. van den Herik and F.W. Takes

Data Fusion for Joining Income and Consumption Information using Different Donor-Recipient Distance Metrics
Florian Meinfelder and Jannik Schaller

Total Process Error: An Approach for Assessing and Monitoring the Quality of Multisource Processes
Fabiana Rocci, Roberta Varriale and Orietta Luzi

Some Thoughts on Official Statistics and its Future (with discussion)
Yves Tillé, Marc Debusschere, Henri Luomaranta, Martin Axelson, Eva Elvers, Anders Holmberg and Richard Valliant

Iterative Kernel Density Estimation Applied to Grouped Data: Estimating Poverty and Inequality Indicators from the German Microcensus
Paul Walter, Marcus Groß, Timo Schmid and Katja Weimer

Data Collection Expert Prior Elicitation in Survey Design: Two Case Studies
Shiya Wu, Barry Schouten, Ralph Meijers and Mirjam Moerbeek

Rejoinder: Measuring Inflation under Pandemic Conditions
W. Erwin Diewert and Kevin J. Fox

Book Review
Ann-Marie Flygare and Ingegerd Jansson

Survey Research Methods

Journal of the European Survey Research Association

Vol 16 No 1 (2021)

https://ojs.ub.uni-konstanz.de/srm/issue/view/226

How to enhance web survey data using metered, geolocation, visual and voice data?
Melanie Revilla

Nonresponse analysis in a longitudinal smartphone-based travel study
Peter Lugtig, Katie Roth, Barry Schouten

Measurement Equivalence in Sequential Mixed-Mode Surveys
Joseph Sakshaug, Alexandru Cernat, Richard J. Silverwood, Lisa Calderwood, George B. Ploubidis

Non-Compliance with Indirect Questioning Techniques: An Aggregate and Individual Level Validation
Thomas Krause, Andreas Wahl

Survey Participation in the Time of Corona an Empirical Analysis of an Effect of the COVID-19 Pandemic on Survey Participation in a Swiss Panel Study
Rolf Becker, Sara Möser, Nora Moser, David Glauser

Postscriptum to "Survey Participation in the Time of Corona"
Rolf Becker, Sara Möser, Nora Moser, David Glauser

Accounting for cross-country-cross-time variations in measurement invariance testing. A case of political participation
Piotr Koc, Artur Pokropek

An Evaluation of the quality of interviewer and virtual observations and their value for nonresponse bias reduction
Weijia Ren, Tom Krenzke, Brady West, David Cantor

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
MS	Ana	Abdelbasit	Bosnia and Herzegovina
DR.	Owens	Akpojaro	Nigeria
MR.	Salah	Barnawi	Saudi Arabia
MR.	Jean-François	Beaumont	Canada
DR.	Victor Alfredo	Bustos y de la Tijera	Mexico
DR.	Pablo	Cabrera Álvarez	United Kingdom
MRS	Jennifer	Daniels	United States
DR.	Carolina	Franco	United States
DR.	Alexander	Kowarik	Austria
MS	Leonor	Laguna	Peru
MR.	Clifford	Lesmoras	Philippines
PROF. DR.	Volker	Mammitzsch	Germany
DR.	Amitava	Mukherjee	India
DR.	Faustino	Oguan	Philippines
DR.	Noboru	Ohsumi	Japan
DR.	Walter J.	Radermacher	Germany
MR.	Dušan	Radovanovic	Serbia
MRS	Agbogidi	Rioborue Bess	Nigeria
PROF. DR.	Tobias	Schoch	Switzerland
MR.	Joseph Saidu	Sesay	Sierra Leone
MR.	Antoine	Simonpietri	France
PROF	Catherine	Vermandele	Belgium
PROF	Kirk M.	Wolter	United States
MR.	Leo Chun-keung	Yu	Hong Kong, SAR China

IASS Executive Committee Members

Executive officers (2022 – 2024)

President:	Monica Pratesi (Italy)	monica.pratesi@unipi.it
President-elect:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
Vice-Presidents:		
Scientific Secretary:	M. Giovanna Ranalli (Italy)	maria.ranalli@unipg.it
VP Finance	Jairo Arrow (South Africa)	jairo.arrow@gmail.com
Liaising with ISI EC and ISI PO plus administrative matters	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
Chair of the Cochran-Hansen Prize Committee and IASS representative on the ISI Awards Committee:	Nikos Tzavidis (UK)	n.tzavidis@soton.ac.uk
IASS representatives on the World Statistics Congress Scientific Programme Committee:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
IASS representative on the World Statistics Congress short course committee:	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
IASS representative on the ISI publications committee	M. Giovanna Ranalli (Italy)	maria.ranalli@unipg.it
IASS Webinars Representatives 2021-2023	Andrea da Silva (Brazil)	andrea.silva@ibge.gov.br
Ex Officio Member:	Ada van Krimpen	an.vankrimpen@cbs.nl

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

IASS LinkedIn Account https://www.linkedin.com/company/international-association-of-survey-statisticians-iass

Institutional Members

International organisations:

- Eurostat (European Statistical Office)

National statistical offices:

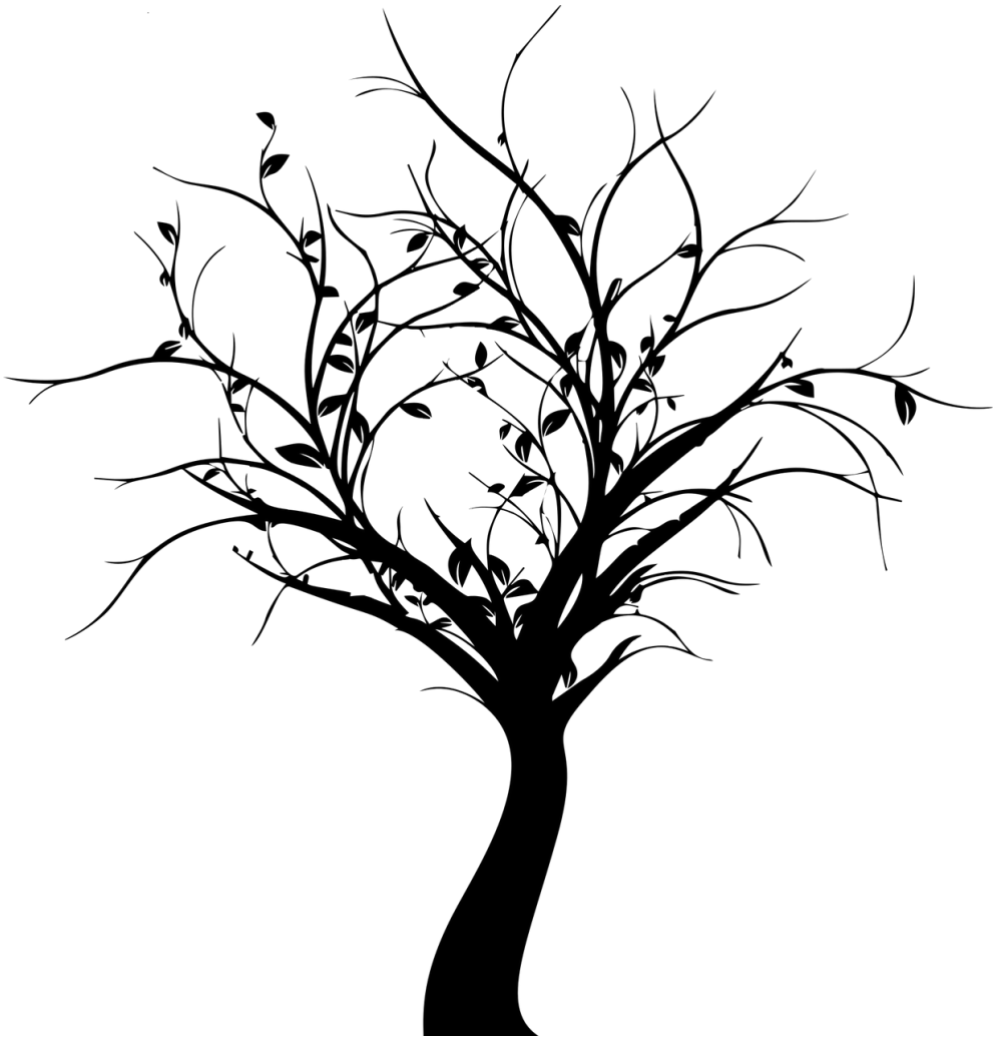
- Australian Bureau of Statistics, Australia
- Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistisches Bundesamt (Destatis), Germany
- Israel Central Bureau of Statistics, Israel
- Istituto nazionale di statistica (Istat), Italy
- Statistics Korea, Republic of Korea
- EC Eurostat – Unit 01: External & Interinst.
- Direcção dos Serviços de Estatística e Censos (DSEC), Macao, SAR China
- Statistics Mauritius, Mauritius
- Instituto Nacional de Estadística y Geografía (INEGI), Mexico
- Statistics New Zealand, New Zealand
- Statistics Norway, Norway
- Instituto Nacional de Estatística (INE), Portugal
- Statistics Sweden, Sweden
- National Agricultural Statistics Service (NASS), United States
- National Center of Health Statistics (NCHS), United States

Private companies:

- Westat, United States

Save a tree!

Read *the Survey Statistician* online!



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