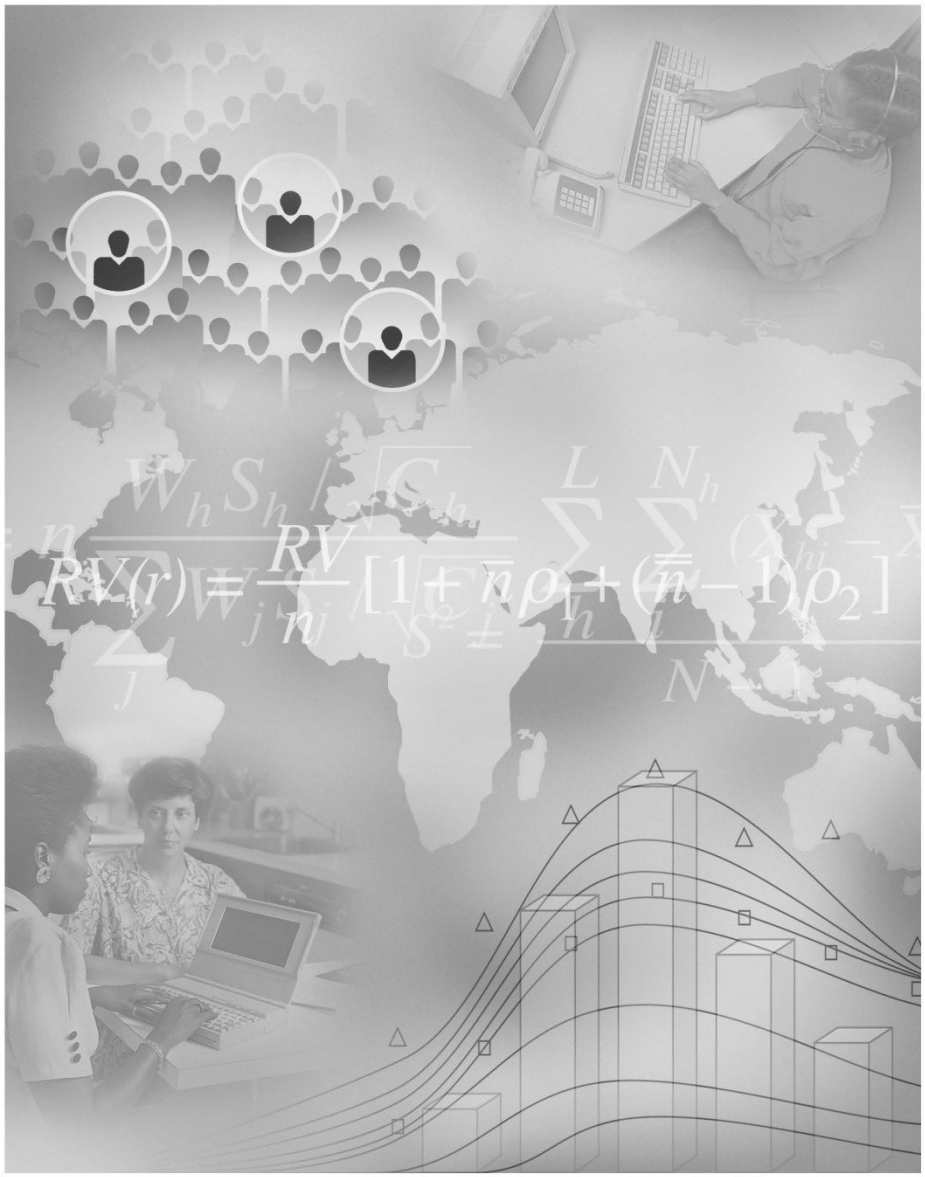
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

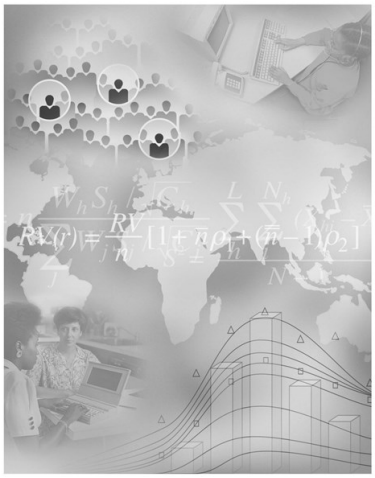
No. 80 **July 2019**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

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



Editors:
Danutė Krapavickaitė and Eric Rancourt

Section Editors:

Peter Wright	Country Reports
Eric Rancourt	Ask the Experts
Risto Lehtonen	New and Emerging Methods
Danutė Krapavickaitė	Book & Software Review

Production and Circulation:
Mārtiņš Liberts (*Central Statistical Bureau of Latvia*), Nicholas Husek (*Australian Bureau of Statistics*), and Olivier Dupriez (*World Bank*)

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

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

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

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

In this Issue:

Letter from the Editors

The July 2019 issue of The Survey Statistician includes all the traditional sections.

Letter from the President. The IASS President Peter Lynn in his letter highlights the reasons why the IASS exists. Ending his Presidential tenure Peter Lynn thanks the supporters and collaborators, and wishes the best to the incoming President.

Report from the Scientific Secretary. The scientific secretary Risto Lehtonen ends his tenure as well. He makes an overview of the scientific events that have been supported by IASS during his two years and presents some reflections of the survey statistics activities in the program of the 62th World Statistics Congress.

The News and Announcements section presents two valuable journal issues based on the presentations at the conference devoted to J. N. K. Rao's 80th birthday. The announcement on the 63rd World Statistics Congress in 2021 is postponed.

In the **Ask the Experts** section, Paul Biemer answers the question of whether estimates obtained from small sample size data are more accurate than estimates obtained from big data. He compares the accuracy of the estimates based on a sample survey of 6000 records and estimates based on 100 million administrative data records.

New and Emerging Methods section. Model-based methods are usually used for Small Area Estimation (SAE). Nikos Tzavidis, James Dawber, and Raymond L. Chambers present an overview of the literature devoted to the quantile-type models used in SAE. The models take into account domain heterogeneity but do not rely on an a-priori specification of the domain structure. Therefore such models possess good properties and have some application possibilities. Methods have been proposed for both continuous and discrete outcomes but they are not trivial. A forthcoming R package that implements the quantile-type approaches to SAE is announced.

The Book & Software review section presents a Proceedings volume of selected papers from the 2017 European Establishment Statistics Workshop (EESW17). The volume consists of 19 chapters on issues in business survey methodology written by different authors. As books dedicated to methodology for producing business statistics are few, this book is expected to attract the attention of statisticians.

Country Reports include 9 information letters from countries all over the world: Algeria, Argentina, Canada, Estonia, Hungary, India, Nepal, New Zealand and Ukraine. The reports cover new surveys and new solutions to problems arising when using multiple data sources and teaching innovations.

Finally, we would like to thank the IASS President Peter Lynn and Scientific Secretary Risto Lehtonen for their Letters and Reports, for the help given to the editors, and wish them success in all their activities.

As always, we want to recognize everyone working hard in putting *The Survey Statistician* together and in particular Margaret A. de Ruiter-Molloy of Statistics Netherlands, Nicholas Husek at the Australian Bureau of Statistics and Olivier Dupriez from the World Bank for their assistance.

Please let Risto Lehtonen (risto.lehtonen@helsinki.fi) know if you want to contribute to the *New and Emerging Methods* section in the future. If you have any questions which you would like to be answered by an expert, please send them to Eric Rancourt at Statistics Canada (eric.rancourt@canada.ca). If you are interested in writing a book or software review or suggesting a source to be reviewed, please get in touch with Danutė Krapavickaitė of the Vilnius Gediminas

Technical University, Lithuania (danute.krapavickaite@vgtu.lt). The country reports should be sent to Peter Wright of Statistics Canada (peter.wright2@canada.ca).

The Latex template for a TSS section paper is now available. We thank Mārtiņš Liberts (Central Statistical Bureau of Latvia) and Dalius Pumputis (Vilnius Gediminas Technical University, Lithuania) for preparation of this template. Please ask the editors for the template if you like to write a paper for TSS in Latex.

If you have any information about the conferences, events or just the ideas you would like to share with other statisticians – please 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

Eric Rancourt

Letter from the President

In my day job, three of the four projects that currently pay my salary are international collaborations between partners from many countries. These projects involve overseeing sampling and weighting for the European Social Survey, developing EuroCohort – a new longitudinal study of the wellbeing of children and young people – and investigating non-sampling errors in the European Statistics on Income and Living Conditions (EU-SILC). What these projects have in common – apart from survey methodology – is that very diverse groups of people with different backgrounds and working under different constraints work together, listen to each other, and achieve fantastic results that could not possibly be achieved any other way. While these particular projects only encompass a mere 35 or so countries, mainly within one continent, statistical surveys are a global undertaking. Researchers in all continents have similar objectives and face similar challenges. That is why the IASS exists, to provide the means to make connections and to share our successes, our challenges and our knowledge.

However, we are constantly questioning whether we are going about this the right way. If you have any thoughts about what you would like to see the IASS doing that it is not currently doing, or what it could do differently, please share them with us. Send an email to any of the Executive Committee listed elsewhere in this issue of the Survey Statistician, or tweet to us at @iass_isi.

Tempus fugit! This is my last letter as President. I feel that I have presided over a period of stability. The IASS has continued to do well the things that it does well. And both our finances and our membership are stable. Nothing disastrous has happened under my watch. That is a relief, but if I am honest I must say that I am a little disappointed that I have no exciting new initiatives to report. I'd like to help to put that right once I am divested of the responsibilities of presidency, which brings me back nicely to my request that you send in your ideas. A couple of modest recent achievements were that for the first time in several years one of the workshops that we sponsored was in Africa and that our Twitter account has seen a fair amount of traffic and has attracted followers from beyond the IASS membership.

Denise Silva will be our President for the next two years, and I know that she will do an excellent job. I first met Denise when we were both involved in running a residential training course in a beautiful location just outside Southampton, UK, more than twenty years ago. Since then we have mostly been on opposite sides of the globe, but as President-elect Denise has provided me with great support these last two years and she knows the IASS as well as anybody. I wish her a fruitful presidency.

I would also like to thank the other members of the Executive Committee, who have fulfilled their roles diligently and enthusiastically these last two years: Risto Lehtonen (scientific secretary), Jean Opsomer (vice-president, finances), Anders Holmberg (Chair of Cochran-Hansen prize committee), Cynthia Clark (IASS rep on the ISI programme committee). As always, the staff at the ISI permanent office have given us excellent support when needed and have unfailingly nagged us to do things that we might otherwise have overlooked.

With that it is time for me to sign off and to wish you (us) all successful survey endeavours in the coming months and years!

Peter Lynn

Report from the Scientific Secretary

The IASS has been active in many areas of scientific interest for the society. As the scientific secretary for the current operating period (2017-2019) I will discuss some areas briefly. These include the supporting of scientific conferences and workshops, activities related to the ISI WSC 2019, the Cochran-Hansen Prize competition, and The Survey Statistician journal.

Promoting and supporting scientific conferences and workshops has been one of the key activities of the IASS. The society has opened at the IASS website calls for application of financial support for workshops, conferences and similar events. Since the previous ISI WSC of 2017, the following 12 scientific events have received co-financing by the IASS. Decision-making for some of them goes back to the previous executive committee.

- Small Area Estimation Conference 2017 (SAE 2017). The conference took place in 10-12 July 2017 in Paris, France, as a satellite meeting for the ISI WSC 2017, and was organized by Ensai (Ecole Nationale de la Statistique et de l'Analyse de l'Information), the CREST (Centre de Recherche en Economie et Statistique) and the ILB (Institute Louis Bachelier). At the conference, the first SAE Award was given to Professor J.N.K Rao for his outstanding contribution to the research, application, and education of small area estimation.
- Workshop on Survey Statistics Theory and Methodology 2017. The workshop was arranged in 21-24 August 2017 in Vilnius, Lithuania, and was devoted to celebrate the 80th birthday of Professor Carl-Erik Särndal. The event was organized by the Baltic-Nordic-Ukrainian Network on Survey Statistics and was hosted by the Vilnius Gediminas Technical University.
- EESW17, the fifth biennial European Establishment Statistics Workshop. The event was organized by the European Network for Better Establishment Statistics (ENBES) in August 30 - September 1, 2017 in Southampton, UK, and was hosted by the University of Southampton.
- The 4th International Workshop on Surveys for Policy Evaluation and the 5th Brazilian School on Sampling and Survey Methodology (ESAMP V). This event was held in 17-20 October 2017 in Mato Grosso, Brazil. ESAMP is a biennial scientific meeting aimed at sharing knowledge and experiences regarding methodological aspects in planning and analyzing data from sample surveys. The 4th International Workshop on Surveys for Evaluation of Public Policies was held as part of the 5th ESAMP.
- The SAE 2018 Conference - Small Area Estimation and Other Topics of Current Interest in Surveys, Official Statistics, and General Statistics. The event was arranged in 16-18 June 2018 in Shanghai, China, as a celebration of Professor Danny Pfeffermann's 75th Birthday, and was hosted by the East China Normal University (ECNU).
- Second International Conference on the Methodology of Longitudinal Surveys (MoLS2). The conference took place in 25-27 July 2018 in Essex, UK, and was hosted by the Institute for Social and Economic Research (ISER) at the University of Essex.
- Workshop on Survey Statistics Theory and Methodology 2018. The event took place in 21-24 August 2018 in Jelgava, Latvia and was organized by the Baltic-Nordic-Ukrainian Network on Survey Statistics, the University of Latvia, the Latvia University of Life Sciences and Technologies, and the Central Bureau of Statistics of Latvia.

- The Francophone Survey Sampling Colloquium 2018. The conference was held in 24-26 October 2018 in Lyon, France. A plenary session was arranged for the celebration of the 2018 Waksberg Award laureate Jean-Claude Deville. The conference was hosted by the University of Lyon.
- Survey Process Design Workshop 2019. The event was organized on 16 January 2019 at the Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.
- The 6th ITAlian COnference on Survey Methodology (ITACOSM 2019). The conference was arranged on 5-7 June 2019 in Florence, Italy, and was organized by the Survey Sampling Group of the Italian Statistical Society and the Department of Statistics, Computer Science, Applications "G.Parenti" of the University of Florence.
- The 5th Baltic-Nordic Conference on Survey Statistics (BaNoCoSS-2019). The conference was arranged in 16-20 June 2019 in Örebro, Sweden, and was organized by the Baltic-Nordic-Ukrainian Network on Survey Statistics, University of Örebro, and Statistics Sweden.
- EESW19, the sixth biennial European Establishment Statistics Workshop. The workshop is organized by the European Network for Better Establishment Statistics (ENBES) and is scheduled for 24-27 September 2019 in Bilbao, the Basque Country, Spain.

The scientific programme of *the 62th World Statistics Congress of the ISI* includes two Special Invited Sessions (SIPS) organized by the IASS. The IASS President's Invited Session (SIPS167) features Gero Carletto of World Bank, and the IASS Special Invited Lecture Session (SIPS182) introduces two speakers, Frauke Kreuter (University of Maryland, USA & University of Mannheim, Germany) and Diego Andrés Pérez Ruiz (University of Manchester, UK), the winner of the Cochran-Hansen Prize 2019. Pedro Silva (IBGE, Brazil) will act as discussant for the session. The programme of invited paper sessions (IPS) contains a reasonable number of topics related to survey statistics, for example IPS-64 (Recent advances in statistical data integration), IPS-79 (Use of unconventional big data for official statistics: some current cases), IPS-101 (Advances in survey statistics for treating non-ignorable nonresponse), IPS-175 (Methodological development of global SDG indicators), and IPS-234 (Big data and small areas in monitoring sustainable development goals), just to mention a few. Also the list of Special Topic Sessions (STS) includes a number of interesting topics. In the programme for Short Courses, there many important and timely course topics on statistics, including for example a course by David Haziza on imputation methods for the treatment of item nonresponse in surveys, and a course entitled "Integrity of official statistics. Independence of official statisticians" by Jean-Louis Bodin.

The Conference on Current Trends in Survey Statistics 2019 should also be mentioned. It provides a satellite conference to the ISI WSC and takes place in 13-16 August 2019 at the National University of Singapore. Keynote speakers are Danny Pfeffermann, Partha Lahiri and Mark Handcock.

Hopefully many of us will have an opportunity to attend and contribute to the WSC and its satellite events. The 62th World Statistics Congress of the ISI will be held in 18-23 August 2019 in Kuala Lumpur, Malaysia, see http://www.isi2019.org/.

The winner of the *2019 Cochran-Hansen Prize* has been recently announced by the prize committee, chaired by Anders Holmberg. The winner is *Diego Andrés Pérez Ruiz* from Mexico. He will present his paper entitled "Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach" in the IASS Special Invited Lecture Session (SIPS182) of the ISI WSC mentioned above. We congratulate warmly the prize winner! The Cochran-Hansen Prize of the IASS is awarded every two years for the best paper on survey research methods submitted by a young statistician from a developing or transition country.

The society has published four issues of *The Survey Statistician* (TSS) journal during the operating period. In the section *New and Emerging Methods*, the following articles were published. "On some

reweighting schemes for nonignorable unit nonresponse", by Alina Matei (January 2018), "Empirical likelihood approaches in survey sampling", by Yves G. Berger (July 2018), "Calibration methods for small domain estimation", by Risto Lehtonen and Ari Veijanen (January 2019), and "Quantile-type methods for small area estimation", by Nikos Tzavidis, James Dawber and Raymond L. Chambers (the current issue). As the editor of the New and Emerging Methods section I want to give credit to the volunteers who kindly contributed to the Journal by publishing an article in the section.

Since the service as scientific secretary will be soon transferred to my successor, I close this report by wishing all success to the incoming scientific secretary. I want to thank Denise Silva, the former scientific secretary, for her excellent briefing and help, Peter Lynn for support, and all members of the executive committee for nice and fruitful cooperation.

Risto Lehtonen

News and Announcements

Special Issues: Contemporary Theory and Practice of Survey Sampling: A Celebration of Research Contributions of J. N. K. Rao

International Statistical Review, 2019, Vol. 87, Issue S1

Survey methodology, 2019, Vol.45, no 1

The conference *Contemporary Theory and Practice of Survey Sampling* was held in Kunming, China, 24–27 May 2017, celebrating 80th birthday of J. N. K. Rao. This conference was sponsored by the Research Institute of Big Data, Yunnan University. The organizing committee was chaired by Professor Jiahua Chen. The special issue of the *International Statistical Review* consists of 14 papers based on plenary talks presented at the conference. The special issue of the *Survey methodology* contains 8 papers which are a subset of the remaining papers which were presented at the conference.

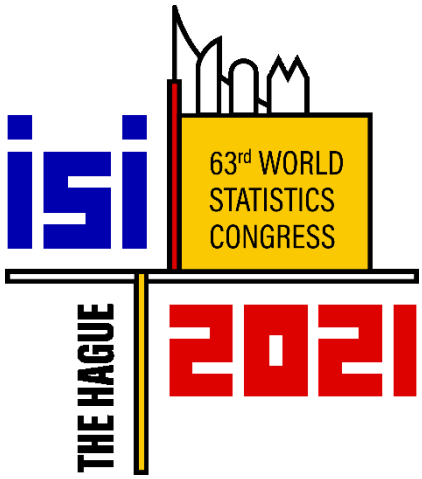
Both issues start with the article by J. N. K. Rao tracing his life as a statistician over the past 60 years. He writes: “my chancy life as a statistician has been very rewarding and satisfying. It was a great pleasure to work with many excellent researchers and graduate students”.

The Editors Jiahua Chen, Changbao Wu, Song Cai, and Mahmoud Torabi write introducing the issues:

> For more than 50 years, he has been a driving force in the development of unequal probability sampling methods, small sample approximations, analysis of complex survey data, empirical likelihood-based inferences, variance estimation techniques and resampling methods and missing data solutions with sound design-based properties. His abiding effort in meeting real-world needs led to another prolific area of his research on small area estimation, highlighted by his book *Small Area Estimation* (1st edition in 2003 and 2nd edition with Isabel Molina in 2015) published by Wiley.
>
> In addition to his phenomenal research impact, Professor Rao has had a significant influence on official statistics agencies through his participation on advisory boards and panels and his role as advisor and consultant. He has also inspired several generations of survey statisticians through his teaching, mentoring and research collaboration.

We congratulate J. N. K. Rao with his 82nd birthday and wish him healthy years, fruitful ideas and good collaborators!

International Statistical Institute, 63rd ISI World Statistics Congress



This major biannual conference of the International Statistical Institute will take place from 11-15 July 2021 in the World Forum in The Hague.

More than 2500 participants are expected from over 130 countries. They can choose from over 1300 presentations.

This is a unique chance for statisticians from the Netherlands to acquaint with new colleagues and meet some the world's leading experts in our science.

Location: The Hague, The Netherlands
Information: ISI Permanent Office, P.O. Box 24070,
2490 AB The Hague, The Netherlands.
E-mail: isi@cbs.nl
Website: https://isi-web.org/
Phone: +31–70–3375737

Ask the Experts

Can a Survey Sample of 6000 Records Produce More Accurate Estimates than an Administrative Data Base of 100 Million? (The Answer May Surprise You)

Paul Biemer

RTI International and University of North Carolina at Chapel Hill

Introduction

"Conducting a survey should be the last resort" said Tom Smith in his keynote presentation at the BigSurv18 Conference in Barcelona, Spain last year. Tom is the Managing Director of the Data Science Campus for the Office for National Statistics in the UK and he knows a thing or two about data. His point is that given the seemingly infinite supply of data available from such disparate sources as administrative records, internet transactions, social media and commercially available databases, policy makers searching for answers should first mine these exististing data sources before launching new surveys. Because of their costs, surveys should be considered only when other data sources are decidedly not "fit for purpose".

Tom notes that survey costs are ever-increasing as are the error risks. Survey nonresponse rates have been on the rise for decades. Measurement errors due to survey fatigue, disclosure insecurity, privacy concerns and so on add to the inaccuracy of survey estimates. However, Big Data often come with "big errors" as well. This article provides an example where a sample survey of about 6,000 households provides a more accurate estimate than one derived from a commercial data base consisting of over 100 million records. Survey quality can, and I suspect often does, trump data quantity.

In my chapter (co-authored with Ashley Amaya) in the BigSurv18 edited volume (Chapter 5 in Hill, et al, in press), we derive an expression for the total mean squared error (MSE) of the mean of a generic data set. The expression is valid for probability samples, nonprobability or convenience samples, or censuses. The expression contains components for essentially all error sources including those associated with the sample recruitment (or selection) process as well as the data encoding (e.g., measurement) process. In this article, I briefly describe the framework and illustrate how it can be used for assessing and comparing the accuracy of estimators from alternative data sources.

Total Error

The simplest decomposition of total error (TE) involves only two subcomponents which we refer to as data encoding error (DEE) and the sample recruitment error (SRE). DEE may be considered a generalization of the concept of "errors of observation" (Groves, 1989). It is a catch-all term for the combined error due to specification, measurement, data processing and other processes that affect the content of a data set. Likewise, SRE may be considered a generalization of the concept of errors of non-observation. It includes coverage error, sampling error, nonresponse/missing data and other processes that affect the representativity of a data set. Letting DEE = $(\bar{y}-\bar{x})$ and SRE =

$\bar{x} - \bar{X}$ leads to the identity $\bar{y} - \bar{X} = (\bar{y} - \bar{x}) + (\bar{x} - \bar{X})$ where $\bar{X}$ is mean of the population of interest, $\bar{y}$ is the mean of the data set, and $\bar{x}$ is mean of the data set assuming no encoding errors. In other words, $\bar{x}$ is the mean of true values or constructs for the records in the data set. Thus, total error (TE) of a data set mean can be defined as $\mathrm{TE} = \mathrm{DEE} + \mathrm{SRE}$.

In Biemer and Amaya (in press), we use this identity to derive an expression for the relative MSE (i.e., $\mathrm{MSE} / \bar{X}^2$) of the data set mean. It is the sum of components associated with DEE and SRE processes as follows:

$$\mathrm{RelMSE}(\bar{y}) = \mathrm{RelMSE}_{\mathrm{DEE}} + \mathrm{RelMSE}_{\mathrm{SRE}},$$

where

$$\mathrm{RelMSE}_{DEE} = RB_{\varepsilon}^2 + \frac{CV_X^2}{n}\left[\frac{1-\tau_y}{\tau_y}\right] + CV_X RB_{\varepsilon}\sqrt{\frac{N-n}{n}}\mathrm{E}_R(\rho_{RX})$$

$$\mathrm{RelMSE}_{\mathrm{SRE}} = \frac{CV_X^2}{n}\left[(N-n)\mathrm{E}_R(\rho_{RX}^2)\right] + CV_X RB_{\varepsilon}\sqrt{\frac{N-n}{n}}\mathrm{E}_R(\rho_{RX})$$

In these expressions, RB_{ε} is the relative bias associated with the data encoding process, CV_X is the population coefficient of variation of the true values, τ_y is the DEE reliability ratio (i.e., 1 minus the ratio of DEE variance to the total variance), *N* is the population size and *n* is the number of records in the data set. Finally, $\rho_{RX} = Corr(R_j, X_j)$ where R_j is the sample recruitment (or "response") indicator, which is 1 if the *j*th population member is present in the data set and 0 otherwise, X_j is the true value of the *j*th population unit and $E_R(\,)$ denotes expectation with respect to the sample recruitment mechanism, which may be random or deterministic. This expression builds on results found in Meng (2018). I now illustrate how these results can be used to answer the question posed in the title of this article.

Illustration

Data users who ignore nonsampling errors would say that a data set of almost 100 million records should be more accurate than one of only 6,000. Using data from the 2015 Residential Energy Consumption Survey (RECS) we show that this may not be true when the *total* error is considered. The RECS is a sample survey of U.S. households that collects energy characteristics, energy usage patterns, and household demographics. This survey has been conducted by the Energy Information Agency (EIA) since 1978. In 2015, the RECS, which was historically conducted by computer assisted personal interviewing (CAPI), experienced a very low CAPI response rate as well as other field issues. So, about halfway through data collection, all CAPI cases that had not yet been completed were moved to the mail/web mode (EIA, 2018). For the purposes of this illustration, only data from the CAPI survey will be used.

For the next RECS (scheduled for 2020), mail/web will be the primary mode of data collection according to EIA. This mode change has raised concerns as to whether some characteristics measured in the survey can be accurately obtained by self-reports. One such characteristic is housing unit square footage – a critical variable for household energy consumption modeling and the target variable for this illustration. Measuring the square footage of a dwelling can be complex and is often miscalculated. A better approach might be to rely on external sources. For example, commercial databases are available that collect housing square footage data from U.S. households, including Zillow, Acxiom, and CoreLogic. Of particular interest in this illustration is the Zillow data. Zillow obtains its data from county, municipal, and jurisdictional records. In addition, Zillow allows realtors and owners to edit these entries.

For the CAPI component of the 2015 RECS, interviewers were trained in the proper methods for collecting and, if necessary, estimating the square footage of housing units. Housing unit square footage was obtained from both respondents and interviewers who estimated the square footage from actual measurements of the housing unit. For purposes of this illustration, we consider interviewer square footage estimates as truth and estimate the error in the respondent and Zillow reports for a house unit as their differences from the interviewer estimate for that unit.

Suppose a researcher is interested in estimating the average square footage (using EIA's definition) of housing units in the U.S. The researcher has a choice between using the entire Zillow database, which contains about 100 million records but only covers about 82% of all housing units in the U.S. or a sample survey of n = 6,000 completed interviews (45% response rate) but nearly full coverage of all U.S. housing units. To help answer the question of which of these two approaches would provide the most accurate estimate, we use some of the results from EIA (2017), which appear in Table 1.

EIA (2017) provides an estimate of the nonresponse bias in the RECS square footage estimate. We show in Biemer and Amaya that this bias is equivalent to ρ_{RX} = -0.000295. For the Zillow data, it was not possible to estimate ρ_{RX} from the available data. Rather, we obtained a range of plausible estimates of ρ_{RX} (see Table 1) using the bounds provided in Meng (2018).

Table 1. Parameters used for computing mean squared errors (MSEs) for survey and Zillow estimates of average housing unit square footage.

MSE Component	Survey	Zillow
Relative bias (RB_ε)	-0.082	-0.14
Population Coef. of Variation (CV_X)	0.64	0.64
Reliability (τ)	0.59	0.66
Selection correlation (ρ_{RX})	-0.000295	[-0.27,0.22]
Population size (N)	118,208,250	118,208,250
Sample size (n)	6,000	96,930,765
Response rate	55.4%	82%
Coverage rate	≈ 99%	
Selection rate	0.009%	

In Figure 1a, we consider the two MSEs assuming there is no error due to data encoding — that is, we set relative encoding error bias (RB_ε) and reliability (τ) to 0. Clearly, the survey estimate has uniformly greater RelMSE over the entire feasible range of ρ_{RX} for Zillow data. This is not unexpected given that the Zillow sample size is more than 16,000 times larger than the survey and its coverage rate is over 80%. Figure 1b repeats the comparison, this time setting RB_ε and τ to their values in Table 1. Now, as shown in Figure 1b, the comparison changes dramatically. The RelMSE for Zillow is larger than its counterpart for the survey except for extreme, positive values of ρ_{RX}. What is happening is that the Zillow MSE_{SRE} and MSE_{DEE} components have the same sign for negative values of ρ_{RX}, which causes the relative MSE to be large when ρ_{RX} is in this range. When $\rho_{RX} = 0$, the RelMSE only reflects the DEE component. As ρ_{RX} becomes increasingly positive, the MSE_{DEE} component is being offset by the SRE component until, at approximately $\rho_{RX} = 0.2$, the two components offset one another to the point where the survey and Zillow MSEs are about

equal. As ρ_{RX} increases further, the RelMSE for Zillow becomes smaller than its survey counterpart.

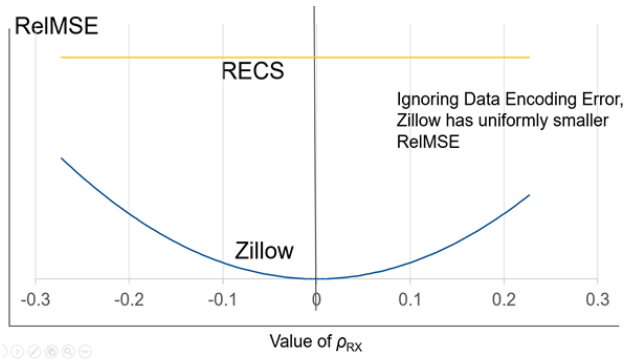


Figure 1a. RelMSE of survey and Zillow data for range of values for ρ_{RX} ignoring data encoding error

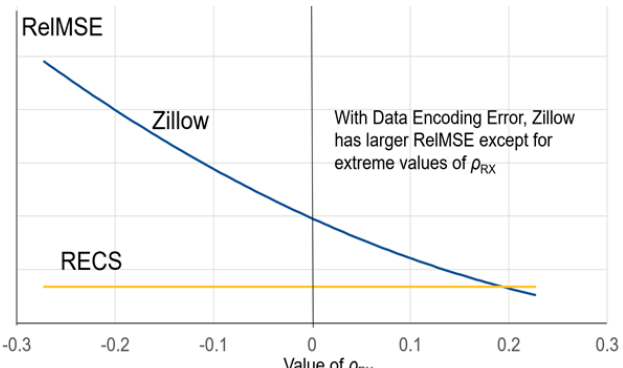


Figure 1b. RelMSE of survey and Zillow data for range of values for ρ_{RX} with data encoding error.

Of course, the situation where the total MSE is small because some error components offset others, is untenable. An unwitting methodologist trying to improve accuracy by reducing the DEE component, for example, might actually increase the total MSE by upsetting the delicate balance of errors. We suspect that the cause of the inaccuracy in the Zillow data may be specification error – i.e., the definitions of housing unit square footage for Zillow and EIA may be different.

This illustration emphasizes the need for evaluating the *total* error in a comparison of estimates from alternate sources rather than just focusing on a subset of MSE components. It also illustrates that the massive size of a data set may not matter when it comes to estimation accuracy.

Final Remarks

Note that our illustration considered the data set mean; however, the accuracy of aggregated data may not be relevant for EIA purposes (i.e., energy modeling) which relies more on micro-data. Thus, despite its data encoding flaws, might the Zillow data still be fit for energy modeling purposes? How accurate do the data need to be for these purposes? This question might be explored through a sensitivity analysis of the modeling process. Further, rather than choosing one data source over another, why not combine data sources in such a way as to maximize the strengths and minimize the weaknesses of each source – a process called *data integration*? For example, interviewer observations of square footage for a relatively small sample of housing units could be used to correct the measurement bias in the Zillow data. Such "hybrid" estimates retain some of the benefits of the massive data set and the measurement accuracy of the survey to produce estimates that are better than either single source estimate.

However, quite often Big Data cannot answer specific questions posed by researchers and data users. This is because, unlike survey data, Big Data are "found" not "designed." In those cases, researchers will continue to resort to surveys for a wide range of questions than cannot be answered by any found data set. However, as shown in several chapters of Hill, et al. (in press), Big Data are well-suited for answering many "unposed" questions – i.e., questions that are discovered in the process of mining Big Data. This only requires imagination – as well as perhaps a research team with expertise in domain science, data science, computer science and statistics. Data are not free.

References

Biemer, P. (in press). “Chapter 5: Total Error Frameworks for Found Data,” in Hill, et al (eds). *Big Data Meets Survey Science: A Collection of Innovative Methods*, John Wiley & Sons, Hoboken, NY

EIA (Energy Information Administration). (2018). Residential Energy Consumption Survey (RECS) 2015 consumption and expenditures technical documentation summary, downloaded on 12/20/2018 from https://www.eia.gov/consumption/residential/reports/2015/methodology/pdf/2015C&EMethodology.pdf

EIA (Energy Information Administration). (2017). 2015 RECS square footage methodology. Retrieved December 20, 2018, from https://www.eia.gov/consumption/residential/reports/2015/squarefootage/pdf/2015_recs_squarefootage.pdf

Groves, R. (1989). *Survey Errors and Survey Costs*, John Wiley & Sons, NY

Hill, C., Biemer, P., Buskirk, T., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. (in press). *Big Data Meets Survey Science: A Collection of Innovative Methods*, John Wiley & Sons, Hoboken, NY

Meng, X. (2018). “Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 U.S. Presidential Election. *Annals of Applied Statistics*12, 2, 685–726.

New and Emerging Methods

Quantile-type methods for small area estimation

Nikos Tzavidis[1], **James Dawber**[2] **and Raymond L. Chambers**[3]

[1]University of Southampton, UK, n.tzavidis@soton.ac.uk
[2]University of Southampton, UK, j.p.dawber@soton.ac.uk
[3]University of Wollongong, Australia, ray@uow.edu.au

Abstract

Small area estimation typically requires the use of model-based methods. One popular class of model-based methods uses random area effects. Alternatively, one can use a quantile-type ensemble model that assigns scores to sample individuals characterising the heterogeneity in the data. These scores are then used for estimating area/domain effects and hence for small area estimation. The aim of this article is to present a review of quantile-type methods for small area estimation. In doing so we consider a range of response data types, including continuous, binary, count and overdispersed data. We further describe areas of current research interest.

Keywords: Domain estimation; influence function; official statistics; outlier robust estimation; quantile regression; survey statistics.

1 Introduction

Sample surveys are commonly designed to measure characteristics of a population at national and large sub-national levels. Due to cost constraints the sample size is usually not large enough to allow for direct estimation of acceptable precision in planned or unplanned domains. Careful use of model-based methods can then be useful for producing estimates of acceptable precision in domains of interest. Here we prefer using the term domain, instead of area, to define a broader group structure comprising geographical and other characteristics. Hence, from now on we will be using the terms area and domain interchangeably. A plethora of small area methods (SAE) have been proposed in the small area literature over the years. To start with, the focus of this research was on the use of direct estimation that utilises only domain-specific data for estimation. This was in line with the tradition in the types of survey estimation methods used for the production of survey and official statistics. Although direct estimators have appealing features, for example design consistency, small domain-sample sizes can lead to imprecise small area estimates. Model-assisted methods for example, regression synthetic estimators have also been extensively studied in the literature. For reviews of SAE methods see Lehtonen and Veijanen (2009), Pfeffermann (2013), Rao & Molina (2015) and Tzavidis et al. (2018).

The present paper focuses on model-based methods that have been at the centre of research developments in recent years. The use of model-based methods is advocated on the basis of the potential improvement in the precision of small area estimation in particular, when working with small sample sizes. Generally speaking, model-based SAE models are classified into two broad categories, namely unit-level and area (domain)-level models. The models used in the latter case utilise domain-level covariates for model fitting and estimation. In contrast, models used in the former case use survey micro-data for model fitting and estimation. Although area-level models have many advantages, including the fact that the use of data for fitting the models are easier to gain access to, in this article we focus on unit-level models which are a more natural approach for quantile-type models.

A common approach to model-based SAE is via a random effects specification, with random effects characterising the heterogeneity between domains. Random effects models are based on the assumption that units that belong to the same domain are more similar than units that belong to different domains. There are, however, alternative approaches to SAE that do not require the use of a random effects model. One such approach utilises an ensemble approach based on fitting quantile-type (in particular M-quantile) regression models. M-quantiles and M-quantile regression were introduced by Breckling and Chambers (1988) and are a generalised form of "quantile-like" estimators which include quantiles and expectiles as special cases. Using ensemble models for SAE offers a different way of characterising between-domain heterogeneity. A suitable ensemble regression function that covers the full spectrum of variability for the characteristic of interest is first used to index the population. Domain heterogeneity is present if the unit-level indices (scores) cluster within domains. SAE for a particular domain is then based on the regression function within the ensemble that corresponds to an "average" index (score) for that domain. Under this approach there is no random domain effect, with consequent distributional assumptions while at the same time estimators are outlier robust. However, this comes at the cost of estimating the domain-specific "average" index by using only the domain-specific unit-level indices for the sampled individuals within the domain.

This paper focuses on describing quantile-type approaches to SAE. The literature generated by these approaches has led to renewed interest in outlier-robust model-based methods with applications in survey and official statistics. In doing so, it has utilised and linked relevant literature from statistics (survey statistics in particular) and econometrics. The paper is organised as follows. In Section 2 we present quantile-type regression models. We start by describing regression models for a location parameter at the centre of a distribution before showing how these ideas can be extended to modelling location parameters for other parts of a distribution. The contribution of this section is that it presents a unified framework for understanding how quantile, M-quantile and expectile regression are connected. In Section 3 we review SAE based on unit-level random effects models. Section 4 then presents small area estimation using quantile-type models. We start by defining the concept of quantile-coefficients that are fundamental to defining measures that are alternative to random effects and then describe how these coefficients are used in small area estimation. In this section we also provide a review of relevant literature that covers methods for continuous, binary, count and overdispersed data. Finally, in Section 5 we summarise the key points and describe current research on the topic.

2 Quantile-type models

2.1 Regression using influence functions

In this section we introduce a framework for estimation using influence functions, which forms the basis for quantile-type SAE estimation. Generally speaking, estimating a location parameter θ for the distribution of a random variable y involves minimisation of a loss function $\rho(\cdot)$. Indexing by i the units (for example in the sample data) and by n the sample size, the estimator for this location is, $\hat{\theta}$,

$$\hat{\theta} = \arg\min_{\theta} \left(n^{-1} \sum_{i=1}^{n} \rho(y_i - \theta) \right). \tag{1}$$

For solving (1) it is easier to use a differentiable and convex $\rho(\cdot)$ function, with corresponding influence function defined as $\psi(y;\theta) = \frac{\partial}{\partial\theta}\rho(y;\theta)$. In this case, the estimator is the solution to the following estimating equation,

$$n^{-1} \sum_{i=1}^{n} \psi(y_i - \hat{\theta}) = 0. \tag{2}$$

Common examples include the sample mean which corresponds to setting $\rho(y;\theta) = (y-\theta)^2$ and the sample median which corresponds to setting $\rho(y;\theta) = |y-\theta|$. Hence, defining ρ to be the absolute value loss defines the median of the corresponding distribution whereas defining ρ to be the squared loss defines the mean of the corresponding distribution. An alternative, and popular choice for the influence function that leads to the so-called M-type ('maximum likelihood type') estimator of the location parameter is the Huber influence function (Huber, 1981). This influence function depends on a tuning constant k and is defined by

$$\psi_k(u) = \begin{cases} -k, & \text{if } u \leq -k \\ u, & \text{if } -k < u < k \\ k, & \text{if } u \geq k. \end{cases} \tag{3}$$

The framework we present in this section can be extended to the regression case, which is of greater interest for small area estimation. Define by $\boldsymbol{x_i}$ a vector of covariates for unit i. In the simplest case where we assume a linear model, regression estimators are defined by solving

$$n^{-1} \sum_{i=1}^{n} \psi(y_i - \boldsymbol{x_i^T}\boldsymbol{\beta})\boldsymbol{x_i} = \mathbf{0}, \tag{4}$$

where $\boldsymbol{\beta}$ denotes the regression parameters. Depending on the choice of the loss function and the corresponding influence function we can define different regression models. For example, using the squared loss function in (1) leads to ordinary least squares estimates for the regression parameters. Using the absolute value loss in (1) leads to median regression estimates whereas using the Huber loss function and the corresponding influence function (3) in (1) leads to M-type regression estimates. We return to this point in the next section to further clarify the links between the different regression types.

2.2 Quantile regression and its variants

In this section we extend the results in the previous section to quantile-like parameters and clarify the links between types of quantile-type regression models. The quantiles of a distribution of a random variable y can be viewed as location parameters of an appropriately weighted transformed distribution. The weights are determined depending on the loss/influence function used. In particular, quantile estimates are found by minimising the following loss function where Q_q denotes the quantile of order q,

$$\rho_{Q_q}(y - Q_q) = \left[(1-q)I_{y \leq Q_q} + qI_{y > Q_q}\right] |y - Q_q|, \tag{5}$$

where $q \in (0, 1)$, and I is the indicator function. Using this framework, quantile regression was proposed by Koenker and Bassett (1978). For the linear regression case where $Q_q(\boldsymbol{x_i}) = \boldsymbol{x_i^T \beta_q}$ the quantile regression coefficients, $\boldsymbol{\beta_q}$, are estimated by

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{q}} = \arg\min_{\boldsymbol{\beta_q}} \left(n^{-1} \sum_{i=1}^{n} \rho_{Q_q}(y_i - \boldsymbol{x_i^T \beta_q}) \right). \tag{6}$$

Newey and Powell (1987) proposed the use of a smooth loss function that led to so-called regression expectiles as an alternative to regression quantiles. Expectiles are defined by minimising the following squared loss function, where E_q denotes the corresponding expectile,

$$\rho_{E_q}(y - E_q) = \left[(1-q)I_{y \leq E_q} + qI_{y > E_q}\right] (y - E_q)^2. \tag{7}$$

For the linear regression case where $E_q(\boldsymbol{x_i}) = \boldsymbol{x_i^T \beta_q}$ the expectile regression coefficients, $\boldsymbol{\beta_q}$, are estimated by solving

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{q}} = \arg\min_{\boldsymbol{\beta_q}} \left(n^{-1} \sum_{i=1}^{n} \rho_{E_q}(y_i - \boldsymbol{x_i^T \beta_q}) \right). \tag{8}$$

Just as quantiles are a generalisation of the median, expectiles are a generalisation of the mean. Hence, expectile regression is the L_2 version of quantile regression. Although expectiles do not have the same intuitive interpretation as quantiles, expectiles are easier to estimate and can be useful for prediction purposes as is the case in SAE.

Breckling and Chambers (1988) proposed an alternative approach to quantile-type regression, namely M-quantile regression. M-quantile regression is an extension of M-type regression that was described in the previous section. The regression M-quantile of order q for a random variable y is the value $m_q(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}_q$ satisfying the estimating equation

$$n^{-1} \sum_{i=1}^{n} \psi_{m_q}(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_q) \boldsymbol{x}_i = \boldsymbol{0}, \tag{9}$$

where $\psi_{m_q}(u) = 2\left[(1-q)I_{u \leq 0} + qI_{u>0}\right]\psi(u)$. A common choice for $\psi(u)$ in M-quantile regression is Huber's influence function (3). It is now easy to see that quantile and expectile regression can be defined under a unified framework based on the use of influence functions. In particular, using $\psi(u) = sgn(u)$, leads to quantile regression. Conversely, using $\psi(u) = u$, leads to "expectile" regression and using $\psi(u) = \psi_k(u)$ leads to M-quantile regression.

The Huber influence function is often preferred as it depends on a tuning constant k which provides a balance between robustness and efficiency. It also provides an intuitive middle ground between quantile regression (Koenker and Bassett, 1978) and expectile regression (Newey and Powell, 1987). In particular we obtain the regression expectile by setting $k \to \infty$ and the regression quantile by

setting $k \to 0$. With any finite choice of k, the Huber influence function remains bounded, and so estimation remains outlier robust. Furthermore, the continuity of $\psi_k(u)$ for $k > 0$ guarantees the existence of a unique solution to the M-quantile functional equation for every value of $q \in (0, 1)$. We therefore focus on this definition of ψ from now on. Throughout the remainder of the article the term "M-quantile" will imply a Huber M-quantile with $k > 0$ unless otherwise stated.

3 Small area estimation using random effects models

In this section we review model-based small area estimation using unit-level models before focusing on the use of quantile-type models. We start by assuming that the variable of interest is continuously distributed for example, income which has been the focus of many recent applications. The industry standard for unit-level model-based small area estimation is the approach of Battese, Harter and Fuller (Battese et al., 1988), which assumes a linear mixed effects model also known as the nested error regression model or the random effects model. In this paper we will use these terms interchangeably.

Let y_{ij} denote the value for the i-th unit in area/domain j and assume that we have D domains in total, with sample in each domain. The vector $\boldsymbol{x}_{ij}$ denotes the vector of covariates defining the fixed part of the model, u_j denotes the domain j random effect, assumed to be independently distributed between domains, and ϵ_{ij} denotes the unit-level error. Here we consider the simplest version of a random effects model, namely the random intercepts model. The model is defined as follows,

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_j + \epsilon_{ij}, \tag{10}$$

where it is common to assume that $u_j \sim N(0, \sigma_u^2)$ and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ although other distributional assumptions are also possible. The domain effect u_j can be seen to adjust the intercept in the linear specification to allow the domain conditional mean for y_{ij} to deviate from its population average. As a consequence it makes sense to refer to u_j in (10) as a parameter that characterises group heterogeneity. It is also important to remember that prediction of the random effects u_j uses data from all domains, the estimated fixed effects parameters $\boldsymbol{\beta}$ and the variance components σ_u^2 and σ_ϵ^2.

We now focus on how the mixed effects model is used for SAE. Assume that we are interested in estimating a set of population parameters for each domain of interest $\boldsymbol{\theta} = (\theta_1, ..., \theta_D)$. Most commonly interest is in estimating the domain-specific means of y for example, the average income. In this case in order to produce small area estimates of the mean we need to have access to survey micro-data measuring y and $\boldsymbol{x}_{ij}$ and population average data for the same set of covariates, $\bar{\boldsymbol{X}}_j$. The survey micro-data are used for fitting the mixed effects model and estimating the regression coefficients, the variance components and the random effects. Using the estimated parameters, we can then compute the Empirical Best Linear Unbiased Predictor (EBLUP) of the small area means,

$$\hat{\theta}_j^{EBLUP} = \bar{\boldsymbol{X}}_j'\hat{\boldsymbol{\beta}} + \hat{u}_j. \tag{11}$$

As a brief aside we note that recent interest has been on estimating non-linear indicators for small areas. Applications focus for example, on the estimation of the at risk of poverty rate, the Gini coefficient and the quintile share ratio. A popular estimation method in this case is the Empirical Best Predictor (EBP) (Molina and Rao, 2010). In the case where interest is in estimating non-linear indicators, current methodology requires access to population-level micro-data for the covariates of interest. This is in contrast to the less onerous requirement for access to aggregate level population data when interest is in estimating domain-specific averages. Access to population-level micro-data is challenging due to confidentiality concerns and part of the current research effort is in reducing the

dependency on such data. We return to this point later in this article. An important aspect of SAE is estimating the Mean Squared Error (MSE) of the small area estimates. MSE estimation has been an area of intensive research. Here we refer to the Prasad and Rao (1990) analytic estimator under the unit-level nested error regression model, which is a popular method for MSE estimation. Alternatively, when interest is in estimating more complex parameters than the small area averages, one can use parametric bootstrap methods (see Molina and Rao, 2010).

4 Small area estimation using quantile-type models

Before describing the use of quantile-type models for SAE, we start by defining the so-called quantile-type coefficients. Quantile-type coefficients form the basis of this approach in the sense they attempt to quantify between domain heterogeneity. One of the earliest applications of quantile-type modelling for predictive purposes can be found in Kokic et al. (1997). The authors used M-quantile regression to calculate a performance measure for measuring the productivity of Australian dairy farms. The M-quantile regression model the authors used regressed the farm gross return against covariates predictive of gross return. The authors then defined the performance measure q_i^* for farm i as follows,

$$\hat{m}_{q_i^*,k}(\boldsymbol{x}_i) = y_i.$$

Aragon et al. (2006) applied this idea for computing q_i^* used to identify drug overprescription by doctors in France. The q_i^* values are referred to as M-quantile coefficients or q-scores. These q-scores can be thought of as ordered indices, between 0 and 1, that carry information about the position of the corresponding unit (e.g. farms and doctors in the examples above) on the conditional distribution $f(y_i|\boldsymbol{x}_i)$ that is, after controlling for effect of covariates.

The q-scores can be computed under different versions of quantile-type regression. For example, when the influence function used is the $sgn(u)$ (so M-quantile regression is just quantile regression), this q-score is the order of corresponding quantile of the conditional distribution $f(y_i|\boldsymbol{x}_i)$. It immediately follows that in this case q_i^* is uniformly distributed over $(0, 1)$. A q-score derived from the fitted regression M-quantiles and the use of the Huber influence function also defines an indexing over the interval $(0, 1)$ but not necessarily one with a uniform distribution over this interval. Since most of the developments in the use of quantile-type models for SAE are based on the use of M-quantile regression and the use of the Huber influence function, from now on we assume that the q-scores are computed by using M-quantile regression.

In practice the q-scores are estimated using the sample data. In particular, a grid G of q values for example, $G = (0.001, \ldots, 0.999)$, with a step that defines how many points of G are selected is defined and M-quantile regression models are fitted using each q value in this grid using the sample data. In general, the collection of these fitted regression M-quantile models is referred to as an ensemble M-quantile regression model. Such an ensemble fit allows calculation of a fitted regression M-quantile value $\hat{m}_{q,k}(\boldsymbol{x}_i)$ for each value of q on the grid at each $\boldsymbol{x}_i$. The value of q_i^* can then be found quite simply by selecting the grid value of q such that $\hat{m}_{q,k}(\boldsymbol{x}_i)$ is closest to y_i. In some instances when q is close to 0 or 1 the fit of the M-quantile model may not converge, in which case the grid of q values can be narrowed. In general, the estimation of q_i^* will be affected by how refined the grid of q values is.

Chambers and Tzavidis (2006) exploited the fact that q-scores characterise the heterogeneity in the conditional distribution of y given $\boldsymbol{x}$ in the sample. They argued that if the domain structure explains this heterogeneity, then the q-scores would tend to be more similar within domains than between domains. They then proceeded by "averaging" the unit q-scores to obtain what they call group ("domain-specific") q-scores.

Assuming a linear M-quantile model for a continuous random variable y and the set of covariates $\boldsymbol{x}$, the M-quantile estimator of the small area means is defined as

$$\hat{\theta}_j^{MQ} = \bar{\boldsymbol{X}}'_j \hat{\boldsymbol{\beta}}_{\hat{q}_j^*}, \tag{12}$$

where $\hat{q}_j^*$ is the q-score for domain j computed by averaging the unit-level q-scores that belong to the domain, $\hat{q}_j^* = n_j^{-1} \sum_{i \in s_j} q_i^*$, and $\hat{\boldsymbol{\beta}}_{\hat{q}_j^*}$ is the corresponding estimate of the vector of M-quantile regression parameters.

The domain q-scores play the same role as the domain random effects in the mixed effects model, but without the need to pre-specify the grouping structure. The reason for this is that q-scores are computed at unit level and can be aggregated to any grouping structure of interest without the need to refit the quantile-type regression model. However, we should note also the following points that in our view are important. Firstly, we note that the unit-level q-scores and the regression parameters, $\hat{\boldsymbol{\beta}}_{\hat{q}_j^*}$, are computed by using all the sample data and not only domain-specific data. Secondly, we note that domain-specific predictions are differentiated by the fact that $\hat{q}_j^*$ is used for each domain similarly to the use of a random effect in the mixed effects model. This means that the quantile-based small area estimator is not synthetic. Finally, although the unit-level q-scores are estimated using the entire sample, the domain q-scores are computed by using only the domain-specific q-scores. This is in contrast to the approach used for predicting the random effects under the linear mixed model which is using the entire sample and the shrinkage factor. Hence, we expect that the domain q-scores will be unstable if the domain sample sizes are very small.

A comprehensive treatment of analytic MSE estimation for the M-quantile estimator was presented in Chambers et al. (2011) while Marchetti et al. (2012) studied the use of the non-parametric bootstrap.

Quantile-type SAE methods offer a natural approach to outlier robust estimation. The paper by Chambers and Tzavidis (2006) created renewed interest in outlier-robust SAE that extended beyond the use of quantile-type models with Sinha and Rao (2009) proposing outlier-robust SAE under the unit-level linear mixed model. Chambers and Tzavidis (2006) noted that the plug-in M-quantile estimator of the domain mean can be biased. To correct this problem, Tzavidis et al. (2010) proposed a general framework for robust small-area estimation based on representing a small-area estimator as a functional of a predictor of the small-area cumulative distribution function. The authors use a non-parametric smearing-type estimator of the distribution function, namely the Chambers and Dunstan (1986) estimator. This approach leads to new estimator of the small area mean that includes a third term in (12) which depends on the domain-specific model residuals. This estimator resembles a model-based GREG estimator that aims to trade-off bias for variance. The residual correction term corrects for bias but at the cost of increased the variance depending on the size of the domain-specific sample. Based on this idea, Chambers et al. (2014a) defined a general framework for outlier-robust SAE both under quantile-based models and random effects models. These authors refer to the plug-in SAE estimator as robust-projective whereas the bias-corrected estimator is referred to as robust-predictive. The authors discuss analytic MSE estimation and approaches to controlling the impact of the residual-based correction term by using influence functions and selecting an appropriate tuning constant. Dongmo et al. (2013) proposed a robust predictive SAE estimator under the linear mixed model that uses a global, instead of a local (domain-specific), bias correction term which improves the stability of the estimator. Finally, the methodology in Tzavidis et al. (2010) naturally leads to integrated estimation of small-area means, quantiles and non-linear indicators for example, inequality and income deprivation indicators. However, in this case SAE requires the use of population-level microdata for the covariates as is the case also with the EBP approach under the linear mixed model.

4.1 Quantile-type SAE estimation for discrete outcomes

In the previous section we provided a review of quantile-type SAE estimation when the outcome of interest is continuous. Extending the use of quantile-based SAE methods to discrete outcomes is challenging because defining quantiles, M-quantiles and expectiles in this case is not clear. Chambers et al. (2016) propose the use of M-quantile regression for small area estimation with binary outcomes, discuss different definitions of the M-quantile coefficients and apply the methodology for estimating unemployment rates in UK local authority districts. The authors argue that M-quantiles and the use of a continuous influence function such as the Huber one allows for the unique definition of M-quantiles. This is in contrast to quantile regression under which the definition of quantiles of a binary random variable is not unique. The paper by Chambers et al. (2016) further describes the links between the statistical literature and the econometric literature on binary quantiles (see for example Manski,1985) and with the asymmetric maximum likelihood estimator (see Efron, 1992), which can be viewed as a version of expectiles for discrete random variables.

Tzavidis et al. (2016) proposed an M-quantile small area predictor when the response is a count by extending the ideas in Cantoni and Ronchetti (2001). The proposed small area predictor can be viewed as an outlier robust alternative to the more commonly used empirical plug-in predictor that is based on a Poisson generalised linear mixed model with Gaussian random effects. Finally, Chambers et al. (2014b) proposed the use M-quantile regression for overdispersed count outcomes with applications to disease mapping.

5 Concluding remarks and brief summary of emerging methods

This present article reviews a body of literature that proposes an alternative approach to small area estimation that captures cluster (domain/area) heterogeneity via quantile-type models. This approach is inherently outlier-robust and offers additional flexibility since it does not rely on an a-priori specification of the grouping structure. Methods have been proposed both for continuous and discrete outcomes.

More recently, there has been renewed interest in this literature that aims to extend already existing methods in several directions. To start with, as part of his PhD thesis Dawber (2017) researched the use of the M-quantile approach to SAE with multi-category outcomes. This research complements previous work on binary and count-type outcomes and has important applications for example, in producing small area official statistics for labour market activity.

Another area of recent research activity attempts to combine quantile regression with random effects models (Weidenhammer et al., 2018). This research exploits the well-known link between the Asymmetric Laplace distribution and maximum likelihood estimation for quantile regression to define a quantile mixed effects model (see Geraci and Bottai, 2014), which is then used for small area estimation. One advantage of this approach is that random effects, instead of quantile coefficients, are used for measuring the between domain heterogeneity and hence all sample data are involved in predicting the random effects. Secondly, the approach can be extended to modelling count outcomes using ideas about jittering from the econometric literature (see Machado and Santos Silva, 2005). Finally, a further advantage is that, in theory, by fitting an ensemble of quantile random effects models one can obtain an estimate of the entire distribution function of the data, which can then be used for micro-simulating synthetic populations for deriving small area estimates of any target parameter of interest. This is similar in spirit to the EBP approach of Molina and Rao (2010), and the use of the Chambers-Dunstan estimator by Tzavidis et al. (2010). However, whereas in the latter cases it is guaranteed that we estimate a proper distribution function, in the former case this is not

strictly true. For example, there is nothing to prevent quantile cross-over occurring when fitting the ensemble of a quantile random effects model. One approach to tackling this issue is to impose simple constraints in the fitting process. An alternative, more complex, approach is simultaneous estimation of multiple quantile-random effects models. This is an area of current research. Another area of research, currently at its infancy, focuses on the use of unconditional quantile regression for predictive purposes.

Before concluding this paper we must refer to research in M-quantile regression that is peripheral, albeit important, for small area estimation. Bianchi et al. (2018) studied model specification and selection for M-quantile regression. Among other developments, the authors propose a pseudo$R2$ goodness of fit measure along with likelihood ratio and Wald type tests for model specification. In addition, these authors propose a test for assessing the presence of domain heterogeneity in the data. This is similar in spirit to testing for the presence of significant area effects, which has been the focus on research in the small area literature (see among others Datta et al., 2011). As part of his doctoral research Dawber (2017), studied alternative scale estimators and optimal tuning constants for M-quantile regression. The use of this research for predictive purposes, for example in small area estimation, is an open area for research. Last but not least, a number of researchers are working on developing an *R* package that implements the various quantile-type approaches to small area estimation.

Acknowledgements

Tzavidis would like to acknowledge support from the InGRID 2 infrastructure grant funded via the European Commission Horizon 2020 programme (`http://www.inclusivegrowth.eu`) and the MAKSWELL grant (`https://www.makswell.eu`) also funded via the European Commission Horizon 2020 programme.

References

Aragon, Y., Casanova, S., Chambers, R. and Leconte, E. (2006). Conditional ordering using non-parametric expectiles. *Journal of Official Statistics*, **21**, 617-633.

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

Bianchi, A., Fabrizi, E., Salvati, N. and Tzavidis, N. (2018). Estimation and testing in M-quantile regression with applications to small area estimation. *International Statistical Review*, **86**, 541-570.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, **75**, 761-771.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**, 1022-1030.

Chambers, R. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**, 597-604

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

Chambers, R., Chandra, H. and Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, **37**, 153170.

Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2014a). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B*, **76**, 4769.

Chambers, R., Dreassi, E. and Salvati, N. (2014b). Disease mapping via negative binomial regression M-quantiles. *Statistics in Medicine*, **33**, 4805-4824.

Chambers, R., Salvati, N. and Tzavidis, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *Journal of the Royal Statistical Society: Series A*, **179**, 453-479.

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

Dawber, J. (2017). Advances in M-quantile estimation. *Doctor of Philosophy Thesis, University of Wollongong*

Dongmo, Jiongo, V., Haziza, D. and Duchesne, P. (2013). Controlling the bias of robust small-area estimators. *Biometrika*, **100**, 843858,

Efron, B. (1992). Poisson Overdispersion estimates based on the method of asymmetric maximum likelihood. *Journal of the American Statistical Association*, **87**, 98-107.

Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, **24**, 461-479.

Huber, P. J. (1981). *Robust statistics*. New York: John Wiley & Sons, Inc.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Kokic, P., Chambers, R., Breckling, J. and Beare, S. (1997). A measure of production performance. *Journal of Business & Economic Statistics*, **15**, 445-451.

Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. *In Rao, C.R. and Pfeffermann, D. (Eds.) Handbook of Statistics, Vol. **29B**. Sample Surveys: Inference and Analysis. Elsevier, Amsterdam*, 219249.

Machado, J. and Santos Silva, J. M. C. (2005). Quantiles for counts. *Journal of the American Statistical Association*, **100**, 12261237.

Manski, C. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics*, **27**, 313333.

Marchetti, S., Tzavidis, N. and Pratesi, M. (2012). Non-parametric bootstrap mean squared error estimation for M-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics and Data Analysis*, **56**, 2889-2902.

Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, **38**, 369385.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819-847.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 4068.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, **85**, 163171.

Rao, J. N. K. and Molina, I. (2015). *Small area estimation, Second Edition*. John Wiley & Sons, Inc.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. *Canadian Journal of Statistics*, **37**, 381399.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust estimation of small-area means and quantiles. *Australian & New Zealand Journal of Statistics*, **52**, 167-186.

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E. and Chambers, R. (2016). Robust small area prediction for counts. *Statistical Methods in Medical Research*, **24**, 373-395.

Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T. and 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*, **181**, 927-979.

Weidenhammer, B., Schmid, T., Salvati, N. and Tzavidis, N. (2018). A unit-level quantile nested error regression model for domain prediction with continuous and discrete outcomes. *Discussion Paper, Freie Universität Berlin, available via* `https://refubium.fu-berlin.de/bitstream/handle/fub188/19588/discpaper2016_12.pdf?sequence=1&isAllowed=y`

Book and Software Review

The Unit Problem and Other Current Topics in Business Survey Methodology (2018), edited by Boris Lorenc, Paul A. Smith, Mojca Bavdaž, Gustav Haraldsen, Desislava Nedyalkova, Li-Chun Zhang and Thomas Zimmermann Cambridge Scholars Publishing

Aleksandras Plikusas

Vilnius University, Lithuania

This book contains 19 chapters (articles from different authors) on several topics of business survey methodology. The idea of the volume belongs to the Scientific Committee of the European Establishment Statistics Workshop (EESW17).

In the context of business surveys, defining well the elements of the population as standard units is a challenging process. So, the unit problem (treated in a wide sense) is a major topic covered by the authors. In the preface of the book; it is declared, that "*the unit problem is a major new one*". Probably this is not true. Survey statisticians and practitioners have met such kind of problems long ago, at least, for example, since statistical business registers have been used for business surveys. The coverage of the sampling frame (over-coverage, under-coverage), the changes of unit characteristics (e.g. change of the kind of economic activity) are well known problems for survey methodologists of many countries. Nevertheless, problems considered in the book are important and may have a significant impact on survey quality.

Let us briefly discuss the content of the book by chapters.

Chapter 1. Introduction. The topics of the chapters are introduced, and the relations between the topics are highlighted.

Chapter 2. The unit problem and unit error are described. The sources of the unit problem are that the hierarchy of units can be complex; the actual business structures are difficult to relate to units for sampling and reporting; business populations are dynamic and many others.

Chapter 3 and 4 concern unit problems from the point of view of the statistical business registers of Germany and France. Which unit (kind of activity, legal, enterprise or enterprise group) to choose for which statistical purposes, is an important and complex problem. And so, one cannot expect any definite theoretical answer. For example, the business statistics program in France moved from legal units to the enterprises. A simulation study is presented in chapter 4 to illustrate the impact of the choice of statistical units. When the statistical unit is changed, the distribution of the study variable changes and this could influence the quality of statistics. Since auxiliary variables and variables from other data sources are used for estimation, the data linking problems become of interest.

Similar problems are considered in chapter 5 on integration of the data sets with different unit types. An advanced and interesting set of case studies is presented.

Chapters 6 and 7 consider producing statistics for various territorial domains and enterprise profiling. The European economy tends to be integrated and business statistics should follow the economic picture. One must take into account the territorial problems when an enterprise group operates in several territories or countries.

Chapters 8, 9, 10 and 11 are devoted to the sampling problems in various business surveys of France, Germany and Netherlands.

The sampling design of the French structural business survey is considered in chapter 8. The sampling design of this survey has been changed. Enterprises are selected as sampling units instead of legal units. All legal units within selected enterprise are surveyed. The sampling design is a two-stage cluster sampling. The drawback of cluster designs is the loss of precision and random sample size. The optimization of sample design is also presented.

The alternative sampling designs of the German structural survey in the service sector are presented in chapter 9. Their approach is to avoid take-all strata as much as possible to spread the response burden among the largest enterprises. Calibrated estimators that use different sources of auxiliary information are also considered.

The advanced methodologies presented in chapters 8–11 should attract the interest of the survey statisticians from of the EU and other EU countries.

The response process through electronic questionnaires is investigated in Chapter 12. Results observed may help to improve questionnaires and the reporting process. A similar topic is considered in Chapter 13: How para-data can be used in data collection. It seems to be a topic of the future.

A study on embedded data validation when filling electronic questionnaires is presented in Chapter 14. The embedded data validation refers to the fact that the electronic data collection instrument contains embedded validation rules. If they are violated, the data provider is informed. Which rules to use and how many is a question. The useful result of the study is that an increase in the number of validation rules may not result in the improvement of data quality and may increase the response burden.

Chapter 15. The French structural business survey is considered (see also chapter 8) from the point of view of estimation problems. It is well known that many of the economic variables have a skewed distribution and the situation could be worse when elements with a small inclusion probability have large values of the survey variable. Such values are called influential. In this case, the variance of the expansion estimators is large. Two alternative estimation methods are considered in the chapter. The first one is winsorization when the influential values are changed (winsorized) by smaller ones if they exceed some threshold. The threshold suggested by Kokic and Bell is used. Robust estimators that use winsorized variables are compared by simulation with the estimates based on the conditional bias. Robust estimators using winsorized variables and estimates based on conditional bias are compared. The simulation results show that estimators based on the winsorized variables perform better. It would be interesting to examine calibrated estimators in such situations as well.

Chapters 16–19 are not in the mainstream of the unit problem but they could be of interest for those who are interested in producing price index. Especially when additional data sources such as scanner data and Web scraping are used.

Finally, Chapter 19 presents an overview of data visualization.

To conclude, the book covers a wide range of the topics and it should be of interest for survey statisticians, methodologists, computer specialists working with statistical registers, statistical data linking and for a wider statistical community.

ALGERIA

Reporting: **M.Z. Rahmani**

Modernizing the MICS program in Algeria

Algeria recently conducted the Multiple Indicator Cluster Survey (MICS 6). The survey lasted just over three months in the field, from late December 2018 to early April 2019. Algeria had already conducted four MICS surveys (1995, 2000, 2006 and 2012). The survey was led by the "Population" Directorate of the Ministry of Health, Population and Hospital Reform (MSPRH) with financial and technical support from UNICEF. This was a national household survey with seven large geographic areas being represented. The overall sample was 31,325 households distributed over 1,253 clusters covering the 48 wilayas (provinces) in the country. Collection was done over one month (October to November 2018) by 170 medical and paramedical staff members divided into 44 teams and trained for the task. The survey involved completing five questionnaires (households, women aged 15 to 49, children under 5, children aged 5 to 17, and one questionnaire on water quality). The last questionnaire was a first in Algeria. About 40 female interviewers were trained by subject-matter specialists, under the auspices of UNICEF.

The main innovation was the use of tablets to collect data, a first in Algeria. All the supervisors had a connection key that allowed them to transfer the collected data at any time onto the MSPRH server via the Ministry's intranet. That is how the Central Survey Bureau (BCE), mandated to monitor the MICS, was able to obtain the data collected daily and the team in charge of Data Processing could intervene in real time, directly on the tablets of the survey staff, in the event of technical problems. Similarly, the supervisors received all the updates from the collection program, which they transferred to their interviewers via Bluetooth.

Control tables were reviewed weekly. The data collected were then dispatched to the six computers at the BCE level for second-level control.

The operational phase of the survey has currently begun and will last until September–October 2019.

For any additional information, please contact BCEMICS6@gmail.com.

For any information on data processing, contact Mr. M.Z. Rahmani (mzrahmani@yahoo.fr).

ARGENTINA

Reporting: **Veronica Beritich**

Fourth national survey of risk factors

The preliminary results of the 4th National Survey of Risk Factors (ENFR) were released on April 22, 2019. This survey was conducted by the National Institute of Statistics and Census (INDEC), the Ministry of Health and Social Development of the Nation, and the regional statistical offices.

This fourth edition of the survey presents, for the first time, the results of physical measurements (blood pressure, weight, height and waist circumference) and biochemical (capillary glycemia and total cholesterol) based on a subsample of people throughout the country, following the STEPS standardized design proposed by the World Health Organization (WHO). In this opportunity, digital devices were used in the operation, which allowed monitoring the development of fieldwork in real time and streamlined data processing. The 4th ENFR was implemented in 49,170 homes located in all jurisdictions of the country.

Preliminary results indicate that:

- The prevalence of low physical activity reaches 6 out of 10 individuals.
- Only 6% of the population meets the recommended consumption of at least 5 daily servings of fruits or vegetables.
- 6 out of 10 adults reported as having excess weight (overweight or obesity).
- The prevalence of tobacco consumption continues its downward trend since 2005. In this edition, it reached 22.2% of the population. Additionally, electronic cigarettes consumption was measured for the first time: 1.1% of the population said they consumed it.
- 12.7% of the population self-reported suffering from diabetes or elevated glycemia. It represents a significant increase with respect to the 3rd ENFR.
- When performing the physical measurements, among people who reported as being hypertensive, 6 out of 10 people had high blood pressure levels. Of those who did not self-report as hypertensive, 3 out of 10 obtained records of high blood pressure. Additionally, 30.7% of the individuals registered elevated cholesterol (greater or equal to 200 mg/dl) in the phase of collecting biochemical measurements.
- With regard to road safety, 15.2% of people said they had driven a car, motorcycle or bicycle having drunk alcohol in the last 30 days, which represents a significant increase compared to the 3rd ENFR.

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

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

CANADA

Reporting: **Steven Thomas**

A new approach for disclosure control: Random Tabular Adjustment

The Government of Canada is investing in making more data available to Canadians. Statistics Canada is also investing in this initiative by looking at the way that it assesses and treats the risk of disclosure. Traditionally to protect the confidentiality of economic data, cell suppression has been applied. New in 2019, Statistics Canada has produced tables of results using a perturbation technique called Random Tabular Adjustment (RTA). RTA allows users to apply statistical inference to all cells of a table, all the while keeping the values of individual contributors confidential. The RTA process involves adding random noise to estimates where disclosure risks are apparent.

The motivation to seek an alternative to cell suppression has been strong. The suppression of complementary cells, which are often perfectly safe in their own right, is a major weakness of the traditional approach. Moreover, the traditional assessment of sensitivity (using a PQ rule, say) does not account for the inherent quality of the estimate in terms of its variance. So as to address both of these concerns, RTA formalizes the idea of controlling the risk of precise inference on

individual contributions through the addition of random noise to the survey estimate. More specifically, a random value is drawn from a normal distribution with mean zero and a certain variance. This variance is set as a function of both the variance already inherent in the cell-level estimate, as well as the variance needed to protect each contributor to a given proportion. The adjustment is applied to cells at the most detailed level in the table. Marginal subtotals are then updated to restore additivity.

The main advantage of RTA is that complete tables are released without suppressions. A second advantage is that RTA accounts for the protective effect of the variance of the estimate when assessing the sensitivity of the contributions that are at risk. A third advantage is that RTA does not add random noise if the variance of the estimate is already sufficient to protect the contributors to the cell. A fourth advantage is that RTA is less sensitive to small changes in the contributions. In contrast, using a traditional assessment such as a PQ rule, small changes in the microdata may cause cells to be deemed sensitive, and the cell suppression pattern may change considerably.

In March 2019 Statistics Canada published tables of results of the Survey of Innovation and Business Strategy (SIBS). Using RTA, complete tables of SIBS results without suppressions were published. Importantly, each cell was assigned a letter grade that represented an interval of values of the coefficient of variation of the estimate. For example, an estimate with a coefficient of variation of 18% was assigned the letter ‘D’. A high coefficient of variation may have been due to sampling variance, due to the variance of the random noise that RTA provided, or due to both reasons. Data users cannot therefore infer with high precision the value of any contributor to a cell.

The decision to apply RTA instead of traditional cell suppression will be made on a case by case basis, mindful of the needs and expectations of the data users. Both RTA and the traditional approach were applied to SIBS. After comparing the results, it was determined that RTA enabled the release of more useful data without compromising the confidentiality of each contributed value. Statistics Canada continues to apply and evaluate RTA with an increasing number of its economic statistical programs.

The successful application of the RTA method on SIBS was an important step forward for Statistics Canada as it researches options to increase access to data while still ensuring the confidentiality of sensitive personal information. The use of perturbation techniques for confidentiality will continue to be developed and will be applied on other surveys where appropriate as a contrast to the traditional suppression techniques used in the past.

ESTONIA

Reporting: **Helle Visk**

Correcting the place of residence: the partnership index

Estonia is developing a methodology called the partnership index to correct for biases induced by inaccurate place of residence data in Population Register.

When place of residence data is used to form households and family nuclei, the number of married and cohabiting partners is underestimated, whereas the number of lone parents is unrealistically high. For example, in 2016, the number of lone parents was estimated to exceed 2011 census results by 67%. The differences are mainly caused by family members registering in different dwellings--a common practice in Estonia. The aim of the partnership index is to reunite these families. The key challenge is finding partners registered at different addresses.

The partnership index uses administrative data that links two persons and influences the probability of partnership. Some examples of 'signs of partnership' (SOPs) are marriage, having mutual children, sharing a car, or taking joint house loans. Also, divorce is informative since

partnership is unlikely in divorced couples. Altogether, 17 SOPs from 9 registers were used to create a logistic regression model that finds actual partners among couples that share at least one SOP. The model parameters were estimated on Estonian Social Survey and Estonian Labor Force Survey data.

For external validation, the Comparative Survey of Household and Place of Residence (CSHPR) data from 2018 was used. Using the partnership index, correct partnership status was assigned to 96% couples with SOPs. Of CSHPR couples, 84% were correctly classified as partners, 7% had no SOPs, 9% had SOPs but were misclassified into non-partners. When using index-based partners as basis for family formation, the family nuclei and family status distributions were close to actual families in CSHPR. However, the register data fails to capture youth leaving parental home and their early relationships--this is a challenge we must keep working on.

For more details, please contact Helle Visk helle.visk@stat.ee.

HUNGARY

Reporting: **Judit Szigeti**

Use of online cash register data (OPG) to estimate retail turnover

For the purpose of reducing abuses committed during the use of cash registers, the government of Hungary decided to introduce an online connection of cash registers with the National Tax Authority in October 2014. As such, cash machines involved in the online cash register system send online retail chain sales information to the Hungarian Tax Office. These data contribute to reducing the reporting burden of the retail trade. Since not all retailers are obliged to use an online system, a part of the retail trade needs to be estimated without these data. The new methodology, based on a source with similarities to big data, is expected to reduce reporting burden by 75%.

Observation of retail turnover is currently carried out by an electronic questionnaire. Legal units are data suppliers, while stores are observed units. Due to the new legislation, receiving data on retail sales from the online cash registers generated an obvious demand to change the methodology of retail trade statistics in such a way that exploits the potential in these data. According to the new methodology, the Hungarian Central Statistical Office combines two data sources: the turnover of small (representative) units are determined entirely from the OPG data, while a simplified data collection via questionnaire is still maintained. This is because: (1) there is no other source of data about online orders and home deliveries (they are out of scope of OPG); (2) there are a few companies outside the OPG system; and (3) in some cases (firmware error, other activity, etc.) there is too much of a difference between the two data sources (maintaining only temporarily manner). Combining the two data sources – the administrative data and the current survey data – is necessary for several reasons:

- First, it is obvious that the data coming from the online cash registers cannot be the one and only data source, since the legislation does not cover the whole population of the retail trade statistics.
- Also, we need to track the data coming from the tax office in order to build into information in the estimation.
- Nevertheless, the case when something is going wrong with the reception of administrative data should not be excluded, so we need to establish a balance between response burden and information loss.

We have changed the sampling design according to the above.

- The population is divided into two parts: the full-scope part involves the significant enterprises with all their shops, while the rest of the shops are surveyed by simple random sampling.
- We still keep this division while there will be two types of questionnaires:
    - a “simplified” questionnaire for those enterprises whose shops (all of them) are in the scope of the legislation and do not carry out any other activity;
    - the original “full” questionnaire for the rest of the full-scoped enterprises and for the simple random sample of the part of the population not covered by the legislation.
- Data for the rest of the population (for which the simple random sampling was used before) are transmitted by the Tax Authority.

For more information please contact: Csaba Gilyán, Head of Business Statistics Department, Hungarian Central Statistical Office (Csaba.Gilyan@ksh.hu).

INDIA

Reporting: **Dr. Gayatri Vishwakarma**

Initiatives at the Indian Spinal Injury Centre

The Indian Spinal Injury Centre (ISIC), New Delhi is the most advanced Spine, Orthopedic and Neuromuscular Surgical centre in India with the latest state of the art diagnostics and surgical equipment and a highly qualified team of specialists recognized internationally who have been trained in leading institutes of India and abroad.

ISIC is much more than a hospital. It is also considered a training institute and teaching hospital affiliated to a leading university of the country. One the most coveted programs is the Masters of Prosthetics and Orthotics among many that include the provision of on the job training, thus attracting students from all over the nation.

Besides that ISIC features a full-fledged research department as well as a recently opened Biostatistics division to strengthen research. With the growing population of researchers in India, ISIC aims to provide a platform for researchers, clinicians and statisticians to meet & learn and to share knowledge and experience. The Biostatistics division of ISIC has launched various capability building programs including hands-on workshops and internship program on “Research Methodology” for young researchers. Henceforth the mission of Biostatistics division of ISIC, which is the new initiative, is to promote statistical applications to minimize errors in clinical and medical research. The Biostatistics division also provides career options for students/interns and to enable the professional development of young researchers by organizing training sessions, meetings and conferences relating to statistical techniques used in medical research. The format of these conferences is very interactive and facilitates networking and learning from experts and peers.

NEPAL

Reporting: **Jishnu Mohan Bhattarai**

Domestic tourism consumption survey

The Central Bureau of Statistics is the custodian agency responsible to collect, compile and disseminate the economic statistics such as GDP, GNI, as well as saving, consumption and other

macroeconomic indicators in the country. CBS is going to conduct the Domestic Tourism Consumption Survey in the coming fiscal year. The questionnaire, concepts and definition are prepared according to guideline of the United Nations Recommendation of Tourism Statistics. The survey will be conducted over 12 months so it will cover the seasonal nature of the tourism characteristics in Nepal. The ultimate aim of this survey is to develop the Tourism Satellite Accounts (TSA) in the country.

User satisfaction survey 2018

The Central Bureau of Statistics conducted the User Satisfaction Survey in order to assess the satisfaction level of data users of the Bureau. The survey included 1200 participants. About 90 percent were from capital city Kathmandu and rest were from the district level. Very few (5%) were interested in economic statistics in contrast to 64 percent who were interested in population statistics followed by education literacy (9.4%) and health (4%). Data were collected from users and suppliers of statistics in seven varying domains such as from scholars and academics, researchers and statisticians, specialists in political and civil societies, media and press, government and semi-government organizations, private organizations and households.

NEW ZEALAND

Reporting: **Soon Song and Nancy Wang**

SoLinks – a tool for integrating data and assessing data linkage errors

Data integration is becoming increasingly important at Stats NZ. Linked data in Stats NZ's Integrated Data Infrastructure (IDI) is widely used for academic and policy research. Its role in the production of Official Statistics is also increasing.

To ensure the quality of the linked data is suitable for these uses, we need to be able to assess linkage errors. Two types of errors can occur when data is integrated:

1. two records are linked but do not belong to the same person or unit (false positives),
2. two records are not linked but do belong to the same person or unit (false negatives).

In the IDI, we estimate false positive rates from a clerical review of a sample of links. These clerical reviews require significant resources (in time and people), and their subjectivity can lead to inconsistent results. False negatives are much harder to estimate because they require an assessment of links that have not been made. Because of this we have not routinely assessed false negatives in the IDI, focusing instead on maximising link rates while minimising false positives.

Over the past year we have developed an in-house linking module, called SoLinks (system of links), to automatically estimate false positive and false negative rates. The tool can also be used for data integration because it uses an independent process to find record pairs that were missed (the false negatives) in the initial integration.

SoLinks is applied in SAS, and uses name, sex, date of birth, and address as linking variables in its data linking process. The linking methodology uses a logistic regression model, where the model parameters are based on data from historical IDI false-positive clerical reviews. SoLinks also produces linkage error estimates such as false positive rates, false negative rates, precision, recall, F-measure, and balance indicators.

We are now using SoLinks to estimate linkage error rates in the IDI, saving significant labour costs as there is no need for clerical reviewers. SoLinks also ensures consistency, so we can be sure that linkage error estimates that change over time are not caused by clerical variations.

Additionally, SoLinks also outputs a link-quality indicator at an individual link level. Previously our clerical reviews reported false positive estimates at the overall dataset level only. The implementation of SoLinks means we have the ability to produce link quality indicators for specific groups of interest. Our hope is that, in the future, researchers will be able to use the outputs from this tool to better understand the link quality of their research populations, and create populations that match their preferred link quality.

We will continue enhancing the model parameters in SoLinks by incorporating more record pairs in the training data.

For more information on SoLinks, please contact Nancy.Wang@stats.govt.nz.

UKRAINE

Reporting: **Tetiana Ianevych**

Ukrainian competition of student works "Social Data Analysis 2019"

On April 20, 2019 the third Ukrainian competition of student works "Social Data Analysis 2019" was held at the Taras Shevchenko National University of Kyiv. It was organized by the Faculty of Sociology in cooperation with the KANTAR TNS Company, whose representatives took an active part in the evaluation of the works and provided the participants with data and prizes.

The competition consists of two stages. At the first stage, the jury selects up to ten finalists based on their papers and invites them to present their results in Kyiv. Students must meet one essential requirement. They are to use the data from either the European Social Survey (https://www.europeansocialsurvey.org/) or data from the online monthly social and political tracking survey of the urban population of Ukraine aged 18–55 (TNS). At the second stage the selected students present their papers and answer the questions asked by the jury. The jury consists of experts from the Faculty of Sociology, the KANTAR TNS Company and other research institutes. The scientific supervisors of the finalists cannot be the members of the jury.

Within the past three years, participants have come from Kyiv, Kharkiv, Lviv, Lutsk, Sumy and Odesa and specializing in Sociology, Psychology or Mathematics.

This competition encourages young researchers to rely on real data for their analyses. The data from the European Social Survey helps to study different European countries and compare them based on some sociological indicators. The data provided by KANTAR TNS make it possible to study the current internal Ukrainian situation. The competition brings together the university teachers and the practitioners from private companies.

Our big thanks to the organizers and wish their work to be continued and rewarded!

Conference on Current Trends in Survey Statistics 2019

A satellite conference to the 62th World Statistics Congress, Kuala Lumpur, Malaysia

This conference on the “Current Trends in Survey Statistics” will showcase the recent progress in the broad field of analysis of survey data, by putting a special emphasis on emerging areas dedicated to solving problems posed by the advances in data collection, and computational techniques. It will also investigate the future directions of growth in these areas of interest. A partial list of subtopics discussed would include, Small area estimation, Data confidentiality, Record linkage, and entity resolution, Synthetic data and Statistical disclosure limitation, Big data and survey sampling, Big data in official statistics, Multiple imputation techniques, Computational social science, and digital humanities, longitudinal survey, Poverty mapping, Microsimulation models, Social networks, Survey in the developing world etc. The conference is part of a broader programme on "Statistical Data Integration", to be held in the Institute for Mathematical Science, National University of Singapore, from 5th to 16th August 2019. The broader programme would also include a "Workshop on Statistical Data Integration" to be held from 5th to 8th August 2019. The conference is a satellite to the 62nd ISI World Statistics Congress, to be held in Kuala Lumpur from 18th to 23rd August 2019.

Institute for Mathematical Sciences

The broader programme is partially supported Institute for Mathematical Science, National University of Singapore and is endorsed by the International Association for Survey Statisticians and co-sponsored by the International Chinese Statistical Association.

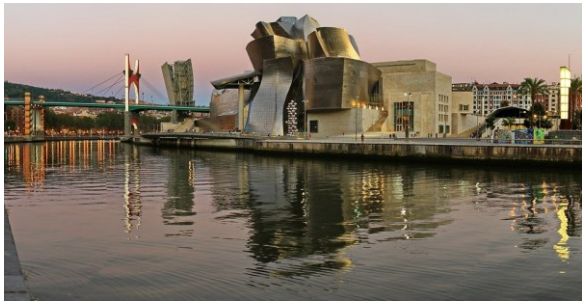
- Organizer: Institute for Mathematical Sciences, National University of Singapore, Singapore
- When: 13–16 August 2019
- Where: Singapore
- E-mail: stasc@nus.edu.sg
- Homepage: https://ims.nus.edu.sg/orgsites/2019data/

International Statistical Institute, 62nd ISI World Statistics Congress

Includes meetings of the Bernoulli Society, the International Association for Statistical Computing, the International Association of Survey Statisticians, the International Association for Official Statistics, the International Association for Statistics Education, the International Society for Business and Industrial Statistics, and The International Environmetrics Society.

- Organiser: International Statistical Institute
- When: 18–23 August 2019
- Where: Kuala Lumpur, Malaysia
- E-mail: secretariat@isi2019.org
- Website: http://www.isi2019.org/

EESW19, 6th biennial European Establishment Statistics Workshop



EESW19, the sixth biennial European Establishment Statistics Workshop, will be held in Bilbao, the Basque Country, Spain, on 24–27 September 2019.

The workshop aims to promote the exchange of results and developments on methodology, practices, approaches and tools in the field of business statistics. The number of participants is limited to 55, with priority given to presenting participants.

Colleagues who are interested to develop and organize a topic for the workshop are invited to contact us. A novelty of this workshop round is that a number of short courses will be given on the 24th September 2019 followed by the 2½-day workshop.

- Organiser: European Network for Better Establishment Statistics
- When: 24–27 September 2019
- Where: Bilbao, Spain
- E-mail: info@enbes.org
- Website: https://statswiki.unece.org/display/ENBES/EESW19

UNECE Statistical Data Collection Workshop "New Sources and New Technologies"

UNECE Statistical Data Collection Workshop "New Sources and New Technologies" will take place at the Palais des Nations in Geneva on 14-16 October 2019.

The objective of this workshop was to identify innovative ways and best practices in statistical data collection, and to provide a platform for practitioners to exchange experiences and foster collaboration in this area. In addition to the more traditional presentations, the agenda of the workshop

included target-driven small group discussions to identify best practices and new opportunities. The target audience for the workshop includes senior and middle-level managers responsible for data collection activities and new data sources, across all statistical domains from Statistical Offices and other agencies from national and international statistical systems.

The programme of the workshop will aim to cover the following areas:

- Promising new technologies and their application
- New data sources and their use
- New technical and people skills needed and how to obtain them
- New management tools and practices needed in the modern data landscape

A non-exhaustive list of examples of suggested topics is provided below.

- Communication with respondents and data providers
- Online/Internet and electronic Data Collection (CAWI)
- The changing data landscape and official statistics. Innovative data collection

The workshop is part of the Conference of European Statisticians' work programme for 2019, within the context of the High-Level Group for the Modernisation of Official statistics.

Participants must register online using the link given in part IV of the Information Notice, by 14 Augustus 2019.

- Organiser: United Nations Economic Commission for Europe
- When: 14–16 October 2019
- Where: Geneva, Switzerland, Palais des Nations in Geneva
- E-mail: taeke.gjaltema@un.org
- Website: https://statswiki.unece.org/display/Collection/2019+Data+Collection+Workshop

2nd Conference on Statistics and Data Science

The purpose of the CSDS 2019 is to bring together researchers and practitioners, from the academy and from the industry, that develop and apply statistical and computational methods for data science. This conference will provide a forum to share and discuss ways to improve the access to knowledge and promote interdisciplinary collaborations. The scientific program will be very appealing for most statisticians and data scientists interested in quantitative methods for decision making and will include plenary talks, invited sessions, short courses, round tables, contributed papers and contributed posters.

- Organizer: Department of Statistics at the Federal University of Bahia, Brazil
- When: 18–20 November 2019
- Where: Salvador, Bahia, Brazil
- E-mail: paulocanas@gmail.com
- Website: http://www.csds2019.ime.ufba.br/

Sampling Methodologies for Monitoring SDG Indicators

The Institute of Statistical Research and Training (ISRT), University of Dhaka is organizing an ISI sponsored workshop on 'Sampling Methodologies for Monitoring SDG Indicators' in Dhaka, Bangladesh. Topics to be discussed include: introduction to SDGs, targets and indicators; the statistical quality assurance framework; role of statisticians in SDG monitoring, data sources and challenges; role of surveys and commonly used sampling designs in monitoring SDG indicators; case studies; new technologies for survey data collection, analysis and visualization. Academicians, students, employees of national statistical offices, government and non-government organizations of SAARC member countries are invited to participate in this event. Travel support, accommodation, meals and workshop kit will be provided.

- Organizer: Institute of Statistical Research and Training (ISRT), University of Dhaka
- When: 17–19 December
- Where: Dhaka, Bangladesh
- E-mail: workshop@isrt.ac.bd
- Website: https://www.isrt.ac.bd/workshop/

Women in Statistics and Data Science Conference

The 2019 Women in Statistics and Data Science Conference in Bellevue, Washington, aims to bring together hundreds of statistical practitioners and data scientists.

WSDS 2019 will highlight the achievements and career interests of women in statistics and data science. Senior, mid-level, and junior stars representing industrial, academic, and government communities will unite to present their life's work and share their perspectives on the role of women in today's statistics and data science fields.

Through formal sessions and informal networking opportunities, the conference will empower and challenge women statisticians and biostatisticians to do the following:

- Share knowledge by offering technical talks about important, modern, and cutting-edge research
- Build community by encouraging discussions establishing fruitful multidisciplinary collaborations, supporting mentoring relationships, and sharing strategies for resolving problems
- Grow influence by providing advice for establishing and sustaining successful careers, showcasing the accomplishments of successful women professionals, and supporting the development of leadership skills

Celebrate your success and *find unique opportunities to grow your influence, your community, and your knowledge*.

- Organizer: American Statistical Association
- When: 3–5 October 2019
- Where: Bellevue, Washington, Hyatt Regency Bellevue on Seattle's Eastside
- E-mail: meetings@amstat.org
- Website: https://ww2.amstat.org/meetings/wsds/2019/

6th International Conference on Establishment Statistics

The Sixth International Conference on Establishment Statistics (ICES VI) will be held in New Orleans, Louisiana, USA, June 15–18, 2020. Continuing in the traditions of ICES -I to ICES -V, ICES VI will explore new areas of establishment statistics, as well as reflect state-of-the-art methodology at the time of the conference.

Participants from all over the world are invited to discuss emerging issues and improved techniques related to business, farm, and institution data.

Topics will include statistical techniques, technologies, and survey methods and feature data from sources such as censuses, sample surveys, and administrative records.

Participation is open to all who are interested in establishment surveys. Whether you're excited by estimation strategies, frame development, questionnaire design, data collection, dissemination, or data visualization, you will find something to like at ICES-VI!

- Organizer: the American Statistical Association
- When: 15–18 June 2020
- Where: New Orleans, LA, USA
- E-mail: asainfo@amstat.org
- Website: http://ww2.amstat.org/meetings/ices/2020
- Website: https://ec.europa.eu/eurostat/cros/content/sixth-international-conference-establishment-statistics-ices-vi_en

Symposium on Data Science & Statistics

- Organizer: American Statistical Association
- When: June 3 – 6, 2020
- Where: The Westin Convention Center, Pittsburgh, Pennsylvania
- Abstract submission: late 2019
- E-mail: meetings@amstat.org
- Website: https://ww2.amstat.org/meetings/sdss/2020/

In Other Journals

Journal of Survey Statistics and Methodology

Volume 7, Issue 1, March 2019

https://academic.oup.com/jssam/issue/7/1

Survey Statistics

Bayesian Nonparametric Functional Mixture Estimation for Time-Series Data, With Application to Estimation of State Employment Totals
Terrance D Savitsky

Survey Methodology

The Effects of Mismatches between Survey Question Stems and Response Options on Data Quality and Responses
Jolene D Smyth; Kristen Olson

The Construction, Maintenance, and Enhancement of Address-Based Sampling Frames
Ned English; Timothy Kennel; Trent Buskirk; Rachel Harter

Simultaneous Estimation of Multiple Sources of Error in a Smartphone-Based Survey
Christopher Antoun; Frederick G Conrad; Mick P Couper; Brady T West

Methods for Exploratory Assessment of Consent-to-Link in a Household Survey
Daniel Yang; Scott Fricker; John Eltinge

Volume 7, Issue 2, June 2019

https://academic.oup.com/jssam/issue/7/2

Survey Statistics

Quantile Regression Analysis of Survey Data Under Informative Sampling
Sixia Chen; Yan Daniel Zhao

Data Fusion for Correcting Measurement Errors
Tracy Schifeling; Jerome P Reiter; Maria Deyoreo

Survey Methodology

Survey Context Effects and Implications for Validity: Measuring Political Discussion Frequency in Survey Research
Mark Boukes; Alyssa C Morey

Evaluating the Utility of Indirectly Linked Federal Administrative Records for Nonresponse Bias Adjustment
Joseph W Sakshaug; Manfred Antoni

The Impact of Interviewer Effects on Regression Coefficients
Micha Fischer; Brady T West; Michael R Elliott; Frauke Kreuter

The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews
Kristen Olson; Jolene D Smyth; Amanda Ganshert

Volume 45, Number 1 (Special issue 2019)

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2019001-eng.htm

This issue of the journal Survey Methodology is a special collaboration with the International Statistical Review in honour of Prof. J.N.K. Rao's contributions.

My chancy life as a Statistician
J.N.K. Rao

Bayesian small area demography
Junni L. Zhang, John Bryant and Kirsten Nissen

Small area estimation of survey weighted counts under aggregated level spatial model
Hukum Chandra, Ray Chambers and Nicola Salvati

Measurement error in small area estimation: Functional versus structural versus naïve models
William R. Bell, Hee Cheol Chung, Gauri S. Datta and Carolina Franco

Small area quantile estimation via spline regression and empirical likelihood
Zhanshou Chen, Jiahua Chen and Qiong Zhang

Development of a small area estimation system at Statistics Canada
Michel A. Hidiroglou, Jean-François Beaumont and Wesley Yung

Weighted censored quantile regression
Chithran Vasudevan, Asokan Mulayath Variyath and Zhaozhi Fan

Empirical likelihood inference for missing survey data under unequal probability sampling
Song Cai and J.N.K. Rao

Improved Horvitz-Thompson estimator in survey sampling
Xianpeng Zong, Rong Zhu and Guohua Zou

Volume 45, Number 2 (June 2019)

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2019002-eng.htm

Volume 35: Issue 1 (Mar 2019)

Volume 35: Issue 2 (Jun 2019)

https://content.sciendo.com/view/journals/jos/35/2/jos.35.issue-2.xml

Remarks on Geo-Logarithmic Price Indices
Jacek Białek

Prospects for Protecting Business Microdata when Releasing Population Totals via a Remote Server
James Chipperfield, John Newman, Gwenda Thompson, Yue Ma and Yan-Xia Lin

Enhancing Survey Quality: Continuous Data Processing Systems
Karl Dinkelmann, Peter Granda and Michael Shove

Measuring Trust in Medical Researchers: Adding Insights from Cognitive Interviews to Examine Agree-Disagree and Construct-Specific Survey Questions
Jennifer Dykema, Dana Garbarski, Ian F. Wall and Dorothy Farrar Edwards

Item Response Rates for Composite Variables
Jonathan Eggleston

Validation of Two Federal Health Insurance Survey Modules After Affordable Care Act Implementation
Joanne Pascale, Angela Fertig and Kathleen Call

Decomposing Multilateral Price Indexes into the Contributions of Individual Commodities
Michael Webster and Rory C. Tarnow-Mordi

Survey Practice

Volume 12, Issue 1 (2019)

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

Impacts of Implementing an Automatic Advancement Feature in Mobile and Web Surveys
Stacey Giroux, Kevin Tharp, Derek Wietelman

Text Mining in Survey Data
Christine P. Chai

Estimation of Survey Cost Parameters Using Paradata
James Wagner

Geographic Inaccuracy of Cellphone Samples
Stephanie Marken, Manas Chattopadhyay, Anna Cahn

Willingness of Online Respondents to Participate in Alternative Modes of Data Collection
Joris Mulder, Marika de Bruijne

Two-Year Follow-up of a Sequential Mixed-Mode Experiment in the U.S. National Monitoring the Future Study
Megan E. Patrick, Mick P. Couper, Bohyun J. Jang, Virginia Laetz, John E. Schulenberg,Lloyd D. Johnston, Jerald Bachman, Patrick M. O'Malley

Sample and Respondent Provided County Comparisons Among Cellular Respondents Using Rate Center Assignments
Carol Pierannunzi, Ashley Hyon, Jeff Bareham, Machell Town

Vol 13 No 1 (2019)

https://ojs.ub.uni-konstanz.de/srm/issue/view/140

Willingness to use mobile technologies for data collection in a probability household panel
Alexander Wenz, Annette Jäckle, Mick P. Couper

Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases
Annette Jäckle, Jonathan Burton, Mick P. Couper, Carli Lessof

Respondent burden in a Mobile App: evidence from a shopping receipt scanning study.
Brendan Read

Tree-based Machine Learning Methods for Survey Research
Christoph Kern, Thomas Klausch, Frauke Kreuter

A Partially Successful Attempt to Integrate a Web-Recruited Cohort into an Address-Based Sample
Phillip S Kott

Hiding Sensitive Topics by Design? An Experiment on the Reduction of Social Desirability Bias in Factorial Surveys
Sandra Walzenbach

Exploring New Statistical Frontiers at the Intersection of Survey Science and Big Data: Convergence at "BigSurv18"
Craig A. Hill, Paul Biemer, Trent Buskirk, Mario Callegaro, Ana Lucía Córdova Cazar, Adam Eck, Lilli Japec, Antje Kirchner, Stas Kolenikov, Lars Lyberg, Patrick Sturgis

Volume 35, issue 1 (2019)

https://content.iospress.com/journals/statistical-journal-of-the-iaos/35/1

Improving health data for indigenous populations: The international group for indigenous health measurement
Chino, Michelle; Ring, Ian; Pulver, Lisa Jackson; Waldon, John; King, Malcolm

Indigenous identification: Past, present and a possible future
Madden, Richard; Coleman, Clare; Mashford-Pringle, Angela; Connolly, Michele

The identification of the Indigenous population in Brazil's official statistics, with an emphasis on demographic censuses
Santos, Ricardo Ventura; Guimarães, Bruno Nogueira; Simoni, Alessandra Traldi; da Silva, Leandro Okamoto; de Oliveira Antunes, Marta; de Souza Damasco, Fernando; Colman, Rosa Sebastiana; do Amaral Azevedo, Marta Maria

First Nations data sovereignty in Canada
The First Nations Information Governance Centre

Identification in a time of invisibility for American Indians and Alaska Natives in the United States
Connolly, Michele; Gallagher, Mehgan; Hodge, Felicia; Cwik, Mary; O'Keefe, Victoria; Jacobs, Bette; Adler, Amy

The identification of Aboriginal and Torres Strait Islander people in official statistics and other data: Critical issues of international significance
Griffiths, Kalinda; Coleman, Clare; Al-Yaman, Fadwa; Cunningham, Joan; Garvey, Gail; Whop, Lisa; Pulver, Lisa Jackson; Ring, Ian; Madden, Richard

Identification of indigenous people in Aotearoa-New Zealand-Ngā mata o taku whenua
Waldon, John

Reflecting back to move forward with suicide behavior estimation for First Nations in Canada
Elias, Brenda

Rethinking health services measurement for Indigenous populations
Mashford-Pringle, Angela; Ring, Ian; Al-Yaman, Fadwa; Waldon, John; Chino, Michelle

Indigenous identity: Summary and future directions
Jacobs, Bette

Volume 35, issue 2 (2019)

https://content.iospress.com/journals/statistical-journal-of-the-iaos/35/2

Beyond code of practice: New quality challenges in official statistics
Sæbø, Hans Viggo; Holmberg, Anders

Quality measures for multisource statistics
de Waal, Ton; van Delden, Arnout; Scholtus, Sander

Automatically generated quality control tables and quality improvement programs
Nguyen, Justin D.; Hogue, Carma R.

Accuracy in contact information for website registrations
Pedlow, Steven; Lickfett, John; Mulrow, Ed; Jamnejad, Cyrus; Erwin, Jared

Development of a complex approach for evaluation of statistical data
Jesiļevska, Svetlana; Šķiltere, Daina

How low response among Latino immigrants will lead to differential undercount if the United States' 2020 census includes a question on sensitive citizenship
Kissam, Edward

An index-based approach to determine partnership in a register-based census
Visk, Helle

Adjusting for linkage errors to analyse coverage of the administrative population
Choi, Hochang

Assessing the quality of life in the European Union: The European Index of Life Satisfaction (EILS)
Maricic, Milica

An estimating parameter of nonparametric regression model based on smoothing techniques
Araveeporn, Autcha

Influence of technologies on the growth rate of GDP from agriculture: A case study of sustaining economic growth of the agriculture sector in Bihar
Sinha, Jitendra Kumar

Integrating the results of a nonresponse follow-up survey into the survey from which its items were selected
Kott, Phillip S.

Determinants of rural household financial literacy: Evidence from south India
Jayanthi, M.; Rau, S.S.

Volume 87, Issue 1, April 2019

https://onlinelibrary.wiley.com/toc/17515823/2019/87/1

Localised Estimates of Dynamics of Multi-dimensional Disadvantage: An Application of the Small Area Estimation Technique Using Australian Survey and Census Data
Bernard Baffour, Hukum Chandra, Arturo Martinez

Semiparametric Regression Analysis of Panel Count Data: A Practical Review
Sy Han Chiou, Chiung-Yu, Huang Gongjun Xu, Jun Yan

Analysing Multivariate Spatial Point Processes with Continuous Marks: A Graphical Modelling Approach
Matthias Eckardt, Jorge Mateu

A Statistical Model to Investigate the Reproducibility Rate Based on Replication Experiments
Francesco Pauli

Distance Metrics and Clustering Methods for Mixed-type Data
Alexander H. Foss, Marianthi Markatou, Bonnie Ray

Why Distinguish Between Statistics and Mathematical Statistics–The Case of Swedish Academia
Peter Guttorp, Georg Lindgren

Extrapolation-based Bandwidth Selectors: A Review and Comparative Study with Discussion on Bivariate Applications
Qing Wang

Confidence Intervals for the Area Under the Receiver Operating Characteristic Curve in the Presence of Ignorable Missing Data
Hunyong Cho, Gregory J. Matthews, Ofer Harel

Volume 87, Issue S1, May 2019

https://onlinelibrary.wiley.com/toc/17515823/2019/87/S1

Special Issue: Contemporary Theory and Practice of Survey Sampling: A Celebration of Research Contributions of J. N. K. Rao

My Chancy Life as a Statistician
J. N. K. Rao

Developments in Survey Research over the Past 60 Years: A Personal Perspective
Graham Kalton

Estimation of Randomisation Mean Square Error in Small Area Estimation
Danny Pfeffermann, Dano Ben-Hur

Modelling Group Heterogeneity for Small Area Estimation Using M-Quantiles
James Dawber, Raymond Chambers

Analysis of Categorical Data for Complex Surveys
Chris Skinner

Combining Data from New and Traditional Sources in Population Surveys
Mary E. Thompson

Some Variants of Constrained Estimation in Finite Population Sampling
Malay Ghosh, Rebecca C. Steorts

Bayesian Analysis of a Sensitive Proportion for a Small Area
Balgobin Nandram. Yuan Yu

Model-Assisted Regression Estimators for Longitudinal Data with Nonignorable Dropout
Lei Wang, Cuicui Qi, Jun Shao

Statistical Analysis with Linked Data
Ying Han, Partha Lahiri

Robust Hierarchical Bayes Small Area Estimation for the Nested Error Linear Regression Model
Adrijo Chakraborty Gauri Sankar Datta Abhyuday Mandal

Sampling Techniques for Big Data Analysis
Jae Kwang Kim, Zhonglei Wang

Recent Developments in Dealing with Item Non-response in Surveys: A Critical Review
Sixia Chen, David Haziza

Small Area Quantile Estimation
Jiahua Chen, Yukun Liu

Some Theoretical and Practical Aspects of Empirical Likelihood Methods for Complex Surveys
Puying Zhao, Changbao Wu

TRANSACTIONS ON DATA PRIVACY
Foundations and Technologies
http://www.tdp.cat

Volume 12, Issue 1, April 2019

http://www.tdp.cat/issues16/vol12n01.php

Bootstrap Differential Privacy
Christine M. O'Keefe, Anne-Sophie Charest

Privacy in Multiple On-line Social Networks – Re-identification and Predictability
David F. Nettleton, Vladimir Estivill-Castro, Julián Salas

Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data
Jingchen Hu

Volume 182, Issue 2, February 2019

https://rss.onlinelibrary.wiley.com/toc/1467985x/2019/182/2

Visualizing spatiotemporal models with virtual reality: from fully immersive environments to applications in stereoscopic view
Stefano Castruccio, Marc G. Genton, Ying Sun

Visualization in Bayesian workflow
Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, Andrew Gelman

Graphics for uncertainty
Adrian W. Bowman

The predictive power of subjective probabilities: probabilistic and deterministic polling in the Dutch 2017 election
Jochem de Bresser, Arthur van Soest

Polling bias and undecided voter allocations: US presidential elections, 2004–2016
Joshua J. Bon, Timothy Ballard, Bernard Baffour

A dynamic inhomogeneous latent state model for measuring material deprivation
Francesco Dotto, Alessio Farcomeni, Maria Grazia Pittau, Roberto Zelli

Embedding as a pitfall for survey-based welfare indicators: evidence from an experiment
Clemens Hetschko, Louisa von Reumont, Ronnie Schöb

Modelling preference data with the Wallenius distribution
Clara Grazian, Fabrizio Leisen, Brunero Liseo

Brexit and foreign investment in the UK
Nigel Driffield, Michail Karoglou

Political rhetoric through the lens of non-parametric statistics: are our legislators that different?
Iliyan R. Iliev, Xin Huang, Yulia R. Gel

Worker absenteeism: peer influences, monitoring and job flexibility
Per Johansson, Arizo Karimi, J. Peter Nilsson

Information-anchored sensitivity analysis: theory and application
Suzie Cro, James R. Carpenter, Michael G. Kenward

Multiperil rate making for property insurance using longitudinal data
Lu Yang, Peng Shi

Experimental evaluation of mail questionnaires in a probability sample on victimization
J. Michael Brick, Sharon Lohr

Bayesian forecasting of mortality rates by using latent Gaussian models
Angelos Alexopoulos, Petros Dellaportas, Jonathan J. Forster

Volume 182, Issue 3, June 2019

https://rss.onlinelibrary.wiley.com/toc/1467985x/2019/182/3

A comparison of sample survey measures of earnings of English graduates with administrative data
Jack Britton, Neil Shephard, Anna Vignoles

A comprehensive approach to problems of performance measurement
N. I. Fisher

A Bayesian semiparametric approach for trend–seasonal interaction: an application to migration forecasts
Alice Milivinti, Giacomo Benini

Spillovers from US monetary policy: evidence from a time varying parameter global vector auto-regressive model
Jesús Crespo Cuaresma, Gernot Doppelhofer, Martin Feldkircher, Florian Huber

A scenario analysis of future Hong Kong age and labour force profiles and its implications
Chris J. Lloyd, Raymond Kwok, Paul S. F. Yip

Multivariate stochastic volatility with large and moderate shocks
Marwan Izzeldin, Mike G. Tsionas, Panayotis G. Michaelides

A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data
Jiancang Zhuang, Jorge Mateu

On probability distributions of the operational law of container liner ships
Yunting Song, Nuo Wang

Adaptive design in surveys and clinical trials: similarities, differences and opportunities for cross-fertilization
Michael Rosenblum, Peter Miller, Benjamin Reist, Elizabeth A. Stuart, Michael Thieme, Thomas A. Louis

Bayesian modelling for binary outcomes in the regression discontinuity design
Sara Geneletti, Federico Ricciardi, Aidan G. O'Keeffe, Gianluca Baio

Spatiotemporal auto-regressive model for origin–destination air passenger flows
Keunseo Kim, Vinnam Kim, Heeyoung Kim

Classifying industries into types of relative concentration
Ludwig von Auer, Andranik Stepanyan, Mark Trede

Pollution state modelling for Mexico City
Philip A. White, Alan E. Gelfand, Eliane R. Rodrigues, Guadalupe Tzintzun

Estimating the changing nature of Scotland's health inequalities by using a multivariate spatiotemporal model
Eilidh Jack, Duncan Lee, Nema Dean

Confidence in risk assessments
Jonathan Rougier

Journal of the American Statistical Association

Volume 113, Issue 524 (2018)

https://www.tandfonline.com/toc/uasa20/113/524?nav=tocList

Applications and Case Studies

A Bayesian Variable Selection Approach Yields Improved Detection of Brain Activation from Complex-Valued fMRI
Cheng-Han Yu, Raquel Prado, Hernando Ombao & Daniel Rowe

Placebo Response as a Latent Characteristic: Application to Analysis of Sequential Parallel Comparison Design Studies
Denis Rybin, Robert Lew, Michael J. Pencina, Maurizio Fava & Gheorghe Doros

Polynomial Accelerated Solutions to a Large Gaussian Model for Imaging Biofilms: In Theory and Finite Precision
Albert E. Parker, Betsey Pitts, Lindsey Lorenz & Philip S. Stewart

An Efficient Surrogate Model for Emulation and Physics Extraction of Large Eddy Simulations
Simon Mak, Chih-Li Sung, Xingjian Wang, Shiang-Ting Yeh, Yu-Hung Chang, V. Roshan Joseph, Vigor Yang & C. F. Jeff Wu

Tracking the Impact of Media on Voter Choice in Real Time: A Bayesian Dynamic Joint Model
Bhuvanesh Pareek, Pulak Ghosh, Hugh N. Wilson, Emma K. Macdonald & Paul Baines

Modeling Random Effects Using Global–Local Shrinkage Priors in Small Area Estimation
Xueying Tang, Malay Ghosh, Neung Soo Ha & Joseph Sedransk

Malware Family Discovery Using Reversible Jump MCMC Sampling of Regimes
Alexander D. Bolton & Nicholas A. Heard

To Wait or Not to Wait: Two-Way Functional Hazards Model for Understanding Waiting in Call Centers
Gen Li, Jianhua Z. Huang & Haipeng Shen

Bayesian Semiparametric Mixed Effects Markov Models with Application to Vocalization Syntax
Abhra Sarkar, Jonathan Chabout, Joshua Jones Macopson, Erich D. Jarvis & David B. Dunson

Theory and Methods

Fast Moment Estimation for Generalized Latent Dirichlet Models
Shiwen Zhao, Barbara E. Engelhardt, Sayan Mukherjee & David B. Dunson

Interpretable Dynamic Treatment Regimes
Yichi Zhang, Eric B. Laber, Marie Davidian & Anastasios A. Tsiatis

Efficient Estimation of the Nonparametric Mean and Covariance Functions for Longitudinal and Sparse Functional Data
Ling Zhou, Huazhen Lin & Hua Liang

Probabilities of Concurrent Extremes
Clément Dombry, Mathieu Ribatet & Stilian Stoev

Linear Hypothesis Testing in Dense High-Dimensional Linear Models
Yinchu Zhu & Jelena Bradic

Optimal Penalized Function-on-Function Regression Under a Reproducing Kernel Hilbert Space Framework
Xiaoxiao Sun, Pang Du, Xiao Wang & Ping Ma

Dynamic Modeling of Conditional Quantile Trajectories, With Application to Longitudinal Snippet Data
Matthew Dawson & Hans-Georg Müller

Modeling Tangential Vector Fields on a Sphere
Minjie Fan, Debashis Paul, Thomas C. M. Lee & Tomoko Matsuo

A Nonparametric Graphical Model for Functional Data with Application to Brain Networks Based on fMRI
Bing Li & Eftychia Solea

Bayesian Estimation and Comparison of Moment Condition Models
Siddhartha Chib, Minchul Shin & Anna Simoni

Reconciling Curvature and Importance Sampling Based Procedures for Summarizing Case Influence in Bayesian Models
Zachary M. Thomas, Steven N. MacEachern & Mario Peruggia

Particle EM for Variable Selection
Veronika Ročková

A Massive Data Framework for M-Estimators with Cubic-Rate
Chengchun Shi, Wenbin Lu & Rui Song

Bayesian Approximate Kernel Regression with Variable Selection
Lorin Crawford, Kris C. Wood, Xiang Zhou & Sayan Mukherjee

Over-Dispersed Age-Period-Cohort Models
Jonas Harnau & Bent Nielsen

A Powerful Bayesian Test for Equality of Means in High Dimensions
Roger S. Zoh, Abhra Sarkar, Raymond J. Carroll & Bani K. Mallick

Tractable Bayesian Variable Selection: Beyond Normality
David Rossell & Francisco J. Rubio

Sparse Pairwise Likelihood Estimation for Multivariate Longitudinal Mixed Models
Francis K. C. Hui, Samuel Müller & A. H. Welsh

Post-Selection Inference Following Aggregate Level Hypothesis Testing in Large-Scale Genomic Data
Ruth Heller, Nilanjan Chatterjee, Abba Krieger & Jianxin Shi

Inference Under Covariate-Adaptive Randomization
Federico A. Bugni, Ivan A. Canay & Azeem M. Shaikh

Sparsity Oriented Importance Learning for High-Dimensional Linear Regression
Chenglong Ye, Yi Yang & Yuhong Yang

Diagnostic Checking in Multivariate ARMA Models with Dependent Errors Using Normalized Residual Autocorrelations
Yacouba Boubacar Maïnassara & Bruno Saussereau

Review

Mixtures of *g*-Priors in Generalized Linear Models
Yingbo Li & Merlise A. Clyde

Volume 114, Issue 525 (2019)

https://www.tandfonline.com/toc/uasa20/114/525?nav=tocList

Applications and Case Studies

Penalized Spline of Propensity Methods for Treatment Comparison
Tingting Zhou, Michael R. Elliott & Roderick J. A. Little

Minimum Mean Squared Error Estimation of the Radius of Gyration in Small-Angle X-Ray Scattering Experiments
Cody Alsaker, F. Jay Breidt & Mark J. van der Woerd

Bayesian Hierarchical Varying-Sparsity Regression Models with Application to Cancer Proteogenomics
Yang Ni, Francesco C. Stingo, Min Jin Ha, Rehan Akbani & Veerabhadran Baladandayuthapani

Spatially Dependent Multiple Testing Under Model Misspecification, With Application to Detection of Anthropogenic Influence on Extreme Climate Events
Mark D. Risser, Christopher J. Paciorek & Dáithí A. Stone

Estimating the Malaria Attributable Fever Fraction Accounting for Parasites Being Killed by Fever and Measurement Error
Kwonsang Lee & Dylan S. Small

Survivor-Complier Effects in the Presence of Selection on Treatment, With Application to a Study of Prompt ICU Admission
Edward H. Kennedy, Steve Harris & Luke J. Keele

Capture-Recapture Methods for Data on the Activation of Applications on Mobile Phones
Mamadou Yauck, Louis-Paul Rivest & Greg Rothman

FreSpeD: Frequency-Specific Change-Point Detection in Epileptic Seizure Multi-Channel EEG Data
Anna Louise Schröder & Hernando Ombao

Marginal Bayesian Semiparametric Modeling of Mismeasured Multivariate Interval-Censored Data
Li Li, Alejandro Jara, María José García-Zattera & Timothy E. Hanson

Theory and Methods

Simulation-Based Bias Correction Methods for Complex Models
Stéphane Guerrier, Elise Dupuis-Lozeron, Yanyuan Ma & Maria-Pia Victoria-Feser

Admissibility in Partial Conjunction Testing
Jingshu Wang & Art B. Owen

Changepoint Detection in the Presence of Outliers
Paul Fearnhead & Guillem Rigaill

Bayesian Graphical Regression
Yang Ni, Francesco C. Stingo & Veerabhadran Baladandayuthapani

Matrix Completion with Covariate Information
Xiaojun Mao, Song Xi Chen & Raymond K. W. Wong

Functional Graphical Models
Xinghao Qiao, Shaojun Guo & Gareth M. James

Least Ambiguous Set-Valued Classifiers with Bounded Error Levels
Mauricio Sadinle, Jing Lei & Larry Wasserman

Confidence Sets for Phylogenetic Trees
Amy Willis

Fisher Exact Scanning for Dependency
Li Ma & Jialiang Mao

Weighted NPMLE for the Subdistribution of a Competing Risk
Anna Bellach, Michael R. Kosorok, Ludger Rüschendorf & Jason P. Fine

Robust Variable and Interaction Selection for Logistic Regression and General Index Models
Yang Li & Jun S. Liu

Principal Component Analysis of High-Frequency Data
Yacine Aït-Sahalia & Dacheng Xiu

Decomposing Treatment Effect Variation
Peng Ding, Avi Feller & Luke Miratrix

A Computational Framework for Multivariate Convex Regression and Its Variants
Rahul Mazumder, Arkopal Choudhury, Garud Iyengar & Bodhisattva Sen

Linear Non-Gaussian Component Analysis Via Maximum Likelihood
Benjamin B. Risk, David S. Matteson & David Ruppert

FSEM: Functional Structural Equation Models for Twin Functional Data
S. Luo, R. Song, M. Styner, J. H. Gilmore & H. Zhu

Optimal Estimation of Genetic Relatedness in High-Dimensional Linear Models
Zijian Guo, Wanjie Wang, T. Tony Cai & Hongzhe Li

Censoring Unbiased Regression Trees and Ensembles
Jon Arni Steingrimsson, Liqun Diao & Robert L. Strawderman

Accurate and Efficient P-value Calculation Via Gaussian Approximation: A Novel Monte-Carlo Method
Yaowu Liu & Jun Xie

Information-Based Optimal Subdata Selection for Big Data Linear Regression
HaiYing Wang, Min Yang & John Stufken

Partially Linear Functional Additive Models for Multivariate Functional Data
Raymond K. W. Wong, Yehua Li & Zhengyuan Zhu

Group SLOPE – Adaptive Selection of Groups of Predictors
Damian Brzyski, Alexej Gossmann, Weijie Su & Małgorzata Bogdan

Modeling Spatial Processes with Unknown Extremal Dependence Class
Raphaël Huser & Jennifer L. Wadsworth

Constructing Priors that Penalize the Complexity of Gaussian Random Fields
Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren & Håvard Rue

Adaptive Bayesian Time–Frequency Analysis of Multivariate Time Series
Zeda Li & Robert T. Krafty

Nonparametric Rotations for Sphere-Sphere Regression
Marco Di Marzio, Agnese Panzera & Charles C. Taylor

Volume 114, Issue 526 (2019)

https://www.tandfonline.com/toc/uasa20/114/526?nav=tocList

Applications and Case Studies

Bayesian Semiparametric Functional Mixed Models for Serially Correlated Functional Data, With Application to Glaucoma Data
Wonyul Lee, Michelle F. Miranda, Philip Rausch, Veerabhadran Baladandayuthapani, Massimo Fazio, J. Crawford Downs & Jeffrey S. Morris

Sequential Nonparametric Tests for a Change in Distribution: An Application to Detecting Radiological Anomalies
Oscar Hernan Madrid Padilla, Alex Athey, Alex Reinhart & James G. Scott

Causal Interaction in Factorial Experiments: Application to Conjoint Analysis
Naoki Egami & Kosuke Imai

Bayesian Semiparametric Estimation of Cancer-Specific Age-at-Onset Penetrance with Application to Li-Fraumeni Syndrome
Seung Jun Shin, Ying Yuan, Louise C. Strong, Jasmina Bojadzieva & Wenyi Wang

Multilevel Matrix-Variate Analysis and its Application to Accelerometry-Measured Physical Activity in Clinical Populations
Lei Huang, Jiawei Bai, Andrada Ivanescu, Tamara Harris, Mathew Maurer, Philip Green & Vadim Zipunnikov

Priors for the Long Run
Domenico Giannone, Michele Lenza & Giorgio E. Primiceri

Batch Effects Correction with Unknown Subtypes
Xiangyu Luo & Yingying Wei

Functional Data Analysis of Dynamic PET Data
Yakuan Chen, Jeff Goldsmith & R. Todd Ogden

Fully Bayesian Analysis of RNA-seq Counts for the Detection of Gene Expression Heterosis
Will Landau, Jarad Niemi & Dan Nettleton

Joint Indirect Standardization When Only Marginal Distributions are Observed in the Index Population
Yifei Wang, Daniel J. Tancredi & Diana L. Miglioretti

Theory and Methods

Stochastic Quasi-Likelihood for Case-Control Point Pattern Data
Ganggang Xu, Rasmus Waagepetersen & Yongtao Guan

Nonparametric Causal Effects Based on Incremental Propensity Score Interventions
Edward H. Kennedy

Parameter Estimation and Variable Selection for Big Systems of Linear Ordinary Differential Equations: A Matrix-Based Approach
Leqin Wu, Xing Qiu, Ya-xiang Yuan & Hulin Wu

Communication-Efficient Distributed Statistical Inference
Michael I. Jordan, Jason D. Lee & Yun Yang

Joint Mean and Covariance Estimation with Unreplicated Matrix-Variate Data
Michael Hornstein, Roger Fan, Kerby Shedden & Shuheng Zhou

Excess Optimism: How Biased is the Apparent Error of an Estimator Tuned by SURE?
Ryan J. Tibshirani & Saharon Rosset

On Sensitivity Value of Pair-Matched Observational Studies
Qingyuan Zhao

Graphical Model Selection for Gaussian Conditional Random Fields in the Presence of Latent Variables
Benjamin Frot, Luke Jostins & Gilean McVean

High-Dimensional Posterior Consistency in Bayesian Vector Autoregressive Models
Satyajit Ghosh, Kshitij Khare & George Michailidis

Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models
Alexandre Belloni, Victor Chernozhukov & Kengo Kato

Large Covariance Estimation for Compositional Data Via Composition-Adjusted Thresholding
Yuanpei Cao, Wei Lin & Hongzhe Li

Variance Change Point Detection Under a Smoothly-Changing Mean Trend with Application to Liver Procurement
Zhenguo Gao, Zuofeng Shang, Pang Du & John L. Robertson

Graph-Guided Banding of the Covariance Matrix
Jacob Bien

Bootstrapping High-Frequency Jump Tests
Prosper Dovonon, Sílvia Gonçalves, Ulrich Hounyo & Nour Meddahi

Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization
Shanika L. Wickramasuriya, George Athanasopoulos & Rob J. Hyndman

Interpretable High-Dimensional Inference Via Score Projection with an Application in Neuroimaging
Simon N. Vandekar, Philip T. Reiss & Russell T. Shinohara

Speeding Up MCMC by Efficient Data Subsampling
Matias Quiroz, Robert Kohn, Mattias Villani & Minh-Ngoc Tran

cmenet: A New Method for Bi-Level Variable Selection of Conditional Main Effects
Simon Mak & C. F. Jeff Wu

Statistical Inference in a Directed Network Model with Covariates
Ting Yan, Binyan Jiang, Stephen E. Fienberg & Chenlei Leng

Testing for Trends in High-Dimensional Time Series
Likai Chen & Wei Biao Wu

A Mallows-Type Model Averaging Estimator for the Varying-Coefficient Partially Linear Model
Rong Zhu, Alan T. K. Wan, Xinyu Zhang & Guohua Zou

Probabilistic Community Detection with Unknown Number of Communities
Junxian Geng, Anirban Bhattacharya & Debdeep Pati

A Cautionary Tale on Instrumental Calibration for the Treatment of Nonignorable Unit Nonresponse in Surveys
Éric Lesage, David Haziza & Xavier D'Haultfœuille

Identifying Cointegration by Eigenanalysis
Rongmao Zhang, Peter Robinson & Qiwei Yao

A Generic Sure Independence Screening Procedure
Wenliang Pan, Xueqin Wang, Weinan Xiao & Hongtu Zhu

Inverse Probability Weighted Estimation of Risk Under Representative Interventions in Observational Studies
Jessica G. Young, Roger W. Logan, James M. Robins & Miguel A. Hernán

Welcome New Members!

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

Title	First name	Surname	Country
MR.	Adekunle	Akande	United Kingdom
MS	Tina	Akande	United Kingdom
MR.	Dayachund	Bundhoo	Mauritius
DR.	Snigdhansu	Chatterjee	United States
MS	Lorie	Dudoignon	France
MR.	Jan	Galkowski	United States
PROF. DR.	Wilfried	Grossmann	Austria
DR.	Paul James	Hewson	United Kingdom
DR.	Taylor	Lewis	United States
DR.	Wendy	Martinez	United States
PROF	Balgobin	Nandram	United States
DR.	Olayiwola	Olayiwola	Nigeria
MS	Dixi	Paglinawan-Modoc	Philippines
MR.	Diego Andres	Perez Ruiz	United Kingdom
MR.	Marcelo Trindade	Pitta	Brazil
MR.	Egi	Prayogi	Indonesia
MRS	Thapelo	Sediadie	Botswana
PROF	Ademola Adeoye	Sodipo	Nigeria
DR.	Boubacar	Sow	Senegal
PROF	Changbao	Wu	Canada

IASS Executive Committee Members

Executive officers (2017 – 2019)

President:	Peter Lynn (UK)	plynn@essex.ac.uk
President-elect:	Denise Silva (Brazil)	denisebritz@gmail.com
Vice-Presidents:		
Scientific Secretary:	Risto Lehtonen (Finland)	risto.lehtonen@helsinki.fi
VP Finance:	Jean Opsomer (USA)	jean.opsomer@colostate.edu
Chair of the Cochran-Hansen Prize Committee and IASS representative on the ISI Awards Committee:	Anders Holmberg, (Norway/Sweden)	anders.holmberg@ssb.no
IASS representative on the 2019 World Statistics Congress Scientific Programme Committee:	Cynthia Clark (USA)	czfclark@cox.net
Ex Officio Member:	Ada van Krimpen	an.vankrimpen@cbs.nl

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

Institutional Members

International organisations:

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

National statistical offices:

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

Private companies:

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

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

CHANGE OF ADDRESS FORM

If your home or business address has changed, please copy, complete, and mail this form to:

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

Name: Mr./Mrs./Miss/Ms. _____ First name: _____

E-mail address *(please just indicate one)*: _____
May we list your e-mail address on the IASS web site?
Yes ☐ No ☐

Home address
Street: _____
City: _____
State/Province: _____ Zip/Postal code: _____
Country: _____
Telephone number: _____
Fax number: _____

Business address
Company: _____
Street: _____
City: _____
State/Province: _____ Zip/Postal code: _____
Country: _____
Telephone number and extension: _____
Fax number: _____

Please specify address to which your IASS correspondence should be sent:
Home ☐ Business ☐

Read the Survey Statistician online!

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