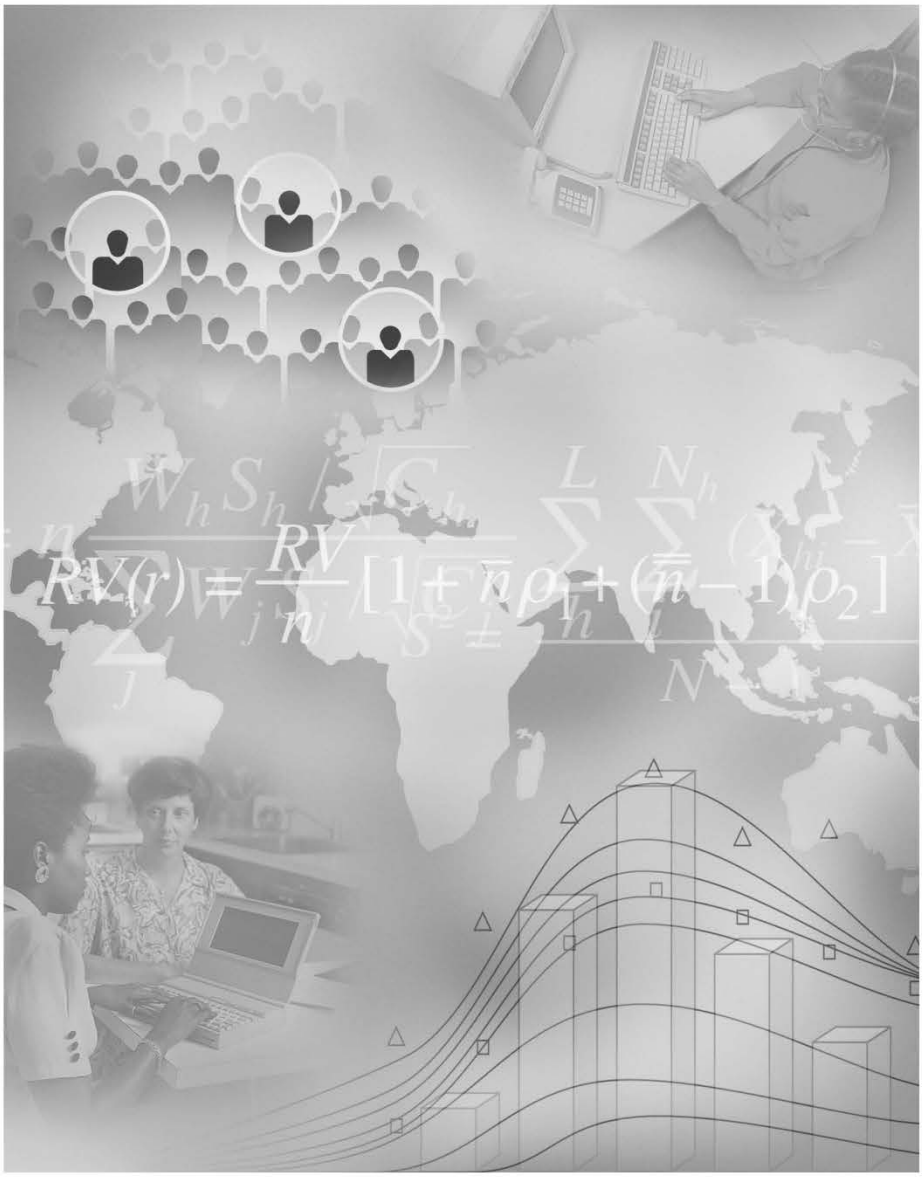
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

No. 72 **July 2015**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

International Statistical Institute

ASSOCIATION INTERNATIONALE DES STATISTICIENS D'ENQUETES

Institut International de Statistique

Editors
Natalie Shlomo and Eric Rancourt

Section Editors

Pierre Lavallée	- Country Reports
Kennon Copeland	- Ask the Experts
Mick Couper	- New and Emerging Methods
Natalie Shlomo	- Book and Software Review

Circulation/Production
Carole Jean-Marie
Henry Chiem
Courtney Williamson
Olivier Dupriez

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.cbs.nl/iass/alluk.htm.

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, Natalie Shlomo (natalie.shlomo@machester.ac.uk) or Eric Rancourt (eric.rancourt@statcan.gc.ca).

ISSN 2521-991X

In This Issue

Letter from the Editors

The July 2015 issue contains articles of interest and important information regarding upcoming conferences, journal contents, updates from the IASS Executive and more. We hope you enjoy this issue, and we would be happy to receive your feedback and comments on how we can make improvements.

In the *New and Emerging Methods* Section (edited by the Scientific Secretary Mick Couper), Trent D. Buskirk from the Marketing Systems Group in the US, has contributed an article titled: The Rise of Mobile Devices: From Smartphones to Smart Surveys. In the article, Trent addresses the rise in mobile internet activity, best practices for designing Smartphone Surveys and the survey question format. Trent concludes with recommended resources for Smartphone Surveys and future challenges.

In the Ask the Experts Section (edited by Ken Copeland), Constance F. Citro, Director, Committee on National Statistics, U.S. National Academies of Sciences, Engineering, and Medicine, has provided a response to the question: What are the threats to the probability survey paradigm and what directions should statistical agencies take to enhance the utility and cost-effectiveness of surveys? In the conclusion, Connie stresses the emerging opportunities to enhance survey data using a multiple data sources paradigm for improving information to the betterment of society.

For the *Book and Software Review* Section, Siu-Ming Tam from the Australian Bureau of Statistics contributes a review of the recent AAPOR Task Force Report on Big Data (Japec *et al.*, 2015):
https://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf
In particular, Siu-Ming provides insight and describes research carried out at the ABS on the inclusion of Big Data in a statistical framework for use in official statistics.

This will be Mick Couper's final issue of the IASS *The Survey Statistician* in his role as editor of the New and Emerging Methods Section, and we wish to thank him very much for the innovative and interesting articles that he has selected in the last two years. The next IASS Scientific Secretary, Denise Silva, will be taking over the editorship of this section and we welcome her to the editorial team. Please let Denise Silva (denise.silva@ibge.gov.br) know if you would like to contribute an article to the New and Emerging Methods Section. If you have any questions which you would like to be answered by an expert, please send them to Ken Copeland (copeland-kennon@norc.org). If you are interested in writing a book or software review, please get in touch with Natalie Shlomo (Natalie.Shlomo@manchester.ac.uk).

The *Country Report* Section has always been a central feature of the IASS *The Survey Statistician* and we thank all the country representatives for their contribution and coordination of the reports. We also thank the editor of the section, Pierre

Lavallee (Pierre.Lavallee@statcan.gc.ca) for his continuing efforts to obtain timely reports from the different countries. We ask all country representatives to please share information on your country's current activities, applications, research and developments in survey methods.

We have added a new section to the newsletter *Contributions from IASS Members* in which we have placed an article written by Seppo Laaksonen from the University of Helsinki titled 'Sampling Design Data File'. If you would like to contribute brief articles or editorials to this new section, please send them directly to the editors of the newsletter, Eric Rancourt and Natalie Shlomo.

This issue of *The Survey Statistician* includes the final letter and updates from our current IASS President, Danny Pfeffermann, and we welcome the incoming IASS President Steven Heeringa as well as congratulate the President-Elect Peter Lynn. The President's Letter includes an overview of achievements over the last two years as well as challenges facing the IASS, in particular the aging of our membership. On behalf of the entire IASS membership, we thank Danny for his hard work over the last two years. In addition, the out-going Scientific Secretary Mick Couper has provided updates in his final letter to the newsletter, including the impressive list of the IASS invited sessions at the 60th WSC in Rio as well as IASS sponsored conferences.

In the *News and Announcement* section we have a report on the work carried out by Geoff Lee (geoff.lee99@bigpond.com) (Chair) and Steve Heeringa, Daniela Cocchi and Natalie Shlomo on developing a proposal for the IASS Strategic Plan as first mentioned in the previous newsletter. In addition, we have a report from the Chair of the Cochran-Hansen Prize, Risto Lehtonen, on the submitted papers to the competition and the results.

We thank Marcel Vieira for putting together the list of conferences for inclusion in the newsletter. Please send to Marcel (Marcel.Vieira@ice.ufjf.br) any conference announcements that you would like advertised in the next *Survey Statistician* to be issued in July 2015. We also thank Carole Jean-Marie from Statistics Canada for collating the advertisements of upcoming conferences and for preparing the tables of contents in the In Other Journals section. This is a very time-consuming and detailed task but the information she gathers is deeply appreciated by IASS members. We also thank Carole for her hard work in collating all the articles into this substantial newsletter that you see before you.

Please take an active role in supporting the IASS newsletter by volunteering to contribute articles, book/software reviews and country reports and/or by making it known to friends and colleagues. We also ask IASS members to send in notifications about conferences and other important news items about their organizations or individual members.

The Survey Statistician is available for downloading from the IASS website at http://isi.cbs.nl/iass/allUK.htm.

Eric Rancourt Eric.Rancourt@statcan.gc.ca

Natalie Shlomo Natalie.Shlomo@manchester.ac.uk

Letter from the President

Dear all,

I am ending my term as IASS President sometime during the WSC in Rio, so this is my last letter to you in this capacity. It was a pleasure serving the IASS and I just wish that I had more time to spend on IASS matters, but this was simply impossible given the few other jobs that I have. I take this opportunity to wish my successor, Steve Heeringa and the continuing and newly elected council members all the very best in leading the IASS to new achievements. I shall be happy to help, as much as I can, if asked for.

So, here is a list of what we were able to achieve during my term. I emphasize "we", because every single task or decision was a team work. I know that I have already mentioned most of these achievements in my past letters and emails, but I thought that it would be nice to summarize them in this last letter.

1- We elected a new president elect, two new vice presidents, a new scientific secretary and six new council members. I sent the names of the new elected officers in a separate email and you can find them also in the 'news and announcements' section of this newsletter, so let me just wish them all again the best of success in their respective roles. Special thanks are due to the members of the nomination committee and of course, to the officers who finished their present duties and council members who continue for another term. Being in the mood of offering my good wishes, it is a great pleasure to congratulate our very active colleague, Pedro Silva, on the occasion of starting his term as President of the ISI. I have not checked it out, but I presume that Pedro is amongst the youngest (if not the youngest) ISI presidents. What an honor and how well deserved.

2- We have a new active, informative and engaging website, which I hope that you visit from time to time. Our thanks are due to Olivier Dupriez, who did it all by himself. We didn't manage to translate the website to other languages, but hopefully this can be done in the future.

3- Following the initial IASS strategic plan that we submitted to the ISI in November of last year, we established a working group composed of Geoff Lee (head), Daniela Cocchi, Natalie Shlomo and Steve Heeringa, with the task of preparing a more comprehensive strategic plan that should take us to a better future. As reported by Geoff in the News and Announcement Section of this newsletter, the working group has completed preparing a draft plan. We shall discuss this plan at the council meeting during the WSC in Rio, and then you will all have the opportunity to express your opinion and propose amendments. I suspect that part of the discussion will focus on the use of "big data" for inference on finite populations, and whether we should change our titles from Survey Statisticians to the more juicy title of Survey Data Analysts (and change the acronym IASS to IASDA).

4- As in previous years, we continued the policy of supporting 2-4 relevant conferences every year, but added two conditions to our support: concession in registration fees for IASS members, and having a special booth with IASS material during the course of the conference, so as to hopefully attract new members. See the report by the Scientific Secretary Mick Couper in this newsletter for the list of

conferences that we have supported in the years 2014-2015 (and already for 2016).

5- With support from the ISI and the World Bank, Pedro Silva and Marcel Vieira gave two courses in Africa: “Analysis of Complex Survey Data Using R” (in Mozambique), and “Analysis of Complex Health Survey Data Using Stata” (in South Africa). Both courses were well attended. I very much hope that we can offer more courses in other countries in the future, and also engage in consulting in developing countries.

6- We supported the participation of four young IASS members from developing and transition countries at the 2014 International Methodology Symposium in Gatineau, Canada, again with help from the ISI and the World Bank. We thank Statistics Canada for waiving the registration fees for the symposium and the cost of a short course for the four members.

7- We continued the tradition of awarding the Cochran-Hansen prize for the best paper on survey research methods submitted by a young statistician from a developing or transition country. This year we had no less than 16 applications, with 8 papers accepted for evaluation. Given the high quality of the papers, we decided to award, for the first time, two papers. Congratulations to the lucky awardees, Kevin Carl Santos from the Philippines and Santanu Pramanik from India. Thanks to the members of the selection committee, Risto Lehtonen (head), Jean Opsomer and Marcel Vieira for heir extra hard work in evaluating all the papers. See the report of Risto Lehtonen in the News and Announcement section of this newsletter. Both winners will present their papers during the WSC in Rio and they would like you all to attend.

8- We reached an agreement with the American Association of Public Opinion Research (AAPOR) and the Survey Research Methods Section of the American Statistical Association, which provides a one-year free on-line subscription to the Journal of Survey Statistics and Methodology (JSSAM) to all IASS members. I hope that many of you took advantage of this agreement.

What about the WSC in Rio? We are going to be very active and visible there.

- We succeeded in having **13** IASS invited paper sessions (IPS). All the credit goes to Christine Bycroft and her committee for this wonderful achievement.

- We shall have a new “IASS President's Invited Speaker Session”, with Jon Rao and Wayne Fuller presenting a joint paper on “Sample Surveys, Past, Present and Future Directions”. This session is an ISI initiative and I presume that it will become a tradition.

We shall have another new “Journal Papers Session”. This session will also have two speakers and a discussant: Alastair Scott will present a joint paper with Thomas Lumley on “AIC and BIC for Modelling with Complex Survey Data”, published in *JSSAM*. Jan ven den Brakel will present a paper on “Design-based analysis of factorial designs embedded in probability samples”, published in *Survey Methodology*. Chris Skinner will discuss the two papers.

I cannot finish this letter without my usual warning about the diminishing size of our association. Since all my initiatives simply failed, I started thinking that perhaps we should try and merge with the International Association of Official Statistics (IAOS). I am examining with my colleague, Mr. Tom Caplan the possible pros and cons of such a merge and if we can come up with a positive proposal, perhaps we can discuss it in Rio before submitting it to the IAOS for their consideration. If anyone has some thoughts about this idea, please let me know.

Well, the list of goodies actually turned out to be longer than I expected, so many thanks to Steve Heeringa (President elect), Geoff Lee and Jairo Arrow (Vice Presidents), Mick Couper (Scientific Secretary), and the council members Christine Bycroft, Ka-Lin Chan, Olivier Dupriez, Natalie Shlomo, Marcel Vieira, Alvaro Villalobos, Michael Brick, Daniela Cocchi, Jack Gambino, Risto Lehtonen, Ralf Munnich and Jean Opsomer, for all their big help and support over the last two years. Finally, I would like to thank also Eric Rancourt for replacing Frank Yu as co-editor of *the Survey Statistician*, (Eric is assisted by Carole Jean-Marie so thanks also to Carole), and Ken Copeland for replacing Robert Clark as editor of the section "Ask the Expert" in *the Survey Statistician*. I hope that I didn't miss anyone in this list of people who deserve our thanks.

VIVA IASS

Danny Pfeffermann,
President (outgoing)

Report from the Scientific Secretary

Members of the IASS Council have been busy the past few months, as we gear up for the biennial World Statistics Congress in Rio de Janeiro. Thanks to the leadership and hard work of Christine Bycroft, the IASS Program Chair for Rio, and the active support of many IASS members, IASS organized or co-organizing 13 invited paper sessions:

- Methodologies relating to Big Data applications (organizer: Siu-Ming Tam)
- Statistical Disclosure Control for official statistics in the 21st century (organizer: Gemma van Halderen)
- Adaptive Survey Design (organizer: Barry Schouten)
- Sampling Frame and Nonsampling Error Issues in Internet Surveys (organiser: Öztas Ayhan)
- New developments in use of model-based methods in official statistics (organizer: Paul Smith)
- Small area estimation for business and economic data (organizer: Susana Rubin-Bleuer)
- What is a Census during times of changing methodologies and technologies? (organizer: Arona Pistiner)
- Recent advances in empirical likelihood approaches under complex sampling (organizer: Yves Berger)
- Using remote sensing for agricultural statistics (organizer: Elisabetta Carfagna)
- Estimation and inference methods based on integrated statistical data (organizer: Li-Chun Zhang)
- Statistical implications of changing ILO international standards for employment and unemployment (organizer: Tite Habiyakare)
- Cross national comparability of national statistics (organizer: Ineke Stoop)
- Bayesian Analysis of complex survey data (organizer: Sahar Zangeneh)

IASS member are also participating in WSC in other ways. Short courses are now organized centrally under ISI leadership, but we were represented on the short course committee, and IASS members are teaching several one- and two-day short courses at the WSC.

The World Bank provided some funds to the ISI to support scholars from developing countries to attend the WSC in Rio. The IASS wrote letters of support for 8 of its members who requested such letters. We don't have the final list of awardees yet (selections were made in several rounds), but we know at least a few IASS members are being supported under this program.

A key activity of the IASS is sponsorship of the Cochran-Hansen Prize for the best research paper on survey methods by a young survey statistician from a developing

country. Risto Lehtonen (chair), Jean Opsomer, and Marcel Vieira served as the prize committee. For the first time, two awards were made this year, reflecting the number and quality of submissions. Please see the News and Announcement Section for a report from Risto and the two winning papers. Both awardees will be presenting their papers at the WSC, and will receive their awards at the ISI Awards ceremony.

For those of you who will be at WSC, we look forward to seeing you there, and hope you are able to attend the General Assembly of the IASS, which will be held at lunch time (12:30-14:00) on Wednesday 29 July.

IASS is also co-sponsoring a satellite meeting on Small Area Estimation in Santiago, Chile, following the WSC (see www.encuestas.uc.cl/sae2015). Other recent or upcoming conferences supported by IASS include:

- ITACOSM 2105, the biannual meeting of the Italian section of the Italian Society of Statistics, held in Rome, June 24-26, 2015 (see http://itacosm15.sta.uniroma1.it/).
- The European Establishment Statistics workshop to be held in Poznan, Poland, in September 2015.
- The Total Survey Error Conference (TSE15) to be held in Baltimore, USA, in September 2015 (see http://www.tse15.org).
- The 9th French Colloquium on Survey Sampling, to be held in Gatineau, Quebec, Canada, from October 14-16, 2016.

Supporting regional and international conferences of interest to survey statisticians is one of the key activities of the Association. Given that the World Statistics Congress is only every two years, this is also an important way for IASS members to keep in touch and to encourage new members to join IASS and participate in the Association's activities.

As many of you already know, the elections for IASS Council were held recently. Newly-elected council members of the IASS are listed in the News and Announcement Section. As Danny Pfefferman noted in his message, nearly 83% of IASS members voted. Congratulations to the newly-elected officers and to those continuing on the IASS Council, especially the incoming IASS President Steve Heeringa. Also heartfelt thanks to Danny Pfeffermann for his excellent leadership of the Association for the past two years, and to outgoing members of Council for their many contributions to the Association.

I look forward to meeting many of you at the 60th WSC in Rio de Janeiro in July. If you have ideas on how to increase the membership (both individual and organizational) of the IASS, or how to provide enhanced services for members, or are willing to volunteer or otherwise engage in IASS activities, please contact the incoming IASS President (Steve Heeringa) or Scientific Secretary (Denise Silva).

With thanks to you all,
Mick P. Couper
mcouper@umich.edu

News and Announcements

60th World Statistics Congress – ISI2015
Rio - IASS meetings

IASS General Assembly meeting on Wednesday, July 29, 2015 from 12:30 to 14:00 in Room 208

IASS Incoming Council meeting on Thursday, July 30, 2015 from 7:30 to 9:00 and from 12:30 to 14:00 in Room 3.

IASS Elections

We now have the results of the elections for new officers. We elected a new President-Elect, two new Vice-Presidents and a new Scientific Secretary for the years 2015-2017, and 6 new Council Members for the years 2015-2019. The new officers will start their term after the 2015 WSC in Rio de Janeiro. The election process was open to all IASS members.

On the closing date of May 12, 2015, a total of **366 out of 443 members casted their votes**, which amounts to a response rate of **82.6 %**, a remarkable response rate given that this is a voluntary survey. **Thank you all** for participating in the election process and casting your votes. There were no election surveys taken during the week before the closing date, and no exit polls!! We did it the old way.

The votes were counted by Gerrit J. Stemerdink, an ISI elected member who works as a volunteer at the ISI Permanent Office. I know that by now you are very eager to see the results, so here they are:

Elected for *President-Elect 2015-2017*: **Peter Lynn**

Elected for *Vice-President 2015-2017*: **Monica Pratesi** and **Imbi Traat**

Elected for *Scientific Secretary 2015-2017*:
There were two candidates for this post, the outcome is: **Denise Silva**

Elected for *Council Members 2015-2019*: **Hukum Chandra**, **Maria Giovanna Ranalli**, **Timo Schmid**, **David Steel, Ineke Stoop** and **Nikos Tzavidis**

CONGRATULATIONS to all the new elected officers and many thanks to all those who agreed to stand for election and didn't make it this time.

I would like to take this opportunity to thank the members of the nomination committee; **Jean Opsomer** (Head), **Ray Chambers**, **Mike Brick**, **Jack Gambino**, **Risto Lehtonen** and **Daniela Cocchi**. You did a marvelous job in assembling such a wonderful group of candidates.

Finally, I would like to thank all the outgoing officers, **Geoff Lee**, **Jairo Arrow**, **Mick Couper**, **Christine Bycroft**, **Ka-Lin Chan**, **Olivier Dupriez**, **Natalie Shlomo**, **Marcel Vieira**, and **Alvaro Villalobos**. You all contributed a lot in running the IASS activities and it was a pleasure working with you.

Danny Pfeffermann

IASS is your association - Have your say about its plans for the future!

The January 2015 edition of the Survey Statistician included a draft strategic plan for the IASS, prepared by our President, Prof Danny Pfeffermann with help from Tom Caplan, and some comments from IASS council members. Since that time a working group, consisting of Geoff Lee (head, Australia), Steve Heeringa (USA), Natalie Shlomo (UK) and Daniela Cocchi (Italy) has worked hard further developing the proposal and producing a discussion paper for consideration by the IASS Council at its meeting in Rio de Janeiro, at the World Statistics Congress in July 2015.

We have proceeded in 3 main stages:

- Stage 1 was to conduct an environmental scan, to review what is happening inside the survey statistics field, and what is emerging outside the field that may impact on the future of our association. This was a "clean sheet" exercise, in which we each independently considered a set of questions designed to elicit our insights into what IASS does well, what it is missing or needs to improve, and what is happening elsewhere that might impact on us. (With a sample size of only 4, independence was crucial to obtain as large an effective sample size as possible!)
- Stage 2 involved reviewing the draft proposal prepared earlier in the light of that environmental scan, plus considerable debate amongst ourselves to firm up on the main opportunities and challenges that any plan for the future will need to consider.
- Stage 3 involved seeking ideas and comments from outside our group, and in particular inviting comments from the network of country representatives who contribute to "The Survey Statistician". While the response was not as large as hoped (we received only 5 responses), the quality was very high. There was pretty well universal support for one of the major themes in the draft proposal (education and capacity building, especially in developing countries), plus some very helpful and insightful specific suggestions about how this might be progressed in practice.

While we have reorganised the objectives and strategies a little, the underlying thrust remains much as developed earlier. What we have principally produced are issues for discussion, digging into each objective and strategy, and teasing out the implications should the IASS decide to commit itself to such a path. The IASS council will consider the paper at its meeting during the WSC in Rio de Janeiro in July 2015. We expect that discussion to be detailed and vigorous. Our work was conducted via email, and ambitious plans for the future, such as contained in the draft proposal, really need to be discussed interactively and subjected to searching questions and challenges. Are the objectives clear, unambiguous and appropriate? Will the strategies actually achieve those objectives? Which have the highest priority? Are there any essential conditions or pre-requisites that must be in place before we begin? And of course, do we have the capacity and the resources to tackle the tasks we are about to set ourselves?

These are not easy questions to answer, and after the IASS Council has deliberated and further improved our work, there will be an opportunity (almost a necessity) for the wider membership to consider the objectives and strategies, comment on them, improve them, and indicate whether we as a worldwide association all believe that the IASS should take this path towards the future. When we reach that stage we urge you all to take the opportunity, and express your views about future directions for the IASS.

To help you prepare your thoughts, here is a 1 page overview of our current thinking, before the IASS council meeting. It loses a bit in the translation to fitting on one page, and the discussion points are missing, but it should help give a **flavour** of what is under consideration and the ambitious nature of possible future plans for the IASS. It will no doubt be further refined after the IASS Council discussions.

The discussion paper identifies a need to refresh the overarching aims and objectives of the IASS to bring them up to date, and includes a context statement describing a range of new challenges (a bit too long to include here) and a proposal for a vision statement, along the lines of:

Vision
The vision of the IASS is that by using modern approaches to survey methodology, including advancements in computing technology that have occurred and will occur, the IASS professionals can collaborate in developing and analysing approaches that will be able to address new challenges.

INTERNATIONAL ASSOCIATION of SURVEY STATISTICIANS

Draft Strategy Map

INTENT

The IASS aims to promote the study and development of the theory and practice of sample surveys and censuses, in conjunction with ongoing developments in technology. It also aims to increase the interest in surveys and censuses among statisticians, among governments and the public in general in the different countries of the world.

Existing Activities

Objective A. Research and Development

1 Encourage research in all areas that could benefit, including new uses of technology; use the WSC and other forums for discussion and collaboration.
2 Make use of The Survey Statistician to include articles on new and emerging methods, ask the expert etc.

Membership and Finances

Objective B. Expand Membership

1 Promote an extensive "a member brings a member" membership campaign
2 Advertise TSS more widely and include membership campaign.
3 Concerted effort to bring in new institutional members.

Objective C. Encourage Young Survey Statisticians

1 Foster a wide competition for the Cochran-Hansen (and other new ?) prizes
2 Encourage current academic members to bring students into the IASS and resource their attendance at WSC and other conferences.
3 Run regional conferences, workshops and seminars aimed at university students on the subject of survey statistics.
4 Establish forums for joint survey statistics research projects by partnerships of IASS members and students.

Exploit modern communication media

Objective H. Become a significantly more participative and responsive organisation

1 Host interactive "Stack Exchange"style question and answer forums on the website, actively and responsively supported by "high reputation" IASS members.

Education and Capacity Building

Objective D. Teaching Materials

1. Develop modular courses in survey statistics for adaptation by universities and colleges.
2. Develop online courses in survey statistics, and encourage universities and colleges to adapt them.

Objective E. Promoting Collaboration and Identifying Need

1. <….. to be added….>. Collaborate (with IAOS, …)

Objective F. Deliver Education (Foster Statistical Capacity Building in developing countries).

1 Develop a multi-year program of teaching sample survey statistics and consulting in developing countries.
2. Encourage senior IASS statisticians to take on assignments which would involve teaching these courses and mentoring local statisticians who could then further develop and teach these courses.

Objective G. Applications

1. Expand the membership to include more academic statisticians as well as more applied statisticians in national statistics institutes and other government ministries and the private sector.
2. Encourage the development of collaborative projects by teams composed of both academic and applied statisticians. These projects could be the basis for future sessions at the WSC and other forums.
3. Encourage statistics students to work with applied survey statisticians. The IASS could develop a program of internship to that end. (see also Objective C Strategy 4)

Report on Cochran-Hansen Prize 2015 Competition for Young Survey Statisticians from Developing and Transitional Countries

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. Participation in competition for the 2015 prize was restricted to young statisticians from developing and transition countries who were living in such countries and were born in 1980 or after. The definition of the target countries was based on the list of countries adhered by the International Statistical Institute. The Cochran-Hansen Prize consists of books and journal subscriptions in the value of EUR 500.

A total of 16 papers were submitted for the 2015 competition. Eight papers from seven different countries (Cameroon, India, Iran, Nigeria, Philippines, Turkey and South Africa) were accepted for review by the members of the Cochran-Hansen Prize Committee appointed by the IASS. The committee members were Risto Lehtonen, Jean Opsomer and Marcel de Toledo Vieira.

The reviewed papers were interesting, timely and covered widely the area of survey research methods. Two papers were ranked highest in the independent review by the jury members. The jury decided to award these two best papers. The winners are Santanu Pramanik (Research Scientist, Public Health Foundation of India) and Kevin Carl P. Santos (Assistant Professor, University of the Philippines-Diliman School of Statistics).

The paper entitled "Selection of Prior for the Variance Component and Approximations for Posterior Moments in the Fay-Herriot Model" by Santanu Pramanik is based on his PhD thesis in statistics completed at University of Maryland. The abstract of the paper summarizes the method as follows. "In the Fay-Herriot model, a prior distribution for the variance component allows posterior moments to be approximated with the Laplace method, avoiding computer intensive Monte Carlo Markov chains. The extremely skewed posterior distribution of the variance component results from the asymmetry of the parameter space with variance parameters constrained to be positive. The prior avoids the extreme skewness of the posterior in contrast to the commonly used uniform prior. With this prior, the mean squared error and coverage in the approximate hierarchical Bayes method are satisfactory when used to estimate small area means. Computation time is shorter than with Monte Carlo Markov chains. The approximations give easy interpretations of Bayesian methods and highlight frequentist properties of the parameters".

The paper entitled "Improving Predictive Accuracy of Logistic Regression Model Using Ranked Set Samples" by Kevin Carl P. Santos is based on his M.S thesis in statistics completed at the School of Statistics of the University of Philippines-Diliman, School of Statistics. As summarized in the abstract of the paper:

"Logistic regression is often confronted with separation of likelihood problem and rare events. We propose to address this issue by drawing sample using ranked set sampling (RSS). Simulation studies illustrated the advantage in terms of predictive ability of logistic regression with RSS in small populations regardless of the distribution of the binary responses. As the sample and population sizes increase, the predictive ability of model from RSS also improves but it becomes comparable to fitted models using simple random samples (SRS). Furthermore, RSS mitigates the problem of separation of likelihood especially when the population size is relatively large. Lastly, even in the presence of ranking errors, RSS still yielded higher predictive power than its SRS counterpart."

The prize winners were invited to present their papers at the 2015 World Statistics Congress of the ISI. The IASS congratulates the winners. The IASS wants to thank all authors who submitted a paper to the competition.

Risto Lehtonen
Chair of Prize Committee

Ask the Experts

What are the threats to the probability survey paradigm and what directions should statistical agencies take to enhance the utility and cost-effectiveness of surveys?

Constance F. Citro, Director, Committee on National Statistics, U.S. National Academies of Sciences, Engineering, and Medicine

1. Probability surveys a major statistical innovation of the 20th century

It is hard to exaggerate the importance of the development and spread of large-scale probability surveys in the United States and around the world beginning in the 1930s (Harris-Kojetin, 2012). Such surveys not only measured phenomena with known precision compared with non-probability surveys, but also provided detailed information at greatly reduced cost and burden and increased timeliness compared with censuses. In 1937, during the Great Depression, a 2 percent sample of U.S. households on nonbusiness postal routes, designed by Calvin Dedrick, Morris Hansen, and others, estimated a much higher—and more credible—number of unemployed than a "complete" voluntary census of all residential addresses. Picking up on that effort, from 1940-1942 the Works Progress Administration fielded the sample-based Monthly Report on the Labor Force, forerunner to the Current Population Survey (CPS), which is the source of official U.S. unemployment statistics. From this beginning, the number and scope of federal surveys ballooned, including a large sample survey embedded in the decennial census.

Sampling was introduced into the 1940 census for six questions asked of a 5-percent sample. The success of sampling, operationally and with the public, led to administering two-fifths of the 1950 census questions on a sample basis and then in 1960 to using separate short and long forms, the latter with the questions asked of everyone plus the sample questions. More recently, following ideas of Leslie Kish, Roger Herriot, and others, the U.S. Census Bureau moved the collection of socioeconomic characteristics for small geographic areas from the long-form sample to continuous measurement via the American Community Survey (ACS) (National Research Council, 2007), which became operational in 2005.

2. Threats to the probability survey paradigm

Beginning without much notice several decades ago, threats to the probability survey paradigm have been snowballing in ways that bode ill for the future. There is a need for official statistical agencies in the United States and elsewhere to not only improve survey methods to counteract these threats, but also move toward a new paradigm that uses multiple sources, including surveys, to improve the relevance,

accuracy, and timeliness of consequential statistics and reduce burden and costs. As part of moving toward a multiple data sources paradigm, statistical agencies are well advised to use quality frameworks (pioneered by Brackstone, 1999; and refined by, e.g., Biemer et al., 2014) to systematically evaluate which data sources are most appropriate to combine for their statistical programs.

Below are summarized eight key threats to U.S. government surveys:

(a) Population coverage in U.S. household surveys, although adjusted for during weighting using census-based population estimates updated with administrative records, has worsened over the past several decades and varies widely across demographic groups. Business surveys also experience coverage error, in particular, more commonly obtaining better coverage of larger compared with smaller businesses and more established versus newer businesses. In household surveys, socioeconomic coverage differences undoubtedly remain even after ratio adjustment.

Coverage Ratios in March 2013 Current Population Survey (CPS)
(before ratio adjustment, as percentage of census-based population estimate)

	Ages 20-24	Ages 65 and older
Total (M/F):	74%	90%
White male:	76	91
Black male:	61	79
Hispanic male:	71	82

(b) Survey (unit) response by households and organizations has steadily declined for several decades in the United States and abroad (National Research Council, 2013).

Screener/Initial Response Rates (National Research Council, 2013:Tables 1-2, 1-4)	**1990/1991**	**2007/2009**
Consumer Expenditure Survey (CE, Diary) (interviewer drops off diary)	83.7%	70.3%
Current Population Survey (CPS) (personal interview)	94.3	90.5
National Health Interview Survey (NHIS) (personal interview)	95.5	82.2
National Household Education Survey (NHES) (RDD, has switched to mail)	81.0	52.5
Survey of Income and Program Participation (SIPP) (personal interview, Wave 1)	92.7	80.8

(c) Item nonresponse by responding units is high and growing for key variables, as evidenced by increasing imputation rates:

Percent of Income Imputed, CPS Annual Social and Economic Supplement and SIPP
(Czajka, 2009, Table A-8)

		1993	**1997**	**2002**
Total Income	–CPS ASEC	23.8%	27.8%	34.2%
	SIPP	20.8	24.0	28.6
Wages/Salaries	–CPS ASEC	21.5	24.8	32.0
	SIPP	17.7	20.5	24.9
Property Income	–CPS ASEC	42.4	52.8	62.6
	SIPP	42.4	42.9	49.7
Welfare Income	–CPS ASEC	19.8	18.1	29.2

SIPP	13.8	31.2	32.8

(d) Measurement error, such as underreporting of transfer income, is often problematic (e.g., Meyer et al., 2015).

Percent of Administrative Benchmarks, CPS ASEC and SIPP
(Czajka, 2009:Table A-5; p. 144)

Aggregate Benefits

		1987	**2005**
SNAP	–CPS	74.2%	54.6%
(food stamps)	SIPP	85.9	76.4
AFDC/TANF	–CPS	74.4	48.7
("welfare")	SIPP	73.0	62.2
OASI	–CPS	89.0	89.7
(Social Security)	SIPP	95.0	97.4

Aggregate Assets – SIPP, 1998-99 – 55% of Survey of Consumer Finances
Aggregate Liabilities – 90% of SCF

(e) Concepts in long-running surveys may become progressively out of date with social reality (e.g., "regular money income" in the CPS ASEC has not accorded for many years with how many people receive retirement and low-income benefits—see Czajka & Denmead, 2012—although pertinent questions have recently been added). Manski (2014) argues that statistical agencies greatly underestimate uncertainty in survey (and other) estimates due to these and other factors.

(f) Perceptions of burden, which may be well-founded in long, complex surveys, can not only result in unit and item nonresponse, but also threaten the political viability of important surveys. At present, the ACS is under attack in the U.S. Congress for this reason.

(g) Survey costs per case appear to be increasing, at least in part to maintain response rates, although there is generally only anecdotal evidence because little systematic analysis of survey costs has been carried out. In the case of the U.S. census, costs per housing unit are known to have increased in real terms by 600 percent from 1960 to 2010 (National Research Council, 2010:Table 2-2).

(h) Alternative sources of information potentially compete with government survey statistics in timeliness and cost. For example, PriceStat, spun off by the MIT Billion Prices Project (bpp.mit.edu), scrapes price data from the Internet to produce daily inflation indexes for over 20 countries, and ADP publishes national and regional (U.S.) monthly employment estimates based on its payroll processing operations (http://www.adpemploymentreport.com/). Both of these indexes acknowledge using Bureau of Labor Statistics data as input to their estimation process, but, in the public eye, the continued need for the underlying official statistics may not be clear.

3. Responses by the survey community

Survey researchers have not been idle in the face of multiple and increasing threats to the survey paradigm. To the contrary, for at least the last 15 years, they have been actively working on ways to reduce or compensate for coverage error, unit and item nonresponse, measurement error, and, more recently, burden on respondents. Some of the steps being taken include:

- Spending more on trying to complete each sample case, although budget constraints facing U.S. statistical agencies undercut the viability of this strategy in the long term.
- Using paradata and auxiliary information for more effective unit nonresponse bias identification and adjustment.
- Employing more sophisticated missing data adjustments that do not assume missing at random (as the commonly used "hot deck" imputation method does).
- Using adaptive design methods to optimize the cost and quality of response.
- Using multiple frames, for example, cell-phone and land-line frames for telephone surveys.
- Using multiple modes to facilitate more cost-effective response—the ACS is a prime example of this strategy, having recently added an Internet response option to its protocol of first mailing out questionnaires, then using CATI follow-up of mail and now Internet nonrespondents, and ending with CAPI follow-up of a sample of remaining nonrespondents.
- Conducting research to address burden by such means as optimizing the numbers of follow-up calls and visits, using matrix sample designs that reduce the burden for an individual respondent, and determining if administrative records or other data sources can substitute for some survey questions.
- Devoting resources to describing the benefits of and needs for the survey data—often, in the United States, data users are recruited to make the case to such stakeholders as members of Congress. For example, the Association of Public Data Users, the Council of Professional Associations on Federal Statistics, and the Population Association of America frequently mobilize data users on behalf of statistical agency programs.

4. New paradigm of multiple data sources

All of the steps listed above to improve surveys are laudable and necessary. I do not believe, however, that they are sufficient to restore a paradigm in which the probability survey is viewed as the primary vehicle for official statistics on households or other types of respondents. In today's environment of constrained budgets and increasing demands for wider, deeper, quicker, better, and cheaper statistics (Holt, 2007), it is necessary to start with understanding what policy makers and the public need from a statistical program and work backwards to the best combination of sources to meet those needs in as cost-effective and least burdensome manner as possible. Such sources may well include not only a survey, but also one or more non-survey sources, such as administrative records, commercial and other transactional data, sensor data, and data "scraped" or otherwise extracted from the Internet. In addition, the use of various modeling techniques will often be needed to make the most effective use of multiple data sources for a statistical program.

5. Uses of administrative records

For U.S. government statistical programs, I argue that greater use of administrative records to bolster household surveys could be very helpful to improve quality, reduce burden, and potentially reduce costs. Administrative records are already used to a considerable extent in business statistics programs and in some components of household survey programs, but they could play a much more important role than at present. In this regard, the U.S., with its federal system of

government and decentralized federal statistical system, which imposes legal and operational barriers to using administrative records, lags behind many other countries. An encouraging step has been taken by the U.S. Office of Management and Budget, which issued a memorandum in February 2014 with the goal of making statistical uses of administrative records government policy (https://www.whitehouse.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf).

Eight ways in which administrative records, suitably evaluated for quality compared with survey responses, can contribute to U.S. household survey data programs are listed below. There are examples of their use in government surveys, but in most instances, there remain many more opportunities to be exploited.

- Assist in evaluation of survey data quality, by comparison with aggregate estimates, appropriately adjusted for differences in population universes and concepts, and by exact matches of survey and administrative records.
- Provide control totals for adjusting survey weights for coverage errors, not only by age, sex, race, and ethnicity, but also by indicators of socioeconomic status.
- Provide supplemental sampling frames for use in a multiple frame design.
- Provide additional information to append to matched survey records to enhance the relevance and usefulness of the data with no additional burden.
- Provide covariates for model-based estimates for smaller geographic areas than the survey can support directly.
- Improve models for imputations for missing data in survey records.
- Replace “no” with “yes” for survey respondents who should have reported an item, replace” yes” with “no” for survey respondents who should not have reported an item, and replace reported values with records values for survey respondents who misreport an item, perhaps with perturbation if required to meet legal requirements for protecting records.
- Replace survey questions and use administrative records values directly, perhaps with perturbation.

6. Uses of other data sources

In addition to administrative records, there are other non-survey data sources that can contribute to survey-based government statistical programs. Statistical agencies, however, face a challenging balancing act. On one hand, they run the risk of looking “out of touch” if they do not strive to incorporate Internet-generated and other “big data” sources (e.g., data streams from recording devices), or what Groves (2011) terms “organic data,” into their work. On the other hand, it is no easy task for statistical agencies to evaluate and learn how to use such data sources in ways that do not compromise the quality and credibility of their series. Couper (2013) and Landefeld (2014) discuss perils in using “big data” for statistical estimation, while Horrigan (2013) gives examples of use. To determine appropriate uses, it would help to develop more illuminating concepts and terms to distinguish various non-survey data sources than the catchword “big data.”

7. Conclusion

Survey researchers, in the United States and elsewhere, face grave challenges to the probability survey paradigm. They also have great opportunities to

enhance the utility and cost-effectiveness of surveys by using them in a multiple data sources paradigm to improve information for policy making and public understanding.

References

[Note: This "Ask the Experts" column draws heavily on: C. F. Citro. (2014). From multiple modes for surveys to multiple sources for estimates. *Survey Methodology*]

Biemer, P., Trewin, D., Bergdahl, H. and Lilli, J. (2014). A system for managing the quality of official statistics, with discussion. *Journal of Official Statistics*, 30(3, September), 381-442.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25(2), 139-149.

Couper, M.P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods* 7(3):145-56

Czajka, J.L. (2009). SIPP data quality. Appendix A in Reengineering the Survey of Income and Program Participation. Panel on the Census Bureau's Reengineered Survey of Income and Program Participation, C.F. Citro and J.K. Scholz, eds. Committee on National Statistics, National Research Council. Washington, DC: The National Academies Press.

Czajka, J.L., and Denmead, G. (2012). Income Measurement for the 21st Century: Updating the Current Population Survey. Final report to the U.S. Department of Health and Human Services. Washington, DC: Mathematica Policy Research. Available at: http://www.mathematica-mpr.com/~/media/publications/pdfs/family_support/income_measurement_21century.pdf

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly* 75(9):861-871. Special 75th Anniversary Issue.

Harris-Kojetin, B. (2012). Federal household surveys. Pp. 226-234 in *Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey*, Second Edition, M. J. Anderson, C.F. Citro, and J.J. Salvo, eds. Washington, DC: CQ Press.

Holt, D.T. (2007).The official statistics Olympics challenge: Wider, deeper, quicker, better, cheaper. *The American Statistician* 61(1, February):1-8. With commentary by G. Brackstone and J.L. Norwood.

Horrigan, M.W. (2013). Big data: A BLS perspective. *Amstat News* 427(January):25-27.
Landefeld, J.S. (2014). Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues. Discussion Paper. International Conference on Big Data for Official Statistics, Beijing, China, October.

Manski, C.F. 2014. *Communicating Uncertainty in Official Economic Statistics*. NBER Working Paper No. 20098. Cambridge, MA: National Bureau of Economic Research.
Meyer, B.D. Wallace, K.C. M., and J.X. Sullivan. (2015). Household surveys in crisis. *Journal of Economic Perspectives* (forthcoming).

National Research Council. (2007). *Using the American Community Survey: Benefits and Challenges.* Panel on the Functionality and Usability of Estimates from the American

Community Survey, C.F. Citro and G. Kalton, eds. Committee on National Statistics. Washington, DC: National Academies Press.

National Research Council. (2010). *Envisioning the 2020 Census.* Panel on the Design of the 2010 Census Program of Evaluations and Experiments, L.D. Brown, M.L. Cohen, D.L. Cork, and C.F Citro, eds. Committee on National Statistics. Washington, DC: National Academies Press.

National Research Council. (2013). *Nonresponse in Social Science Surveys: A Research Agenda.* Panel on a Research Agenda for the Future of Social Science Data Collection. R. Tourangeau and T.J. Plewes, eds. Committee on National Statistics. Washington, DC: National Academies Press.

Ask the Experts - Call for Questions

If you'd like to ask the experts a question, please contact Kennon Copeland at copeland-kennon@norc.org.

New and Emerging Methods

Emerging Technologies: The Rise of Mobile Devices: From Smartphones to Smart Surveys

Trent D. Buskirk, Ph.D.
Marketing Systems Group

The Mobile Device Landscape

Over the past three decades cellular phones have seen an unprecedented development in both penetration and technological capabilities. Since their creation, the number of these mobile devices[1] in use around the world has surpassed the global population. Specifically, most recent estimates from the U.S. Census Bureau puts the worldwide population at just over 7.25 Billion people while the mobile tracking company GSMA's real time tracker estimates that the number of connected mobile devices is just under 7.54 Billion worldwide. Over the past 35 years, cell phones have also evolved from larger car phones to multimedia or flip phones with SMS/text and limited photo and internet capabilities to feature phones with more extended email and web services to smartphones which essentially employ operating systems and support downloadable applications (apps) that can be deployed by the user. Currently, the majority of mobile devices in use today are a combination of feature phones and smartphones (Nielsen, 2013) and the overall usage and penetration of these two types of cell phones varies by region/country. While the smartphone penetration in many industrialized countries exceeds 50% of the population, worldwide it is hovering just above 25% (International Telecommunication Union, 2014). However, estimates published by eMarketer (2014) indicate that the trend in smartphone penetration among the worldwide population will approach 50% by 2018. For some countries, smartphones are nearly ubiquitous or represent the vast majority of cellular phones in use (e.g. US, UK and Sweden) while in other countries, feature phones still represent the vast majority of devices used by subscribers (e.g. India, Russia, Brazil) according to estimates produced by Nielsen (2013).

The Rise in Mobile Internet Activity

With the increases globally in both feature phones and smartphones comes increases in shares of internet traffic from these devices. In a recent report released by comScore (2015), this year marks the first time in history where the share of internet traffic from users who access the internet using only their mobile devices exceeded that of users who accessed the internet using only desktop computers.

[1] Mobile devices will refer to the full collection of internet capable cellular phones and for the purposes of this article *will not* include tablets.

The Statistics Portal from Statista estimates that over 90% of the global online population will access the internet using their mobile devices by 2017. The share of total internet traffic accessed using mobile devices has also increased worldwide. Currently, StatCounter: Global Stats estimates that within the past three months (March – May, 2015) that the total share of internet traffic from mobile phone devices was just under 25% in North America, 21% in Europe, 46% in Asia, 38% in Africa and 18% in South America. A recent report released by Nielsen (2014) estimates that on average U.S. adults spend about 27 hours per month using a computer to access the internet compared to 34 hours per month using a smartphone browsers or apps. The increases in both mobile penetration and mobile internet traffic have also had an impact on online surveys as well. Several studies have reported increased incidence of mobile phones used to access (and complete) online surveys. For example, Courtright, Saunders and Tice (2014) report that the percentage of online surveys that have been accessed via a mobile device for one research company has increased from about 8% to over 27% between 2012 – 2014 and increased from 5% to just under 10% for another. Back in 2010 Kinesis estimated the prevalence of these once called "unintentional mobile respondents" at around 2%. What is clear is that increases in internet access via mobile devices has created a new breed of respondents who intentionally access the internet via mobile devices and that online surveys are no longer completed using a single mode.

New Data Collection Avenues and Data Types for Mobile Devices

Not only do mobile devices add heterogeneity to the modes by which online surveys can be completed, they also offer an array of new data that can be collected from respondents in their native environments as well as new ways to access data from respondents. For example, researchers from medical, health and behavioral sciences are already making use of smartphone related apps for heart rate monitoring (Gregoski et al., 2012), glucose monitoring/diabetes management (Kumar et al., 2012) , flu tracking and for tracking daily activity (Lai et al., 2010). Scagnelli et al. (2012) provide details on other types of data collected beyond answers to typical survey questions including product bar code capture as well as product picture data where access to the phone's camera was facilitated using a smartphone survey app. Michaud, Buskirk and Saunders (2014) also provided examples of using smartphones to capture picture data where the phone's web browser was used to facilitate camera access. They also provided examples of voice-to-text data capture for open-ended survey questions.

While most data survey researchers collect is active in that respondents answer questions or provide information that is requested of them, mobile devices open up the possibility of passively collected information as well. While these data may initially require permission from the respondent, they are generally collected without the respondent having to provide direct answers to survey questions. In this way passively collected data are like paradata, but may themselves represent key survey outcomes of interest or their correlates. Some common examples of passively collected data include Bluetooth connected monitoring devices that transmit data to the smartphone and then to the researcher (e.g. heart rate monitor) or GPS location data that is collected over a specified period of time. For example, Scagnelli et al. (2012) report on collecting GPS coordinates automatically using a survey app that was installed on Android smartphones provided to respondents and these coordinates were used to map out purchase locations for convenience items among Millennials. Olson and Wagner (2015) also evaluated the feasibility of using a GPS logging app provided on mobile devices of field interviewers for tracking field work efforts. As discussed in more detail in AAPOR's Emerging Technologies Task Force Report (2014), passive data collection of location-based data or other similarly

collected information can have great implications for improving measurement error often associated with recall biases or personal estimation biases. Apart from passively collected data, new research has tested the use of smartphones for data collection gigs via crowd sourcing applications including Duan, Lai and Link (2013) and Welbourne et al. (2014). A combination of passively collected crowdsourced data has also been used to track potholes on the streets of Boston since 2012 (see http://www.streetbump.org/) using an app installed on residents' smartphones.

In addition to respondent data that can be captured either actively or passively using Smartphones, new types of paradata can also be tracked. One of the most notable is the user agent string (Callegaro, 2010) that can be used to parse out the type of device used to access online surveys as well as properties of the browser initiating the request. Michaud, Saunders and Buskirk (2014) show how these user agent strings were used to sort out both missing data and possible primacy effects for online mobile surveys. Another important piece of new paradata comes from the information that is tagged on photographs captured and transmitted using their mobile devices. Certainly, respondents could upload similarly tagged photos from their desktop or laptop computer, but what is perhaps new is that the photos that are captured using a mobile device may have time and location data that more closely coincide with where and when survey data collection occurred and are likely to be outside the respondent's home. This paradata does raise some privacy concerns for researchers as outlined in the Emerging Technologies Task Force Report (AAPOR, 2014). In the case of app-based surveys (discussed in more detail later), the number of times a respondent "checks into the app" and the amount of time elapsed from in-app survey request to survey completion are examples of paradata that might be used to examine survey engagement or survey quality. Page load times and the size of transmitted data from respondents' devices are not new as they can be collected already for online surveys – but take on new currency in the mobile space as they relate to respondent burden associated with user data allowances (Buskirk, 2015).

Emerging Best Practices for Deploying and Designing Smartphone Surveys

Over the past five years, survey researchers have begun to explore different ways to conduct surveys using smartphones as well as how those surveys might be optimized to improve measurement and reduce potential mode effects. Smartphones represent a survey mode for which potential respondents have themselves gained extensive experience using – including checking emails, using apps and browsing the web (Link and Buskirk, 2013). In order to develop optimal experiences for respondents using smartphones, respondent expectations regarding intuitive apps, fast loading mobile web pages and simple navigation must be considered in addition to new dimensions of burden including both personal components, such as safety and privacy, as well as technological components, such as data consumption, bandwidth and battery drain (Buskirk, 2015).

While there are yet no definitive best practices for smartphone surveys some general taxonomies for deploying smartphone surveys have been offered (Buskirk and Andrus, 2012) and a growing body of literature has emerged from which best practices and recommendations on smartphone survey designs are beginning to emerge with some consistency. For a good starting place for the "bigger picture" readers should consult the AAPOR Emerging Technologies Task Force Report (AAPOR, 2014). More comprehensive emerging best practice recommendations can be found in Link and Buskirk (2012) or Buskirk (2015); a quicker snapshot of tips can also be found in McGeeney (2015). In what follows we will offer a very top-level overview of some of the current approaches for deployment and design of smartphone surveys.

Survey Deployment Method

The smartphone itself represents a single device that can be used for surveys in more than one mode. For example, the smartphone can be used for traditional telephone surveys, for short text message surveys and for online surveys using either the phone's internet browser or through an app loaded onto the phone. Buskirk and Andrus (2012) provide a more detailed discussion of the primary methods for deploying surveys via smartphones including: surveys via SMS or text message, mobile browser surveys and app/ app-like surveys. In this taxonomy mobile browser surveys are further categorized into two types: passive and active. Passive mobile browser surveys involve no special optimization or consideration for mobile devices and these types of surveys are simply designed and intended for completion on computers but in reality are completed using a mobile device. In comparison, active mobile browser surveys represent surveys that have some degree of optimization for completion on mobile devices. App-like mobile browser surveys, sometimes referred to as "native web apps" represent the greatest degree of mobile optimization in that they make use of heavy scripting to allow the survey completed using a smartphone browser to appear and function more like a smartphone app rather than a web survey. Surveys deployed by apps represent the highest degree of optimization and control for the survey designer but require development for multiple device operating systems such as Android, Windows and IOS and require respondents to download the app prior to survey completion.

To date, passive mobile browser surveys represent one of the more common deployment methods for mobile surveys. Our earliest understanding of potential mode effects for these types of smartphone surveys comes from studies that tracked the device used for accessing online surveys using the user agent string. From these surveys and "natural" experiments a profile of the consequences of not adapting survey content to the mobile environment has emerged to include reports of significantly longer survey completion times (Cunningham et al., 2013 and Peterson, 2012), higher break-off/drop-out rates (Saunders and Kessler, 2015; Poggio, Bosnjak and Weyandt, 2015; Cunningham et al., 2013 and Petersen, 2012) and different demographic profiles of survey completers compared to those computer completers (Wells, Bailey and Link, 2012 and Peterson et al., 2013). Essentially, this growing body of research has shown that a one-size fits all approach for online surveys may not work. As Peytchev and Hill (2010) note "mobile web surveys have unique features, such as administration on small screens and keyboards, different navigation, and reaching respondents in various situations that can affect response processes." Indeed, the emerging research on passive mobile browser surveys seems to indicate that some of the tenets and best practices for online surveys don't automatically translate into the mobile landscape and particular attention to optimizing for mobile devices is needed.

Active mobile browser surveys provide optimization of survey content, questions and response options and formats for mobile devices. While optimization for mobile devices is a key tenet of these smartphone surveys, the degree of optimization can vary from one active mobile browser survey to the next. This variability is a key component to understanding differences across studies in survey completion times, break off as well as differences in response distributions. For example, if an active mobile browser survey uses a company logo on each survey page that is based on a large graphics file (which is typical for computer surveys) rather than a compressed version of the graphics file more suitable for mobile devices, then completion times across these two active mobile browser surveys could be confounded by the amount of time required for downloading the two types of images. A growing number of studies has emerged reporting results of experiments comparing passive and active

mobile browser surveys to online surveys completed via computers including recent work by Peterson and colleagues (2013) as well as Baker-Prewitt and Miller (2013). Peterson et al. (2013) report that active mobile browser surveys fared better than passive browser surveys and produced results that were more consistent with respondents completing the online surveys using computers. Peterson and colleagues also reported that passive mobile browser survey completion times were on average longer than active mobile browser surveys which, in turn, were longer than those for surveys completed via computers. The extended time needed to complete passive mobile browser surveys compared to active moble browser surveys was also consistent to completion times reported by Baker-Prewitt and Miller (2013); however Baker-Prewitt and Miller note that on average, active mobile browser survey completion times were shorter than the completion time for surveys taken via computers. Baker-Prewitt and Miller also reported higher drop-out and higher straight-lining rates for passive mobile browser surveys compared to either active mobile browser surveys or those completed using a computer.

Buskirk and Andrus (2014) reported on one of the early experiments comparing possible mode effects between app-like mobile surveys compared to online surveys completed by computer. They report shorter survey completion times, on average, for app-like smartphone surveys compared to those completed via computers along with some differences in key survey outcomes including greater number of smartphone apps for those completing the app-like survey compared to respondents completing via computers. Mavletova and Couper (2013) also compared app-like smartphone web-based surveys to surveys completed using computers. Different from the findings from Buskirk and Andrus, Maveltova and Couper report longer completion times, on average, for those completing the app-like surveys on smartphones compared to computers. Differences in the completion time effects by mode could be related to differences in optimization for the two app-like surveys as we mentioned previously. Mavletova and Couper also report a greater number of mobile surveys were completed outside of the respondent's home compared to those completed on computers. No differential effects of satisficing were found across the two modes, however.

A recent report by the Pew Research Center (2015) offers a very comprehensive and thoughtful investigation of survey apps versus mobile optimized surveys for experience sampling method surveys (e.g. repeated surveys over short time intervals completed upon signals from researchers). Other work involving the use of a survey app for data collection has also been explored by Lai et al. (2013) and Lai et al. (2014). One aspect of app based surveys that needs careful consideration relates to the "fit for purpose" concept of surveys. For one-time surveys, an app may be too large of a proposition in that respondents must take the extra step of downloading and installing the app prior to completing any survey questions. Lai et al. (2013) report that only 41% of eligible respondents recruited to download a survey app actually did. This loss in response due to app download requirement was also noted by Johnson et al. (2012) who reported that only 37% of eligible mobile respondents willing to participate actually downloaded the survey app. Those who downloaded the app represented only 16% of the total number of online panelists who were invited to participate, overall. Wells, Bailey and Link (2014) compared an app-based mobile survey to one that is completed on computers for a subset of respondents recruited from a national probability panel. They report few mode differences between app-based smartphone surveys and those completed online using computers. So while there is burden associated with app downloads, smartphone apps that administer surveys give researchers the most control over design offer the largest range of features (such as flash video content) that can be deployed on

respective devices. However, separate versions of these apps must be designed for each mobile operating system.

The research on the utility of the other smartphone survey deployment methods described by Buskirk and Andrus (2012) is also beginning to emerge. For example, Schober and colleagues (2015) report on an experiment using text messaging surveys that have either been automated or are interactive with an interviewer who texts survey questions to respondents in real time. Compared to the control conditions of human and automated phone interviews, the text conditions show favorable results on more complete answers to sensitive questions. This finding has also been supported by international work of West and colleagues (2015) who compared simple real-time text surveys and modular text surveys to voice interviews with a live interviewer in Nepal. Schober et al. (2015) also report that respondents in the texting condition reported strong preference for future interviews by text and West et al. (2015) report that an overwhelming majority of respondents found single text questions per day "very easy."

Survey Invitation

Going beyond texting as a means for surveying, several studies have investigated their use for survey invitations compared to the more common email invitation standard for online surveys, in general. Mavletova and Couper (2014) also reported that SMS texts can be more efficient compared to emails in encouraging respondents to complete surveys using mobile devices rather than computers and de Bruijne and Wijnant (2014) found that among known smartphone users, survey response rates were higher for those who received an SMS invitation compared to an email invitation. A recent study has also looked at various ways email invitations can be optimized for smartphones using responsive design techniques specifically applied to email. Saunders and Kessler (2015) compared two versions of survey invitations sent via email to prospective respondents using standard formatting and responsive email designs (see Buskirk, 2015). They also compared using a start button in these emails versus including answer choices for the first survey question as an embedded survey start. They report that responsive email designs lead to higher completion rates relative to the standard emails. They also found that embedding the first survey question as the "start" trigger for the survey lead to twice as many clicks and 65% more completed surveys compared to the standard email invitation with the usual "start" button. By the end of the third quarter of 2014 Experian estimated that roughly 53% of total email opens occurred on a mobile device. While the work is emerging on the relationship between survey invitation and survey completion on mobile devices, more understanding between the interplay between type of invitation and type of mobile survey optimization is needed to develop refined best practices on survey invitations for mobile surveys.

Survey Length

While no definitive best practice on survey length for mobile surveys has been established, the general consensus is to try to make smartphone surveys shorter rather than longer. A recent study found that only 25% of smartphone users are willing to spend more than five minutes completing surveys (Kelly et al., 2013). Keep in mind that smartphones are portable and respondents can be outside the home and may not be free from other distractions for a longer period of time in order to complete the survey. Maveltova and Couper (2013) found that more surveys were completed outside the home for mobile respondents compared to computer respondents. One thing to note here is that user tolerances for longer surveys could be improved by choosing a better deployment method – one that places less burden

on the respondents' time by being easier to navigate (Link and Buskirk, 2012), including images that are optimized for mobile devices to reduce page load times (Buskirk, 2013) and using question types that reduce the number of clicks or taps required to provide answers to posed questions (Buskirk, 2015). In fact, Johnson et al. (2012) report survey completion times that were approximately two minutes shorter for a survey completed using a mobile app designed to optimize survey presentation on mobile devices compared to an identical survey completed via computers. Some researchers have proposed modular surveys as one potential solution for making surveys that are long by design more tolerable for mobile devices by breaking them into a series of shorter surveys (see Kelly et al., 2013 or Johnson et al., 2012). More work is needed to better understand how to optimally split longer surveys into shorter modules and what length is most optimal for these shorter surveys.

Survey Question Format

A growing body of literature has begun to examine how question formats that have traditionally been used in online surveys completed using computers translate into those completed via mobile devices. As you might imagine, some question types that ask respondents to select a single response from a short response set using radio buttons translate fairly easily from online to mobile. Implementing grid questions, traditionally used in online surveys completed by computer, have presented many challenges for mobile surveys. Sterrett et al. (2015) discuss issues with presenting grids on smartphones including increased respondent burden of having to scroll both vertically and horizontally to answer each of the questions in the grid. Thomas et al. (2015) discuss the optimization of grids on smartphones by reducing the number of scale points presented on mobile devices relative to PCs. Research on using slider bars versus radio buttons has been reported by Michaud, Saunders and Buskirk (2014) and Toepoel and Funke (2014). Studies using scrolling versus paging question presentation have also been conducted by Mavelotova and Couper (2014) and de Bruijne and Wijnant (2014). Both of these studies found slightly shorter survey completion times using scrolling survey formats in which more questions are placed on fewer pages and require the respondent to scroll down the mobile survey web pages to complete the survey. Both of these studies also reported that scrolling survey formats on lead to slightly higher item missing rates. Tips for mobile survey design based on these and other studies have been suggested by McGeeney (2015) and Buskirk (2015).

Recommended Resources for Smartphone Surveys

For those researchers who might be new to smartphone surveys or who might want to deploy such surveys in new countries or with different target populations, there are several resources that might be of interest to you. As we have already referenced, the AAPOR Emerging Technologies Task Force report (2014) is a great starting place for a broad overview on this topic. More practical advice on how to create web pages for mobile devices including information about touch interfaces, font size recommendations and color schemes can be found in the Mobile Web Best Practice Guidelines – a report produced by the World Wide Web Consortium (see: http://www.w3.org/TR/mwabp/). The European Society for Opinion and Market Research (ESOMAR) has also issued two reports that might be of interest including the "Guidance for Mobile Market Research" as well as the "Key Requirements for Mobile Research" (see: https://www.esomar.org/knowledge-and-standards/codes-and-guidelines/mobile-guideline.php). There are also three websites of note that offer many resources for understanding mobile device penetration as well as other issues related to mobile web design including: (1) StatCounter Global Statistics

(http://gs.statcounter.com/) allows the user to create graphs by platform comparing various device usage statistics (e.g. browser type, screen resolution, search engine use, social media use) by country, region or for the entire globe and over time ranging from 2008 to present; (2) Our Mobile Planet (http://think.withgoogle.com/mobileplanet/en/) provides cross tabulations of user penetration and device types for a host of countries and specific user related variables; (3) Mobi-Thinking (http://mobiforge.com/mobithinking) provides a host of articles on market penetration statistics for mobile devices and operating systems worldwide along with a host of thought pieces on issues related to mobile web design and implementation.

What's Next for Smartphone Surveys

While the literature and practice of mobile surveys continues to develop at a healthy pace and best practices continue to emerge, there are still areas of design that have yet to be solved completely. Making grid presentations consistent and optimal across mobile and computer versions of online surveys still needs rigorous exploration. While some research posits that grids on mobile devices may not be viable (see McGeeney, 2015 or Sterrett et al., 2015 for example), others have modified grids in various ways to include shortening the number of answer choices that are provided in response scales on mobile devices (see Thomas et al., 2015, for example). Another related issue that needs more attention involves components of mode effects that are related to possibly different presentations of questions for smartphones compared to PC. In some ways, it seems reasonable to think about the question presentation to be as native as possible for the given device on which they will be completed, but this decision then opens the question about device/mode effects versus question presentation effects and how these two might be confounded. For example, if grids are enabled on computer versions of surveys but they are collapsed on mobile devices to be presented as a series of questions that each have the same response set (e.g. the response scale options shown in the grids), then differences across mode will be possibly confounded with the way grid questions were presented. We need experiments that not only compare answer distributions and other aspects of survey completion across devices, but also a new level of experimentation that looks at various design and presentation options both within and across devices to better understand the components of mode effects that are associated with differences in question designs versus those that are generally associated with mode. To adequately move smartphone surveys from emerging to surging we encourage anyone working on experiments regarding question types or other user experiences with mobile survey designs to publish and contribute your work to conferences, journals or other accessible venues so we can all continue to learn how to make our surveys smarter than smartphones!

Acknowledgements: The author would like to thank Mick Couper and The Survey Statistician for the invitation to write this article.

References

AAPOR (2014) “Mobile Technologies for Conducting, Augmenting and Potentially Replacing Surveys.” Report from the AAPOR Emerging Technologies in Public Opinion Research Task Force. Available at: https://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/REVISED_Mobile_Technology_Report_Final_revised10June14.pdf

Baker-Prewitt J and Miller J (2013) "What Happens to Data Quality When Respondents Use a Mobile Device for a Survey Designed for a PC."" Paper presented at the 2013 CASRO Online Research Conference, San Francisco, March, 2013. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/0A81BA94-3332-4135-97F6-6BE6F6CEF475/Paper_-_Jamie_Baker-Prewitt_-_Burke.pdf

Buskirk TD (2013) "Smarter Smartphone Surveys 201: Data Collection Methods and Survey Design Considerations." Webinar presented for the American Association of Public Opinion Research. Available at http://www.aapor.org/source/education/webinar_recordings.cfm#.UsmPYNrnYfg

Buskirk TD (2015) "Going Mobile with Survey Research: Design, Data Collection, Sampling and Recruitment Considerations for Smartphone and Tablet Based Surveys."" Short course presented at the Journal of Official Statistics Anniversary Conference, 2015. Stockholm, Sweden. Available at: http://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/JOS-2015/buskirk-FINAL-participant-JOS2015ShortCourseBuskirkJUNE2015.pdf

Buskirk TD and Andrus C (2012) "Smart surveys for smart phones: Exploring various approaches for conducing online mobile surveys via smartphones." *Survey Practice*, 5. Available at: http://surveypractice.wordpress.com/2012/02/21/smart-surveys-for-smart-phones/.

Buskirk TD, and Andrus C (2014) "Making mobile browser surveys smarter: Results from a randomized experiment comparing online surveys completed via computer or smartphone." *Field Methods*, 26 , 322-342.

Callegaro M (2010) "Do you know which device your respondent has used to take your online survey?" *Survey Practice*. Available at: http://surveypractice.wordpress.com/2010/12/08/devicerespondent-has-used/.

comScore (2015) "Number of Mobile-Only Internet Users Now Exceeds Desktop-Only in the U.S." Available at: http://www.comscore.com/Insights/Blog/Number-of-Mobile-Only-Internet-Users-Now-Exceeds-Desktop-Only-in-the-U.S; accessed on June 20, 2015.

Courtright M, Saunders T and Tice J (2014) "Innovation in Web Data Collection: How 'Smart' Can I Make My Web Survey?" Paper presented at the CASRO Technology and Innovation Event, May, 2014, Chicago. Available at: http://c.ymcdn.com/sites/www.casro.org/resource/collection/97E56036-D4ED-4552-8A5F-E0A75899AEA8/2T1.1_-_T_Saunders_-_Maritz_-_M_Courtright_-_Research_Now_-_J_Tice_-_Decipher.pdf

Cunningham, JA et al. (2013) "Use of mobile devices to answer online surveys: implications for research." *BMC Research Notes,* Vol 6:258 available online: http://www.biomedcentral.com/1756-0500/6/258

DeBruijne M and Wijnant A (2014) "Improving response rates and questionnaire design for mobile web surveys." *Public Opinion Quarterly*, 78, 951-962.

eMarketer (2014) "2 Billion Consumers Worldwide to Get Smart(phones) by 2016." Available at: http://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694; accessed on June 21, 2015.

Gregoski M, Mueller M, Vertegel A, et al. (2012) "Development and Validation of a Smartphone Heart Rate Acquisition Application for Health Promotion and Wellness Telehealth Applications." *International Journal of Telemedicine and Applications*, doi:10.1155/2012/696324. Available at:

http://www.hindawi.com/journals/ijta/2012/696324 /

Johnson A, Kelly F, and Stevens S (2012) “Modular Survey Designs for Mobile Devices.” Paper presented at the 2012 CASRO Online Conference, March 2, 2012, Las Vegas, NV.

Kelly F, Johnson A and Stevens S (2013) “Modular Survey Design: Bite Size Chunks 2.” Paper presented at the 2013 CASRO Online Research Conference, San Francisco, March, 2013. Available at:
http://c.ymcdn.com/sites/www.casro.org/resource/collection/0A81BA94-3332-4135-97F6-6BE6F6CEF475/Paper_-_Frank_Kelly_-_Lightspeed_Research_and_Sherri_Stevens_-_Millward_Brown.pdf

Kumar A (2012) “Pilot study tests wearable sensors and smartphone-based data collection for diabetes,” iMedicalApps, Available at
http://www.imedicalapps.com/2012/11/pilot-study-wearable-sensors-smartphone-diabetes/

Lai JW, Vanno L, Link MW, Pearson J, et al. (2010) “Life360: Usability of Mobile Devices for Time Use Surveys.” *Survey Practice,* 3(1).

Link MW and Buskirk TD (2012) “The role of new technologies in powering, augmenting, or replacing traditional surveys.” Short-course presented at the annual meeting of the American Association for Public Opinion Research, Orlando, FL.

Mavletova A and Couper MP (2014) “Mobile Web Survey Design: Scrolling versus Paging, SMS versus E-mail Invitations.” *J Surv Stat Methodol* Vol 2 (4): 498-518; available at: http://jssam.oxfordjournals.org/content/2/4/498

McGeeney K (June 11, 2015) “Tips for Creating Web Surveys for Completion on a Mobile Device,” Pew Research Center. Available at: Accessed on June 25, 2015.
Nielsen (2014) “Digital Consumer Report, 2014.” available at:
http://www.nielsen.com/content/dam/corporate/us/en/reports-downloads/2014%20Reports/the-digital-consumer-report-feb-2014.pdf

Olson K and Wagner J (2015) “A feasibility test of using smartphones to collect GPS information in face-to-face surveys.” *Survey Research Methods*, Vol. 9(1). DOI:
http://dx.doi.org/10.18148/srm/2015.v9i1.6036

Pew Research Center (2015) “App vs. Web for Surveys of Smartphone Users.” Available at: http://www.pewresearch.org/files/2015/03/2015-04-01_smartphones-METHODS_final-3-27-2015.pdf. Accessed June 24, 2015.

Peytchev A and Hill CA (2010) “Experiments in mobile Web survey design similarities to other modes and unique considerations.” *Social Science Computer Review* 28:3 pp. 319-335.

Poggio T, Bosnjak M and Weyandt K (2015) “Survey Participation via Mobile Devices in a Probability-based Online-Panel: Prevalence, Determinants, and Implications for Nonresponse.” *Survey Practice* Vol. 8(1).

Saunders, T and Kessler A. (2015) "Read Me! Click Me! Innovations in Email Invitation Design for Today’s Digital World." Paper presented at the 2015 CASRO Digital Research Conference, February, 2015, Nashville, TN.
http://c.ymcdn.com/sites/www.casro.org/resource/collection/D3972A6E-AF15-4737-86F3-7DC13535B675/Innovations_in_Email_Invitation_Design.pdf

Scagnelli J, Bailey J, Link MW, Benezra K, Makowska H (2012) On the Run: Using Smartphones to Track Millennial's Purchase Behavior," Paper Presented at the 67th Annual Conference of the American Association for Public Opinion Research, Orlando, FL. Available at:
https://www.amstat.org/sections/SRMS/Proceedings/y2012/Files/400201_500569.pdf

Schober MF, Conrad FG, Antoun C, Ehlen P, Fail S, Hupp AL, et al. (2015) "Precision and Disclosure in Text and Voice Interviews on Smartphones." PLoS ONE 10(6): e0128337. doi:10.1371/journal.pone.0128337 available at:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0128337

Shu D, Lai J, and Link, MW (2013) Crowdsourcing via Mobile: Evaluating Viability of Data Collection "Gigs" with iPhone Users. *Survey Practice*. Vol 6(3).

Sterrett D, Stern M, Rugg G, Raker E, Baek J, and Bilgen I (2015) "The Effects of Grids on Web Surveys Completed on a Mobile Device." Presented at the Annual Conference of the American Association for Public Opinion Research, Hollywood, FL.

Thomas, R, Barlas FM, Graham P and Subias T (2015) "Improving Grids for Mobile Devices," Paper presented at the CASRO Digital Research Conference, February, 2015, Nashville.

Toepoel, V and Funke, F (2014) "Investigating Response Quality in Mobile and Desktop Surveys: A Comparison of Radio Buttons, Visual Analogue Scales and Slider Scales." Paper presented at the 2014 American Association of Public Opinion Research Conference. Anaheim, CA, May, 2014.

Welbourne E, Wu P, Bao X and Munguia-Tapia E (2014) " Crowdsourced Mobile Data Collection: Lessons Learned from a New Study Methodology." Paper presented at HotMobile'14, February, 2014, Santa Barbara, CA. Available at:
http://evan-welbourne.com/CSDC-HotMobile-2014-camera.pdf

Wells T, Bailey J and Link MW (2012) "A Direct Comparison of Mobile vs. Online Survey Modes." Paper presented at the annual conference of the American Association for Public Opinion Research, Orlando, FL.

Wells T, Bailey J and Link MW (2014) "Comparison of Smartphone and Online Computer Survey Administration." *Soc. Sci. Comput. Rev.* Vol. 32(2), 238-255. DOI=10.1177/0894439313505829 http://dx.doi.org/10.1177/0894439313505829

West BT, Ghimire D and Axinn WG (2015) "Evaluating a Modular Design Approach to Collecting Survey Data Using Text Messages: Evidence from Nepal," PSC Research Reports, No. 15-834, Population Studies Center, University of Michigan, Institute for Social Research http://www.psc.isr.umich.edu/pubs/pdf/rr15-834.pdf, accessed on June 22, 2015.

New and Emerging Methods – Call for Volunteers

If you're interested in contributing an article to the "New and Emerging Methods" section of a future edition of *The Survey Statistician*, please contact Denise Silva at denise.silva@ibge.gov.br.

Book and Software Review

A STATISTICAL FRAMEWORK FOR ANALYSING BIG DATA

Dr Siu-Ming Tam
Australian Bureau of Statistics[2]

1. INTRODUCTION

In a 2014 talk to the Victorian Branch of the Australian Statistical Society, Professor Terry Speed, an eminent mathematical statistician and winner of the 2014 Australian Prime Minister's Science Award, expressed surprise about the lack of visibility of statisticians in the Big Data debate, and said "…the absence of statisticians in Big Data activities is striking (to a statistician)". He also observed that there was generally lack of presence of statisticians in national and international conferences on Big Data.

In an article entitled "Big Data or Big Fail? The Good, the Bad and the Ugly and the Missing Role of Statistics", Iacus (2014) echoed Terry Speed's point about the role statistics and statisticians can play in the field of Big Data.

Against this background, I warmly welcome the well written and researched Report by the American Association for Public Opinion Research (AAPOR) Task Force (Japec *et al.*, 2015). The references provided in the Report would be very useful to statisticians who want to use Big Data or make a contribution to the Big Data debate. I particularly like the report's comprehensiveness in raising the many different issues of Big Data, covering not only what it is and why it matters, but also the policy, technical and technology challenges facing users of Big Data in solving business problems or finding answers to societal questions.

As a practicing official statistician, I find Section 7 of the AAPOR Report very interesting, and in particular, Sub-section 7.3 about combining Big Data and Survey Data. I would therefore devote most of my comments on this issue. I would also outline the preliminary work undertaken in the Australian Bureau of Statistics (ABS) to investigate into the business case and validity of harnessing certain Big Data sources for the regular production of official statistics.

2. THRESHOLD CHALLENGES FOR BIG DATA

Whilst the Report has outlined a number of key challenges for Big Data use and analysis, I would contend Business Case, using Big Data in statistically valid ways, i.e. Validity of Statistical Inference (page 22 of the Task Force Report) and Data

2 Views expressed in this paper are those of the author and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the author.

Ownership (page 30) are the threshold challenges confronting official statisticians in the use of Big Data in the regular production of official statistics.

In saying this, I am not downplaying the other challenges such as Data Stewardship, Data Collection Authority, Privacy and Re-identification. National Statistical Offices (NSOs) are generally well set up and have developed capability to address these challenges. For example, many statistical offices have already developed methods, processes and procedures to address privacy and confidentiality issues in their statistical releases – see, for example, the Special Issue of the *Statistical Journal of the International Association of Official Statistics* on "Official Statistics and Micro Data: Access and Confidentiality" released in 2009 – which may be adapted to address releases based on, or supplemented by, Big Data. A detailed discussion of the Big Data challenges faced by NSOs are provided in Tam and Clarke (2015a). My contention is that if the threshold challenges cannot be overcome, i.e. there is no business case for using a particular Big Data source, if the Big Data source cannot provide valid statistical inferences, and if the Big Data source is not available to official statisticians, there is no question of using the Big Data source in regular statistical production, and the other challenges do not arise.

3. BUSINESS CASE

What is the Business Case of Big Data? Business case comprises business need – what business problems we want to solve and can Big Data be part of the solution – and business benefit – whether the benefit of Big Data as a solution does outweigh the costs?

Being a collective term for a diverse range of data sources (page 5), the business case for Big Data does vary from source to source. For example, there is clearly a business case in the use of Administrative Data (page 9) by official statisticians in the production of official statistics, e.g. in the use of birth, death and migration records to complement the data from population censuses to provide contemporary population estimates. Cargo manifests are used to produce trade statistics. Without these sources, it will not be possible to provide population estimates or trade statistics. In other words, these sources provide valuable information to fill a data gap.

However, I have heard of propositions such as "… let's bring all the Big Data into our organisation and then figure out what we want to do with it. And to effectively do this, let's upgrade our computer hardware, or software, because Big Data requires big data processing capabilities …". These propositions worry me as they put the cart (Big Data) before the horse (business problems) and treat "Big Data as a solution in search of a problem".

In my view, Big Data should only be used if it can:

- improve the product offerings of statistical offices e.g. more frequent release of official statistics, more detailed statistics, more statistics for small population groups or areas, or filling an important data gap – business need; or
- improve the cost efficiency in the production of official statistics – business benefit.

The AAPOR Report rightly points out (page 15) that the "costs and risks of realising these (i.e. Big Data) benefits are non-trivial". For example, in the case of satellite data, whilst the risk of not having access to the data is small given that most of these are available free of charge on the internet, the cost associated with creating the ground truth data and marrying them up with satellite data, at the observation unit e.g. a statistical local area the cost of storing, cleaning, processing, quality assuring and software development are substantial. In the case of the Australian Bureau of Statistics (ABS), while the business need for using satellite data, instead of direct data collection, to estimate crop areas and crop yields has been well established, the business benefit has yet to be assessed.

4. A POSSIBLE APPROACH TO USING BIG DATA FOR OFFICIAL STATISTICS

An approach which has recently been actively pursued at the ABS (Tam and Clarke, 2015b) for the use of Satellite data in official statistics production is to consider the $N \times 1$ vector of measurements, $\mathbf{Y}_t$, of interest to the official statistician, e.g. crop areas or yields, at time t as a realisation of a super-population model, with the Big Data augmented with non-Big Data sources, $\mathbf{Z}_t$, treated as a (design) matrix of covariates for the model, i.e.

$$\mathbf{Y}_t = \mathbf{Z}_t \boldsymbol{\beta}_t + \mathbf{e}_t \qquad (1)$$

and allowing the vector of regression parameters, $\boldsymbol{\beta}_t$, to change over time, i.e.

$$\boldsymbol{\beta}_t = \mathbf{H}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t \ . \qquad (2)$$

Here N denotes the size of the finite population e.g. total number of land parcels. Equations (1) and (2) form the well-known State Space Model. Under this formulation, we consider that a sample, s_t, of units is chosen, e.g. a sample of observation units at time t, on which observations of the value of $\mathbf{Y}_{ot}$, where 'o' denotes observed (or responding) units, are obtained. Denote by 'm' the units in s_t on which there is no observation, i.e. missing data, and 'r', the units of s_t not selected in the sample, then the vector $\mathbf{Y}_t$ can be partitioned as $\mathbf{Y}_t = \left(\mathbf{Y}_{ot}, \mathbf{Y}_{mt}, \mathbf{Y}_{rt}\right)'$. State Space Models were used in Tam (1987) for predicting finite population parameters in finite population sampling.

Assuming that we can match these observed units to the corresponding units in the Big Data source and non-Big Data sources available to the statistician e.g. geographic location (in a survey, the linkage is automatic through the questionnaire as a collection instrument), and as can be seen from diagram 5.1 below, for every unit in the sample, s_t, one of the following two conditions will apply, namely, that there is a corresponding set of data from Big Data for the unit, and there is not. Denote by 'B' those units that have Big Data information, and '$\tilde{B}$' those that don't. Then (1) can be re-written as:

$$\begin{bmatrix} \mathbf{Y}_{o_B t} \\ \mathbf{Y}_{m_B t} \\ \mathbf{Y}_{r_B t} \\ \mathbf{Y}_{o_{\bar{B}} t} \\ \mathbf{Y}_{m_{\bar{B}} t} \\ \mathbf{Y}_{r_{\bar{B}} t} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{o_B t} \\ \mathbf{Z}_{m_B t} \\ \mathbf{Z}_{r_B t} \\ \mathbf{Z}_{o_{\bar{B}} t} \\ \mathbf{Z}_{m_{\bar{B}} t} \\ \mathbf{Z}_{r_{\bar{B}} t} \end{bmatrix} \boldsymbol{\beta}_t + \begin{bmatrix} \mathbf{e}_{o_B t} \\ \mathbf{e}_{m_B t} \\ \mathbf{e}_{r_B t} \\ \mathbf{e}_{o_{\bar{B}} t} \\ \mathbf{e}_{m_{\bar{B}} t} \\ \mathbf{e}_{r_{\bar{B}} t} \end{bmatrix} \qquad (3)$$

Note that (3) can be extended to Generalised Linear Models and Generalised Linear Mixed Models – see the penultimate section of this paper.

Let $\mathbf{I}_t$, $\mathbf{R}_t$ and $\mathcal{R}_t$ denote random variables representing sampling, response and Big Data under-coverage processes respectively. These are column vectors whose i-th element is given by $\delta_i^{(\mathbf{I}_t)}$, $\delta_i^{(\mathbf{R}_t)}$ and $\delta_i^{(\mathcal{R}_t)}$ respectively, which is 'one' if the i-th unit is in the sample, responded or covered in the Big Data respectively; and 'zero' otherwise.

The inference problem under the model (2) and (3) can then be stated as follows:

1. The data for inference for the finite population, say the population total, $\mathbf{1}'\mathbf{Y}$, at time t are

$$\mathbf{D}^{(t)} = \left\{ \mathbf{Y}_{o_B 1}, \mathbf{Y}_{o_{\bar{B}} 1}, \mathbf{Z}_{o_B 1}, \mathbf{Z}_{m_B 1}, \mathbf{Z}_{r_B 1}, \ldots, \mathbf{Y}_{o_B t}, \mathbf{Y}_{o_{\bar{B}} t}, \mathbf{Z}_{o_B t}, \mathbf{Z}_{m_B t}, \mathbf{Z}_{r_B t} \right\}$$

   and $$\mathbf{P}^{(t)} = \mathbf{P}_1^{(t)} \cup \mathbf{P}_2^{(t)}$$

   where $$\mathbf{P}_1^{(t)} = \left\{ \mathbf{I}_1, \mathbf{R}_1, \ldots, \mathbf{I}_t, \mathbf{R}_t \right\}$$

   and $$\mathrm{P}_2^{(t)} = \left\{ \mathcal{R}_1, \ldots, \mathcal{R}_t \right\}.$$

2. Model-assisted methods (Särndal *et al.*, 1992) and model-based methods (Chambers and Clark, 2012), including Bayesian methods (Puza, 2013), may be applied for making inference.

3. Whatever method of inference is used, the official statistician needs to understand, or make assumptions, about the processes leading to the missing and non-sample data, i.e. how those highlighted in black in equation (3) come into being; Where missing at random conditions are not met (see Section 5 below), modelling for the missing and non-sample selection processes have to made. For Big Data sources, this can be very challenging, if not insurmountable.

5. VALIDITY OF STATISTICAL INFERENCES

I welcome the attempt by the Task Force to provide a total error framework for Big Data (page 18), and Couper (2013) provides a good description of the types of errors encountered in Big Data.

I cannot agree more strongly with the Report that "… using Big Data in statistically valid ways is challenging and one misconception is the belief that the volume of the

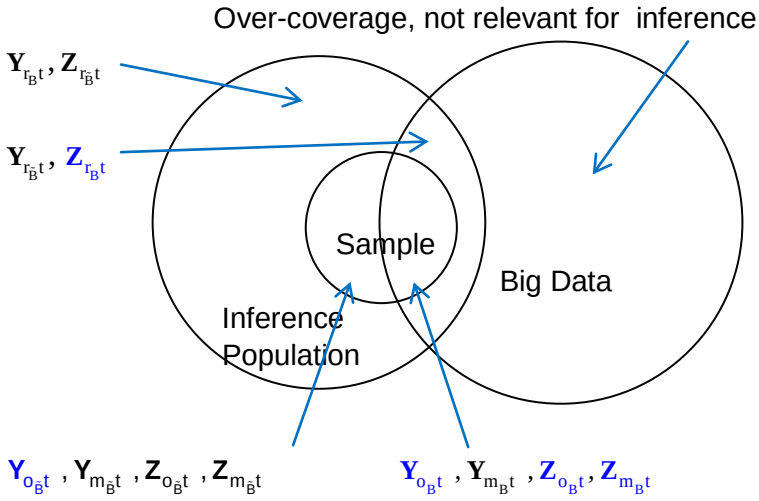
data can compensate for any other deficiency in the data (Big Data Hubris)" (page 22). Unlike sampling errors, non-sampling errors will not be reduced by increasing the sample size. Likewise, correlation is not the same as causality. In a recent article in *Significance*, entitled "Big Data, Big Mistake?", Harford (2014) showed how such a misunderstanding can have fatal consequences. The Report's reference to Fan *et al.* (2014) is particularly valuable to those Big Data enthusiasts who believe that size is everything!

To explore the conditions for validity of statistical inference, we will depict the relationship between a particular Big Data source (e.g. satellite imagery data) and the target population of interest (e.g. the land parcels) to the official statistician, in diagram 5.1 below. As well, we will make the simplified (but not always true, e.g. social media data) assumption that the unit of interest in the target population will appear in the Big Data source, if at all, only once. This is to ensure the possibility of making an unique linkage between the $\mathbf{Y}$ value of a unit in the target population and the corresponding $\mathbf{Z}$ values from Big Data (and non-Big Data sources). (Note that if there are multiple appearances, an approach that may be adopted would be randomly choose one appearance where the appearances are homogeneous, include an additional covariate where there is structured heterogeneity, or use a repeated measures model (Denham *et al.*, 2011) where the appearances are sufficiently heterogeneous.)

The joint areas of the two big circles in diagram 5.1 are divided into three segments – under-coverage, i.e. information of interest to the official statistician but not available from Big Data; over-coverage, i.e. information available from Big Data that is of no interest; and finally, information of interest and available. Also, the 'system' can be described as comprising a data process, state process and censoring process, with prior distributions $f(\varphi)$, $f(\theta)$ and $f(\phi)$ with known hyper-parameters.

Under the approach advocated in this paper, I assume that a probability sample (so as to fulfil the non-informative sampling conditions for descriptive and analytic inferences – see (6) and (10) below) is drawn from the population of interest, from which observations are made. These, combined with the corresponding Big Data for the same observation units, are used to provide the posterior distribution of the model parameters – the Estimation step. The resultant posterior distribution, together with the Big Data for the non-sampled units, are then used to predict the values of these units using the predictive distribution – the Prediction step.

5.1 Integrating designed data with found data



At time $=t$, State Space Process comprising:

Data Process – $f(\mathbf{Y}_t;\varphi)$

State Process – $f(\beta_t,\theta)$ assumed to be Markovian.

'Censoring Processes' – $f(\mathbf{I}_t,\phi)$, $f(\mathbf{R}_t,\phi)$, $f(\mathcal{R}_t,\phi)$

Parameter Models – $f(\varphi)$, $f(\theta)$, $f(\phi)$

Data – $\mathbf{D}^{(t)},\mathbf{P}^{(t)}$

Blue denotes observed/available.

Values of $\mathbf{Y}$ in units denoted by '$\tilde{B}$' not available due to under-coverage.

Values of $\mathbf{Y}$ in units denoted by 'r' not available due to the sampling process, $\mathbf{I}_t$.

Conceptualise missing values due to Non-response Process, $\mathcal{R}_t$, applied to Big Data

Values of $\mathbf{Y}$ in units denoted by 'm' not available due to a missing process applied to the sampled data, $\mathbf{R}_t$.

5.1 Descriptive inferences

Under a Bayesian framework, the predictive inference of $\mathbf{Y}_t$, $f\left(\mathbf{Y}_t \middle| \mathbf{D}^{(t)},\mathbf{P}^{(t)}\right)$, given the data $\mathbf{D}^{(t)}$ and $\mathbf{P}^{(t)}$ – which I shall denote by $\left[\mathbf{Y}_t \middle| \mathbf{D}^{(t)},\mathbf{P}^{(t)}\right]$ to simplify notation – is given by

$$\left[\mathbf{Y}_t \middle| \mathbf{D}^{(t)},\mathbf{P}^{(t)}\right] = \frac{\left[\mathbf{P}_1^{(t)} \middle| \mathbf{Y}_t,\mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right]\left[\mathbf{Y}_t,\mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right]}{\left[\mathbf{P}_1^{(t)} \middle| \mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right]\left[\mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right]}$$
$$= \left[\mathbf{Y}_t \middle| \mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right],$$

provided that $$\left[\mathbf{P}_1^{(t)} \middle| \mathbf{Y}_t,\mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right] = \left[\mathbf{P}_1^{(t)} \middle| \mathbf{D}^{(t)},\mathbf{P}_2^{(t)}\right]. \quad (4)$$

Assuming further that the finite population sampling and non-response processes at time τ_1 and τ_2 are independent for $\tau_1 \neq \tau_2$ and $\tau_1,\tau_2 = 1,\ldots,t$, sufficient conditions for (4) to hold are

$$\left[\mathbf{R}_\tau \,\middle|\, \mathbf{I}_\tau, \mathbf{Y}_\tau, \mathbf{D}_\tau, \mathcal{R}_\tau\right] = \left[\mathbf{R}_\tau \,\middle|\, \mathbf{I}_\tau, \mathbf{D}_\tau, \mathcal{R}_\tau\right] \quad (5)$$

and

$$\left[\mathbf{I}_\tau \,\middle|\, \mathbf{Y}_\tau, \mathbf{D}_\tau, \mathcal{R}_\tau\right] = \left[\mathbf{I}_\tau \,\middle|\, \mathbf{D}_\tau, \mathcal{R}_\tau\right] \quad (6)$$

for $\tau = 1, \ldots, t$.

Equation (6) holds for probability sampling, and Equation (5) holds if the non-response mechanism is missing at random (MAR) (Rubin, 1976). See, for example, Little and Rubin (2002) for response process modelling in which MAR does not hold.

Now

$$\begin{aligned}\left[\mathbf{Y}_t \,\middle|\, \mathbf{D}^{(t)}, \mathbf{P}_2^{(t)}\right] &\propto \int \left[\mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{D}_c^{(t)}, \mathbf{P}_2^{(t)}\right] d\mathbf{D}_c^{(t)} \\ &= \int \left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{D}_c^{(t)}\right] \left[\mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{D}_c^{(t)}\right] d\mathbf{D}_c^{(t)} \\ &= \int \left[\mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{D}_c^{(t)}\right] d\mathbf{D}_c^{(t)} \\ &\propto \left[\mathbf{Y}_t \,\middle|\, \mathbf{D}^{(t)}\right]\end{aligned}$$

provided that

$$\left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{Y}_t, \mathbf{D}^{(t)}, \mathbf{D}_c^{(t)}\right] = \left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{Y}_t, \mathbf{D}^{(t)}\right], \quad (7)$$

where

$$\mathbf{D}_c^{(t)} = \left\{\mathbf{Y}_{m_B 1}, \mathbf{Y}_{r_B 1}, \mathbf{Y}_{m_{\bar{B}} 1}, \mathbf{Y}_{r_{\bar{B}} 1}, \mathbf{Z}_{o_{\bar{B}} 1}, \mathbf{Z}_{m_{\bar{B}} 1}, \mathbf{Z}_{r_{\bar{B}} 1}, \ldots, \mathbf{Y}_{m_B t}, \mathbf{Y}_{r_B t}, \mathbf{Y}_{m_{\bar{B}} t}, \mathbf{Y}_{r_{\bar{B}} t}, \mathbf{Z}_{o_{\bar{B}} t}, \mathbf{Z}_{m_{\bar{B}} t}, \mathbf{Z}_{r_{\bar{B}} t}\right\}$$

represents the set of unobserved response variables and covariates in (3) for time 1 to time t .

Assuming that the under-coverage 'processes' for Big Data at time τ_1 and τ_2 are independent for $\tau_1 \neq \tau_2$ and $\tau_1, \tau_2 = 1, \ldots, t$, sufficient conditions for (7) to hold are:

$$\left[\mathcal{R}_\tau \,\middle|\, \mathbf{Y}_\tau, \mathbf{D}_\tau, \mathbf{D}_{\tau c}\right] = \left[\mathbf{R}_\tau \,\middle|\, \mathbf{Y}_\tau, \mathbf{D}_\tau\right], \quad (8)$$

where

$$\mathbf{D}_\tau = \left\{\mathbf{Y}_{o_B \tau}, \mathbf{Y}_{o_{\bar{B}} \tau}, \mathbf{Z}_{o_B \tau}, \mathbf{Z}_{m_B \tau 1}, \mathbf{Z}_{r_B \tau}\right\}$$

and

$$\mathbf{D}_{\tau c} = \left\{\mathbf{Y}_{m_B \tau c}, \mathbf{Y}_{r_B \tau c}, \mathbf{Y}_{m_{\bar{B}} \tau c}, \mathbf{Y}_{r_{\bar{B}} \tau c}, \mathbf{Z}_{o_{\bar{B}} \tau c}, \mathbf{Z}_{m_{\bar{B}} \tau c}, \mathbf{Z}_{r_{\bar{B}} \tau c}\right\}.$$

Note that $\left[\mathcal{R}_\tau \,\middle|\, \mathbf{Y}_\tau, \mathbf{D}_\tau, \mathbf{D}_{\tau c}\right] = \left[\mathbf{R}_\tau \,\middle|\, \mathbf{Y}_\tau, \mathbf{D}_\tau\right]$, for $\tau = 1, \ldots, t$, may be satisfied for certain Big Data sources e.g. administrative data, but not others e.g. data from social media where participation is self-selected.

Where (4) and (7) are satisfied, $\left[\mathbf{Y}_t \,\middle|\, \mathbf{D}^{(t)}, \mathbf{P}^{(t)}\right] \propto \left[\mathbf{Y}_t \,\middle|\, \mathbf{D}^{(t)}\right]$. In other words, the sampling, missing data and under-coverage processes can be ignored when making inference about $\mathbf{Y}_t$.

Where (7) is not fulfilled, predictive inferences for Big Data will have to be based on $\left[\mathbf{Y}_t \,\middle|\, \mathbf{D}^{(t)}, \mathbf{P}_2^{(t)}\right]$, which in turn requires modelling of $\mathbf{P}_2^{(t)}$.

Prediction with missing covariates can be a very challenging problem. See, for example, Chapter 4 of Wu (2010) for possible methods and references to tackle this issue.

5.2 Analytic inferences

The posterior distribution of the parameters, θ and φ, is given by

$$\begin{aligned}\left[\theta,\varphi \,\middle|\, \mathbf{D}^{(t)},\mathbf{P}^{(t)}\right] &= \int\left[\theta,\varphi,\mathbf{D}_c^{(t)} \,\middle|\, \mathbf{D}^{(t)},\mathbf{P}^{(t)}\right] d\mathbf{D}_c^{(t)} \\ &\propto \int\left[\mathbf{P}^{(t)} \,\middle|\, \mathbf{D}^{(t)},\mathbf{D}_c^{(t)},\theta,\varphi\right]\left[\mathbf{D}^{(t)},\mathbf{D}_c^{(t)},\theta,\varphi\right] d\mathbf{D}_c^{(t)} \\ &\propto \left[\theta,\varphi \,\middle|\, \mathbf{D}^{(t)}\right]\end{aligned}$$

provided that $$\left[\mathbf{P}^{(t)} \,\middle|\, \mathbf{D}^{(t)},\mathbf{D}_c^{(t)},\theta,\varphi\right] = \left[\mathbf{P}^{(t)} \,\middle|\, \mathbf{D}^{(t)},\theta,\varphi\right]. \quad (9)$$

Sufficient conditions for (9) to hold are

$$\left[\mathbf{P}_1^{(t)} \,\middle|\, \mathbf{D}^{(t)},\mathbf{D}_c^{(t)},\mathbf{P}_2^{(t)},\theta,\varphi\right] = \left[\mathbf{P}_1^{(t)} \,\middle|\, \mathbf{D}^{(t)},\mathbf{P}_2^{(t)},\theta,\varphi\right] \quad (10)$$

and $$\left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{D}^{(t)},\mathbf{D}_c^{(t)},\theta,\varphi\right] = \left[\mathbf{P}_2^{(t)} \,\middle|\, \mathbf{D}^{(t)},\theta,\varphi\right]. \quad (11)$$

Where (10) and (11) are satisfied,

$$\left[\theta,\varphi \,\middle|\, \mathbf{D}^{(t)},\mathbf{P}^{(t)}\right] = \left[\theta,\varphi \,\middle|\, \mathbf{D}^{(t)}\right] \propto \left[\mathbf{D}^{(t)} \,\middle|\, \theta,\varphi\right]\left[\theta \,\middle|\, \varphi\right][\varphi].$$

Whilst the above is formulated under a Bayesian framework, I note that the data, $\mathbf{D}^{(t)}$ and $\mathbf{P}^{(t)}$, are ancillary for $\mathbf{Y}_t$ or $(\theta,\varphi)'$ under the assumptions laid out above. Under the conditionality principle, frequentist inference for $\mathbf{Y}_t$ or $(\theta,\varphi)'$ should be based on holding the data, $\mathbf{D}^{(t)}$ and $\mathbf{P}^{(t)}$, fixed (see, for example, Cox and Hinkley, 1974, page 31).

6. DATA OWNERSHIP

I also agree that data ownership and access is a key issue for NSOs and one where there is a generally lack of legislation and a supporting framework (page 30). The challenge is to unlock public good from privately collected data whilst protecting the commercial interests of the data custodians.

In many cases, commercial value is placed on primary and derived non-government data sets by their owners, since either the provision of such data is the basis of their business, or its possession is a significant element of competitive advantage. This raises the issue of how the NSO might acquire commercially valuable or sensitive data for statistical production, particularly if the statistics compete directly with information products created by the data owner or they compromise its market position. This issue is made more complex by the fact that there may be several parties with some form of commercial right in relation to a data set, either through ownership, possession or licensing arrangements.

Much Web content is also unstructured and ungoverned – the metadata describing its usage and provenance (origin, derivation, history, custody, and context) are either incomplete or incongruous. Indeed, the long-term reliability of Big Data sources may be an issue for ongoing statistical production. Reputable statistics for policy making and service evaluation are generally required for extended periods of time, often

many years. However, large data sets from dynamic networks are volatile (and arguable static sources as well) – the data sources may change in character or disappear over time. This transience of data streams and sources does not sit comfortably with the reliability of statistical production and publication of meaningful time series.

With more statistics potentially available from the Web subject to different levels of biases and measurement errors at different points in time, what guidance can statisticians provide to report, connect and compare these statisticians over time and between different sources? As a minimum, the statistical profession should encourage the dissemination of these statistics to be accompanied by relevant meta data, for example, in the form of quality declarations and in accordance with Quality Frameworks (ABS, 2010; Brackstone, 1999; OECD, 2011) widely adopted by official statisticians.

7. A POSSIBLE ANALYSIS OF SATELLITE DATA TO PREDICT CROP YIELDS

To illustrate the potential analysis being developed in the ABS, I shall assume that equations (5), (6) and (7) are fulfilled by satellite data. Equation (6) is satisfied by choosing a random sample of observation units and collecting (ground truth) data on crop yields – the data are then integrated with satellite data to provide the 'training dataset'. Equation (7) is fulfilled as the coverage of satellite data is the same as the coverage for land parcels. Equation (5) may not hold for certain areas in Australia due to persistent cloud cover, as a result of moisture in the atmosphere, which may affect the type of crops being grown, or yields. This issue may, however, be by-passed by using traditional data collections e.g. statistical surveys, instead of using satellite data, for these areas.

Let the $N \times 1$ vectors $\mathbf{M}_t$, $\mathbf{m}_t$ and $\mathbf{Q}_t$ be the column vector of the crop yield, crop type and quantity harvestable respectively, for every observation unit of Australia. Then, $\mathbf{M}_t = \mathbf{Q}_t * \mathbf{m}_t = \mathbf{Exp}(\mathbf{Y}_t) * \mathbf{m}_t$, where $*$ denotes the Hadamard product, the $N \times 1$ vector $\mathbf{Exp}(\mathbf{Y}_t)$ has $exp(Y_{it})$ as its i-th element, $Y_{it} = \log Q_{it}$ and Q_{it} is the i-th element of $\mathbf{Q}_t$. Under the MAR assumptions made above, we can ignore $\mathbf{P}^{(t)}$ for predictive inference. That is,

$$\begin{aligned}\left[\mathbf{Y}_t, \mathbf{m}_t \,\middle|\, \mathbf{D}^{(t)}, \mathbf{P}^{(t)}\right] &= \left[\mathbf{Y}_t, \mathbf{m}_t \,\middle|\, \mathbf{D}^{(t)}\right] \\ &= \left[\mathbf{Y}_t \,\middle|\, \mathbf{m}_t, \mathbf{D}^{(t)}\right]\left[\mathbf{m}_t \,\middle|\, \mathbf{D}^{(t)}\right].\end{aligned}$$

By assuming $\mathbf{m}_t$ and $\mathbf{Y}_t \,|\, \mathbf{m}_t$ can be modelled by Dynamic Logistic Regression and Dynamic Linear models respectively, Tam and Clarke (2015b) provided results for the predictive distributions, $\left[\mathbf{Y}_t \,|\, \mathbf{m}_t, \mathbf{D}^{(t)}\right]$ and $\left[\mathbf{m}_t \,|\, \mathbf{D}^{(t)}\right]$.

To illustrate the idea for predicting quantity, under the assumptions of this Section, (3) becomes

$$\begin{bmatrix} \mathbf{Y}_{ot} \\ \mathbf{Y}_{rt} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{ot} \\ \mathbf{Z}_{rt} \end{bmatrix} \beta_t + \begin{bmatrix} \mathbf{e}_{ot} \\ \mathbf{e}_{rt} \end{bmatrix} \tag{12}$$

in which we have dropped the subscript 'B' to simplify notation. See Section 7.2 below for the choice of covariates and suggestions for improving the model in (12). Assuming that

$$\begin{aligned} \mathbf{Y}_t \mid \mathbf{Z}_t, \beta_t &\sim N\left(\mathbf{Z}_t\beta_t, \Sigma_t\right) \\ \beta_t &= \mathbf{H}\beta_{t-1} + \varepsilon_t \ , \ \beta_t \perp \mathbf{Z}_t \\ \beta_1 &\sim N\left(\beta_0, \Omega_{\beta_0}\right) \\ \varepsilon_t &\sim \text{independent } N\left(0, \Omega_t\right) \ , \ \varepsilon_t \perp \mathbf{D}^{(t)} \end{aligned} \tag{13}$$

and Ω_t and $\Sigma_t = \begin{pmatrix} \Sigma_{oot} & \mathbf{0} \\ \mathbf{0} & \Sigma_{rrt} \end{pmatrix}$ are known, the predictive distribution of the total yield of a particular crop (Tam and Clarke, 2015b) is $\mathbf{1}_o'\mathbf{Q}_{ot} + \mathbf{1}_r'\mathbf{Exp}\left(\hat{\mathbf{Y}}_{rt}\right)$, where

$$\begin{aligned} \hat{\mathbf{Y}}_{rt} &\sim N\left(\mathbf{Z}_{rt}\hat{\beta}_{t|t} \ , \Sigma_{rrt} + \mathbf{Z}_{rt}\Omega_{t|t}\mathbf{Z}_{rt}'\right) \\ \hat{\beta}_{t|t} &= \mathbf{H}\hat{\beta}_{t-1|t-1} + \Omega_{t|t}\mathbf{Z}_{ot}'\Sigma_{oot}^{-1}\left(\mathbf{Y}_{ot} - \mathbf{Z}_{ot}\hat{\beta}_{t-1|t-1}\right) \\ \Omega_{t|t} &= \left(\Omega_{t|t-1}^{-1} + \mathbf{Z}_{ot}'\Sigma_{oot}^{-1}\mathbf{Z}_{ot}\right)^{-1} \\ \Omega_{t|t-1} &= \mathbf{H}\Omega_{t-1|t-1}\mathbf{H}' + \Omega_t \ . \end{aligned} \tag{14}$$

Here $\mathbf{Exp}\left(\hat{\mathbf{Y}}_{rt}\right)$ denotes the vector with $exp\left(\hat{Y}_{irt}\right)$ as its i-th element, $\hat{Y}_{irt}$ is the i-th element of $\hat{\mathbf{Y}}_{rt}$, $\hat{\beta}_{t|t}$ denotes the posterior mean of β_t given $\mathbf{D}^{(t)}$, and $\Omega_{t|t}$ is the variance-covariance matrix of $\hat{\beta}_{t|t}$.

Note the above methodology may be adapted to a 'design-assisted' approach (Särndal *et al.*, 1992) for estimating finite population parameters using the following heuristic argument. From (3), the Generalised Regression Estimator for the total yield, $\mathbf{1}'\mathbf{Y}_t$, is

$$e_{ot}\left(\mathbf{Y}_t\right) + \left\{\mathbf{1}'\mathbf{Z}_t - e_{ot}\left(\mathbf{Z}_t\right)\right\}\bar{\beta}_{Dt}$$

where $e_{ot}\left(\mathbf{Y}_t\right)$, $e_{ot}\left(\mathbf{Z}_t\right)$ are the Horvitz-Thompson estimators of $\mathbf{Y}_t$ and $\mathbf{Z}_t$ respectively, and $\bar{\beta}_{Dt}$ is the design based estimator of β_t at time $\mathbf{t}$ (Särndal *et al.*, 1992). Following Wright (1983), even though $\hat{\beta}_{t|t}$ is not asymptotically design unbiased, we may use it for $\bar{\beta}_{Dt}$.

Likewise, denoting $\sigma\left(\mathbf{Z}_{it}'\gamma_t\right) = \left[1 + exp\left(-\mathbf{Z}_{it}'\gamma_t\right)\right]^{-1}$ as the logistic sigmoid for observation $\mathbf{i}$ at time $\mathbf{t}$, and assuming $\mathbf{m}_{it} \sim$ independent Binomial Logistic $\left(\sigma\left(\mathbf{Z}_{it}'\gamma_t\right)\right)$, or

$$\begin{bmatrix} \mathbf{m}_{ot} \\ \mathbf{m}_{rt} \end{bmatrix} = \begin{bmatrix} \sigma(\mathbf{Z}_{ot}\gamma_t) \\ \sigma(\mathbf{Z}_{rt}\gamma_t) \end{bmatrix}$$

$$\gamma_t = \mathbf{H}\gamma_{t-1} + \varepsilon_t \ , \ \gamma_t \perp \mathbf{Z}_t$$

$$\gamma_1 \sim N(\gamma_0, \Xi_{\gamma_0})$$

$$\varepsilon_t \sim \text{independent } \mathrm{N}(\mathbf{0}, \Xi_t) \ , \ \varepsilon_t \perp \mathbf{D}^{(t)} \qquad (15)$$

where $\mathbf{m}_{ot} = (\mathbf{m}_{1t}, \ldots, \mathbf{m}_{ot})'$ and $\sigma(\mathbf{Z}_{ot}\gamma_t) = (\sigma(\mathbf{Z}'_{1t}\gamma_t), \ldots, \sigma(\mathbf{Z}'_{nt}\gamma_t))'$ etc. and Ξ_t is known, then (Tam and Clarke, 2015b)

$$\mathbf{m}_{it} \mid \mathbf{D}^{(t)} \approx \text{independent Binomial Logistic}\left(\sigma\left(\mathbf{Z}'_{it}\hat{\gamma}_{t|t}\right)\right) \qquad (16)$$

for unobserved units $i = 1, \ldots, r_t$,

where $$\hat{\gamma}_{t|t} = \mathbf{H}\,\hat{\gamma}_{t-1|t-1} + \Sigma^{-1}_{t|t-1}\left\{\mathbf{Z}'_{ot}\mathbf{m}_{ot} - \mathbf{Z}'_{ot}\sigma\left(\mathbf{Z}'_{ot}\hat{\gamma}_{t|t}\right)\right\}$$

and $$\Sigma_{t|t-1} = \Sigma_{t-1|t-1} + \Xi_t \ .$$

7.1 Statistical computing issues

The examples shown above make the unrealistic assumptions that quantities like Σ_t, Ω_t and Ξ_t are known. In reality they are not and have to be estimated by the observed data. To make the estimation task more manageable, one can consider modelling the unknown quantities as follows

$$\Sigma_t = \lambda_t(\Sigma)\,\Sigma$$

$$\Omega_t = \lambda_t(\Omega)\,\Omega$$

$$\Xi_t = \lambda_t(\Xi)\,\Xi$$

where the scalars $\lambda_t(\Sigma)$, $\lambda_t(\Omega)$, $\lambda_t(\Xi) > 0$ follow an uninformative prior,

$$\Sigma \sim \mathrm{W}^{-1}(\Sigma_0, \nu_\Sigma) \ , \ \Omega \sim \mathrm{W}^{-1}(\Omega_0, \nu_\Omega) \ , \ \Xi \sim \mathrm{W}^{-1}(\Xi_0, \nu_\Xi)$$

and W^{-1} denotes the Inverse-Wishart distribution.

Let $\Theta_t = \{\theta, \phi, \varphi, \lambda_t(\Sigma), \lambda_t(\Omega), \lambda_t(\Xi), \Sigma, \Omega, \Xi\}$ and may also include $\mathbf{H}$ if it is not known. Assuming (4) and (9) are fulfilled, then the posterior distribution of Θ_t ,

$$\left[\Theta_t \mid \mathbf{D}^{(t)}\right] \propto \left[\mathbf{D}^{(t)} \mid \Theta_t\right]\left[\Theta_t\right],$$

i.e. likelihood times the prior. 'Maximum a posteriori' estimates of Θ_t can be derived using the EM algorithm – see Haykin (2001, Chapter 5) and also Strickland *et al.* (2009, 2011) for efficient estimation applied to satellite data.

Alternatively, the predictive distribution, $\left[\mathbf{M}_t \mid \mathbf{D}^{(t)}\right]$, where $\mathbf{M}_t = \mathrm{E}(\mathbf{Y}_t) * \mathbf{m}_t$ as before, can be derived using Monte Carlo via the method of composition as follows. From

$$\left[\mathbf{Y}_t, \mathbf{m}_t, \Theta_t \,\middle|\, \mathbf{D}^{(t)}\right] = \left[\mathbf{Y}_t, \mathbf{m}_t \,\middle|\, \mathbf{D}^{(t)}, \Theta_t\right]\left[\Theta_t \,\middle|\, \mathbf{D}^{(t)}\right]$$
$$= \left[\mathbf{Y}_t \,\middle|\, \mathbf{m}_t, \mathbf{D}^{(t)}, \Theta_t\right]\left[\mathbf{m}_t \,\middle|\, \mathbf{D}^{(t)}, \Theta_t\right]\left[\Theta_t \,\middle|\, \mathbf{D}^{(t)}\right],$$

one can use the *LibBi* software as outlined in Murray (2015) to draw J samples $\Theta_t^1, \ldots, \Theta_t^J$ from $\left[\mathbf{D}^{(t)} \mid \Theta_t\right]\left[\Theta_t\right]$. Using these values and equations (14) and (16), we obtain samples $\mathbf{Y}_t^1, \ldots, \mathbf{Y}_t^J$ from $N\left(\mathbf{Z}_{rt}\hat{\beta}_{t|t}, \Sigma_{rrt} + \mathbf{Z}_{rt}\Omega_{t|t}\mathbf{Z}'_{rt}\right)$ and $\mathbf{m}_t^1, \ldots, \mathbf{m}_t^J$ from $\left[\mathbf{m}_t \,\middle|\, \mathbf{D}^{(t)}, \Theta_t\right]$ respectively, where the i-th element of the vector $\mathbf{m}_t$ follows a Binomial Logistic Regression model with logistic sigmoid $\sigma\left(\mathbf{Z}'_{ti}\hat{\gamma}_t\right)$, from which a sample of $\mathbf{M}_t^1, \ldots, \mathbf{M}_t^J$ can be obtained for Monte Carlo inference on $\mathbf{M}_t \mid \mathbf{D}^{(t)}$. Strickland *et al.* (2013) has also developed a Python package, pyMCMC, for fast multivariate state space modelling, which is scheduled for release in June, 2015.

7.2 Choosing covariates and improving the model fit

There is a huge literature in predicting crop yields, see for example, Johnson (2014) and the references therein. A review of the methodology is provided in Lobell (2013). Based on the science of crops, most of these use the Normalised Difference Vegetation Index (NDVI), or Enhanced Vegetation Index (EVI), which are simple functions of the near-infrared radiation and visible radiation, and other variables like soil moisture, land surface temperature etc. available from satellites and other sources are included as covariates. Stress Index as a covariate derived from thermal time and crop phenology both from remote sensing (Idso *et al.*, 1981; Jackson *et al.*, 1983; Rodriguez *et al.*, 2005) as well as directly modelled from a biophysical crop model (Potgieter *et al.*, 2005; Potgieter and Hammer, 2006) has been proposed. In addition, evapotranspiration derived from EVI and Global Vegetation Moisture Index has been suggested as covariates (Guerschman *et al.*, 2009). These covariates can be incorporated in an obvious way into the State Space Model described above, although care has to be exercised to ensure there is no collinearity issue, or model over-fitting.

Becker-Reshef *et al.* (2010) fitted a simple regression model using county yield statistics as response variables and NDVI as explanatory variables, and use it to predict yields. Newlands *et al.* (2014) extends this work by employing a multivariate regression model using NDVI and agro-climate data as covariates. In addition, their model also allows a lag-1 autoregressive term for crop yields and the coefficients to vary over time and space, although no stochastic relationships between these coefficients were exploited. Priors on the parameters of the multivariate regression model were constructed using residual bootstrapping (Bornn and Zidek, 2012). The State Space Modelling advocated in this paper can be regarded as an extension of the methodology developed by Newlands *et al.* (2014).

Where the model defined (13) does not adequately predict crop quantities, the following model may be considered:

$$\mathbf{Y}_t \mid \mathbf{Z}_t, \boldsymbol{\beta}_t, \mathbf{F}, \boldsymbol{\alpha}_t \sim N\left(\mathbf{F}\boldsymbol{\alpha}_t + \mathbf{Z}_t\boldsymbol{\beta}_t, \Sigma_t\right)$$

$$\begin{bmatrix} \boldsymbol{\alpha}_t \\ \boldsymbol{\beta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1\boldsymbol{\alpha}_{t-1} \\ \mathbf{H}_2\boldsymbol{\beta}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix}, \quad \boldsymbol{\alpha}_t, \boldsymbol{\beta}_t \perp \mathbf{Z}_t$$

$$\boldsymbol{\varepsilon}_t = \begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} \sim \text{independent } N\left(\mathbf{0}, \boldsymbol{\Omega}_t\right), \quad \boldsymbol{\varepsilon}_t \perp \mathbf{D}^{(t)}$$

$$\boldsymbol{\Omega}_t = \begin{pmatrix} \boldsymbol{\Omega}_{1t} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{2t} \end{pmatrix}.$$

In other words, the time-variant fixed effects, $\mathbf{F}\boldsymbol{\alpha}_t$, is used to 'sweep' up any missing covariates in the modelling. This approach is akin to using random slopes in multi-level modelling (Snijders and Bosker, 1999 – Chapter 5) and is also known as Generalised Linear Mixed Model. A similar approach may be adopted for the model described in equation (14). The suggested approach, however, would require large sample sizes, as well as longer time series for accurate and precise estimation.

7.3 Concluding remarks

In developing the above models and building the training data set for analyses, I found that I have to involve crop scientists (or more generally "domain experts" – page 26 of the AAPOR Report), statisticians and computer scientists, supporting the comment that a multi-disciplinary team is required to harness opportunities, and addressing challenges, from Big Data. New skill sets are required to integrate ground truth data with satellite data.

Recommendation 1 of the AAPOR Report (page 2) says:

> "Survey and Big Data are complementary data sources and not competing data sources. There are differences between the approaches, but this should be seen as an advantage rather than a disadvantage".

This paper outlines an approach to combine the strength of Big Data with survey data – which has been regarded as the 'gold standard' for collecting data to make valid statistical inference – for predicting crop yields. The basic ideas are to use the Big Data and other auxiliary sources to calibrate the response variables, and to apply State Space Modelling to solve finite population inference problems. However, this is possible because the population covers by satellite imagery is identical to the population of land parcels, and the missing covariates problem is by-passed by relying on the traditional survey methods of estimation in those areas without satellite data e.g. missing data due to clouds. The efficacy of the approach will be tested using the training data set that is being built in the ABS. I hope to be able to report the outcome of the analyses, successful or otherwise, in the future elsewhere.

Once again, I congratulate the AAPOR Task Force for providing an excellent Report.

REFERENCES

Australian Bureau of Statistics (2010a) *The ABS Data Quality Framework.* <https://www.nss.gov.au/dataquality/aboutqualityframework.jsp>

Becker-Reshef, I.; Vermote, E.; Lindeman, M. and Justice, C. (2010) "A Generalised Regression-Based Model for Forecasting Winter Wheat Yields in Kansas and Ukraine Using MODIS Data", *Remote Sensing of Environment*, 114(6), pp. 1312–1323.

Bornn, L. and Zidek, J.V. (2012) "Efficient Stabilization of Crop Yield Prediction in the Canadian Prairies", *Agricultural and Forest Meteorology*, 152, pp. 223–232.

Brackstone, G. (1999) "Managing Data Quality in a Statistical Agency", *Survey Methodology*, 25(2), pp. 139–149.

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

Couper, M.P. (2013) "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys", *Survey Research Methods*, 7(3), pp. 145–156. <https://ojs.ub.uni-konstanz.de/srm/article/view/5751/5289>

Cox, P.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman and Hall, London.

Denham, R.J.; Falk, M.G. and Mengersen, K.L. (2011) "The Bayesian Conditional Independence Model for Measurement Error: Applications in Ecology", *Environmental and Ecological Statistics*, 18(2), pp. 239–255.

Fan, J.; Han, F. and Liu, H. (2014) "Challenges of Big Data Analysis", *National Science Review*, 1(2), pp. 293–314.

Guerschman, J.P.; Van Dijk, A.I.; Mattersdorf, G.; Beringer, J.; Hutley, L.B.; Leuning, R.; Pipunic, R.C. and Sherman, B.S. (2009) "Scaling of Potential Evapotranspiration with MODIS Data Reproduces Flux Observations and Catchment Water Balance Observations Across Australia", *Journal of Hydrology*, 369(1–2), pp. 107–119.

Harford, T. (2014) "Big Data: A Big Mistake?", *Significance*, 11(5), pp. 14–19.

Haykin, S.S. (ed.) (2001) *Kalman Filtering and Neural Networks*, John Wiley & Sons, Inc., New York.

Iacus, S.M. (2014) "Big Data or Big Fall? The Good, the Bad and the Ugly and the Missing Role of Statistics", *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*, 5(1), pp. 4–11.

Idso, S.B.; Reginato, R.J.; Jackson, R.D. and Pinter, P.J. (1981) "Measuring Yield-Reducing Plant Water Potential Depressions in Wheat by Infrared Thermometry", *Irrigation Science*, 2(4), pp. 205–212.

Jackson, R.D.; Slater, P.N. and Pinter, P.J. (1983) "Discrimination of Growth and Water Stress in Wheat by Various Vegetation Indices Through Clear and Turbid Atmospheres", *Remote Sensing of Environment*, 13(3), pp. 187–208.

Japec, L.; Kreuter, F.; Berg, M.; Biemer, P.; Decker, P.; Lampe, C.; Lane, J.; O'Neil, C. and Usher, A. (2015) *American Association for Public Opinion Research (AAPOR) Report on Big Data.*
<http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf>

Johnson, D.M. (2014) "An Assessment of Pre- and Within-Season Remotely Sensed Variables for Forecasting Corn and Soybean Yields in the United States", *Remote Sensing of Environment*, 141, pp. 116–228.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, Second Edition, Wiley, New York.

Lobell, D.B. (2013) "The Use of Satellite Data for Crop Yield Gap Analysis", *Field Crops Research*, 143, pp. 56–64.

Murray, L.M. (2015) *Bayesian State-Space Modelling on High-Performance Hardware using LibBi.*
<http://arxiv.org/pdf/1306.3277v1.pdf>

Newlands, N.K.; Zamar, D.S.; Kouadio, L.A.; Zhang, Y.; Chipanshi, A.; Potgieter, A.; Toure, S. and Hill, H.S.J. (2014) "An Integrated, Probabilistic Model for Improved Seasonal Forecasting of Agricultural Crop Yield Under Environmental Uncertainty", *Frontiers in Environmental Science*, 2, pp. 1–21.
<http://journal.frontiersin.org/article/10.3389/fenvs.2014.00017/full>

OECD (2011) *Quality Dimensions, Core Values for OECD Statistics and Procedures for Planning and Evaluating Statistical Activities*.
<http://www.oecd.org/std/21687665.pdf>

Potgieter, A.B.; Hammer, G.L.; Doherty, A. and de Voil, P. (2005) "A Simple Regional-Scale Model for Forecasting Sorghum Yield Across North-Eastern Australia", *Agriculture and Forest Meteorology*, 132(1–2), pp. 143–153.

Potgieter, A.B. and Hammer, G.L. (2006) *Oz-Wheat: A Regional-Scale Crop Yield Simulation Model for Australian Wheat*, Information Series, Queensland Department of Primary Industries and Fisheries, Brisbane.

Puza, B. (2013) *Lectures on Bayesian Statistics*, Unpublished manuscript, Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University.

Rodriguez, D.; Fitzgerald, G.J.; Belford, R. and Christensen, L. (2006) "Detection of Nitrogen Deficiency in Wheat From Spectral Reflectance Indices and Basic Crop Eco-Biophysiological Concepts", *Australian Journal of Agricultural Research*, 57(7), pp. 781–789.

Rubin, D.B. (1976) "Inference and Missing Data", *Biometrika*, 63(3), pp. 581–592.
Särndal, C.-E.; Swensson, B. and Wretman, J. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Snijders, T. and Bosker, R. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, Sage, London.

Snijders, T.A.B. and Bosker, R.J. (1999) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, SAGE Publications Ltd, London.

Speed, T. (2014) *Data Science, Big Data and Statistics: Can We All Live Together?* Chalmers Initiative Seminar on Big Data.
<http://www.chalmers.se/en/areas-of-advance/ict/events/Documents/Terry%20Speed_Data%20Science,%20Big%20Data%20and%20Statistics%20-%20Can%20We%20All%20Live%20Together.pdf>

Strickland, C.M.; Turner, I.W.; Denham, R.J. and Mengersen, K.L. (2009) "Efficient Bayesian Estimation of Multivariate State Space Models", *Computational Statistics and Data Analysis*, 53(12), pp. 4116–4125.

Strickland, C.M.; Simpson, D.P.; Turner, I.W.; Denham, R.J. and Mengersen, K.L. (2011) "Fast Bayesian Analysis of Spatial Dynamic Factor Models for Multi-Temporal Remotely Sensed Imagery", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60, pp. 109–124.

Strickland, C.M.; Denham, R.J.; Alston, C.L. and Mengersen, K.L. (2013) "PyMCMC : A Python Package for Bayesian Estimation using Markov Chain Monte Carlo", in

C.L. Alston, K.L. Mengersen and A.N. Pettitt (eds.), *Case Studies in Bayesian Statistical Modelling and Analysis,* John Wiley, London, pp. 421–460.

Tam, S.M. (1987) "Analysis of Repeated Surveys Using a Dynamic Linear Model", *International Statistical Review*, 55, pp. 67–73.

Tam, S.M. and Clarke, F. (2015a) "Big Data, Official Statistics and Some Initiatives of the Australian Bureau of Statistics", *International Statistical Review* (to appear).

Tam, S.M. and Clarke, F. (2015b) "Big Data, Statistical Inference and Official Statistics", *Methodology Research Papers*, cat. no. 1351.0.55.054, Australian Bureau of Statistics, Canberra.
<http://www.abs.gov.au/ausstats/abs@.nsf/mf/1351.0.55.054>

Wright, R.L. (1983) "Finite Population Sampling with Multivariate Auxiliary Information", *Journal of the American Statistical Association*, 78(384), pp. 879–884.
Wu, L. (2010) *Mixed Effect Models for Complex Data*, CRC Press, Florida.

We are interested in fostering review of books and software in the area of survey methods. This would include standard review of individual books or software packages. This may also include broader reviews of groups of text and monographs in specific sub-areas; or similarly broad reviews of available software. Of particular interest are some of the new R libraries that have been developed recently for survey methods. If you are able to write a review for this section, please contact Natalie Shlomo (natalie.shlomo@manchester.ac.uk).

ARGENTINA

Reporting: Veronica Beritich

14th Meeting of the Washington Group on Disability Statistics

On October 8-10, 2014 was held in the City of Buenos Aires (for the first time in Argentina) the 14th Meeting of the Washington Group on Disability Statistics, organized by the Instituto Nacional de Estadística y Censos (INDEC).

The Washington Group operates within the framework of the United Nations Statistical Commission. It was created in 2001 with the objective of developing the conceptual and methodological elements necessary to implement and update a set of harmonized questions for measurement of disability in censuses, surveys and administrative records in the context of demographic and social statistics.

During these three days more than 50 specialists in this field, representing various countries from all continents, shared their experiences in the production of data on disability.

On the other hand, INDEC technical teams integrating the Washington Group provided information on the population with permanent physical and/or mental difficulties or limitations (PDLP) in our country, based on an analysis of data arising from the 2010 National Census of Population, Households and Dwellings.

The categories incorporated in the Census have allowed for obtaining data on the different types of limitations: visual, hearing, upper motor, lower motor and cognitive, by sex and age.

Some of the data from the census indicate that:

- The percentage of PDLP is 12.9% of the people living in private dwellings.
- The majority of people in this group declared having only one permanent difficulty or limitation.
- From those who declared having only one permanent difficulty or limitation, approximately 60% declared to be only visually impaired.
- Motor difficulties (upper and lower) affect less than 24% of this population group.
- The impaired auditory and cognitive, in turn, represent something more than 8%, in each case.

- Please note also that the prevalence of permanent difficulty or limitation increases as the population ages.
- Starting from the age of 15, the percentage of affected women is higher than that for men.
- Cognitive difficulties predominate in the earliest ages.
- The PDLP with 14 years and older only 47.7% corresponds to the Economically Active Population (EAP) and 44.6% is occupied.

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

For further information, please contact ces@indec.mecon.gov.ar.

BOSNIA AND HERZEGOVINA

Reporting: Edin Šabanović

Pilot Income and Living Conditions Survey in Bosnia and Herzegovina 2015

Statistics on Income and Living Conditions Survey (SILC) in European Union is produced by the survey based on a sample of households that aims to gather both household- and person-level information. This survey has rules that apply to both types of statistical units and it collects data on inome and living conditions, employment, health and material deprivation of households and their members. The overall objective of the SILC survey is to collect, produce and disseminate information on income level and structure, as well as to measure poverty and living standard in the country, which are calculated according the EU methodology and regulations.

Pilot ILC survey in Bosnia and Herzegovina was conducted in the first quarter 2015 as a pilot survey whose main objective is to test survey methodology, data collection method and field work organization. For the first time, statistical institutions in Bosnia and Herzegovina implemented Computer Assisted Personal Interview (CAPI) data collection method. Pilot survey results will be only used for the preparation of the full-scale survey in the nearest future.

It was realized on the sample of 340 households in the whole country, and 17 interviewers and 6 supervisors have done the field work. CAPI application was made in Blaise software, while the data analysis will be done in SPSS. The micro data file will be transmitted to Eurostat, as well as the Quality Report.

Pilot ILC survey in Bosnia and Herzegovina is fully funded by IPA 2012 Multi-beneficiary program on statistical cooperation and supported by technical assistance of Eurostat experts and GOPA consultants.

For more information, contact Edin Šabanović (edin.sabanovic@bhas.ba), Sector for Statistical Methodology, Standard, Planning, Quality and Coordination, Agency for Statistics of Bosnia and Herzegovina.

CANADA

Reporting: François Brisebois

A Redesigned Health Survey – the Canadian Community Health Survey

Created in 2000, the annual component of the **Canadian Community Health Survey** (CCHS) is a cross-sectional survey that provides information related to health status, health care utilization and health determinants for the Canadian population. It relies upon a large sample of respondents and is designed to provide reliable estimates at the health region level.

In 2015, the survey underwent a full-scale redesign for the first time since 2008 when it had adopted a continuous collection design. The 2015 redesign had seven specific objectives including the complete revision of the questionnaire content, the adoption of Statistics Canada's new Household Surveys Frame Service, the revision of the sample allocation strategy, the expansion of the coverage of the CCHS to children under 12 years old, and the adoption of an internet data collection mode. Although some developmental work was done for the coverage of younger children and the adoption of the internet for data collection, operational and budget constraints postponed the implementation for these two objectives.

The resulting redesigned CCHS was implemented in January 2015. In anticipation of the expansion of the coverage to children under 12 years, two distinct methodologies have been developed, one for children 12 to 17 years old, and one for adults 18 years and older. For the 12-17 population, children were randomly selected from an administrative file available at Statistics Canada and offering a very extensive coverage of the population under 18 years old. Using the contact information available on the file, CATI interviewing is used for the survey. The plan is to add an internet option once the development of a new infrastructure supporting this collection mode is completed. For the adult population, the survey design now relies on one single dwelling-based survey frame, extracted from the new Household Surveys Frame Service, and replacing the complex and less efficient dual frame approach used in the past. The collection strategy has also been adapted to optimise the use of contact information also available through the Frame Service. The strategy will build on the CATI and CAPI modes, where CATI will be promoted to minimize collection costs. CAPI will mainly be used to contact the dwellings for which the frame does not offer a phone number, or as a last resort tool to convert nonresponse in health regions showing poor performance in this matter.

For more information, contact François Brisebois, Chief methodologist, Household Survey Methods Division, Statistics Canada, at francois.brisebois@statcan.gc.ca .

FIJI

Reporting: M.G.M. Khan

Household Income and Expenditure Survey (HIES) & Household Listing Exercise (HLE) 2013-14

The Household Survey Unit Division in Fiji Bureau of Statistics (FBoS) is updating the household and population specifically targeting the small area data by locality within the Enumeration Area. This will enable FBoS to release accurate data by locality when requested by the users. It will also enable FBoS to compare the result with the 2007 Census and plan for the best way to improve the data collection for the next Census in 2017.

Data cleaning processes for both surveys are ongoing and a report to be released at the end of this year or early next year.

Use of R Software in FBoS

A team from FBoS attended a one and half weeks training on the use of R software in June last year at the University of the South Pacific which was facilitated by Dr. MGM Khan, the Associate Professor in Statistics of the School of Computing, Information and Mathematical Science. The training was very informative and well conducted.

Since the software is more user friendly and free of charge it was decided by FBoS to use "R" for analysis purposes. After the training in June, FBoS formed an R-User Group within the organization. Currently, the group is using the software in generating tables for both the above surveys. In the long term FBoS will be replicating SAS programmes with R as a form of backup.

FBoS and Stakeholders

Last year FBoS had internal presentations on how each section collected and compiled statistics. The objectives are for a better understanding and improving statistics compilation and reporting and awareness of the different data collection portals within the Bureau.

The next steps include engaging with stakeholders and highlighting FBoS data requirements that would encourage improved data sharing arrangements.

For further information on Fiji's External Debt Statistics Release contact: Ms Sashee Nath (Statistician Balance of Payments) on email: snath@statsfiji.gov.fj

ISRAEL

Reporting: Tom Caplan

IASS Country Report: Israel April 2015

Since our last country report, Israel has seen significant developments in statistical activity. The activity has come as a result both of meeting national needs and because of international requirements and cooperation. In Israel, since the last report there have been two major internationally based events that have impacted significantly on Israel's national statistical program . The first is Israel's accession, in the year 2010 to the Organization of Economic Cooperation and Development (OECD) and the second is the implementation of a large encompassing Twinning program between the Israel Central Bureau of Statistics (ICBS) and Statistics Denmark, under the auspices of the European Union. In this report we provide an overview of some of the new developments. In future reports we may examine some of the new programs in more detail.

Countries accessing to the OECD are faced with conditions and requirements in order to be accepted for accession. These requirements include the statistical area. These requirements not only serve the statistical needs of the OECD but they enhance and develop the statistical program of the accessing country. As a result of the OECD accession, Israel saw the expansion of its statistical program. The following describes some of the new programs. Among the new programs, a new Job Vacancy Survey (JVC) was implemented with a sample of 4300 businesses, of which 1000 businesses are surveyed each month coming to 3000 surveyed each quarter. On the basis of this survey there is enhanced information about the demand side of the labour market. This survey provides needed new information and combining its results with the results of the Labour Force Survey allows the publication of a quarterly combined report on labour supply and demand.
Two other new surveys that have been implemented in conjunction with joining the OECD are the Business Tendency Survey and the Consumer Confidence Survey. The Business Tendency Survey surveys a subsample of the JVC sample of businesses and asks them about the state of the business's current financial situation, the company's business experience in the previous quarter and expected changes in the company, in the national and the international economy. The Consumer Confidence Survey surveys monthly, members of the Social Survey sample (spread over twelve months) and asks them their views of their personal and family financial situations as well as theirs views of the country's economic situation and expected economic situations.

A very important result of joining the OECD is the implementation of the PIAAC survey in Israel. PIAAC is the Program for the International Assessment of Adult Competencies. It focusses on examining the skill levels of adults in a range of areas. The emphasis is on key cognitive and workplace skills (literacy, numeracy and problem solving technology). There is a consistent approach to the assessment across all OECD countries.

The European Union – Israel Twinning Project on Statistics began officially on May 23 2013 and the closing ceremony was held on December 17, 2014. Under the project, the Israel Central Bureau of Statistics (ICBS) was twinned with Statistics

Denmark and a program was set up toward the improvement/development of statistical programming in five key component areas: National Accounts, Education Statistics, Coordination of the National Statistics System and Strategic Planning, Survey Methodology and Dissemination and the Israel Central Bureau of Statistics Internet Website. Throughout the program there were approximately 40 activities between the two countries involving either experts from Denmark coming to Israel to provide expert advice or Israel statisticians traveling to Denmark on study visits or other forms of activities. The expert advice was provided by means of workshops, seminars and one on one consultations. The focus was on reaping from the Danish expertise in order to develop or advance the Israeli situation in the five key areas. At the outset of the program practical objectives were set out for each component and these were all successfully met. The full final report of the Twinning Project can be seen on the ICBS website. Some of the major results of the Twinning project include:

- Improvements in survey methodology, more specifically in management of surveys and quality assurance of the work of interviewers in the field and by telephone, measurement and improving response burdens, support in designing and writing internet questionnaires, etc.
- The improvement and further development of many important parts of the National Accounts, (according to EU and other international guidelines and recommendations*)* including Government Accounts, Financial Accounts and Balance of Payments.
- Planning of new statistics on Education as well as Culture and Sports and further development of existing statistics specifically in the areas of dropouts, higher education and adult education
- The recognition that there is a National Statistics System. Seminars were held producers of official statistics and focus groups with users. There are many producers of official statistics in addition to the ICBS and they welcomed the coordination of the Government Statistician and the Public Council for Statistics.
- The writing of a Strategic Plan for the ICBS
- The setting up a team for the development and writing of metadata according to a standardized, internationally recognized approach.

In addition to the new or developed statistical activities that have come as a result of these international processes there are other significant developments that have occurred since the last report. One is the implementation of a Longitudinal Survey of Families that focusses on economic social and financial topics. It is now collecting data for its third wave. There is a new survey of health and as part of trials for using biometric personal identification systems there is a new survey on satisfaction concerning international border crossing using biometric passports. All of these will be reported on in more detail in future country reports.

NEW ZEALAND

Reporting: Felibel Zabala

The 'spine' is the new linking model for our Integrated Data Infrastructure

Statistics New Zealand has improved the linking model for the Integrated Data Infrastructure (IDI). The IDI is a linked longitudinal dataset of NZ Government administrative data used for research purposes. The data is held in a controlled, protected environment ensuring individuals are not identified.

Under the new linking model, each dataset links to a central prototype 'spine', which includes New Zealand birth information, tax information from Inland Revenue, and information on visas for people entering New Zealand (excluding visitor and transit visas). It replaces the previous model which was limited primarily to linking New Zealand Government data to Inland Revenue data.

The target population for the prototype spine is all people who have been in New Zealand for a length of time (who potentially have significant interaction with government services). The aim of the spine is to have a complete list of uniquely identified members of the target population and a minimum of duplicate records (ie minimised coverage error).

One of the deficiencies in the previous model was the low coverage of children in tax data. This has been improved by adding births to the spine. We have also started looking at the quality of the spine, in terms of false positives (incorrect links), false negatives (missed links), and missing populations. This will allow incremental improvement in spine quality and coverage.

See Integrated Data Infrastructure on www.stats.govt.nz for further information or contact Andrew Black at andrew.black@stats.govt.nz

Census Transformation in New Zealand – Exploring the feasibility of producing census information from administrative data sources

Statistics New Zealand's Census Transformation strategy has two strands: to modernise the current census model in the short- to medium-term; and to investigate alternative ways of producing small-area population and socio-demographic statistics in the longer term. This includes the possibility of changing the census frequency from every five to every 10 years and exploring the feasibility of producing census information from administrative data sources. The main focus of investigations to date is on understanding the potential for using administrative sources to produce census information.

Overview of progress on the potential use of administrative data for census information in New Zealand summarises findings from work carried out during 2013 that were presented to government in February 2014. The progress report concluded that while existing administrative data sources cannot at present act as a replacement for the current census, early results have been sufficiently promising that it is worth continuing investigations into the use of administrative sources for producing census information.

See Future approaches to social and population statistics on www.stats.govt.nz for Census transformation research papers, including:

- Evaluating the potential of linked data sources for population estimates: The Integrated Data Infrastructure as an example, which describes a method to construct population estimates from linked administrative data sources. These are compared at an aggregate level against the official estimated resident population.
- An initial investigation into the potential for administrative data to provide census long-form information is a high-level assessment of the potential for administrative sources to replace the current census attribute (or long form) information, based on metadata comparisons.
- Coverage assessment in an administrative census: A progress report on issues and methods is an initial identification of issues and potential methods for coverage assessment and population estimation for an administrative census

We are primarily focused on the potential of linking multiple existing administrative data sources, since New Zealand does not have a national population register. Assessments are now being undertaken using more detailed individual-level comparisons with the 2013 Census linked to the IDI specifically for the purposes of this research.

Methodological work underway includes:

- investigating population coverage and address misclassification of New Zealand residents identified in the IDI using rules based on recent 'activity', and researching and developing methodologies for coverage assessment, content validation, and population estimation
- investigating attribute data in administrative sources for both the 2018 Census or a potential future administrative census. Topics include family and household, education and training, ethnicity, income, and work, as well as housing/dwelling information.
- preparing a preliminary sample design and costings for a large-scale continuous attribute sample survey, along the lines of the American Community Survey.

As well as the methodological work, we are engaging with key users of census information to develop specific quality criteria for future censuses as well as holding discussions with wider government on impacts on the electoral system and with other cross-government initiatives that are driving changes to government systems. The next major report to government will be in October 2015, with the aim of agreeing a preferred future direction for the New Zealand census.

See Census transformation in New Zealand on www.stats.govt.nz for further information or contact Tracey Savage at tracey.savage@stats.govt.nz.

PALESTINE

Reporting: Abdulhakeem Eideh

Dr. Abdulhakeem Eideh- Best paper award in the field of Sampling

The Indian Society of Agricultural Statistics has instituted prizes for the best papers in different fields published in the Journal of the Indian Society of Agricultural Statistics. Accordingly, the papers published in the Journals, Volume 66 (2012) and Volume 67 (201 3) have been evaluated for judging the best papers. The paper entitled Estimation and Prediction under Nonignorable Nonresponse via Response and Nonresponse Distributions by Abdulhakeem AH Eideh (Department of Mathematics, Al-Quds University, Palestine) published in Volume 66, o. 3, December, 2012, pp.359-380 has been selected for best paper award in the field of Sampling.

Energy Consumption Survey in Palestine

The Palestinian Central Bureau of Statistics (PCBS) start conducting the survey on energy consumption in transport sector, under the **project "**Strengthening Statistical Capacity of Arab Countries in Producing Energy Statistics and Energy Consumption in Transport Sector Surveys". The overall objective of this project, funded by the Islamic Development Bank for one year, is to strengthen the capacity of National Statistical Offices (NSOs) in three member countries Egypt, Jordan, and Palestine, of the Economic and Social Commission for Western Asia (ESCWA) in improving the information on energy products consumption in the transport sector in order to assist governments in more effectively managing energy consumption in the countries.

Contributions from IASS Members

Sampling design data file

Seppo Laaksonen
University of Helsinki
E-mail: Seppo.Laaksonen@Helsinki.Fi

Keywords: Auxiliary variables, calibration, data quality, fieldwork, inclusion probability, non-response

Abstract: The paper first determines the term 'sampling design file' that is not commonly used in survey sampling literature. The methodology behind this term is, of course, used to some extent, but only implicitly. Its explicit determination facilitates many things in survey practice and also gives a clear target for one big part of a survey, that is, sampling, fieldwork and finally for estimation. The sampling design file consists of all the gross sample units and its variables include those that give opportunity to create sampling weights, to analyse the survey quality, and to estimate. The file is possible to complete after the fieldwork. Its most important characteristics, including sampling design variables and weights, will be finally merged together with the real survey variables at respondent level, and then the survey analysis is ready to begin.

1. Introduction

The term 'sampling file' or more broadly 'sampling design data file' is rarely used in standard survey literature. One of this first users is the sampling expert panel of the European Social Survey (ESS) that was established in 2001 (see more information about this survey that initially started in 2002, europeansocialsurvey.org). The document of the panel says: "The Sampling design data file **(**SDDF) is routinely generated by an ESS country's National Coordinator after fieldwork has finished. It includes information on the implemented sample design such as inclusion probabilities and clustering. As such, it serves the sampling team with the data required for computation of design weights, design effects and as a general basis for benchmarking the quality of sampling. The ESS analyst may use it for several purposes such as incorporating cluster information in her/his analyses."

A SDDF is required for all types of surveys, thus for surveys from households, individuals, businesses and corporations. Here we concentrate on surveys of individuals who are members of households.

2. Basic targets of sampling file

The statistical units of the sampling file should ideally cover all the gross sample units of the survey. Such units are selected addresses in the case of address-based samples (but there are individuals behind these addresses or dwelling units), and selected individuals in the case of individual-based samples. In the end, the file of these statistical units thus covers the respondents, the non-respondents and the in-eligibles. It might be difficult to completely numerate in-eligibles for the file, since any contact for some individuals/addresses cannot be made and hence the file may be inaccurate, but all efforts to complete the file with appropriate information should be done. It follows that such a unit may thus be either an in-eligible or a non-respondent. Correspondingly, some bias in estimates necessarily follows.

The *first-order sampling file* is good to create while the gross sample has been drawn. In this case, the file includes:

- Non-confidential and confidential identifier
- Sampling frame variables and respective statistics
- Stratification variables, explicit strata in particular
- Implicit strata if they include useful information; implicit stratification specifies the order of the units selected by equidistance or other systematic selection but basically this design corresponds to a simple random selection
- Inclusion probabilities of each stage within explicit strata.

In the case of multi-stage sampling, all inclusion probabilities may not be available before the end of the fieldwork. This is typical in a three-stage sampling if the primary sampling units (PSU) are small-areas and the secondary sampling units (SSU), respectively, are addresses or households, but the third stage units are individuals. This missingness for the third stage units is due to the problem of contacting a dwelling unit or an address in order to know how many target population members there exists. Even in register-based countries such information is hard to get correctly, since the register is not up-to-date for a survey period.

It is possible and also useful to calculate the gross sample design weights immediately when the first-order file is available. This gives opportunity to check basic figures and the quality of the sampled file of this phase. For example, when summing up these design weights we should obtain the correct target population statistics that represent the final target population if no missingness occurs. In contrast, if the third-stage units, for instance, are missing, the target population of the households or addresses can only be computed.

The above variables derived from a sampling frame are minimal requirements but not sufficient. It is rational at the same occasion to download other useful information for the sampling file from the sample frame that we call here the *second-order sampling file*. For example, in register-based countries, the sampling frame has been created from the population register, that is reasonably up-to-date. The sampling design only requires aggregate population statistics by large region, age group and gender, for example. But the same information can be matched at micro level into gross sample units too. In addition, the same data source consists of many other variables that are beneficial to download to the second-order sampling file at the same time since it is basically free of charge. It is not common even in Finland to distend over the minimum although it is possible to expand the file with the following auxiliary variables, among others: marital status, year of marriage, multi-marriage, number of children, house size, type of house, citizenship, mother tongue, coordinates of the house and municipality at birth.

The second-order sampling file can further be completed from other sources at the same time as the first-order file has been created. This usually may require some additional administration and paper work but it is best to do as soon as possible since the data sources cannot be up-to-date for long, or even some data are destroyed. Typical other sources are: formal education, tax register information on income and wealth, jobseekers' register. Section 4 presents a Finnish example on this issue in more details.

The *third-order sampling file* can be created as soon as the fieldwork has been completed. In this case, the most important new variable is the outcome of the fieldwork that indicates who is a unit respondent, and who is a non-respondent and an in-eligible, respectively. As said above, the last two categories are often hard to definitely determine with accuracy. This seems to be a worsening problem in Europe due to more or less permanent absence of the official address (home). A reason for

this is working outside the country over several months, or using a second home in another country, respectively.

A drawback, in many countries as said already above, is that all inclusion probabilities cannot be known after the fieldwork. In the ESS, the selection of one individual within the selected household or address is a good example. As a consequence, it is not possible to calculate a complete inclusion probability for the individuals of the gross sample, but only for the second stage address/household. The sampling weight for the respondent can be, nevertheless, calculated, assuming, for example, that the response mechanism for the third stage is ignorable within strata.

After the fieldwork, the sampling file can be further reinforced with other data on fieldwork. Opportunities for that are dependent also on the survey mode used. Face-to-face interviewers can collect information about the quality of the location where a potential respondent lives. For example, an interviewer can classify the quality of the living area or the type of house. This indicator is of course useful only if a valid measurement is available and the same information is available both for the respondents and for the non-respondents. Moreover, the interviewer information (e.g. their basic characteristics, attitudes toward this survey) can be added to the sampling file too.

3. What to do with sampling file?

The sampling file is necessary in order to calculate the sampling weights for the respondents. This requires that the inclusion probabilities are available in the file. Naturally, the identifiers of the respondents should be available in the sampling file in order to match the sampling weights and other sampling design variables into the survey data file of the respondents.

The narrowest correct sampling file is such that the sampling design is simple random sampling. In this case, the file consists only of an identifier and one constant inclusion probability, and the survey outcome variable that identifies the respondents, the non-respondents and the in-eligibles. These data allows the calculation of a single sampling weight for each respondent. No real non-response analysis can be done due to completely missing auxiliary data.

If a two- or three-stage design has been used, there are more variables, including PSU's as clusters, and SSU's, respectively. Even though there are no other strata or auxiliary variables, it is possible to review non-response by PSU and SSU, respectively. This gives the opportunity to adjust the weights to some extent since non-response may vary by SSU conditional to PSU. Hopefully, all PSU's are still in the file. Otherwise, the fieldwork has failed.

The sampling file is primarily needed to create the weights for the respondents although it is best to first create weights for the gross sample. Secondarily, the file is for analysing the success of the fieldwork. It is possible that a particular survey may use more than one survey mode, like in the case of a mixed-mode design. The sampling file naturally must include the mode used in data collection for all individuals. If two or more modes are used for one individual, this should be coded at variable level as well.

A good sampling file is naturally very useful to analyse survey quality. Auxiliary variables particularly are needed for this purpose. Also, we would be happy if some variables of the fieldwork file would be merged with the sampling file.

4. Auxiliary variables in the sampling file

We have above given examples of auxiliary variables of a good sampling file. Now, we concretise this issue. It is good to recognize that all such variables are given for individual gross sample units whatever they are. In the case of multi-stage sampling, such variables can be more problematic since they are first concerned with clusters of the target units. There can thus only be such variables that are related to clusters. If the clusters are small-areas, regional information is available. However, it is more difficult to know, for example, about the education of all cluster persons. This may not be necessary since it is more important to gather information about the education of the respondents and the non-respondents within this cluster.

Auxiliary variables can thus be either **macro** or **micro**. Both of these variables are useful and even for the same purpose; they can be derived from the same basis. For example, age of an individual can be used in non-response analysis in several forms, such as individual ages or as groups. However, the same variable is useful as target population statistics and thus a macro auxiliary variable would indicate how many target population members are in each age group. This is an example of the benchmarking information, and they can be used in calibration methods that require macro auxiliary data, that is, known population margins (e.g. Deville and Särndal 1992). There can be several population margins in calibration at the same time. And if such information is available in the sampling file, it is easy to compute the calibrated weights, respectively, using the French software Calmar 2, among others (see Le Guenne & Sautory 2005).

Macro auxiliary variables can thus be margins of known population figures giving opportunity to use these in calibration. They can also be relative frequencies of small areas like PSU's, concerning for instance register unemployment rates, rates of highly educated people, or crime and poverty rates. Such variables could be used for analysing reasons of nonresponse.

The richness of the auxiliary variables in the sampling file facilitates in analysing the success of the fieldwork. For example, unit non-response can be assessed against these variables and the multivariate response propensity model estimated as a result. This model may respectively be a good starting point for adjusting the sampling weights to take into account the variation in non-response (e.g. Laaksonen 2007, Laaksonen and Heiskanen 2013)).

The sampling file should be explicitly available, that is, for all gross sample units, and all inclusion probabilities should be in the file. Sometimes, these probabilities are only implicitly available. For simple random sampling it is most common since the inclusion probabilities are unique and only requires one population statistics figure and the gross sample size. Hence it is impossible to check based on this probability that everything has been done correctly. Thus, is SRS really good or not?

Another difficult situation is two-stage sampling when the equal absolute sample sizes are used in the second stage. This leads to final inclusion probabilities in which the PSU sizes of the first stage clusters will disappear. It means that this size is not necessarily needed in the formula of the inclusion probability. Unfortunately, there exists sampling files where there is only one 'final' probability of this kind. One example is in Burnham et al (2006) that Laaksonen (2008) criticised due to missing inclusion probabilities and in particular that all concrete information about first stage sampling is missing. So, it is possible that everything has not been done correctly since sampling design information is lacking. This is true for all designs, even in simple random sampling, since the sampling data file is so restricted that very little

can be checked. Good auxiliary data (macro and micro) also lever confidence in the survey data and hence it should be recommended to collect.

5. A Finnish example

In the end, an example from the Finnish security survey (FSS) 2010 is presented (Aromaa 2010, Laaksonen and Heiskanen 2014). The characteristics of its sampling data file are given below.

The number of statistical units of the FSS is 7933. They are thus gross sample units. Table 1 illustrates the variables of the sampling file.

This list is rather long, and good in many meanings. It gave opportunity to analyse non-response by various auxiliary variables. Based on the data, we also created the so-called adjusted sampling weights. This first exploits the response propensity modelling and finally the stratification based on such calibration that the known population statistics match with our gross sample design weights by strata.

Naturally, we used the data also for survey quality including the analysis of problems in the fieldwork. This was possible for two reasons: (i) based on paradata, we were able to follow the interviewing time that was shortening during the fieldwork; the response time vary by mode as well so that web took least time and face-to-face the most time, (ii) we made a special survey for the interviewers and found that the point (i) was in telephone interviewing due to the busy call schedule at the end of the fieldwork. Naturally, the results were not ideal.

Our sampling file thus is rich but it is not common everywhere. The file content also depends on the survey practice. Our European Social Survey team has found various interesting contents that should be included in the sampling file. One is the so-called reserve sample that is initially created to guarantee that enough respondents will ultimately be found. It is clear that this reserve should be probability based, but if the reserve part is not included in the sampling file, it will be hard to follow the fieldwork well and even to calculate correct response rates. This reserve sample option is now in our template. It is interesting that a certain country found this option in our sampling file and incorrectly wanted to take a reserve sample even though this was not in their sampling file.

6. End notes

I sincerely hope that survey organisers will pay attention to create as good a sampling data file as possible and such that it would help in getting improving estimates from the survey. Unfortunately, this concept is not currently in standard literature. Hopefully it will be so in a future.

References

Burnham, G., Lafta, R, Doocy, S. and Roberts, L. (2006) Mortality After the 2003 Invasion of Iraq: a Cross-sectional Cluster Sample Survey. *The Lancet* **368,** 1421–1428.

Deville, J-C. & Särndal, C-E. (1992) Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 376-382.

Laaksonen, S. (2007) Weighting for Two-Phase Surveyed Data. *Survey Methodology,* December Vol. 33, No. 2, pp. 121-130, Statistics Canada.

Laaksonen, S. (2008) Retrospective Two-Stage Cluster Sampling for Mortality in Iraq. *International Journal of Market Research* 50, 3, 403-417.

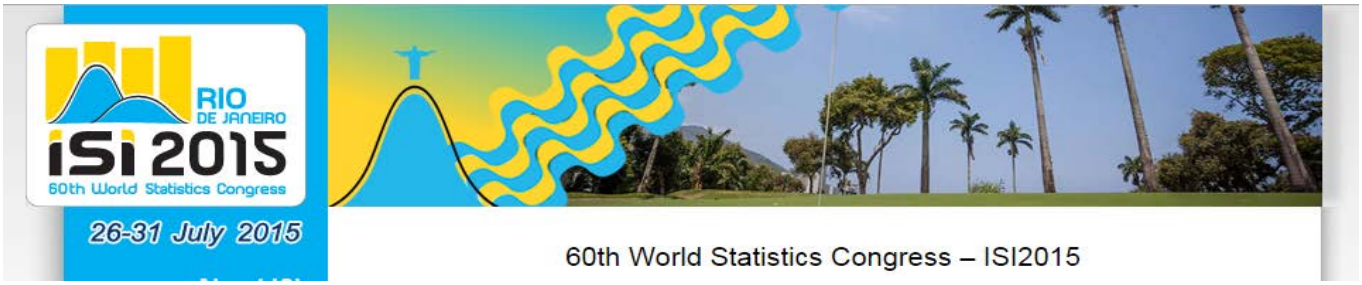
Laaksonen, S. and Heiskanen, M. (2014). Comparison of Three Modes for a Crime Victimization Survey, *Journal of Survey Statistics and Methodology 2 (4): 459-483 doi:10.1093/jssam/smu018*

Le Guenne, J. & Sautory, O. (2005) CALMAR 2 : Une Nouvelle Version de la Macro Calma de Redressment D'Échantillion Par Calage. http://vserver-insee.nexen.net/jms2005/site/files/documents/2005/327_1-JMS2002_SESSION1_LE-GUENNEC-SAUTORY_CALMAR-2_ACTES.PDF

Table 1. **The key variables of the sampling data file for the Finnish Security Survey.** Symbols in the column 'Source': SO = Created by survey organisation, M = Computed by methodologist, PR = Population Register, FER = Formal Education Register, ER = Employment Register, TR = Tax Register. The alternatives for use: R = Merging with respondent data, S = Sampling, U = Unit non-response, W = Weighting, E = estimation

Variable	**Source**	**Use for**
Identifier	SO	R
Survey mode (face-to-face, telephone or web)	SO	W E
32 Explicit strata, anonymous code	M	S U W E
PSU's, anonymous code	M	S U E
PSU size, Stratum size (incl. size of the target population)	M	S E
1st stage inclusion probabilities for PSU's and 2nd stage probabilities for households, 3rd stage probabilities for individuals	M	S W E
Age in years Age group respectively, both micro codes and macro statistics	PR	S U W
Gender, code and macro statistics	PR	S U W
Regional variables including municipality, postal code, co-ordinates of home, code and for some also macro	PR	S U W
Marital status with different options, year of marriage, number of marriages, code	PR	U W
Native language and citizenship	PR	U W
Occupational or socio-economic status (fairly rough only available)	PR	U
Household composition including number of children at different age groups	PR	U W
House variables such as size, number of rooms and type of kitchen	PR	U W
Level and field of education	FER	U W
Unemployed or not, number of months unemployed	ER	U W
Taxable income	TR	U W
Fieldwork outcome (respondent, non-respondent, in-eligible)	SO	M U W E
Neighbourhood variables in face-to-face surveys (option)	SO	U W
Reserve sample indicator if used, responsive design indicator respectively	SO	U W
Reason for non-response (well for face-to-face, badly for web)	SO	U
Para data, e.g. interviewing time, responding time	SO	E

Upcoming Conferences and Workshops



60th ISI World Statistics Congress

Organized by: International Statistical Institute
Where: Rio de Janeiro, Brazil
When: 26.07.2015 to 31.07.2015
Homepage: http://www.isi2015.org

We are delighted to invite you to the 60th ISI World Statistics Congress (WSC), which will take place in Rio de Janeiro, Brazil, during 26–31 July 2015.

The WSC is the flagship conference of the International Statistical Institute (ISI) and its seven associations. It is a biennial conference with a rich tradition, and IBGE is pleased to host and organize ISI2015 in Brazil.

The congress will bring together members of the statistical community to present, discuss, promote and disseminate research and best practice in every field of Statistics and its applications. The Scientific Programme of the 1512015 will include a wealth of activities that will cover stimulating topics and will offer delegates innovative and well-balanced presentations, as well as plenty of opportunities for discussion and exchange.

A rich and exciting Social Programme is also being developed, with plenty to see and enjoy for participants and their accompanying persons, hoping to make your trip to Rio and taking part in ISI2015 a truly unforgettable experience.

The venue - Riocentro - is located in Barra da Tijuca, a district surrounded by natural beauty but also many sophisticated bars, restaurants and several malls and close to a variety of historical and cultural programs that only the Wonderful City can offer.

We are confident that all the ingredients are in place to ensure that the 60th ISI World Statistics Congress will be a memorable statistical event!

For further information please email Francisco Samaniego fjsamaniego@ucdavis.edu

Centro UC
Encuestas y Estudios
Longitudinales

FIRST LATIN AMERICAN ISI SATELLITE MEETING
ON SMALL AREA ESTIMATION
3-5 AUGUST 2015, SANTIAGO, CHILE

SAE 2015 santiago

First Latin American ISI Satellite Meeting on Small Area Estimation (SAE 2015)

Organized by: International Statistical Institute (ISI), the International Association of Survey Statisticians (IASS), the Sociedad Chilena de Estadística (SOCHE), the Instituto Nacional de Estadísticas (INE), the Ministerio de Desarrollo Social (MDS), and the Universidad Católica de Chile (Departamento de Estadística, Departamento de Salud Pública e Instituto de Sociología)

Where: Santiago, Chile

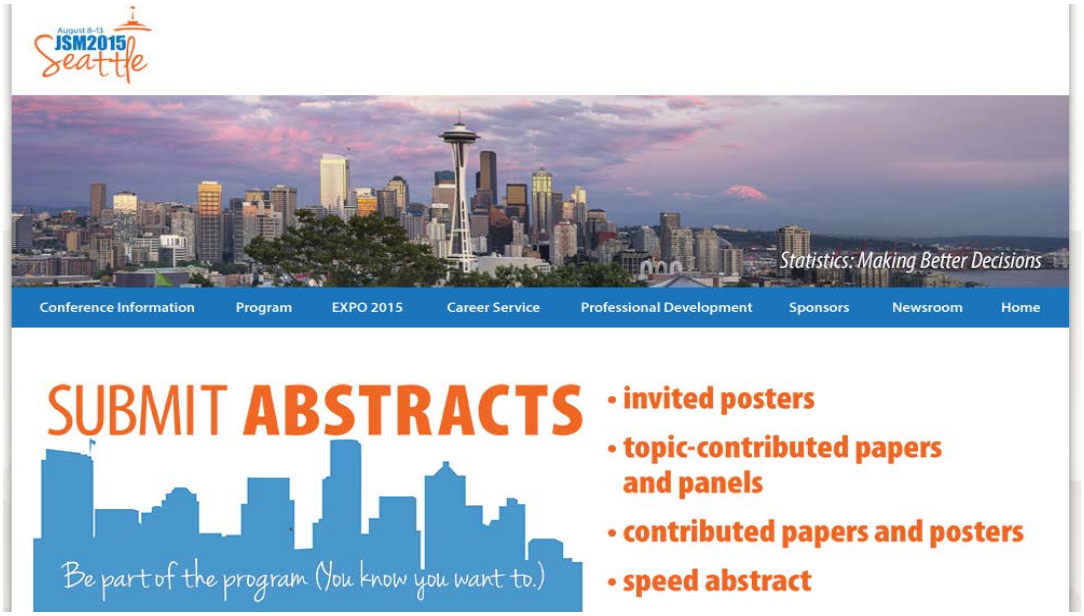
When: 03.08.2015 to 05.08.2015

Homepage: http://www.encuestas.uc.cl/sae2015/index.html

Welcome to the website of the First Latin American ISI Satellite Meeting on Small Area Estimation (SAE 2015), to be held at the Pontifical Catholic University of Chile on August 3-5 of 2015.

The SAE 2015 conference is co-sponsored by the International Statistical Institute (ISI), the International Association of Survey Statisticians (IASS), the Sociedad Chilena de Estadística (SOCHE), the Instituto Nacional de Estadísticas (INE), the Ministerio de Desarrollo Social (MDS), and the Universidad Católica de Chile (Departamento de Estadística, Departamento de Salud Pública e Instituto de Sociología).

For more information, please visit the homepage or contact sae2015@uc.cl.



2015 Joint Statistical Meetings

Organized by: American Statistical Association, International Biometric Society (ENAR and WNAR), Institute of Mathematical Statistics, Statistical Society of Canada, International Chinese Statistical Association, International Indian Statistical Association, Korean International Statistical Society, International Society for Bayesian Analysis, Royal Statistical Society, and International Statistical Institute
Where: Seattle, USA
When: 08.08.2015 to 13.08.2015
Homepage: http://www.amstat.org/meetings/jsm/2015/index.cfm

JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the *American Statistical Association, *International Biometric Society (ENAR and WNAR), *Institute of Mathematical Statistics, *Statistical Society of Canada, International Chinese Statistical Association, International Indian Statistical Association, Korean International Statistical Society, International Society for Bayesian Analysis, Royal Statistical Society, and International Statistical Institute. Attended by more than 6,000 people, meeting activities include oral presentations, panel sessions, poster presentations, professional development courses, an exhibit hall, the Career Placement Service, society and section business meetings, committee meetings, social activities, and networking opportunities

Seattle, Washington, the host city for JSM 2015, offers a wide range of options for sharing time with friends and colleagues or sightseeing with family. For information, contact meetings@amstat.org.

The 2015 Joint Statistical Meetings will be held August 8–13, at the Washington State Convention Center, 800 Convention Place, Seattle, WA 98101.

The Fourth Baltic-Nordic Conference on Survey Statistics -
BaNoCoSS-2015 -
will be held on 24-28 August 2015 in Helsinki, Finland

Fourth Baltic-Nordic Conference on Survey Statistics BaNoCoSS 2015

Organized by: University of Helsinki
Where: Helsinki, Finland
When: 24.08.2015 to 28.08.2015
Homepage: https://wiki.helsinki.fi/display/banocoss2015/4th+Baltic-Nordic+Conference+on+Survey+Statistics

The Fourth Baltic-Nordic Conference on Survey Statistics - BaNoCoSS-2015 - will be held on 24-28 August 2015 in Helsinki, Finland.

BaNoCoSS-2015 is a scientific conference presenting developments on theory, methodology and applications of survey statistics in a broad sense.

The conference provides a platform for discussion and exchange of ideas for a variety of people. These include, for example, statisticians, researchers and other experts of universities, national statistical institutes, research institutes and other governmental bodies, and private enterprises, dealing with survey research methodology, empirical research and statistics production. University students in statistics and related disciplines provide an important interest group of the conference.

BaNoCoSS-2015 is organized by the Baltic-Nordic-Ukrainian Network on Survey Statistics, University of Helsinki, Statistics Finland and The Finnish Statistical Society.

RSS 2015 International Conference

Organized by:	Royal Statistical Society
Where:	Exeter, United Kingdom
When:	07.09.2015 to 10.09.2015
Homepage:	http://www.statslife.org.uk/events/annual-conference

Location The Forum, University of Exeter, Exeter, Devon UK

A feature of all RSS annual conferences is the breadth and variety of the programme of talks and workshops - and the 2015 conference is no different. There will be eight session streams appealing to theoretical and applied statisticians, data scientists and statisticians working in the public and private sectors, people working with data more generally and those with a general interest in the topic.
Information about the RSS 2015 Conference is available from the **conference website**

Confirmed plenary speakers include: Dame Julia Slingo, Scott Zeger, Peter Hall and Alberto Nardelli.
Deadline for contributed talk submissions: 31 March
Deadline for contributed poster submissions: 30 June
Early registration discount closes 5 June

Contact	http://www.rssconference.org.uk/
Organiser Name	Paul Gentry
Email Address	conference@rss.org.uk
Organising Group(s)	Royal Statistical Society

EESW15, The Fourth European Establishment Statistics Workshop

Where: Poznan, Poland
When: 07.09.2015 to 09.09. 2015
Homepage: http://enbes.wikispaces.com/EESW15

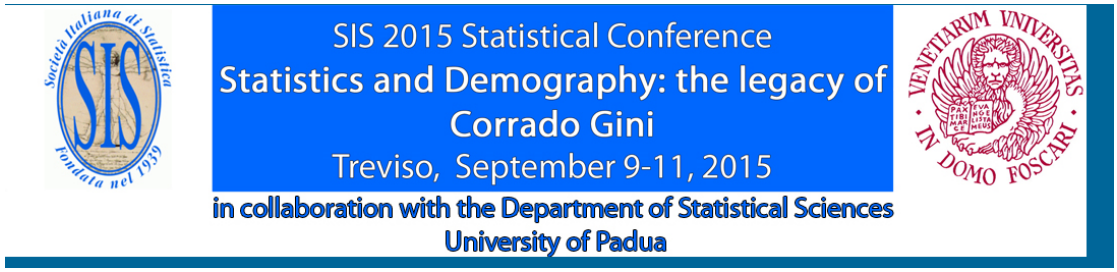
We invite you to a workshop aiming to provide a strong offering in latest results on methods and practices for producing business, economic and organisational statistics. The workshop gives the opportunity for official statistics methodologists, academic researchers and practitioners from the business sphere to interact with colleagues and exchange experiences on the topics of common interest, both through ample allocated discussion times and on the informal fringes of the workshop. The workshop covers best methodologies and practices for all stages of the statistical production process: specifying needs, designing sampling and estimation procedures, data collection instruments, building systems, collecting data, processing (including editing, imputation and estimation), analysis, dissemination, creating and using process data, quality measures, and so on.

We especially invite contributions on new developments such as:

- supra-national integration of business statistics
- improving consistency of business statistics over unit types
- design and maintenance of business registers from a statistical perspective
- business profiling: process and effects
- methodology for coordinated sampling frames
- estimation based on the combination of surveys and administrative data
- changes and trends in structure and organisation of businesses
- business statistics and national accounts.

Proposals addressing issues concerning statistical units are especially welcomed.

The workshop will be held at the Poznań University of Economics, Poland. Previous EESWs have been held in Stockholm in 2009, in Lausanne in 2011, and in Nuremberg in 2013.



Statistics and Demography: the Legacy of Corrado Gini
Organized by: Italian Statistical Society
Where: Treviso, Italy
When: 09.09.2015 to 11.09.2015
Homepage: meetings.sis-statistica.org/index.php/ginilegacy/

Statistics and Demography: the Legacy of Corrado Gini
Treviso - Ca' Foscari University of Venice
September 9, 2015 – September 11, 2015

ORGANIZED BY: Italian Statistical Society (SIS), Ca' Foscari University of Venice in collaboration with the Department of Statistical Sciences University of Padua.

The Italian Statistical Society (SIS) promotes an international specialized statistical conference on the legacy of Corrado Gini. This conference is an occasion to investigate and to present themes of research in Statistics, Economics, Demography, Biology, Sociology and Official Statistics.

CALL FOR PAPERS

Statisticians, demographer, economists and sociologists are invited to participate to this conference by submitting a paper for an oral or poster presentation.

The conference is structured into plenary sessions with general interest contributions and invited speakers, parallel sessions of specific interest, round tables and a poster session.

After the SIS 2015 Meeting**, a selection of extended papers will be published by**:
Studies in Theoretical and Applied Statistics - Selected Papers of the Statistical Societies – Springer
Quality and Quantity – Springer
Social Indicators – Springer
Metron, journal founded by Corrado Gini – Springer
Genus, journal founded by Corrado Gini – Springer

The deadline for paper submission is the 25th of May 2015.

General information: sis2015info@gmail.com

2015 International Total Survey Error Conference

Where: Baltimore, USA
When: 19.09.2015 to 22.09.2015
Homepage: https://www.tse15.org/ehome/90248/TSE15home/?&

About the TSE15 International Conference

TSE15 is an international conference focused on survey quality and the challenges of big data. The Total Survey Error concept summarizes the ways a survey estimate may deviate from the corresponding value in the population.

The conference is for statisticians, survey managers and methodologists, pollsters, public opinion researchers, and marketing research professionals from around the world who are concerned about data quality and analytics. Presentations and short courses cover tools and approaches for understanding the sources of error and developing methods for reducing total error. We look forward to your participation at TSE15 in Baltimore.

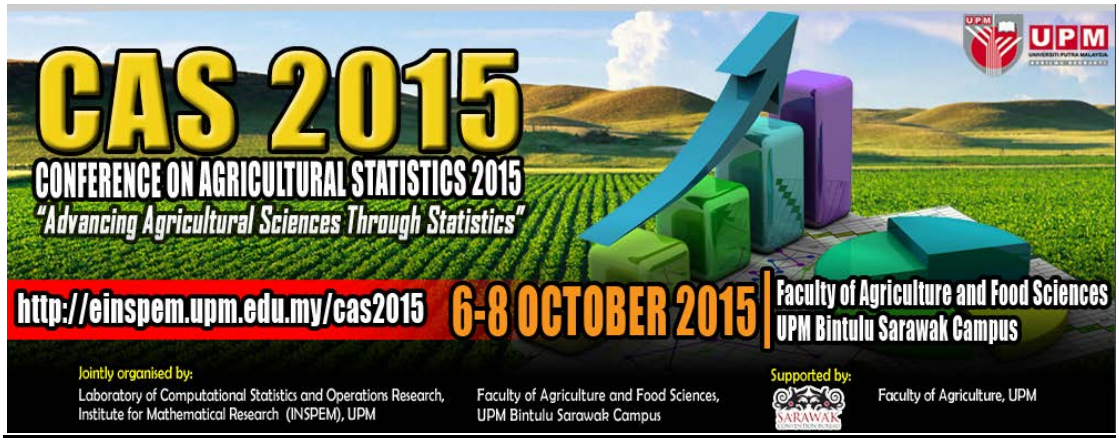
— Dr. Stephanie Eckman and Brad Edwards (TSE15 co-chairs)

Contact Us

The mailing address for the **2015 International Total Survey Error Conference** is:
TSE15 Conference
c/o Kellen (formerly The Sherwood Group Inc.)
111 Deer Lake Road, Suite 100
Deerfield, IL 60015 USA
Ph: +1-847-480-9712

For general info about TSE15, email info@tse15.org.
If you are interested in exhibiting at or sponsoring the TSE15 conference, email exhibits@tse15.org.

Questions about TSE15 program content or speakers, email TSE15 Co-chairs Brad Edwards OR Stephanie Eckman.



Conference on Agriculture Statistics 2015

Organized by: International Statistical Institute (ISI) Committee on Agricultural Statistics

Where: Sarawak, Malaysia
When: 06.10.2015 to 08.10.2015
Homepage: http://einspem.upm.edu.my/cas2015

Important Dates

Abstract Submission Deadline	**31 July 2015**
Notification of the Abstract Acceptance	**15 August 2015**
Fullpaper Submission Deadline	**31 August 2015**
Registration and Payment Deadline	**31 August 2015**

For further information or enquiry, please contact the conference secretariat at:

CAS 2015 Conference Secretariat
Institute for Mathematical Research
Universiti Putra Malaysia
43400 UPM Serdang
Selangor MALAYSIA
Tel: +6 03 8946 8923 / 7254 / 6878
Fax: +6 03 8946 6973
e-mail: cas2015upm@gmail.com
Webpage: http://einspem.upm.edu.my/cas2015/

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MICHIGAN

Celebrating Rod Little's 65th Birthday: Advances in Causal Inference, Survey Statistics, Disclosure Risk, and Missing Data

Organized by: Department of Biostatistics and the Survey Research Center
Where: **University of Michigan**
When: **30-31 October 2015**
Homepage: http://www.sph.umich.edu/biostat/events/rod-little-event.html

Rod Little, Richard D. Remington Professor of the Department of Biostatistics at the University of Michigan, has been a leader throughout his career in the development and application of statistical methods in public health research. To celebrate his wide range of accomplishments and services to science, as well as his tenure at the Department of Biostatistics since 1993, the Department of Biostatistics and the Survey Research Center at the Institute for Social Research are proud to host a symposium in honor of his 65th birthday entitled "**Celebrating Rod Little's 65th Birthday: Advances in Causal Inference, Survey Statistics, Disclosure Risk, and Missing Data**." The Symposium will include fourteen invited speakers (former students and colleagues) that are leaders in these areas, as well as a contributed poster session and a banquet in his honor.

The Conference will be held Friday October 30 and Saturday October 31 at Palmer Commons on the the University of Michigan campus, Ann Arbor, MI. Registration is $150 ($50 for students); registration for the banquet is separate. Details on registration, housing, location, and transportation are below.

All events will take place on the fourth floor of Palmer Commons 100 Washtenaw Avenue

Organizing Committee

- Michael Elliott: mrelliot@umich.edu (chair of the organizing committee)
- Veronica Berrocal: berrocal@umich.edu
- Timothy Johnson: tdjtdj@umich.edu
- Trivellore Raghunathan: teraghu@umich.edu
- Jeremy Taylor: jmgt@umich.edu

If you have any questions about the conference, please contact Michael Elliott at mrelliot@umich.edu or 734-647-5160.

Statistics Canada 2016 International Methodology Symposium

Organized by: Statistics Canada

Where: the Palais des congrès de Gatineau (in Gatineau, Québec, Canada)

When: March 22 to 24, 2016

Homepage: http://www.statcan.gc.ca/eng/conferences/symposium2016/index

Growth in Statistical Information: Challenges and Benefits

Call for Contributed Papers

Statistics Canada's 2016 International Methodology Symposium will take place at the Palais des congrès de Gatineau (5 minutes from downtown Ottawa) from **March 22 to 24, 2016**.

The title of the Symposium is **"Growth in Statistical Information: Challenges and Benefits".** In recent years, the amount of data available for potential use in producing statistical information has grown by leaps and bounds. Terms such as Big Data, Data Science and Data Mining are becoming more and more common in the literature and the media. But what does it all mean for official statistics and what is the impact on how they are collected, compiled, analyzed and presented?

We are soliciting contributed papers that examine **methodological issues related to the sustained growth in statistical information**.

Topics may include the following:

- Administrative data
- Big Data
- Data on the Web
- Paradata
- Record linkage and statistical matching
- Data mining
- Disclosure Control
- Data warehousing
- Database systems
- Legal and operational access to Big Data
- Data quality
- Measurement errors and Estimation
- Analyses of large data set
- Non-standard analyses of data
- Representativeness
- Dissemination
- Visualisation of multidimensional complex data

Please submit your proposal by email to STATCAN.Symposium2016-Symposium2016.STATCAN@statcan.gc.ca by **September 14, 2015**. It must include the following: title, an abstract of approximately 250 words (in English or French), three to six keywords and your full contact information.

Please visit our Web site regularly in order to get more detailed and updated information

http://www.statcan.gc.ca/eng/conferences/symposium2016/index

The Fifth International conference on Establishment Surveys (ICES-V)

Where: Geneva, Switzerland
When: June 20-23, 2016
Homepage: http://www.portal-stat.admin.ch/ices5/

The Fifth International Conference on Establishment Surveys (ICES-V) will be held in Geneva, Switzerland, on June 20-23, 2016. Continuing in the traditions of ICES-I to ICES-IV, ICES-V intends to explore new areas of establishment statistics as well as to reflect state-of-the-art at the time of holding the conference.

Situated for the first time in Europe, in the beautiful surroundings of the canton and city of Geneva, ICES-V is expected to be attractive to professionals and researchers in the area of statistics on businesses, farms and institutions throughout the world.
The conference is planned to include :

- Strong offering of short courses on different levels (introductory, intermediate, advanced)
- Introductory overview lectures on important and timely topics
- Selection of invited and contributed papers
- A keynote speaker and reception
- Poster sessions and software demonstrations

This site will be updated with new information as we progress in our steps towards the conference, so please do visit it occasionally. Alternatively, send an email to ices-v@bfs.admin.ch with the subject line “Please add to ICES list” to be kept abreast per email of events related to the ICES conference series.

Looking forward to welcoming you in Geneva,
Boris Lorenc and Jean-Pierre Renfer, Conference Co-Chairs

Second International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts (3MC)

Organized by: CSDI (Comparative Survey Design and Implementation)
Where: Chicago, Illinois
When: July 2016
Homepage: http://csdiworkshop.org/v2/index.php/3mc-2016

As part of an ongoing effort to promote quality in multipopulation surveys and to raise the level of methodological expertise in various applied fields of comparative survey research, the Second International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts will be held July 2016 in Chicago (3MC 2016).

This conference will bring together researchers and survey practitioners concerned with survey methodology and practice in comparative contexts. It will provide a unique opportunity to discuss and present research that contributes to our understanding of survey needs and methods in multi-cultural, multi-national, and multi-lingual contexts. Conference contributions will help document current best practices and stimulate new ideas for further research and development.

On behalf of the 2016 3MC Conference organizing committee: Timothy Johnson, Beth-Ellen Pennell, Lars Lyberg, Peter Ph. Mohler, Alisu Schoua-Glusberg, Tom W. Smith, Ineke Stoop, Christof Wolf.

9th French Colloquium on Survey Sampling

Organized by: Societe Francaise de Statistique
Where: l'Universite du Quebec en Outaouais, Quebec, Canada
When: 11-14 October 2016
Homepage: http://sondages2016.sfds.asso.fr/en/

The Ninth French Colloquium on Survey Sampling (*Colloque francophone sur les sondages*) will take place on **October 12-14, 2016,** on the main campus of the *Université du Québec en Outaouais* (UQO), in Gatineau (Canada). It will be preceded by training workshops on **October 11, 2016,** also on the main campus of UQO. This ninth Colloquium is organized by the *Société Française de Statistique* (SFdS) and its *Enquêtes, Modèles et Applications* group, and by UQO. UQO is considered a university in which the human aspect promotes learning, thought and creation, UQO is strong as a part of the *Université du Québec* network and can rely on the educational resources and shared services of the largest university network in Canada.

The Gatineau Colloquium will capitalize on two synergies, namely the synergy from the meeting of several continents and the synergy from the meeting of specialists from various communities and disciplines: statisticians and statistics users (for example, sociologists, demographers and political scientists) from academia, governments and the private sector.

Looking forward to see you!

In Other Journals

Journal of Survey Statistics and Methodology

VOLUME 3 / ISSUE 2 / JUNE 2015
http://jssam.oxfordjournals.org/content/current

A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation
Hukum Chandra, Nicola Salvati, and Ray Chambers

Observed Best Prediction for Small Area Counts
Senke Chen, Jiming Jiang, and Thuan Nguyen

Bayesian Inference for the Finite Population Total from a Heteroscedastic Probability Proportional to Size Sample
Sahar Z. Zangeneh and Roderick J. A. Little

Third-Party Presence Effect with Propensity Score Matching
Abdoulaye Diop, Kien T. Le, and Michael Traugott

Distractions: The Incidence and Consequences of Interruptions for Survey Respondents
Stephen Ansolabehere and Brian F. Schaffner

The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth
Brady T. West, James Wagner, Frost Hubbard, and Haoyu Gu

Survey Methodology

Statistics Canada Statistique Canada Canada

June 2015, VOL 41, NO 1
http://www.statcan.gc.ca/pub/12-001-x/12-001-x2015001-eng.htm

Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys
Piero Demetrio Falorsi and Paolo Righi

An efficient estimation method for matrix survey sampling
Takis Merkouris

JOS

Journal of Official Statistics

June 2015, Vol. 31 Issue 2
http://www.jos.nu/entry.asp

Variance Estimation of Change in Poverty Rates: an Application to the Turkish EU-SILC Survey
Oguz Alper, Melike ; Berger, Yves G.

ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys
Zardetto, Diego

Dwelling Price Ranking versus Socioeconomic Clustering: Possibility of Imputation
Fleishman, Larisa ; Gubman, Yury ; Tur-Sinai, Aviad

Quality Assessment of Imputations in Administrative Data
Schnetzer, Matthias ; Astleithner, Franz ;/ Cetkovic, Predrag ; Humer, Stefan ; Lenk, Manuela ; Moser, Mathias

Big Data as a Source for Official Statistics
Daas, Piet J.H. ; Puts, Marco J. ; Buelens, Bart ; Hurk, Paul A.M. van den

Small Area Model-Based Estimators Using Big Data Sources
Marchetti, Stefano;/ Giusti, Caterina ; Pratesi, Monica ; Salvati, Nicola ; Giannotti, Fosca ;/ Pedreschi, Dino ; Rinzivillo, Salvatore ; Pappalardo, Luca ; Gabrielli, Lorenzo

Sentiments and Perceptions of Business Respondents on Social Media: an Exploratory Analysis
Torres van Grinsven, Vanessa ; Snijkers, Ger.

Measuring Disclosure Risk and Data Utility for Flexible Table Generators
Shlomo, Natalie ; Antal, Laszlo ; Elliot, Mark

Statistical Metadata: a Unified Approach to Management and Dissemination
Signore, Marina ; Scanu, Mauro ; Brancato, Giovanna

Survey Practice
Practical Information for Survey Researcher

VOL 8, NO 3 (2015)
http://www.surveypractice.org/index.php/SurveyPractice/issue/view/60

Racial Attitudes and Race of Interviewer Item Non-Response
Matt Barreto, Loren Collingwood, Chris Parker, Francisco Pedraza

Understanding Participation in a Web-Based Measurement Burst Design: Response Metrics and Predictors of Participation
Jamie Griffin, Megan E. Patrick

Considerations for and Lessons Learned from Online, Synchronous Focus Groups
Sarah G. Forrestal, Angela Valdovinos D'Angelo, Lisa Klein Vogel

The Alberta Inpatient Hospital Experience Survey: Representativeness of Sample and Initial Findings
Kyle Kemp, Nancy Chan, Brandi McCormack

Probabilistic Web survey methodology in education centers: An example in Spanish schools
Jesus Alberto Tapia, Jose Antonio Menéndez

Assessing the cognitive validity of the Women's Empowerment in Agriculture Index instrument in the Haiti Multi-Sectoral Baseline Survey
Kiersten Blair Johnson, Pablo Diego-Rosell

SRM Survey Research Methods

Vol. 9, No. 1 (2015)
https://ojs.ub.uni-konstanz.de/srm/

A feasibility test of using smartphones to collect GPS information in face-to-face surveys
Kristen Olson, James Wagner

Dependent Interviewing and Sub-Optimal Responding
Johannes Eggs, Annette Jäckle

The effect of events between waves on panel attrition
Mark Trappmann, Tobias Gramlich, Alexander Mosthaf

Tackling city-regional dynamics into a survey using grid sampling
Seppo Laaksonen, Teemu Kemppainen, Matti Kortteinen, Mats Stjernberg, Mari Vaattovaara, Henrik Lönnqvist

Going Online with a Face-to-Face Household Panel: Effects of a Mixed Mode Design on Item and Unit Non-Response
Annette Jäckle, Peter Lynn, Jon Burton

Statistical Journal of the IAOS: Journal of the International Association for Official Statistics

VOL 31, NO 2 (2015)
http://content.iospress.com/journals/statistical-journal-of-the-iaos/31/2?rows=50

Interview with Mr. shigeru Kawasaki
Kirsten West

International collaboration to understand the relevance of Big Data for official statistics
Vale, Steven

Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation
Polidoro, Federico | Giannini, Riccardo | Conte, Rosanna Lo | Mosca, Stefano | Rossetti, Francesca

The production of salary profiles of ICT professionals: Moving from structured database to big data analytics
Ramasamy, Ramachandran

“Re-make/Re-model": Should big data change the modelling paradigm in official statistics?
Braaksma, Barteld | Zeelenberg, Kees

Measuring output quality for multisource statistics in official statistics: Some directions
Agafiţei, Mihaela | Gras, Fabrice | Kloek, Wim | Reis, Fernando | V\hat{a}ju, Sorina

Assessing coverage of the 2010 Brazilian Census
da Silva, Andréa Diniz | de Freitas, Marcos Paulo Soares | Pessoa, Djalma Galvão Carneiro

A guide to social media emergency management analytics: Understanding its place through Typhoon Haiyan tweets
Graham, Cat | Thompson, Chris | Wolcott, Michiko | Pollack, Joseph | Tran, Minh

Curbstoning and culture
Kennickell, Arthur

Quality assessment in production systems with registers and sample surveys
Gren, Anders | Wallgren, Britt

Micro data integration for Labour Market Account
Stender, Pernille | Thorsen, Thomas | Henrik Andersen, Hans

Reconciliation of labour market statistics using macro-integration
Mushkudiani, Nino | Daalmans, Jacco | Pannekoek, Jeroen

Selection bias and the statistical patterns of mortality in conflict
Price, Megan | Ball, Patrick

Role of official statistics in situations of conflict and post-conflict
Abu-Libdeh, Hasan

Meeting national information needs on homelessness: Partnerships in developing, collecting and reporting homelessness services statistics
Neideck, Geoff | Siu, Penny | Waters, Alison

Beyond traditional customer surveys: The reputation analysis
Willand, Ilka

Understanding customer needs
Ross, Paul F.

A strategy to test questionnaires at a national statistical office
Persson, Andreas | Björnram, Anette | Elvers, Eva | Erikson, Johan

An empirical examination of the relationship between nonresponse rate and nonresponse bias
Wright, Graham

Recognition and indigenizing official statistics: Reflections from Aotearoa New Zealand and Australia
Kukutai, Tahu | Walter, Maggie

International Statistical Review
Revue Internationale de Statistique

VOL 83, ISSUE 1 (April 2015)
http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1751-5823

New Co-Editor-in-Chief of the International Statistical Review
Victor Perez-Abreu

A Conversation with Jean-Louis Bodin
Gilbert Saporta

Ladislaus von Bortkiewicz—Statistician, Economist and a European Intellectual
Wolfgang Karl Härdle and Annette B. Vogt

Varying Coefficient Regression Models: A Review and New Developments
Byeong U. Park, Enno Mammen, Young K. Lee and Eun Ryung Lee

Discussions on Varying Coefficient Regression Models: A Review and New Developments
Jianqing Fan and Wenyang Zhang

Discussion
Toshio Honda

Discussion
Jianhua Z. Huang and Ya Su

Rejoinder to Varying Coefficient Regression Models: A Review and New Developments
Byeong U. Park, Enno Mammen, Young K. Lee and Eun Ryung Lee

Statistical Approaches for Non-parametric Frontier Models: A Guided Tour
Léopold Simar and Paul W. Wilson

Statistical Issues in Assessing Forensic Evidence
Karen Kafadar

Applications of Statistics in the Field of General Insurance: An Overview
Yves L. Grize

Journal of Privacy and Confidentiality

Volume 6, Issue 2 (2014)
http://repository.cmu.edu/jpc/

TRANSACTIONS ON DATA PRIVACY

Foundations and Technologies
http://www.tdp.cat

Volume 8, Issue 2August 2015
http://www.tdp.cat/issues11/vol08n02.php

Journal of the Royal Statistical Society SERIES A Statistics in Society

June 2015, Vol 178 Issue 3
http://onlinelibrary.wiley.com/doi/10.1111/rssa.2015.178.issue-3/issuetoc

From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects
Erin Hartman, Richard Grieve, Roland Ramsahai and Jasjeet S. Sekhon

Obituary: A. H. Halsey, FBA
Anthony Heath

Journal of the American Statistical Association

ASA
AMERICAN STATISTICAL ASSOCIATION

Volume 110, Issue 509 (2015)
http://www.tandfonline.com/toc/uasa20/current

Why Your Involvement Matters
Nathaniel Schenker

A Spatio-Temporal Point Process Model for Ambulance Demand
Zhengyi Zhou, David S. Matteson, Dawn B. Woodard, Shane G. Henderson & Athanasios C. Michaeas

Risk-Adjusted Cumulative Sum Charting Procedure Based on Multiresponses
Xu Tang, Fah F. Gan & Lingyun Zhang

Estimating a Strutctured Covariance Matrix From Multilab Measurements in High-Throughput Biology
Alexander M. Franksk, Gábor Csárdi, D. Allan Drummond & Edoardo M. Airoldi

A Flexible Bayesian Approach to Monotone Missing Data in Longitudinal Studies With Nonignorable Missingness With Application to an Acute Schizophrenia Clinical Trial
Antonio R. Linero & Michael J. Daniels

Power Curve Estimation With Multivariate Environmental Factors for Inland and Offshore Wind Farms
Giwhyun Lee, Yu Ding, Marc G. Genton & Le Xie

Calibration of Computational Models With Categorical Parameters and Correlated Outputs via Bayesian Smoothing Spline ANOVA
Curtis B. Storlie, William A. Lane, Emily M. Ryan, James R. Gattiker & David M. Higdon

What Happens Depends on When It Happens: Copula-Based Ordered Event History Analysis of Civil War Duration and Outcome
Kentaro Fukumoto

A Dynamic Directional Model for Effective Brain Connectivity Using Electrocorticographic (ECoG) Time Series
Tingting Zhang, Jingwei Wu, Fan Li, Brian Caffo & Dana Boatman-Reich

Simulating and Analyzing Order Book Data: The Queue-Reactive Model
Weibing Huang, Charles-Albert Lehalle & Mathieu Rosenbaum

An Analysis of an Incomplete Marked Point Pattern of Heat-Related 911 Calls
Matthew J. Heaton, Stephan R. Sain, Andrew J. Monaghan, Olga V. Wilhelmi & Mary H. Hayden

Robust Principal Component Analysis for Power Transformed Compositional Data
J. L. Scealy, Patrice de Caritat, Eric C. Grunsky, Michail T. Tsagris & A. H. Welsh

Multi-Agent Inference in Social Networks: A Finite Population Learning Approach
Jianqing Fan, Xin Tong & Yao Zeng

Bayesian Inference of Multiple Gaussian Graphical Models
Christine Peterson, Francesco C. Stingo & Marina Vannucci

Homogeneity Pursuit
Zheng Tracy Ke, Jianqing Fan & Yichao Wu

A Unifying Model for Capture–Recapture and Distance Sampling Surveys of Wildlife Populations
D. L. Borchers, B. C. Stevenson, D. Kidney, L. Thomas & T. A. Marques

Bahadur Efficiency of Sensitivity Analyses in Observational Studies
Paul R. Rosenbaum

R-Estimation for Asymmetric Independent Component Analysis
Marc Hallin & Chintan Mehta

Model-Robust Designs for Quantile Regression
Linglong Kong & Douglas P. Wiens

Quantile Correlations and Quantile Autoregressive Modeling
Guodong Li, Yang Li & Chih-Ling Tsai

Tuning Parameter Selection for the Adaptive Lasso Using ERIC
Francis K. C. Hui, David I. Warton & Scott D. Foster

Regularization Methods for High-Dimensional Instrumental Variables Regression With an Application to Genetical Genomics
Wei Lin, Rui Feng & Hongzhe Li

SPReM: Sparse Projection Regression Model For High-Dimensional Linear Regression
Qiang Sun, Hongtu Zhu, Yufeng Liu & Joseph G. Ibrahim for the Alzheimer's Disease Neuroimaging

Inference for Subgroup Analysis With a Structured Logistic-Normal Mixture Model
Juan Shen & Xuming He

Semiparametric Relative-Risk Regression for Infectious Disease Transmission Data
Eben Kenah

Multivariate Meta-Analysis of Heterogeneous Studies Using Only Summary Statistics: Efficiency and Robustness
Dungang Liu, Regina Y. Liu & Minge Xie

Varying Index Coefficient Models
Shujie Ma & Peter X.-K. Song

A Unified Family of Covariate-Adjusted Response-Adaptive Designs Based on Efficiency and Ethics
Jianhua Hu, Hongjian Zhu & Feifang Hu

Size and Shape Analysis of Error-Prone Shape Data
Jiejun Du, Ian L. Dryden & Xianzheng Huang

On the Prediction of Stationary Functional Time Series
Alexander Aue, Diogo Dubart Norinho & Siegfried Hörmann

Risk Classification With an Adaptive Naive Bayes Kernel Machine Model
Jessica Minnier, Ming Yuan, Jun S. Liu & Tianxi Cai

Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data
Nadja Klein, Thomas Kneib & Stefan Lang

Sufficient Reductions in Regressions With Elliptically Contoured Inverse Predictors
Efstathia Bura & Liliana Forzani

Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective
P. Richard Hahn & Carlos M. Carvalho

BIOMETRIKA

Volume 102, issue 2 (2015)
http://biomet.oxfordjournals.org/content/current

Testing differential networks with applications to the detection of gene-gene interactions
Yin Xia, Tianxi Cai, and T. Tony Cai

Hierarchical recognition of sparse patterns in large-scale simultaneous inference
Wenguang Sun and Zhi Wei

On random-effects meta-analysis
D. Zeng and D. Y. Lin

Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator
A. Doucet, M. K. Pitt, G. Deligiannidis, and R. Kohn

A useful variant of the Davis–Kahan theorem for statisticians
Y. Yu, T. Wang, and R. J. Samworth

Information-theoretic optimality of observation-driven time series models for continuous responses
F. Blasques, S. J. Koopman, and A. Lucas

On the dependence structure of bivariate recurrent event processes: inference and estimation
Jing Ning, Yong Chen, Chunyan Cai, Xuelin Huang, and Mei-Cheng Wang

A Möbius transformation-induced distribution on the torus
Shogo Kato and Arthur Pewsey

Maximum projection designs for computer experiments
V. Roshan Joseph, Evren Gul, and Shan Ba

Automatic structure recovery for additive models
Yichao Wu and Leonard A. Stefanski

Jump information criterion for statistical inference in estimating discontinuous curves
Zhiming Xia and Peihua Qiu

A validated information criterion to determine the structural dimension in dimension reduction models
Yanyuan Ma and Xinyu Zhang

Effective dimension reduction for sparse functional data
F. Yao, E. Lei, and Y. Wu

Envelopes and reduced-rank regression
R. Dennis Cook, Liliana Forzani, and Xin Zhang

On the degrees of freedom of reduced-rank estimators in multivariate regression
A. Mukherjee, K. Chen, N. Wang, and J. Zhu

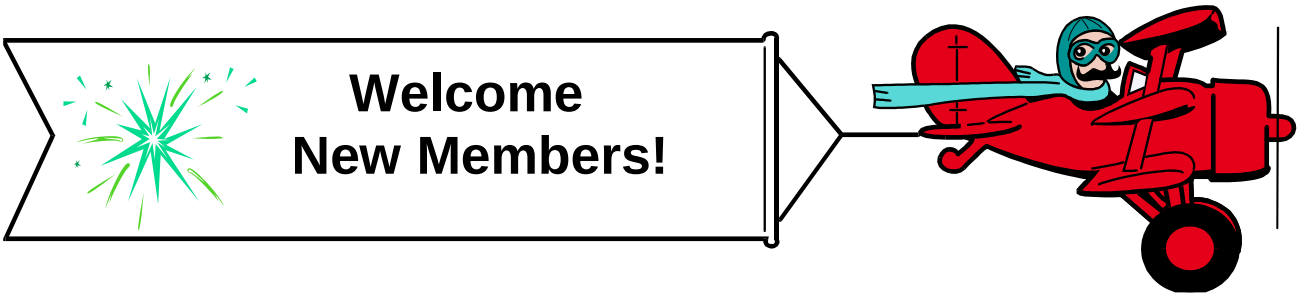
Miscellanea

Effective degrees of freedom: a flawed metaphor

Lucas Janson, William Fithian, and Trevor J. Hastie

Semiparametric exponential families for heavy-tailed data
William Fithian and Stefan Wager

Optimum designs for two treatments with unequal variances in the presence of covariates
A. C. Atkinson



We are very pleased to welcome the following new members!

	Member	Country
Dr.	Carolina Casas-Cordero	Chile
Ms.	Monica Castillo	France
Ms.	Mariia Chebanova	Ukraine
Ms.	Leesha Delatie-Budair	Jamaica
Prof.	Yongmao Dong	China
Dr.	Jason Hsia	United States
Dr.	Tetiana Ianevych	Ukraine
Mr.	Nkandu Kabibwa	Zambia
Mrs.	Saeideh Kamgarsangari	Iran, Islamic Republic of
Dr.	Paul Kiff	United Kingdom
Mr.	Zhilong Il	China
Dr.	Stefano Marchetti	Italy
Mr.	Moviele Mentor	Haiti
Mr.	Raymond Muyovwe	Zambia
Ms.	Flavia Naiga Oumo	Uganda
Prof.	Alessandra Petrucci	Italy
Prof.	Samir Safi	Palestine, State of
Mr.	Bayram Samet Sahin	Turkey
Mr.	Ghulam Muhammad Shah	Nepal
Dr.	Emilio Lopez Excobar	Mexico

IASS Officers and Council Members

Executive Officers

President:	Danny Pfeffermann (Israel)	msdanny@huji.ac.il
President-elect:	Steve Heeringa (USA)	sheering@isr.umich.edu
Vice-Presidents:	Jairo Mabil Oka Arrow (South Africa)	jairo.arrow@gmail.com
	Geoffrey Lee (Australia)	geoff.lee99@bigpond.com
Scientific Secretary:	Mick Couper (USA)	mcouper@umich.edu
Council Members (2011-2015):	Christine Bycroft (New Zealand)	christine.bycroft@stats.govt.nz
	Ka-Lin Karen Chan (China)	klchan@censtatd.gov.hk
	Olivier Dupriez (Belgium/USA)	odupriez@worldbank.org
	Natalie Shlomo (UK)	natalie.shlomo@manchester.ac.uk
	Marcel de Toledo Vieira (Brazil)	marcel.vieira@ufjf.edu.br
	Alvaro Gonzalez Villalobos (Argentina)	alvarun@gmail.com
Council Members: (2013-2017)	J. Michael Brick (USA)	mikebrick@westat.com
	Daniela Cocchi (Italy)	daniela.cocchi@unibo.it
	Jack Gambino (Canada)	gambino@statcan.ca
	Risto Lehtonen (Finland)	risto.lehtonen@helsinki.fi
	Ralf Münnich (Germany)	muennich@uni-trier.de
	Jean Opsomer (USA)	jopsomer@stat.colostate.edu

Committee Chairs

Chair of the Rio 2015 Programme Committee	Christine Bycroft (New Zealand)	christine.bycroft@stats.govt.nz
Chair of the committee for the Cochran Hansen prize award	Risto Lehtonen (Finland)	risto.lehtonen@helsinki.fi

Ex Officio Members

ISI Operation Manager/Assistant Director, supporting the IASS	Shabani Mehta	s.mehta@cbs.nl
Transition Executive Director:	Catherine Meunier (France)	katherine.meunier@orange.fr
Treasurer:	Ada van Krimpen (The Netherlands)	an.vankrimpen@cbs.nl
Finance:	Michael Leeuwe	
Webmaster:	Mehmood Asghar and Olivier Dupriez	odupriez@worldbank.org
IASS Secretariat Membership Officer	Margaret de Ruiter-Molloy (The Netherlands)	m.deruitermolloy@cbs.nl

Institutional Members

2 International Organisations

AFRISTAT
EUROSTAT

15 Bureaus of Statistics

AUSTRALIA – AUSTRALIAN BUREAU OF STATISTICS
BRAZIL – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE)
CANADA – STATISTICS CANADA
DENMARK –DANMARKS STATISTIK
FINLAND – STATISTICS FINLAND
GERMANY –STATISTICHE BUNDESAMT
ITALY –INSTITUTO NAZIONALE DI STATISTICA (ISTAT)
KOREA, REPUBLIC OF – STATISTICS KOREA
LUXEMBOURG – EUROPEAN COMMISSION – EUROSTAT
MACAO – DIREÇCAO DOS SERVIÇOS DE ESTATISTICA E CENSOS
MALI – AFRISTAT
MAURITIUS – STATISTICS MAURITIUS
MEXICO –INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA (INEGI)
MEXICO – NUMÉRIKA-MEDICION Y ANALISIS ESTAD. AVANZADO, SC
NEW ZEALAND – STATISTICS NEW ZEALAND
NORWAY – STATISTICS NORWAY
PORTUGAL –INSTITUTO NACIONAL DE ESTADÍSTICA (INE)
SWEDEN – STATISTICS SWEDEN
UNITED STATES – RESEARCH TRIANGLE INSTITUTE
UNITED STATES - SURVEY RESEARCH CENTER
UNITED STATES – NASS, RESEARCH & DEVELOPMENT DIVISION
UNITED STATES – NATIONAL CENTER FOR HEALTH STATISTICS
UNITED STATES – WESTAT INC.

5 Universities, Research Centers, Private Statistics Firms

USA – CENTERS FOR DISEASE CONTROL AND PREVENTION
USA – RESEARCH TRIANGLE INSTITUTE
USA – SURVEY RESEARCH CENTER, UNIVERSITY OF MICHIGAN
USA – U.S. DEPARTMENT OF AGRICULTURE
USA – WESTAT

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/