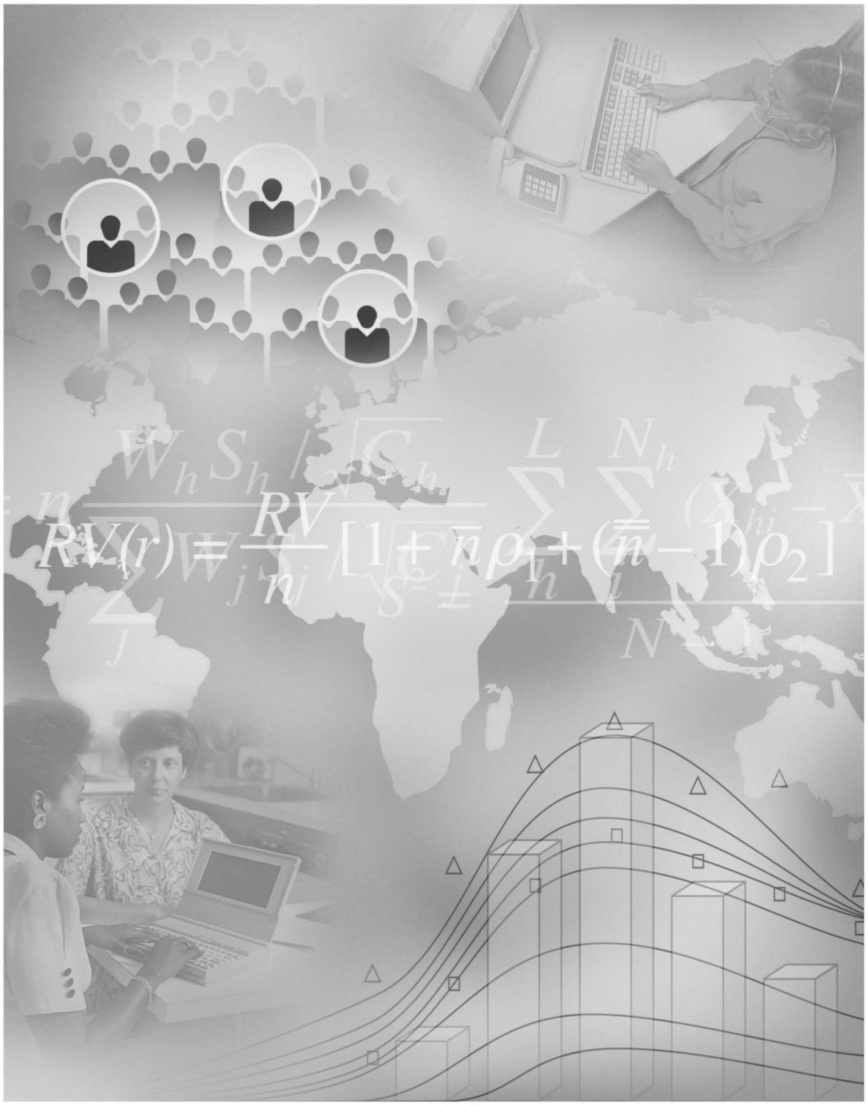
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

No. 76 July 2017



$RV(r) = \frac{RV}{n}[1+\bar{n}\rho_1+(\bar{\bar{n}}-1)\rho_2]$

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

Editors:
Natalie Shlomo and Eric Rancourt

Section Editors:
Peter Wright – Country Reports
Ken Copeland – Ask the Experts
Denise Silva – New and Emerging Methods
Natalie Shlomo – Book & Software Review

Circulation/Production:
Lori Young
Nick Husek

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@manchester.ac.uk) or Eric Rancourt (Eric.Rancourt@canada.ca).

ISSN 2521-991X

In this Issue:

Letter from the Editors

The July 2017 issue contains articles of interest and important information regarding upcoming conferences, contents of relevant journals, updates from the IASS Executive and more. We hope you enjoy this issue. Please send us your feedback and comments on how we can make improvements.

In the *New and Emerging Methods Section* (edited by the Scientific Secretary Denise Silva), Marco Di Zio (Istat), Li-Chun Zhang (University of Southampton and Statistics Norway) and Ton De Waal (Statistics Netherlands and Tilburg University) have contributed an article titled: Statistical Methods for Combining Multiple Sources of Administrative and Survey Data. In the article, the authors classify statistical tasks of data integration which have recently been developed under the research project "Estimation methods for the integration of administrative sources" funded by Eurostat (Eurostat 2017). The statistical tasks form the 'building blocks' of a process for producing statistics and carrying out inference based on multiple data sources.

The *Ask the Experts* Section (edited by Ken Copeland) also deals with data integration of two or more data files in the response provided by Edward J. Mulrow (NORC, University of Chicago) to the question: 'What does the term Data Linkage mean?' The author defines and distinguishes between 'data linkage' and 'statistical matching'.

This July 2017 issue of the newsletter includes a review of software in the *Book and Software Review Section* (edited by Natalie Shlomo). We have two contributed articles which demonstrate the use of 'Shiny' R in official statistics. 'Shiny' is an open source R package that forms the basis for interactive web applications. The first article is written by Thijs Benschop (Humboldt University, Berlin) and Matthew Welch (World Bank) titled: 'A Shiny application for statistical disclosure control: sdcApp opening the power of sdcMicro to non-R users'. sdcMicro was developed for the preparation of anonymised microdata files. The second article is written by Tobias Schoch (ECOPLAN AG, Switzerland) titled: Monitoring Statistical Indicators Using R/Shiny: A Case Study on Public Transport Safety in Switzerland'. The Shiny tool was developed to produce indicators for monitoring safety performance in public transportation in Switzerland.

We would like to thank very much the outgoing Scientific Secretary, Denise Silva, for her role as editor of the *New and Emerging Methods Section.* She will be replaced by the incoming Scientific Secretary following the 2017 World Statistics Conference in Marrakech. If you would like to contribute an article to *New and Emerging Methods Section*, please contact Natalie Shlomo (Natalie.Shlomo@manchester.ac.uk). 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). You are also welcome to submit your own questions with an answer if you are aware of an important topic of interest. If you are interested in writing a book or software review, please contact Natalie as well. Finally, if you would like to contribute brief articles or editorials to the newsletter, please send them directly to the editors of the newsletter, Eric Rancourt and Natalie Shlomo.

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 contributions. We also thank the new editor of the section, Peter Wright (peter.wright2@canada.ca) of Statistics Canada for contacting all country representatives and coordinating the country reports. Please get in touch with Peter if there has been a change in the country representative in order to keep our contact list up-to-date. We ask all country representatives to contribute articles on your country's current activities, applications, research and developments in survey methods. This is of great interest to our IASS membership and the editorial board would like to see the number of country reports grow.

This issue of *The Survey Statistician* includes a letter and updates from our IASS President, Steven Heeringa and from our Scientific Secretary, Denise Silva. This will be their last updates as their terms will end at the 2017 World Statistics Conference in Marrakech. We congratulate the incoming IASS council for 2017-2019: President-elect Denise Silva and the four Vice-Presidents: Cynthia Clark, Anders Holmberg, Risto Lehtonen and Jean Opsomer.

The *News and Announcement* section includes an announcement on the results of the election to the 2017-2019 IASS council and we wish to express our congratulations to the incoming President, Denise Silva, and Vice-presidents. In addition, there is an announcement about the 2017 Cochran-Hansen prize. We congratulate the winning recipient, Girish Chandra, and look forward to hearing the paper that will be presented at the 2017 WSC Conference in Marrakech: Session IPS134 - IASS President's Invited Lecture and Cochran-Hansen Prize Lecture, Wednesday, July 19th, 2017 at 14:00 -15:40, Room TBA. We would also like to congratulate the winners of the 2017 IAOS Young Statistician Prizes: First Prize: The Dissemination Game: How to communicate official statistics to non-expert users by: Hannah Thomas (United Kingdom); Second Prize: Telematics Data for Official Statistics: An Experience with Big Data by: Nicholas Husek (Australia); Third Prize: Improving Seasonal Adjustment by Accounting for Sample Error Correlation Using State Space Models by: Andreas Mayer (Australia).

We thank Lori Young 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 the members. We also thank Lori for her hard work in collating all the articles into this newsletter that you see before you. In addition, we would like to thank Nick Husek from the Australian Bureau of Statistics for final editing of the newsletter and distribution via post.

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@canada.ca

Natalie Shlomo Natalie.Shlomo@manchester.ac.uk

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS (IASS)

Dear IASS colleagues,

Greetings from Ann Arbor. My two-year term as IASS President has passed quickly and as we approach the World Statistics Congress (WSC) meetings in Marrakech, Morocco and the transition to a new IASS Executive I would like to take this opportunity to review the important activities and accomplishments of our Association and also reflect briefly on work that remains undone.

The 2017 WSC meeting (http://payment.isi2017.org/) that will take place July 16-21, in Marrakech, Morocco remains the premier opportunity for the IASS membership to gather, meet old (and new) friends face-to-face, network with international colleagues and share in the instructional and scientific program activities related to our professional interests in survey statistics and survey methodology. With the strong support of Marcel Vieira (IASS Representative) and other members of the Scientific Program Committee (SPC), a total of 15 IASS sponsored Invited Paper Sessions (IPS) have been scheduled for the 2017 World Statistics Congress (WSC). Many other session proposals from the IASS member community have been organized as special topic sessions. I want to acknowledge Ralf Muennich for his service on the ISI's 2017 Short Course Committee. IASS members submitted several successful proposals for short courses to be offered at the 2017 World Statistics Congress (http://payment.isi2017.org/scientific-programme-2/short-courses/).

There are several special IASS meetings/sessions at the 2017 WSC that I would like to draw your attention to. The first of these is the IASS General Assembly meeting that is scheduled for 12:30-2:00 pm on Wednesday, July 19. All IASS members who are able to attend the 2017 WSC are encouraged to participate in this annual meeting. On the 19th, immediately following the General Assembly meeting (2:00-3:40pm), I will chair a special invited session that will include the IASS President's Invited Lecture and a presentation of the paper that earned the 2017 IASS Cochran-Hansen Prize. The President's Invited Lecture will be presented by my career-long colleague and friend, Colm O'Muircheartaigh. Colm will be presenting on the topic, *Next Generation Data: Surveys in a Transformed Environment.* The winner of the 2017 Cochran-Hansen prize is Girish Chandra (please see the report of the C-H Prize committee in this issue of *The Survey Statistician*). In the special Wednesday afternoon session, Girish will present a talk based on his paper titled, "*Ranked Set Sampling Approach for Estimating Response of Developmental Programs with Linear Impacts under Successive Phases*".

IASS is an official sponsor of the 2017 Small Area Estimation (SAE) satellite conference (http://sae2017.ensai.fr/) that is scheduled for July 10-12, 2017 (the week prior to the 2017 WSC) in Paris. In recent years the SAE satellite conferences have proved to be very popular with IASS members. Please take a moment to review the scientific program on the conference website and consider a visit to Paris to participate in this satellite meeting.

As in the cases of the 2017 SAE satellite conference, the IASS has rarely played the primary role as the organizer of a major conference, choosing instead to provide sponsorship and financial support to a diverse set of regional conferences, symposia, workshops and training activities devoted to survey methodology and survey statistics. The Report of the IASS Scientific Secretary included in this issue of TSS provides more specific information on the 2017-2018 conferences and meetings that the IASS Executive officers have selected to receive IASS sponsorship and support.

This past February and March, the ISI Permanent office oversaw the election of officers for 2017-2019. As I reported in an e-mail message in early April, the results are as follows. Denise do Nascimento Silva was chosen as President-Elect for 2019-2021. Cynthia Clark, Anders Holmberg, Risto Lehtonen, and Jean Opsomer were elected as 2017-2019 Vice-Presidents. Along with 2017-2019 President-Elect, Peter Lynn, and IASS Executive Director, Ada vanKrimpen (ex-officio), these newly elected officers will form the IASS Executive for 2017-2019. As President, Peter will designate one of the elected VPs as the Scientific Secretary of the IASS for 2017-2019 and another VP to have the duties of Financial Officer for this same period. I want to take this opportunity to again congratulate the candidates who were elected to office but also extend my thanks to all of the candidates who agreed to stand for election and to serve if chosen for office. A note of appreciation is also due to Jairo Arrow (chair) and the other members of the 2017 IASS Nominations Committee.

As I reflect back on my time in official roles in the IASS leadership, I recognize that despite good intentions much work remains undone and many promising initiatives were not pursued as aggressively as they might have been. One such initiative that the 2015-2017 leadership will leave in the hands of the newly elected and very capable 2017-2019 IASS Executive is the refinement of the strategic plan for the future of our Association. We now have the benefit of the recently released ISI draft strategic plan, *ISI Strategy 2017 to 2019* that ISI President, Pedro Silva, shared with ISI members in early May. Many of the prime issues in the ISI plan are common to the future of our Association: How do we encourage communication and networking among the members of our association?, What role should we play in statistical capacity building related to survey statistics and survey methodology?, Beyond our traditional strong contributions to the WSC Scientific and Short Course Programs how should IASS expertise and financial resources be used in such statistical capacity building efforts? and finally, How can IASS attract, retain and engage the young survey statisticians and methodologists that are the future of our discipline?

In response to the last of these questions, I would like to repeat a suggestion that I have offered in previous letters. I know from personal experience (both as a beneficiary and a benefactor) that it can be highly effective for senior, established members of the IASS network to sponsor the membership of students and more junior statisticians, encouraging them to apply for IASS membership and paying their initial year’s membership dues. Leslie Kish, my mentor when I came to this profession in 1975, insisted that I join IASS as a student. As I have said before, not all of us can be as persuasive as Leslie was. But if each of us committed annually to introduce a student or junior colleague in our work place to the IASS (and in cases of financial need paying their membership fee) we would quickly begin to bring much-needed “youth” to our international association. For myself and I assume many of you, joining the IASS early in our career and remaining a member through the years did not bring a long list of tangible “perks” or “goods”. Its greatest benefit, one that could not be fully realized through memberships in national statistical societies or even regional networks, was that IASS membership brought us into the global network of survey statisticians and methodologists. I, for one, am forever grateful for that network and the way in which the shared knowledge and personal acquaintances have enriched my career. So thank you to each of you for being part of our international network.

IASS is an organization of volunteers. I will not attempt to list each by name but in closing, I would like to thank each of the IASS members who during the past two years has served in an IASS leadership role, as an association representative, as chair or a member of an IASS committee or in some other role. I would like to extend a special thanks to Natalie Shlomo and Eric Rancourt who have worked behind the scenes to edit and produce *The Survey Statistician*. Also, I want to specifically recognize Olivier Dupriez for managing the IASS website. Thank you all for carrying the real burden of the work of our Association.

Yours in the science and practice of survey research,

Steve Heeringa
IASS President
sheering@umich.edu

Report from the Scientific Secretary

An important activity of the IASS is to co-sponsor conferences and meetings, fostering opportunities to bring together survey statisticians and methodologists for the development of good survey practices and related methods. For 2017, IASS is committed to sponsor six great events in which IASS members are entitled to a reduced registration fee.

The first of the year was a conference in honour of Jon Rao on the occasion of his 80th birthday, hosted by the School of Mathematics and Statistics at Yunnan University (http://www.raoconference2017.com/). This was followed by the 5th Italian Conference on Survey Methodology (ITACOSM 2017), hosted by the Department of Statistical Sciences of the University of Bologna, and promoted by the Survey Sampling Group (S2G) of the Italian Statistical Society (SIS). The conference website is https://events.unibo.it/itacosm2017. The upcoming events for this year are:

- The **61st ISI World Statistics Congress Satellite Meeting on Small Area Estimation (SAE 2017)** to take place in Paris from 10-12 July 2017 (http://sae2017.ensai.fr/). The conference is organized by **ENSAI** (Ecole Nationale de la Statistique et de l'Analyse de l'Information), the CREST (Centre de Recherche en Economie et Statistique) and the ILB (Institute Louis Bachelier);

- The **Baltic-Nordic-Ukrainian Network Workshop on Survey Statistics 2017** that will take place in Vilnius, Lithuania, from 21-25 August 2017, celebrating the 80th birthday of Carl-Erik Särndal http://vilniusworkshop2017.vgtu.lt/;

- The **European Establishment Statistics Workshop (EESW) 2017** to be held in the University of the Southampton (UK) on August 30 – September 1. For more information on EESW17 and the European Network for Better Establishment Statistics (ENBES), visit www.enbes.org.;

- The **4th International Workshop on Surveys for Policy Evaluation** and the **5th Brazilian School on Sampling and Survey Methodology – ESAMP V**, from 20-17 October, in Mato Grosso, Brazil (http://www.redeabe.org.br/esamp2017/).

The semester was also marked by the 2017 edition of our biennial Cochran-Hansen Prize competition for the best paper on survey methods by a young statistician from a developing or transition country. Monica Pratesi (chair), Hukum Chandra and Timo Schmid served as the prize committee. The recipient of the 2017 C-H Prize is Girish Chandra, from India. He will present the paper titled: Ranked Set Sampling Approach for Estimating Response of Developmental Programs with Linear Impacts under Successive Phases, co-authored by Rajiv Pandey and Dinesh Bhoj, in an IASS Invited Paper Session at the World Statistics Congress 2017 (WSC2017).

In addition, IASS has also been involved with preparations for the WSC2017. We are pleased to report that 23 sessions submitted by the IASS community were accepted as

Invited Paper Sessions (IPS) or Special Topic Sessions (STS), as listed next. Some sessions were endorsed by more than one association. The WSC2017 programme also includes a session for the IASS President's Invited Lecture and the Cochran-Hansen awardee paper plus a short course proposed by IASS members.

We like to thank the IASS colleagues for making an active contribution to the development of the WSC2017 scientific programme. Although no IPS, STS or short course have an association label, it is good to know that all these activities constitute great opportunities to place survey statistics subjects in the spotlight of the 2017 congress.

WSC2017 Invited Papers Sessions, Special Topic Sessions and Short Course

IPS011	Synthetic datasets for statistical disclosure control: research and applications around the world
IPS043	The future of the social sciences and statistics
IPS049	Old and new frontiers in sampling
IPS057	Small area estimation in developing countries
IPS068	New challenges to disseminate statistical literacy and problem-solving skills among people by Statistics Bureaus of Governments in "Open Data and Big Data" era
IPS071	Agri-environmental statistics
IPS073	Innovative approaches for agricultural and rural censuses
IPS074	Practical and innovative approaches using master sampling frames for agricultural surveys
IPS078	Innovative approaches to addressing (or avoiding) survey nonresponse
IPS090	Methods for analysis of public use survey data reporting replicate weights
IPS094	Approaches to promote comparability in household surveys across countries: results of recent testing by international organizations in follow-up to the 19th International Conference of Labour Statisticians (ICLS)
IPS096	The role of surveys in a 21st century census
IPS100	Development and application of international guidelines for statistical business registers
IPS108	Keeping official statistics relevant: adaptive designs to reduce bias and improve timeliness
IPS115	Measuring income and well-being through longitudinal household surveys
IPS134	IASS President's Invited Lecture and Cochran-Hansen Award Lecture
IPS135	Multiple imputation methods for survey nonresponse
IPS136	Big Data and big opportunities: how the minds of statisticians can help
IPS140	Accounting for statistical dependence in socio-economic data
STS005	Statistical Disclosure Control - methods and applications
STS006	Recent advances in statistical methods for large scale correlated phenomena

STS013 Methodology and practice of earning and income from employment statistics

STS099 Statistical thinking through student work with real-world data leading to an understanding of evidence-based policy decision and the role of statistics in civil society

SC16 Construction of weights and treatment of influential units in surveys

Once again, I invite all members to contribute to our newsletter by volunteering to send articles, book/software reviews and country reports **to the next issue of *The Survey Statistician* to be published in January 2018**. You may also suggest a subject, or write an article, to the ***New and Emerging Methods*** section. This would not be a traditional scientific paper but an informative article of at most 8-10 pages, introducing the challenges, the methods, their uses and applications, and also the relevance of the subject for the development of survey methods. Our newsletter will really benefit from your ideas and collaboration.

It is now time to thank you all for the opportunity to act as IASS scientific secretary, to welcome the newly elected officers who will form the IASS Executive for 2017-2019, to thank the IASS officers and colleagues with whom I got the opportunity to liaise and share the IASS activities in these last two years, and to give special thanks to Steve Heeringa for his leadership and commitment to the Association.

I hope to meet some of you at the WSC2017 in Marrakesh. If you attend the congress, please join us for the IASS General Assembly, the IASS President's Invited Lecture and the Cochran-Hansen Prize awardee presentation, and also for the other great sessions supported by IASS.

Warm wishes,

Denise Silva

News and Announcements

2017 Cochran-Hansen Prize of the IASS

Report: Monica Pratesi, President of the jury for the 2017 Cochran-Hansen Prize

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 prize is restricted to young statisticians from developing and transition countries, which are living in such countries. They are asked to develop innovative methods and models with the goal of answering scientific questions arising from a problem in survey methods.

This prize aims at playing as a stimulator of research and it represents a great opportunity for interaction, sharing ideas and promoting the discussion. As a result, methodological and applied contributions to the survey methods will be expected following the prize.

Early career researchers in Statistics and Survey Methods have been invited to apply for attending the prize posting their online application by February 15, 2017. A paper submitted for the competition consisted of original work which was either unpublished or has been published after January 1st, 2017.

We received a total of 9 papers. A total of 3 papers were accepted for evaluation (one applicant was not from target countries, 5 papers were not focused on survey research methods).

The winning paper will be presented during the 2017 WSC Conference in Marrakech at the session: **IPS134 - IASS President's Invited Lecture and Cochran-Hansen Prize Lecture which will be held on Wednesday, July 19th, 2017 at 14:00-15:40, Room TBA.**

By the evaluation of the jury, the best paper is: “Ranked Set Sampling Approach for Estimating Response of Developmental Programs with Linear Impacts under Successive Phases” by Girish Chandra [1], co-authored by Rajiv Pandey[1] and D S Bhoj[2]([1]Division of Forestry Statistics, Indian Council of Forestry Research and Education, Dehradun, India, Email: gchandra23@yahoo.com [2]Department of Mathematical Sciences, Rutgers University, Camden, USA) and the prize will be awarded to Girish Chandra.

Development programs are implemented by the government and nongovernment organisations for upgrading or improving the desired characteristic over time on the desired unit in a region or community. This is the case of India's Developmental program entitled "Education for all towards quality with equity" which was launched with the aim of providing free and compulsory education to all children of India. The Gross Enrolment Ratio of Scheduled Tribes (ST) students in secondary education (targets) during the period 2004-05 and 2008-09 was recorded by the government based upon the estimates given by the schools of the country. In practice it is difficult to get the information from all the schools of India in time. To estimate the impact of the programme the paper proposes to use a Ranked Set Sampling strategy, assuming that the program has been implemented over the various phases and that the impacts are linearly proportional to the phases.

The paper addresses an important topic in survey methods that is data collection and sampling in the evaluation of policy actions. As President of the jury of the Cochrane-Hansen prize 2017, composed by Hukum Chandra and Timo Schmid, I would like to personally welcome all IASS members to the presentation of the winning paper at the session listed above. The early career researchers are truly our greatest assets today and tomorrow, and we could not accomplish what we do without their work and the support of IASS senior members, especially in developing countries.

The world of Survey methods is an exciting area in which to work/study/play, and we'll continue to meet and bring inspired people together in sessions like this, to ensure our Association remains at the cutting edge.

Report on the 2017 IASS Officers’ Elections, 6 April 2017

The 2017 IASS Officers’ Elections were organized in order to choose the President-Elect for 2017-2019 and four Vice-Presidents 2017-2019. Their term of office will start after the WSC 2017 in Marrakech.

The Election was open to all IASS members.

On 13 February 2017 an electronic ballot form was sent to all 417 members. From these, 5 ballots bounced due to various reasons - such as ‘full e-mailbox’ or ‘invalid e-mail address’. A reminder was sent to all members on 23 March 2017.

On the closing date, 29 March 2017, a total of 190 electronic votes was received. This means that 45.6 % of the 417 members cast their votes.

The votes were counted by Gerrit J. Stemerdink, an ISI Elected Member who works as a volunteer at the ISI Permanent Office. They were checked by Jelke Bethlehem, Professor Emeritus of Statistical Methodology and long-time ISI Elected Member. The results are as follows:

Denise do Nascimento Silva is elected as President-Elect.

Cynthia Clark; Anders Holmberg; Risto Lehtonen; Jean Opsomer are elected as Vice-Presidents.

Along with President-Elect, Peter Lynn, and IASS Executive Director, Ada van Krimpen (ex-officio) these newly elected officers will form the IASS Executive for 2017-2019. As President-Elect, Peter will designate one of the elected VPs as the Scientific Secretary of the IASS for 2017-2019 and another VP to have the duties of Financial Officer for this same period.

Ask the Experts

"What does the term Data Linkage mean?"

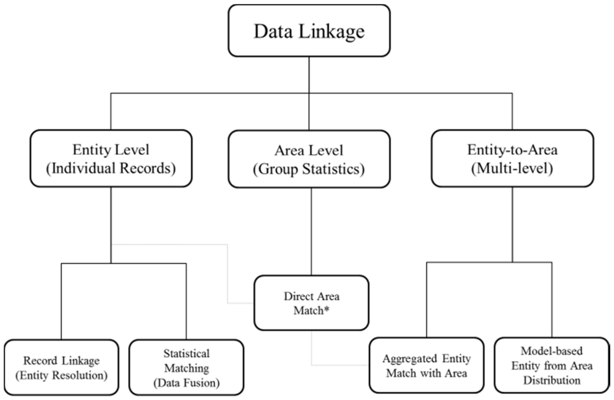
Edward J. Mulrow, NORC at the University of Chicago

May 2017

Data linkage is a general term for the integration of two or more data files. It is often carried out to exploit as much as possible the information contained in each source. In the face of limited resources, related information from different sources can be combined to take advantage of varying strengths of multiple data sources (Schenker and Ragunathan 2007). Commonly used terms for this integration are "linking" and "matching." In the scientific literature, these terms tend to be associated with the solutions to two related but different problems.

Exhibit 1 is an adaptation of a figure used by Fox and Stratychuk (2010) which provides a very broad categorization of the techniques used for data linkage based on the characteristics of the records in the files that are to be combined. If the records in the files represent entities, e.g. individuals,[1] then the integration process combines entity-level records. If there are a large number of entities with records in each file, record linkage is used to combine records for the same entity across data files. On the other hand, statistical matching refers to entity-level matching in which the linkage of data for the same entity across files may not be possible because there is little to no overlap of entities in the files. Statistical matching can also be used in situations where exact linkages are not essential to the desired analysis. For a comparative discussion of exact and statistical matching see the FCSM Statistical Policy Working Paper 5 (Federal Committee on Statistical Methodology, 1980).

Exhibit 1: Combining Data Techniques



* Entity Level methods can be used if area identifiers are uncertain or there is little overlap of area records in the files.

[1] An entity might be a business, a household, a person, or some other type of unique unit that is represented by a record in a database.

Exhibit 1 are related to area level summary statistics for groups of entities. Often the grouping factor is a geographic unit such as a county. Linking data at the area level refers to situations in which the data files of interest contain only area summary statistics. On the other hand, combining entity level and area level data (multi-level) occurs when one data file contains entity level records and another file contains area level summary statistics. In one sense, linking area-level data is straightforward because the grouping factor is usually a well-defined and known area. Estimates from data file 1 are easily matched to estimates from data file 2 for the same area. However, care must be taken in linking the files because the area estimates within each file may differ in terms of estimation methodology and uncertainty measurement. When this occurs the data from each source need to be harmonized. We will not discuss harmonization methods in this note. See Ruggles (2005) for examples of harmonization.

Statistical matching can be used to construct “complete” entity level data files in situations where two or more surveys are based on different samples with low likelihoods of overlap. Information on a common set of variables is available in both files; however each file contains information on variables not observed in the other. Statistical matching is used to “fill in” the unobserved variable information in each file. This is called a “micro” approach and the resulting data set is referred to as synthetic data because the data are not direct observations of units in a population. Rather the complete data are obtained by modelling of partially observed information.

In statistical matching the linkages are based on similar characteristics rather than unique identifying information. Linked records need not correspond to the same unit. In a statistical match each observation in one microdata set (the "base" set) is assigned one or more observations from another microdata set (the "non-base" set); the assignment is based upon similarity in selected characteristics. Conceptually, statistical matching is closely related to imputation. The method relies on the joint distribution of the variables (i.e., the characteristics forming the basis for matching). This can lead to inaccurate analysis if the joint distribution is incorrectly specified. For example, if we want to match units based on the distribution of household income and household type, and we only link single males, the distribution for single persons will be misspecified (Fox and Stratychuk 2010). D'Orazio, Di Zio, and Scanu (2006) provide details on the theory and practice of statistical matching. D'Orazio has created the R package StatMatch for performing statistical matching.

Record linkage is the task of quickly and accurately identifying records corresponding to the same entity from one or more data files and combining them (Herzog, Schreuren, and Winkler, 2007). In other words, an exact match between records for the same entity from different databases is desired. In the absence of unique identifiers, the basic methods for creating the linkage compare identifiers such as name and address information of entity records across data files of interest to determine those sets of records within and across files that are associated with the same entity. Probabilistic record linkage identifies a match between records based on a formal probabilistic model. The advantage of probabilistic record linkage is that it uses all available identifiers to establish a match (e.g., name, sex, date of birth, SSN, race, address, phone number) and does not require identifiers to match exactly. Identifiers that do not match exactly are assigned a “distance” measure to express the degree of difference between files. Each identifier is assigned a weight and the total weighted comparison yields a score, which is used to classify records as linked, not linked, or uncertainly linked according to whether the probability of a match exceeds a certain threshold. Herzog, Scheuren, and Winkler (2007) describe the key principles of probabilistic record linkage.

Although Newcombe et al (1959) and Newcombe and Kennedy (1962) introduced the use of the frequency ratio for record linkage in earlier work, Fellegi and Sunter (1969) are recognized as the first researchers to rigorously present the mathematical model and theoretical foundation for probabilistic record linkage. Their framework groups possible pairs into three sets, referred to as links (L), non-links (N), and possible links (P), based on objective criteria. Each set (L, P, and N) has associated error rates. An optimal linkage rule is defined as one that minimizes the probability of classifying a pair as belonging to set P for fixed error levels in L and N. The decision rules in the Fellegi-Sunter model are optimal in the sense that, given fixed upper bounds on the rate of false matches and failed matches, the decision rules minimize the size of the possible (indeterminate) links (P). A number of software products exist for probabilistic record linkage. Choi et al (2017) provides a review of some public domain and open source packages.

Borg, A. and Sariyar, M. (2016). RecordLinkage: Record Linkage in R. R package version 0.4-10. https://CRAN.R-project.org/package=RecordLinkage.

Choi, S., Lin, Y., and Mulrow, E. (2017) “Comparison of Public-Domain Software and Services for Probabilistic Record Linkage and Address Standardization,” available at https://www.researchgate.net/profile/Sou_Cheng_Choi/publication/312166409_Comparison_of_Public_Domain_Software_and_Services_for_Probabilistic_Record_Linkage_and_Address_Standardization/links/5873d23708ae329d621d09dc.pdf.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006) *Statistical Matching: Theory and Practice*. Wiley Series in Survey Methodology.

D'Orazio, M. (2017). StatMatch: Statistical Matching. R package version 1.2.5. https://CRAN.R-project.org/package=StatMatch.

Federal Committee on Statistical Methodology (1980), “Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques,” Washington, DC: Office Federal Statistical Policy and Standards, U.S. Department of Commerce. Available at http://www.fcsm.gov/working-papers/wp5.html.

Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, 64, pp. 1183-1210.

Fox, K. and Stratychuk, L. (2010), Proceedings of the Statistics Canada Symposium 2010, Workshop1: Record Linkage Methods.

Herzog, T. N., Scheuren, F., and Winkler, W.E., (2007), *Data Quality and Record Linkage Techniques*, New York, N.Y.: Springer.

Newcombe, H. B., Kennedy, J. M. Axford, S. J., and James, A. P. (1959), "Automatic Linkage of Vital Records," *Science*, 130, 954-959.

Newcombe, H.B. and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" *Communications of the Association for Computing Machinery*, 5, 563-567.

Schenker, N., and Raghunathan, T.E. (2007), “Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health,” *Statistics in Medicine, 26, 1802-1811.*

New and Emerging Methods

Statistical methods for combining multiple sources of administrative and survey data

Marco Di Zio (Istat), Li-Chun Zhang (University of Southampton & Statistics Norway), Ton De Waal (Statistics Netherlands & Tilburg University)

1. Introduction

Combining multiple sources of administrative and survey data is becoming increasingly common in the production of official statistics, as it presents great opportunities for generating statistics of greater scope and richer details, with less burden on the respondents and reduced cost of the producers. On the one hand, there is an established tradition to use administrative data for the creation of sampling frames and for providing auxiliary information in data collection, editing, imputation and estimation. On the other hand, more and more often, the data from administrative sources are being transformed directly into statistical data, either in place of purposefully collected census or survey data, or to be combined with the latter on an equal footing in estimation. Such *secondary statistical uses* of administrative data, whether or not together with *primary* statistical data collected in census and sample surveys, open up new methodological questions.

There is currently much on-going research and development regarding the principles of statistical inference based on multi-source data, the relevant statistical methods and quality assessment of the resulting statistical products. Zhang (2012) outlines the potential errors in integrated statistical data from a total error perspective. The approach is based on a two-phase extension of the life-cycle schema of administrative data described in Bakker (2010), which in turn was an adaption of the total-survey-error framework of Groves et al. (2004), De Waal et al. (2017) categorise the most typical situations involving multiple data sources, and list many related statistical methods. The categorisation provides also a basis for the on-going development of a quality assessment framework for multi-source statistics (Komuso, 2017).

The purpose of this paper is to provide a *short* overview of statistical methods relevant for combining multiple administrative and survey (including census) data sources, in the context of the aforementioned secondary statistical uses of administrative data. The statistical methods are grouped in terms of a classification of the generic *statistical tasks* of data integration*,* which has recently been developed in the research project “Estimation methods for the integration of administrative sources” funded by Eurostat (Eurostat 2017). The statistical tasks can be considered as the ‘building blocks’ of a process, in order to produce statistics based on multiple sources. Each statistical task specifies *what is to be done* to the data, whilst the relevant statistical methods address the question *how it can be done in various ways.* In this way, the statistical tasks provide the connection between the statistical methods for integrated data and the statistical usage of integrated data, and they can serve as an economic means to a synopsis. In a sense we place ourselves in the position of a statistician at a statistical institute who is confronted with a number of tasks in order to process and analyse multi-source data, and who wants to know the appropriate methods for carrying out these tasks.

Due to the space limit, the description of methods will have to be brief, and only selected references are provided here. We refer to e.g. De Waal et al. (2017) and Eurostat (2017) for more extensive lists of references. It should be noticed that some of the tasks and methods are as well relevant to the traditional *survey-assisting uses* of administrative data, which also require combining data sources. However, in the description here we will focus on the aspects that are most prominent in the context of secondary statistical uses of administrative data.

The synopsis leads naturally to the question of gaps in the statistical methods and the inference approach for producing official statistics from multiple primary and secondary data sources. We discuss the matter on several different levels, albeit shortly, and point to some broader issues that can both affect and are dependent on the methodological development.

2. Statistical tasks and related statistical methods

In general, the statistical tasks refer to:

1. Transformation of administrative data for secondary statistical uses, i.e. transforming objects and attributes to statistical units and measurements needed for the target statistics.

2. Integration of the different data sources to join initially separate and possibly overlapping statistical information, either at the level of micro-data or aggregated estimates.

The generic statistical tasks and associated methods are described below.

I. Data editing and imputation

Errors are virtually always present in the data files used by producers of statistics, also when data originate from external data sources. In order to produce statistical output of sufficient accuracy, it is important to detect and treat these errors. Integration of data sources at micro-level may give rise to *composite records* that consist of a combination of values obtained from different sources that may lead to consistency problems because the information is conflicting, in the sense that edit rules that involve variables obtained from the different sources are violated, or there may be conflicting values of the same variable in different sources. The purpose of editing conflicting micro-data is to achieve numerical consistency in the first instance and, ultimately, statistical consistency of the resulting aggregates. It is worth noting that editing conflicting micro-data is used to resolve apparent numerical inconsistencies between *values* of the same variable from different data sources. While variable harmonisation (described later on) is applied to situations where there exist *similar versions* of the *same target* variable, for instance due to different definitions across the sources.

Methods

For a general description of the editing and imputation methods we refer to De Waal et al.

(2011) and Memobust (2014). Methods concerned with the reconciliation of conflicting micro-data are often based on the idea that the values in the record with inconsistent microdata are changed, as little as possible, such that the modified record is consistent in the sense that it satisfies all edit rules. It can be described as minimising a chosen distance between the original (inconsistent) record and the adjusted record, subject to the constraint that all edit rules are satisfied by the adjusted record. The optimisation approach resolves inconsistencies in data records with numerical variables that are required to adhere to a set of specified linear edit rules. Methods discussed in literature include prorating, minimum adjustment methods, generalised ratio adjustments for continuous data – see e.g.: Pannekoek and Zhang (2015), and Canceis (Bankier et al., 2000) for categorical data.

II. Creation of joint micro data

- Multiple datasets are composed of the same units.

Methods

Record linkage procedures generate links based on the comparison of individual identifiers among the available datasets. If the link is based on a deterministic rule, e.g., if all or some identifiers are identical, the procedure is named deterministic record linkage. It is a viable approach when the entities in the datasets are identified by common identifiers that are observed without errors. In case unique identifiers are not available or are affected by errors, probabilistic record linkage procedures are suggested. They take into account the uncertainty of links (e.g., the probability of being a match and a non-match) and decision rules are based on this information. References for the classic probabilistic record linkage procedure are Fellegi and Sunter (1969) and Jaro (1989). An alternative more recent approach is Steorts et al (2016), under which records are directly linked to latent true individuals and only indirectly linked to other records.

- Multiple datasets are composed of different units.

When data sources share some common variables but not common units, statistical matching (SM) techniques are used. In the micro approach, SM aims at creating a synthetic dataset in which all the variables are available by exploiting the information residing in the common variables.

Methods

Statistical methods are essentially based on the idea of predicting (or imputing) the missing information in one dataset chosen as reference dataset. Imputation methods exploiting the information observed in the common variables are generally used to complete the dataset. A typical example is given by nearest-neighbour imputation where the record of a (reference) dataset is enriched with information of a similar unit observed in another dataset, and the similarity is computed by measuring the distance between the values observed in the set of common variables. More methods can be found in D’Orazio et al. (2006).

III. Alignment of statistical data

- Alignment of units: Harmonisation of relevant units, creation of target statistical units.

When administrative data are used, we sometimes need to transform available objects into statistical units referring to the target population. This is called harmonisation or alignment of units. This means that the objects observed in the administrative data sources must be transformed into relevant - possibly several - types of units and the target statistical unit may then need to be created on the basis of them. For example, to create *living household* as the ideal unit for household income statistics, one needs to make use of units such as person, family, residence address, study or workplace address, etc. The different units need first to be aligned with each other, in order to improve the accuracy of the living household created based on them.

Methods

No general statistical methods are available at the moment. Rule-based deterministic methods are commonly used in practice. For instance for business statistics, statistical units are derived from legal units using certain deterministic derivation rules that take account of ownership relations (Memobust, 2014a).

- Alignment of measurements: Harmonisation of relevant variables, derivation of target statistical variable.

When administrative data are used, we sometimes need to transform attributes of the objects into measurements referring to the target concepts, i.e. the target variables. This is called harmonisation or alignment of measurements from different sources.

Differences in definition can occur between variables in different sources. In particular, variables in an administrative data source are defined according to the administrative purposes of the register owner. These definitions may differ from those of the target variables for statistical purposes. For example, a tax authority collects data of value-added tax (VAT) declarations from businesses which contain turnover values. Since the administrative purpose of these data is to levy taxes on turnover, the tax authorities will be interested only in the amount of turnover of each business that is derived from taxable economic activities. Depending on the specific tax regulations that apply, for some sectors these administrative turnover values will differ from the turnover values that a statistical institute needs: some economic activities that are relevant for economic statistics may be exempt from taxes, and vice versa.

Discrepancy can also arise because of source-specific measurement, despite a variable having the same definition everywhere. An example is when the variable ‘age-group’ is observed in two data sources with different groups. A task for obtaining a unique variable ‘age’ is needed, which can involve a statistical estimation method (e.g. classification techniques) whenever a well-defined mapping is not known.

In cases such as those above, the relevant variables need to be harmonised during data integration: for each unit in the integrated dataset, the values of the target variable according to the desired definition need to be estimated from the observed values.

Methods

The alignment of measurements is commonly performed by means of ad-hoc procedures. However, in recent papers (see Eurostat 2017a), a more general approach is applied based on *latent variable models*. Such models have a long history in psychology, sociology, econometrics, etc. However, validation of latent variable models may require different approaches for descriptive official statistics than statistical analysis that is of a theoretical nature. See discussion of gaps later on.

IV. Multisource estimation at aggregated level

This task refers to producing estimates at aggregated levels based on multi-source data including possibly multiple administrative datasets, with or without survey or census data in addition. One of the most important problems to deal with is the consistency of estimates, i.e., numerically and statistically consistent estimation of variables. Macro-integration is the general process of reconciling statistical figures on an aggregate level.

- Consistency of estimates of common variables.

When using a mix of administrative and survey sources, estimation may result in different estimates concerning the same variable, if one does not take special precautions. For instance, one may get different estimates of the same variable based on two samples, because different units and different weights are used in the two samples. In the ideal case, these differences are merely caused by “noise” in the data, such as the sampling errors. Therefore, from a statistical point of view, different estimates concerning the same variables are to be expected and do not constitute a problem on the face it. However, different estimates may cause problems for the users when trying to interpret these estimates.

A similar problem of macro-integration arises when times series data of the same variable exist with different frequencies in different sources, e.g. quarterly (or yearly) administrative data vs monthly survey data. The data of the lower frequency may be based on a larger or more reliable set of data, in which case one may adjust the higher-frequency time series accordingly, while keeping the adjustments as small as possible in some sense.

Methods

Several approaches have been developed. For instance, in the repeated weighting approach a separate set of weights is assigned to sample units for each table of population totals to be estimated in order to achieve numerical consistency between tables; in the repeated imputation approach separate imputations are assigned to all population units for each table of population totals to be estimated. More generally, mass imputation (i.e., all variables and all population units missing are imputed) guarantees the numerical consistency of estimates. The quality of mass-imputed data depends on the ability to capture the relevant variables and relations between them in the imputation model, and to estimate the model parameters accurately.

A practical problem is how to avoid misuses of a mass imputed dataset. For instance, if some variables are not included in the imputation model, future analysis studying relations with these variables can be misleading. Let us imagine that a model for income is estimated at the NUTS2 level of aggregation and the analyst computes estimates at a finer level (NUTS3). In this case the analyst will observe independence between the imputed income with NUTS3 within the NUTS2-stratum, which is entirely due to the imputation model and not the data (Eurostat, 2017).

Macro-integration with times series is not a new problem, it is well studied and methods are developed to this aim, see for instance (De Waal et al., 2011).

- Consistency of variables that relate to each other in terms of constraints (e.g., accounting equations).

The estimates involved in a system of accounting equations, such as the supply-and-use table for National Accounts, are often derived based on different sources, and initially do not satisfy the accounting equations directly. This is a problem of coherent estimation, whereby the initial estimates need to be adjusted in order to satisfy the accounting equations. The resulting final accounts are affected by both the initial estimators and the method of adjustment. Appropriate summary of the “accounting uncertainty” can thus point to the most effective improvements needed for the initial estimators, as well as help to choose the efficient adjustment method.

Methods

This problem studied in National Accounts is known as a balancing problem. Early works are Stone (1942) and Byron (1978). These algorithms can be applied also to reach consistency of demographic figures (Di Zio et al., 2017).

- Multiple lists with imperfect coverage of the target population: Population size estimation.

The problem of an imperfect frame which needs to be estimated before anything else is not a new one. It has nevertheless generated quite a lot of interest in recent years, especially in the context of replacing traditional censuses with administrative registers. When each one of the multiple sources has imperfect coverage of the target population, including both under- and over-coverage, statistical methods for population size estimation are needed. The most common approaches are based on capture-recapture methods where each unit in a data source is considered a capture. In estimating the population size based on several administrative sources, the misalignment between the scope of the administrative data and that of the statistician poses several methodological challenges and sets us apart from a classical capture-recapture setting. For instance, it is often useful to develop methods taking into account the dependency among data sources, the fact that some data sources refer to a specific subpopulation, and that data may contain units outside of the target population that are not deterministically identifiable.

Methods

Log-linear models are generally used to deal with dependency of lists, latent variables are used for modelling in and out of scope units, and algorithms to estimate missing data are introduced to cope with the problem of lists observing specific subsets of populations only. A model taking simultaneously into account these three components is discussed in Di Cecco et al. (2017).

3. Discussion of gaps

Gaps in statistical methods

There is obviously much potential in further developing the statistical methods already introduced. As an example we consider here harmonisation of measurements based on latent variable models, which has received much attention recently. In principle the approach offers an attractive general formulation compared to traditional deterministic derivation rules, because the latter rely strongly on subjective decisions. However, a drawback is that the existing models often require strong assumptions that are unlikely to hold in many practical situations, e.g. normally distributed data, independent errors between sources, etc. Future research should therefore pay attention to two aspects simultaneously. On the one hand, more realistic models should be developed for the data encountered in official statistics. On the other hand, viable model selection and validation approaches need to be developed. The investigation of sensitivity towards (minor) departures from the model assumptions must be a systematic part of the model building. Moreover, unlike when statistical models are used for theoretical analysis, many official statistics are of a descriptive nature, which places different requirements on model validation, including the latent variable models. Audit sample survey may be necessary; but the design and analysis of audit sample survey for validation of latent variable models is another gap that needs to be filled.

A related broader issue is the conceptualisation of the target construct which the measurement is intended to capture. For instance, an often studied application of latent variable models is economic activity status, for which both survey and register data are available. But we are not aware of any official statistics based on such an approach. One could e.g. ask the question: what might be a meaningful and judicious definition of Employment, when the majority of the data is to be obtained from the related administrative sources? With respect to both temporal and spatial disaggregation register data will be superior to anything that can possibly be achieved by periodic sample surveys. Can we adjust the definition of Employment (and relevant metadata) to take advantage of this, instead of keeping the existing definition and resorting to a latent variable approach?

Gaps in statistical tasks

A statistical task where relatively little has been accomplished in terms of statistical methods is unit harmonisation. As noted above, in practice rule-based deterministic procedures are generally used. Research to develop a stochastic approach should be planned, in which respect the latent entity resolution approach to record linkage may potentially provide inspiration. For example, the allocation of persons into living households may be represented by links between the unobserved living households and the observed persons. The links can be introduced and removed based on statistical models; each realisation of all the links would then yield a different living household population, making unit harmonisation stochastic in nature. Unit errors are unavoidable, which can be evaluated and taken into account when producing statistics, see e.g. Zhang (2011).

Indeed, one may expect much more research and development in the coming years on the various and many problems that may be referred to as *entity ambiguity*. For instance, linkage errors are ultimately due to the lack of entity identification, erroneous enumeration can be caused by the failure to detect out-of-scope units, convolute entity may be the case when multiple elements are wrongly identified as a single one, missing units may be due to imperfect alignment among the relevant entities, etc. All these entity ambiguity problems become unavoidable in general when combining multiple datasets, and are often critical to the subsequent statistical uses.

Gaps in inference framework

Notwithstanding the various issues of combining multiple sources, there is still clearly a lack of statistical theories for assessing the uncertainty of statistics based solely on administrative data, i.e. the so-called register-based statistics. Making transformed administrative data the main body of statistical data seems necessary, in order to overcome the challenges that are eroding the foundation of traditional *survey-based register-assisted* outlook. Such challenges include limited capacity to satisfy user demands of statistics with greater scope and richer details, decreasing quality due to non-sampling survey errors despite the ever-increasing costs that are being demanded to counter the trend. Whilst statistical modelling will be the primary means of estimation for register-based statistics, it seems important to develop efficient audit sampling methods, both to inform the model building and for model validation over time. This may result in a gradual transition towards a sustainable *register-based survey-assisted* approach. However, both the basic principles and details of such an inference framework remain to be clarified. For instance, suppose capture-recapture models are used to produce detailed population statistics based on multiple administrative registers, and a coverage survey of sorts is conducted. Should the survey data only be used to validate the model-based estimates and possibly affect the choice of the estimation model, or should it affect the estimates directly? Should the uncertainty be evaluated only with respect to the model, even when the coverage survey data is directly built into the population estimates? How to incorporate in the uncertainty assessment the sampling error of the coverage survey, when the data is only used to facilitate the model selection?

Gaps in statistical needs

What are the statistics/parameters of interest? In the discussion above we have implicitly in mind the traditional needs of statistics. But part of the historical origin of the existing needs is the data availability itself. So what are the newly arising possibilities of integrated data sources? Sooner or later one should be asking this question more seriously. The answer will depend on the available statistical methods and, at the same time, point to directions of methodological development. For instance, linked census-like datasets of household and business populations would e.g. make it possible to represent the data in networks of nodes instead of simply in different lists of elements. Various network parameters can be defined to capture the statistical needs, beyond what is immediately apparent based on the list-representation of relevant populations. Statistical methods for modelling and sampling networks are e.g. needed to address the elevated statistical needs.

4. Conclusion

Concluding this paper, we would like to say that we feel that we are just at the start of multisource statistics and that the landscape of official statistics is likely to change rapidly. It seems hardly sustainable to rely on sample surveys and censuses as the only or even main source of statistical data in future, whilst the potential benefits of high-quality statistical data from administrative (and other) sources are too numerous to be ignored. This means that we will need to face and overcome many theoretical and practical challenges of methodology in the coming years.

References

Bakker B. F. M. 2010. Micro-Integration: State of the art. In Report WP1: State-of-the-art on Statistical Methodologies for Data Integration, ESSNET on Data Integration, http://www.cros-portal.eu/content/wp1-state-art.

Bankier, M., Poirier, P., Lachance, M. and Mason, P. (2000) A Generic Implementation of the Nearest Neighbour Imputation Methodology (NIM). *Proceedings of the second International Conference on Establishment Surveys*, Buffalo, pp. 571-578.

Byron, R., (1978) The estimation of large social account matrices, Journal of the Royal Statistical Society A, Vol. 141(3), pp. 359-367.

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

De Waal, T., A. Van Delden and S, Scholtus (2017) Multi-source Statistics: Basic Situations and Methods. Discussion paper, Statistics Netherlands.

Di Zio M., Fortini M., Zardetto D., (2017). Balancing Methods for Ensuring Time and Space Consistency of Demographic Estimates in the Italian Integrated System of Statistical Registers. Proceeding of the NTTS conference, Eurostat, March 14 - 16, 2017, Bruxelles.

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

Di Cecco D., Di Zio M., Filipponi, D., Rocchetti I. (2016). Estimating Population Size from Multisource Data with Coverage and Unit Errors. Proceeding of ICES-V, Geneva, Switzerland, June 20-23, 2016.

Eurostat 2017. Deliverable 5b of "Estimation methods for the integration of administrative sources" research project funded by Eurostat.

Eurostat 2017a. Deliverable "Method: Variable harmonisation based on latent variable models" of "Estimation methods for the integration of administrative sources" research project funded by Eurostat.

Fellegi I.P. and Sunter A.B. (1969). A theory for record linkage, Journal of the American Statistical Association, Vol. 64: 1183-1210.

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Tourrangeau (2004), Survey methodology, Wiley, New York.

Jaro M.A. (1989), Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 84, 414–420.

Komuso (2017), ESSnet on Quality of Multisource Statistics Final report WP 3 "Framework for the quality evaluation of statistical output based on multiple sources", Deliverable of the ESSnet on Quality of Multisource Statistics.

Memobust (2014). *Memobust Handbook on Methodology of Modern Business Statistics.* https://ec.europa.eu/eurostat/cros/content/memobust_en
Memobust (2014a) Theme module: Derivation of Statistical Units.

https://ec.europa.eu/eurostat/cros/content/memobust_en
Pannekoek J., Zhang L-.C (2015). Optimal adjustments for inconsistency in imputed data. Survey Methodology, Vol. 41, No. 1, pp. 127-144.

Steorts, R.C., Hall, R., Fienberg S.E. (2016). A Bayesian Approach to Graphical Record Linkage and De-duplication, Journal of the American Statistical Association, Theory and Methods, Vol. 111 (516), pp. 1660-1672.

Stone, R., Champernowne, D.G., and Meade, J.E., (1942) The Precision of National Income Estimates, Review of Economic Studies (1942), vol. 9 (2), pp. 111-125.

Zhang, L-C. (2011). A Unit-Error Theory for Register-Based Household Statistics. Journal of Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 415–432.

Zhang, L-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica, vol. 66, pp. 41-63.

Book and Software Review

A Shiny application for statistical disclosure control: sdcApp opening the power of sdcMicro to non-R users

T. Benschop[2], M. Welch[3]

1. R and Shiny applications for Statistical Disclosure Control (SDC)

Templ and Todorov [1] discuss the usefulness of the open-source statistical software R for the tasks of statistical offices. TThey also provide a comprehensive overview of the R packages[4] related to different topics in official statistics and survey statistics. The package sdcMicro [2], which is developed and maintained by Alexander Kowarik, Bernhard Meindl and Matthias Templ, includes the necessary methods for the preparation of anonymized microdata files. The sdcMicro package is a powerful application when used by an R expert from the command line. While a growing number of staff at statistical agencies are using R, the majority are users of the well-known commercial products and do not have the R skills necessary to use sdcMicro effectively from the command line. With its ability to facilitate the creation of interactive web applications, Shiny, from RStudio[5], provides a way to bring the power of R to non-R users. Shiny essentially allows the creation of a graphic user interface for R applications. It allows the development of R applications which run locally in a web browser, but frees the developer from the need to write HTML or JavaScript code themselves.

Faced with an increased demand from statistical agencies to provide support, which would allow agencies to grant greater access to anonymized microdata, the World Bank sponsored the development of a Shiny application for sdcMicro - sdcApp.[6] Funding for the development of sdcApp was provided by the World Bank Group[7] and the United Kingdom Department for International Development (DfID)[8]. Development of the application was carried out by the developers of sdcMicro. The application is available as part of the sdcMicro package in the R CRAN repositories as well as in GitHub.

[2] Thijs Benschop is a research associate at Humboldt University Berlin and a consultant for the World Bank.

[3] Matthew Welch is a Senior Statistician in the Survey Unit of the Development Data Group at the World Bank.

[4] R and the packages available in the CRAN repositories are open-source and can be downloaded free of charge from: https://cran.r-project.org.

[5] See https://shiny.rstudio.com/ for more information on Shiny.

[6] The sdcApp was introduced into the sdcMicro core package in version 5.0.0. Before the release of version 5.0.0. of the sdcMicro package, a GUI with limited functionality was available in a separate package, sdcMicroGUI. This package is now discontinued with the Shiny application embedded as part of the sdcMicro core package.

[7] http://www.worldbank.org

[8] https://www.gov.uk/government/organisations/department-for-international-development

The remainder of this review is organized as follows. Section 2 gives an overview of the relevant SDC methods and tools in the sdcMicro package. Section 3 describes how to install and load the sdcApp and provides an overview of the structure of the application.
Section 4 describes the structure and functionality in more detail as well as a selection of its features. Section 5 summarizes and concludes.

2. Microdata anonymization in R with the sdcMicro package

The functionality of the sdcMicro package can be categorized into four parts: 1) anonymization methods, 2) risk measures, 3) utility measures, and 4) data manipulation functions. All the common anonymization methods for microdata used in statistical offices are implemented in sdcMicro. These include, amongst others, global recoding, local suppression, post-randomization (PRAM), top- and bottom-coding, microaggregation, noise addition and rank swapping. Measures for estimating disclosure risk for both categorical and numerical variables are available in sdcMicro and include k-anonymity, individual and global risk, l-diversity, SUDA scores and proximity measures. To measure information loss due to the anonymization process standard measures such as differences in eigenvalues and tabulations are computed.

Users wanting to compute custom utility measures before and after anonymization would typically fall back to other packages developed in R. R and sdcMicro allows users to read data from files in “foreign” data formats, such as SAS, SPSS, STATA and CSV files. Data manipulation required for the SDC process can be carried out with R and sdcMicro functions.

3. sdcApp - a Shiny interface for sdcMicro

The use of R for many users, especially in statistical agencies, is relatively new. Many staff in agencies would like to apply SDC methods but do not have the necessary R skills to use sdcMicro. These users would benefit from a friendly GUI for sdcMicro. The aim being to provide a GUI that removes the need for proficiency in R, but still allows access to all the features of sdcMicro. The Shiny [4] application, sdcApp, built into the sdcMicro package fulfills that need. The sdcApp is started from R after loading the sdcMicro package and is launched in a web-browser running locally on the user’s machine.

sdcApp implements all the main functionality available in the sdcMicro package. In addition, and in an effort to help users not familiar with R, a number of additional features are brought in from other R packages. These make measuring utility and visualizing the data and changes made in the SDC process easier. The aim being to allow users to complete the whole SDC process without leaving the sdcApp environment and without the need to revert to command line R.

To use the application, the user needs to install the latest version of the R software from the CRAN website: https://cran.r-project.org. R is available for free for the following platforms; Linux, Windows and Mac OS X. After installing R, the user installs the sdcMicro add-on package as well as other packages used by sdcMicro (so-called dependencies). This is done automatically by running the command *install.packages (“sdcMicro”)* in the R console. After installing sdcMicro, the package sdcMicro needs to be loaded with the command *library (sdcMicro).* sdcApp is then launched with the command (*sdcApp)*.

The application launches in the default web browser of the system. The user can interact with the application by using control inputs such as buttons, drop-down menus, sliders, radio buttons or text input. No further interaction with the R console is required.

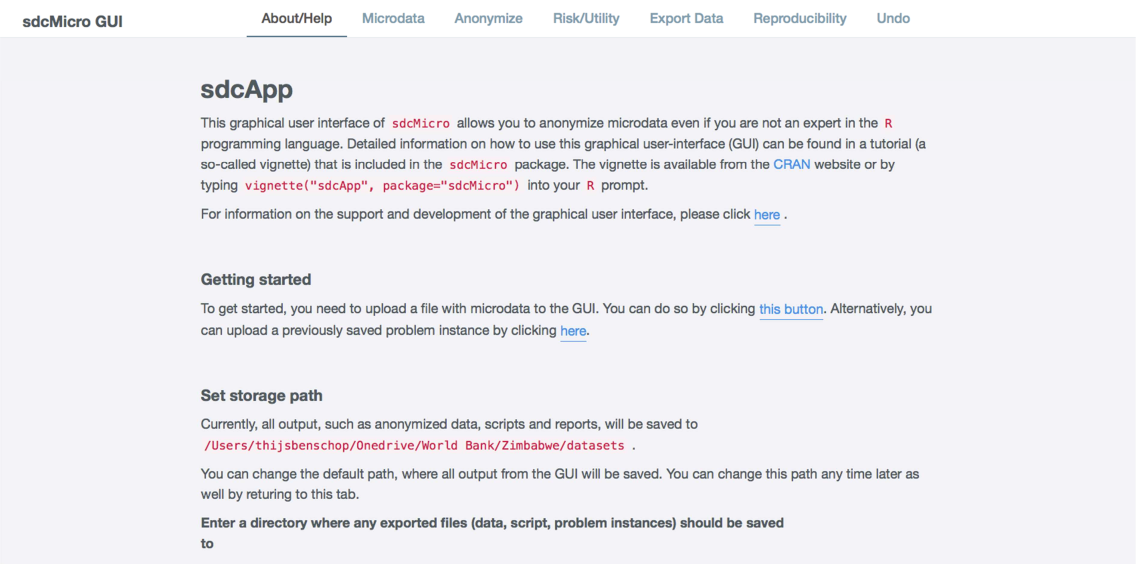
SdcApp consists of seven tabs that can be navigated using the top navigation bar. Table 1 gives an overview and description of the tabs, which are further described in the next section.

Screen tab name	Description
About/Help	Help and general settings
Microdata	Load and prepare dataset
Anonymize	Anonymization methods
Risk/Utility	Risk and utility measures
Export Data	Export data, reports
Reproducibility	Generate R script
Undo	Undo steps in anonymization process

Table 1 Overview of screen tabs in sdcMicro application

4. Description of sdcApp

The application opens in the tab *About/Help*, which is shown in Screenshot 1. This screen provides basic information on using the application as well as some options to specify where output will be saved. Also, being an open source project, information and links are provided for providing feedback and contributing code to the project through GitHub.



sdcApp

This graphical user interface of `sdcMicro` allows you to anonymize microdata even if you are not an expert in the `R` programming language. Detailed information on how to use this graphical user-interface (GUI) can be found in a tutorial (a so-called vignette) that is included in the `sdcMicro` package. The vignette is available from the CRAN website or by typing `vignette("sdcApp", package="sdcMicro")` into your `R` prompt.

For information on the support and development of the graphical user interface, please click here .

Getting started

To get started, you need to upload a file with microdata to the GUI. You can do so by clicking this button. Alternatively, you can upload a previously saved problem instance by clicking here.

Set storage path

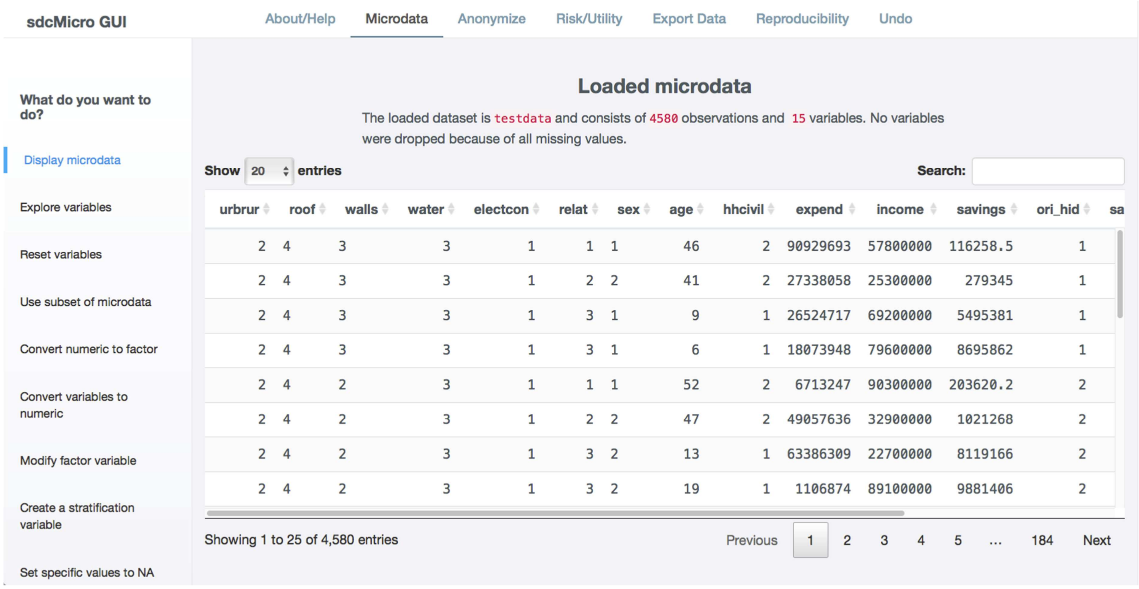
Currently, all output, such as anonymized data, scripts and reports, will be saved to `/Users/thijsbenschop/Onedrive/World Bank/Zimbabwe/datasets` .

You can change the default path, where all output from the GUI will be saved. You can change this path any time later as well by returing to this tab.

Enter a directory where any exported files (data, script, problem instances) should be saved to

Screenshot 1 The application opens in the tab *About/Help*

The first step to start using the application for anonymization of microdata is to load the microdata. This is done from the *Microdata* tab. In addition to the standard R data format, the application supports several other file formats such as, SPSS (.sav), SAS (.sas7bdat), CSV (.csv) and STATA (.dta) files. For testing and demonstration purposes, the application includes two test datasets (testdata and testdata2). All data are loaded locally into R and the web browser is only used to communicate with R. An internet connection is not required during the anonymization process. After loading the Dataset into the application, the *Microdata* tab shows the loaded dataset and allows the user to explore, manipulate and prepare the data for the anonymization process. This is shown in Screenshot 2. The available functions for data exploration depend on the variable type. Examples are tabulations, histograms, mosaic plots and standard summary statistics. Examples of data preparation are variable type conversion or setting missing values to R system missing values. This tab also provides functionality to deal with datasets with a hierarchical structure, such as household surveys as well as to use only a subset of the full dataset for quicker testing.



Loaded microdata

The loaded dataset is testdata and consists of 4580 observations and 15 variables. No variables were dropped because of all missing values.

Show 20 entries

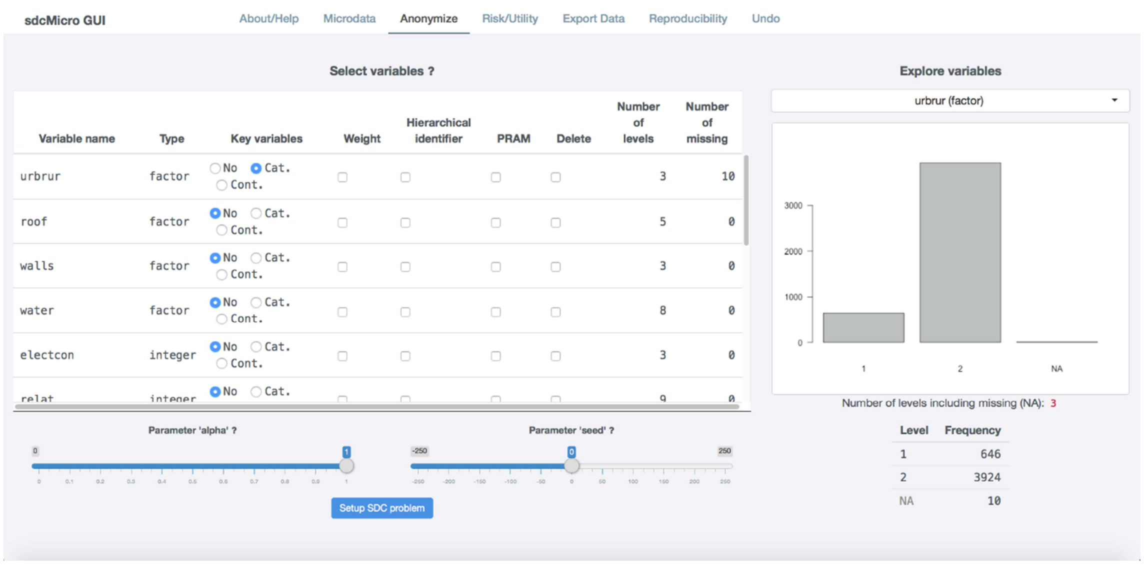
Search:

urbrur	roof	walls	water	electcon	relat	sex	age	hhcivil	expend	income	savings	ori_hid	sa
2	4	3	3	1	1	1	46	2	90929693	57800000	116258.5	1	
2	4	3	3	1	2	2	41	2	27338058	25300000	279345	1	
2	4	3	3	1	3	1	9	1	26524717	69200000	5495381	1	
2	4	3	3	1	3	1	6	1	18073948	79600000	8695862	1	
2	4	2	3	1	1	1	52	2	6713247	90300000	203620.2	2	
2	4	2	3	1	2	2	47	2	49057636	32900000	1021268	2	
2	4	2	3	1	3	2	13	1	63386309	22700000	8119166	2	
2	4	2	3	1	3	2	19	1	1106874	89100000	9881406	2	

Showing 1 to 25 of 4,580 entries

Screenshot 2 *Microdata* tab after loading the testdata dataset

After loading and preparing the dataset, the user can navigate to the *Anonymize* tab to select the key variables and create the anonymization problem instance. This is shown in Screenshot. In an interactive table the user can select categorical and continuous key variables, the sampling weight, a hierarchical identifier, variables suitable for the PRAM method as well as variables to be removed from the dataset before release. If an invalid choice is made, e.g., the user selects a variable both as key variable and sampling weight, the application provides feedback in the form of a pop-up window stating the variable and the error. A number of parameters can also be set; these include setting the parameter alpha used for the k-anonymity calculation as well as setting a seed for the random number generator for use in probabilistic methods. In the right sidebar on this screen, the user can browse summary information of the variables, such as frequency counts, histograms and summary statistics.



Select variables ?

Variable name	Type	Key variables	Weight	Hierarchical identifier	PRAM	Delete	Number of levels	Number of missing
urbrur	factor	○ No ● Cat. ○ Cont.	□	□	□	□	3	10
roof	factor	● No ○ Cat. ○ Cont.	□	□	□	□	5	0
walls	factor	● No ○ Cat. ○ Cont.	□	□	□	□	3	0
water	factor	● No ○ Cat. ○ Cont.	□	□	□	□	8	0
electcon	integer	● No ○ Cat. ○ Cont.	□	□	□	□	3	0
relat	integer	● No ○ Cat.	□	□	□	□	9	0

Parameter 'alpha' ? 0 – 1: 1

Parameter 'seed' ? -250 – 250: 0

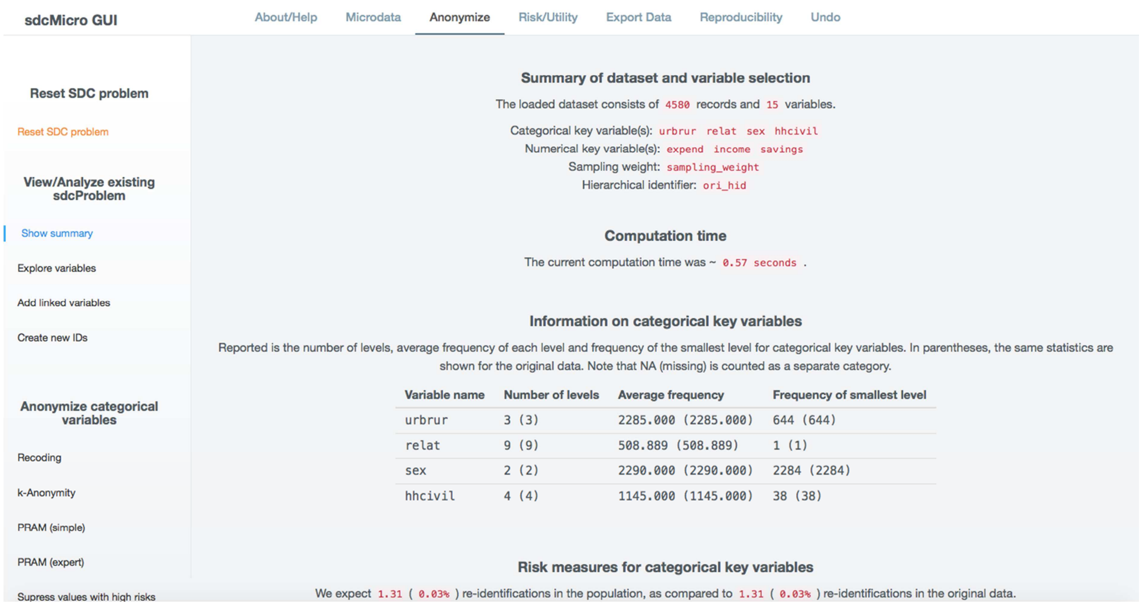
Explore variables

Number of levels including missing (NA): 3

Level	Frequency
1	646
2	3924
NA	10

Screenshot 3 *Anonymize* tab with the interactive variable selection table

Once the user has made the selection of variables and clicked the button ‘Setup SDC problem’, the *Anonymize* tab shows a summary view of the SDC problem, including the variable selection and selected risk measures as shown in Screenshot . In the left sidebar the anonymization methods can be selected. They are grouped by variable type. For categorical variables, the methods global recoding, local suppression to achieve k-anonymity and PRAM can be selected. For numerical variables, the methods top-/bottom-coding, microaggregation, rank swapping and noise addition are available.



Summary of dataset and variable selection

The loaded dataset consists of 4580 records and 15 variables.

Categorical key variable(s): urbrur relat sex hhcivil
Numerical key variable(s): expend income savings
Sampling weight: sampling_weight
Hierarchical identifier: ori_hid

Computation time

The current computation time was ~ 0.57 seconds .

Information on categorical key variables

Reported is the number of levels, average frequency of each level and frequency of the smallest level for categorical key variables. In parentheses, the same statistics are shown for the original data. Note that NA (missing) is counted as a separate category.

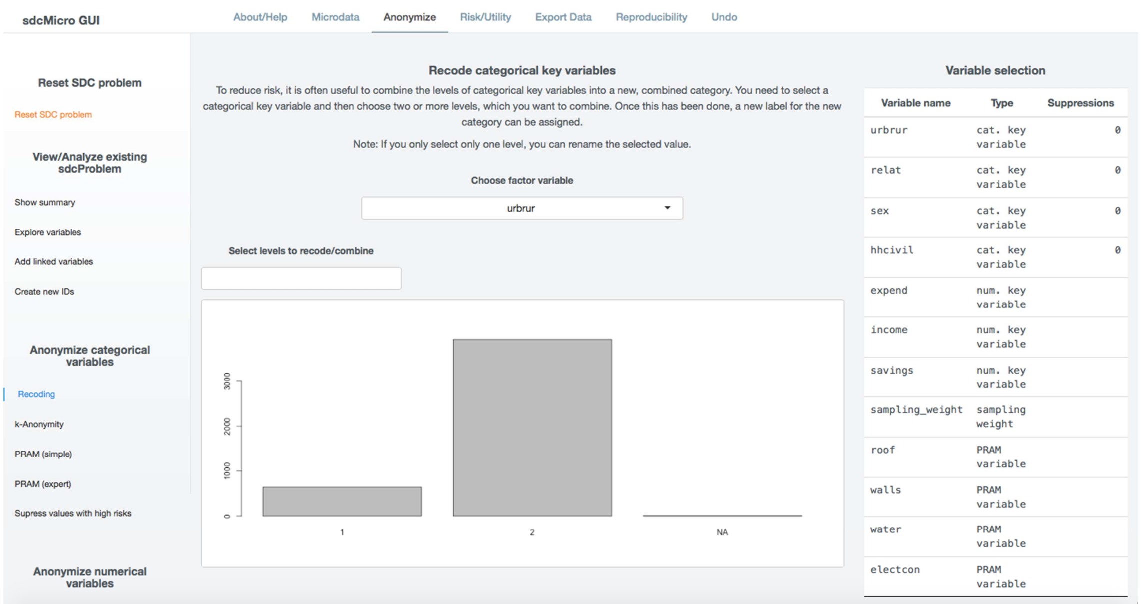
Variable name	Number of levels	Average frequency	Frequency of smallest level
urbrur	3 (3)	2285.000 (2285.000)	644 (644)
relat	9 (9)	508.889 (508.889)	1 (1)
sex	2 (2)	2290.000 (2290.000)	2284 (2284)
hhcivil	4 (4)	1145.000 (1145.000)	38 (38)

Risk measures for categorical key variables

We expect 1.31 (0.03%) re-identifications in the population, as compared to 1.31 (0.03%) re-identifications in the original data.

Screenshot 3 *Anonymize* tab after setting up the SDC problem with summary

After selecting a method, the user is presented with a three column page as shown in Screenshot 5: on the left the methods that can be selected, on the right a summary overview of the current problem is shown, the main part presents options and parameter settings for applying the currently selected method. The parameters and options available to the user in the sdcApp are the same as those available from the command line in sdcMicro. For each method a brief description is included and for most parameters a help button provides more information.



Recode categorical key variables

To reduce risk, it is often useful to combine the levels of categorical key variables into a new, combined category. You need to select a categorical key variable and then choose two or more levels, which you want to combine. Once this has been done, a new label for the new category can be assigned.

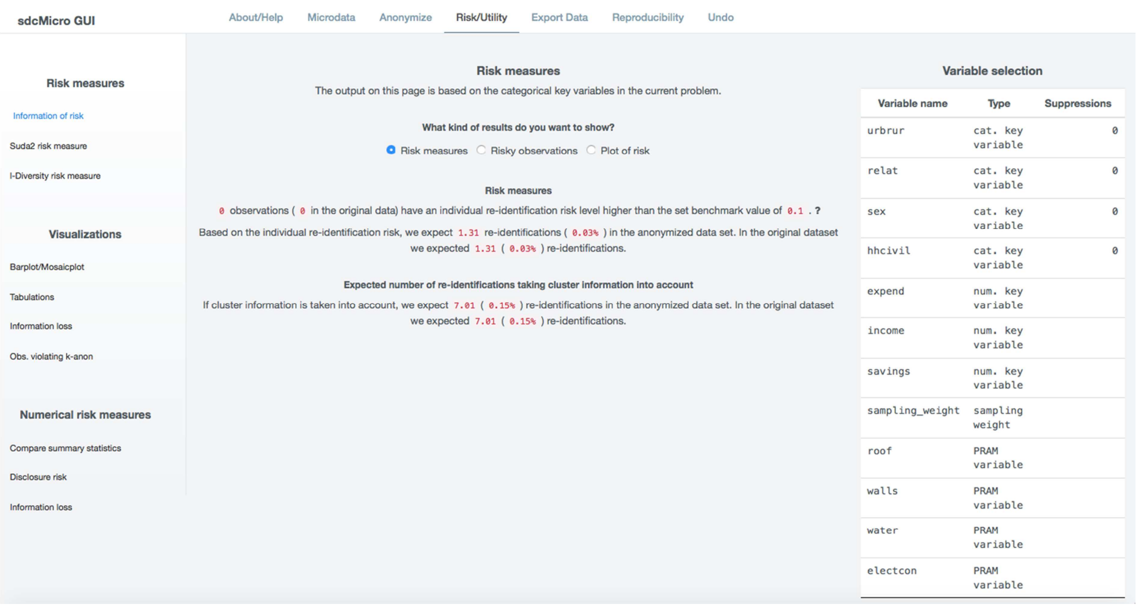
Note: If you only select only one level, you can rename the selected value.

Variable selection

Variable name	Type	Suppressions
urbrur	cat. key variable	0
relat	cat. key variable	0
sex	cat. key variable	0
hhcivil	cat. key variable	0
expend	num. key variable	
income	num. key variable	
savings	num. key variable	
sampling_weight	sampling weight	
roof	PRAM variable	
walls	PRAM variable	
water	PRAM variable	
electcon	PRAM variable	

Screenshot 5 *Anonymize* tab after selecting the method recoding

In order to compare risk and measure utility (information loss) before and after applying a specific method, the tab *Risk/Utility* presents the user with a range of detailed risk measures and selected utility measures; as shown in Screenshot. All the risk and utility measures are automatically updated after an anonymization method is applied.



Risk measures

The output on this page is based on the categorical key variables in the current problem.

What kind of results do you want to show?

● Risk measures ○ Risky observations ○ Plot of risk

Risk measures

0 observations (0 in the original data) have an individual re-identification risk level higher than the set benchmark value of 0.1 . ?

Based on the individual re-identification risk, we expect 1.31 re-identifications (0.03%) in the anonymized data set. In the original dataset we expected 1.31 (0.03%) re-identifications.

Expected number of re-identifications taking cluster information into account

If cluster information is taken into account, we expect 7.01 (0.15%) re-identifications in the anonymized data set. In the original dataset we expected 7.01 (0.15%) re-identifications.

Variable selection

Variable name	Type	Suppressions
urbrur	cat. key variable	0
relat	cat. key variable	0
sex	cat. key variable	0
hhcivil	cat. key variable	0
expend	num. key variable	
income	num. key variable	
savings	num. key variable	
sampling_weight	sampling weight	
roof	PRAM variable	
walls	PRAM variable	
water	PRAM variable	
electcon	PRAM variable	

Screenshot 6 Risk/Utility tab

On the tab *Export Data* the user can browse the anonymized data and export the dataset in the data format of their choice. At any point in the anonymization process, the current data can be exported to perform analyses using the users' software of choice. This is useful, for example, for computing benchmark indicators and for the assessment of information loss. This can be particularly convenient if code for generating indicators and tables is already available in another software format. From the same tab the application will also generate both internal and external reports of the anonymization process.

The tab *Reproducibility* shows a commented and downloadable R script that is ready to run in the R console and can be used to recreate the SDC problem and recreate all the steps. This functionality also provides users who, at some point, wish to move to using the sdcMicro package from the command line, an easy way to learn the methods and commands available. On the same tab, the user can also export and reload the R workspace. The workspace contains the data as well as all settings, selections and results. This feature is useful as a backup as well as to restore and continue working at a later point.

Finally, the tab *Undo* allows the user to undo the last anonymization step. This is useful when exploring the best methods and parameters in a trial-and-error fashion. In order to revert to a previous state more than one step back, the user can import a previously saved workspace.

5. Conclusion

The recently developed Shiny based sdcApp makes the complete set of SDC functions and tools included in the R package sdcMicro available to a wider group of users not proficient in R. sdcApp includes the full functionality of the sdcMicro package, including all the methods, parameters and options, as well as additional functionality from other R packages, such as visualization of results and data manipulation. This makes it a complete solution for the anonymization of microdata. sdcApp is user-friendly and provides the functionality to import and export to all the major statistical package file formats. The Shiny GUI lowers the barriers to entry to sdcMicro for non-R users and brings the ability to apply the most widely used SDC methods to a larger audience. Unencumbered from having to know how to program the methods, it is our hope that this will allow more agencies to apply appropriate SDC methods which will lead to greater and safer release of microdata.

References

[1] Matthias Templ and Valentin Todorov, Official Statistics and Survey Methodology Meets R: An Overview of Corresponding Packages, The Survey Statistician, 66 (2012), 26-34.

[2] Alexander Kowarik and Matthias Templ and Bernhard Meindl, R-Package “sdcMicro” (2017), URL: https://cran.r-project.org/package=sdcMicro

[3] R Core Team, R: A language and environment for statistical computing. (2016), URL: https://www.R-project.org.

[4] Matthias Temple and Alexander Kowarik and Bernhard Meindl, Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro, Journal of Statistical Software 67(4) (2015), 1-36.

[5] Winston Chang and Joe Cheng and JJ Allaire and Yihui Xie and Jonathan McPherson, Shiny: Web Application Framework for R (2016), URL: https://cran.r-project.org/package=shiny

Monitoring Statistical Indicators Using R/Shiny: A Case Study on Public Transport Safety in Switzerland

By
T. Schoch[9]

1. Introduction

The goal of the paper is to present a software tool designed to support *indicator-based* safety monitoring that has been adopted by the safety authority for public transportation in Switzerland. The software tool is written in the R language for statistical computing (R Development Core Team, 2017) and is operated in a web browser (e.g. Mozilla Firefox).

The paper is organized as follows. In Section 2, we introduce the approach adopted in Switzerland to monitor safety performance of public transportation. Section 3 is devoted to the implementation of the tool in the R language of statistical computation. Finally, Section 4 summarizes the key messages.

2. Monitoring Public Transport Safety

The Swiss Federal Office of Transport (FOT) is the national safety authority in charge of monitoring (and regulation of) the safety performance concerning:

- rail transport,
- local transport (bus, tramways),
- cable car (funicular, aerial lift, cableways),
- ship transport / boats (on rivers and lakes).

The safety monitoring effort undertaken by FOT forms an important part of the overall risk management (including audits, inspections, etc.). The aspect of the safety monitoring referred to in this article, that is the *indicator-based* safety monitoring, provides safety intelligence and information on risks to national policy-making bodies (and to the general public).

2.1 Event Database and Indicators

The data on accidents, incidents and other hazardous events are reported to FOT by public and private transport operators in all four sectors of public transportation outlined above. The agency then processes and validates the reported data and complies the *Event Database* on an annual basis (available since 2009).

Each event datum in the Event Database is assigned a set of classifiers or codes due to the event’s characteristics in terms of the “w”-dimensions: “who” (who is affected, who is responsible), “what (happened)”, “why” and “where” did it happen. Based on these classifiers, the events are then assigned a set of base-level safety indicators (within each of the four “w”-dimensions), which serves as the foundation of the hierarchical system of the software tool presented in this article has been developed on behalf of the Swiss Federal Office of Transport (FOT). The information and views set out in this article are those of the author and do not necessarily reflect the official opinion of FOT.

[9] ECOPLAN AG – Research in Economics and Policy Consultancy, Monbijoustrasse 14, CH-3011 Bern, Switzerland. E-Mail: schoch@ecoplan.ch

Responsibility for the information and views expressed therein lies entirely with the author indicators. In total, the *system of safety indicators* is organized along the following three levels of abstraction:

- top-level safety indicators (5 indicators composed of basic- and intermediate level indicators, highest abstraction level),
- intermediate level safety indicators (55 indicators; e.g., "EA11: Train collisions"),
- base-level safety indicators (229 specific indicators; e.g., "EA114: Collisions train with road vehicle").

In addition to the indicators and the four "w"-dimensions, the Event Database stores further event-specific attributes, such as data on the severity of personal injuries and damage to material and the environment, and further contextual information.

2.2. Monitoring Objective and Evaluation

The main objective of the indicator-based monitoring of public transportation safety in Switzerland is to evaluate the safety performance in terms of the criterion: "maintain safety at least at its current level". (Note: In view of growing complexity in public transportation networks, maintaining the current level of safety is already quite a challenge). Therefore, an *intertemporal evaluation* of the safety performance is carried out. That is, the current safety performance is related to the level of safety from a previous period (safety target). Safety performance and target are expressed in terms of the indicators and the following dimensions of analysis:

- frequency / occurrence of events (e.g. number of accidents per month),
- damage to persons (measured in terms of fatality weighted injuries[10]),
- damage to material (in monetary terms).

Besides overall evaluation of the indicators (i.e. whether the current safety levels meet the targets), the software tool also facilitates subgroup comparisons (e.g., person and freight transport). Moreover, the tool provides the means to compare individual transport carriers or companies (e.g. freight operators) with each other instead of intertemporal comparisons. In these cases, the event data are normalized (e.g. by train-kilometre, route length, freight volume, etc.) prior to analysis in order to account for the companies' particular characteristics.

2.3. Statistical Methods in Use

The set of statistical methods implemented in the tool is the result of a committee decision. Besides statistical requirements, convenience and other non-technical considerations (e.g., methods must be easy to convey to a wider audience) played an equally important role in the process of method selection. The chosen methods refer primarily to intertemporal comparison, target examination (i.e. statistical testing), and computation of trends in the event data. The following methods are currently implemented:

[10] Fatality weighted (serious) injuries is a single figure combining the number of fatalities (deaths), major injuries (e.g. fractures, amputations, loss of sight, etc.) and minor injuries through the linear combination: FWI = [fatalities] + 0.1[major injuries] + 0.01[minor injuries]; FWSI is defined in similar manner but does not include minor injuries; see e.g. European Railway Agency (2014): Railway Safety Performance in the European Union 2014.

- target examination using the *overlap between confidence intervals*[11] for the arithmetic mean (current safety level vs. target),
  - parametric approach: the underlying distributional assumptions can be chosen by the analyst (models: Poisson, negative binomial, exponential, gamma, Weibull; also, all models can be fitted as hurdle-models under the assumption of zero inflation);
  - alternatively, the "ABC" approximate bootstrap method can be used to compute confidence intervals; see DiCiccio and Efron (1996).[12]
- *Wilcoxon-Mann-Whitney* test statistic on homogeneity of the distributions: current vs. target safety performance (nonparametric, rank-based statistic and thus not heavily influenced by extreme events; see Kendall, 1975),
- *trend estimation* and evaluation (either Mann-Kendall test statistic or estimating a times-series model with a deterministic trend; see Kendall, 1975),
- *moving averages* for sub-annual analysis (with approximate bootstrap for statistical inference).

Some further comments are in order. Most of the methods do not account for serial or spatial correlation that is inherent (although rather weak) in the event data. Therefore, the methods apply only in an approximate sense. Further, the set of methods does not include computer-intensive statistical methods (such as resampling methods for variance estimation). The rationale behind the exclusion of such methods is to avoid long delays (a major cause of user frustration) between the refresh cycles of the interactively operated tool. Responsiveness is therefore considered an essential usability issue.

The choice of methods is left to the analyst in charge and depends heavily on the indicator under consideration. The tool provides the analyst with graphical diagnostic measures (e.g. quantile-quantile plots) for choosing an appropriate statistical method.

3. Implementation in R / Shiny

The monitoring tool is implemented in the R statistical software (R Development Core Team, 2016) and uses the web application framework shiny (Chang et al., 2016); see Fig. 1. The tool is operated using a web-based graphical user interface and does not require any knowledge of R. To be functional, the tool only requires a local installation of R and a web browser (Mozilla Firefox™, Internet Explorer™, etc.). Moreover, the tool is started using the start menu in Microsoft Windows™ or by double click on the desktop shortcut / icon. The starting procedure (through calling a series of batch files) opens R and the preferred web browser in the background, both unnoticed by the user. We went through that "effort" with the start-up procedure – having in mind an inexperienced computer user – to provide him/her a tool that behaves like any other software.

[11] We use a modified method of overlap between confidence intervals, not the naïve overlap method; see Schenker and Gentleman (2001).

[12] The ABC method is tailored to models with distributions in the exponential family (e.g. exponential or gamma distribution); see DiCiccio and Efron (1992). Our experiences show that the method also gives reasonably good results for distributions not in the exponential family (e.g. Weibull distribution).

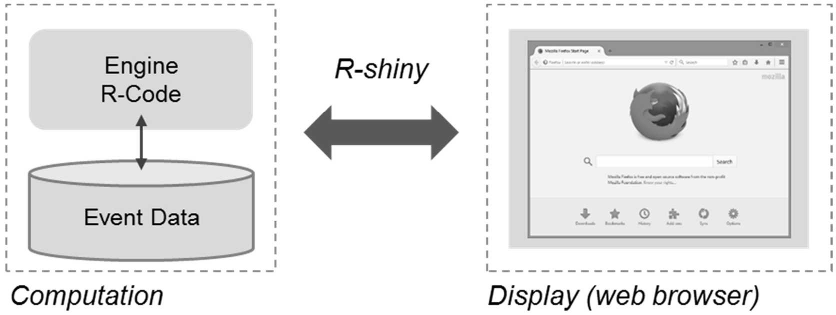


Figure 1. *Schematic display of the interaction between R and a web browser via shiny.*

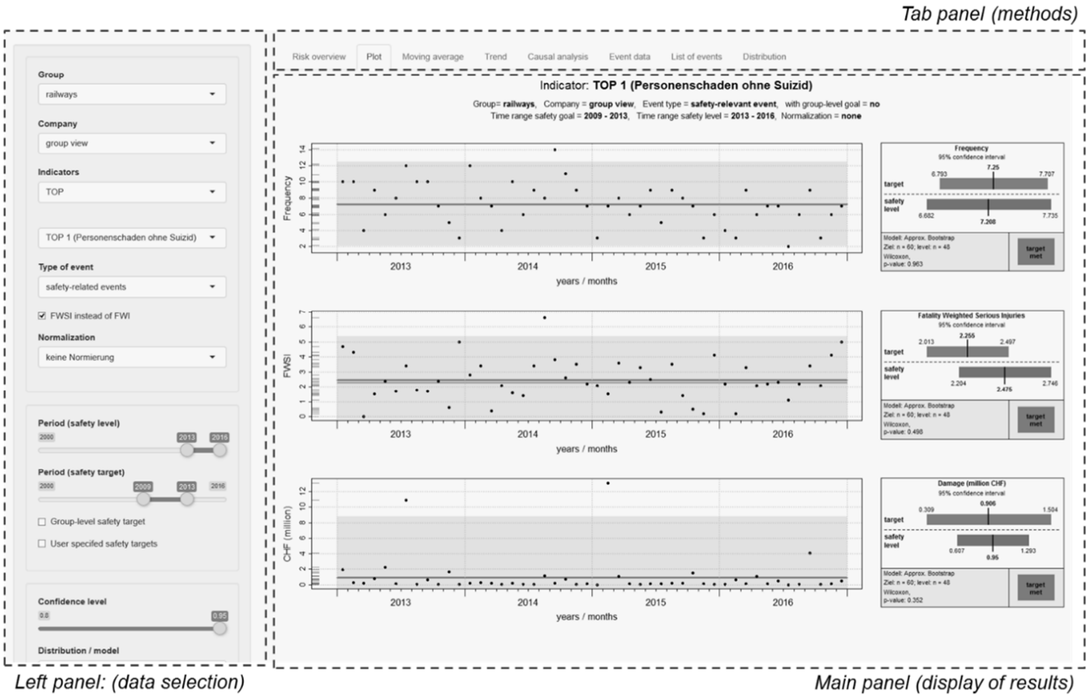
3.1. Design of the Graphical User Interface

The graphical user interface uses a “classical” layout that is composed of three panels: data and indicator selection (left panel), method selection (tab panel), and the main panel for the display of the results; see Figure 2. We have chosen such a simple design because the tool must be easy to use for people working in safety auditing, who utilized the tool only a few times per year (typically at the end of a reporting period). Therefore, a familiar and easily understood design is an essential usability issue for this type of users; this requirement does not apply to safety analysts who have the tool up and running virtually all-day long.

The monitoring workflow is also kept as simple as possible. Every analysis starts with choosing the relevant specs (indicator, data, time period, etc.) in the left panel, followed by a choice of method or visualization type (using the tab panel). The selected data and all other specifications that are accessible through the left-hand panel are kept fix irrespective of the chosen method. Hence, we may switch between, for instance, the tabs “evaluation of the indicators” and “trend analysis” back and forth without changing the selected data. On the other hand, changes to the specs in the left panel affect all tabs and the tool re-draws the current display in a split of a second.

In the current implementation, the main features of the tool (organized as tabs) are:

- visualization and evaluation of the indicators
  - time series plots of the indicators (and evaluation; see Figures 2 and 3; note that the display shows three separate time series plots, one for each dimension of analysis: even frequency, damage to people and material);
  - trend evaluation;
  - moving average plots (monthly / sub-annual data);
  - causal analysis (contingency tables and histograms);
  - dash-board-line risk overview.
- utility features
  - distributional analysis (diagnostic measures to check whether the imposed distributional assumptions are met);
  - tabular display of an event’s characteristics (in-depth analysis of singular events);
  - automatic generation of an evaluation report (output format: Microsoft Excel™).



Group: railways

Company: group view

Indicators: TOP

TOP 1 (Personenschaden ohne Suizid)

Type of event: safety-related events

☑ FWSI instead of FWI

Normalization: keine Normierung

Period (safety level): 2000 2013 2016

Period (safety target): 2000 2009 2013 2016

☐ Group-level safety target

☐ User specifed safety targets

Confidence level: 0.8 0.95

Distribution / model

Indicator: **TOP 1 (Personenschaden ohne Suizid)**

Group= **railways**, Company = **group view**, Event type = **safety-relevant event**, with group-level goal = **no**
Time range safety goal = **2009 - 2013**, Time range safety level = **2013 - 2016**, Normalization = **none**

Frequency
95% confidence interval
target: 6.793 **7.25** 7.707
safety level: 6.682 **7.208** 7.735
Modell: Approx. Bootstrap
Ziel: n = 60; level: n = 48
Wilcoxon,
p-value: 0.963
target met

Fatality Weighted Serious Injuries
95% confidence interval
target: 2.013 **2.255** 2.497
safety level: 2.204 **2.475** 2.746
Modell: Approx. Bootstrap
Ziel: n = 60; level: n = 48
Wilcoxon,
p-value: 0.498
target met

Damage (million CHF)
95% confidence interval
target: 0.309 **0.906** 1.504
safety level: 0.607 **0.95** 1.293
Modell: Approx. Bootstrap
Ziel: n = 60; level: n = 48
Wilcoxon,
p-value: 0.352
target met

Figure 2. *The three-part layout of the tool: In the left panel, data and indicators (and further specs) are selected; the tabs panel provides methods and visualizations; the results (here time series plots and confidence overlap plots) are shown in the main panel.*

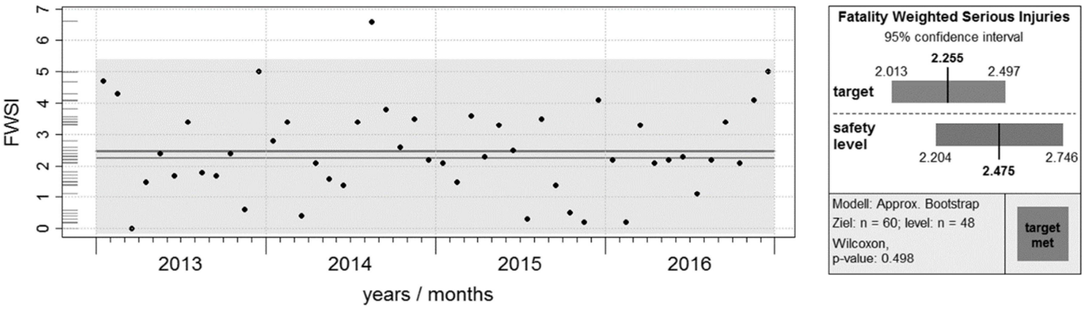


Figure 3. *Time series plot and statistical evaluation whether the actual safety performance meets the safety target (here, occurrence of events; confidence intervals on the right-hand side).*

3.2. Internal Data Management

The tool's data management builds on the functionality of the R-package *data.table* (Dowle et al., 2016), which provides data-base-like functionality and reduces computing time (for data subsets, groups, updates, and joins) tremendously. This marks the crucial difference in comparison with R's standard operations on *data.frames* and contributes to the high responsiveness of the monitoring tool.

4. Summary / Key Messages

▷ The scope of the presented tool is not limited to safety-related monitoring applications; the tool – at least its basic structure – can be applied in the context of virtually any indicator-based monitoring system.

▷ The R/shiny monitoring tool can be operated without knowledge of R. Thus, the tool is open to a wide range of potential users / analysts (safety managers etc.). Tool developers are recommended to put themselves into the shoes of an inexperienced user prior to designing the tool layout / workflow; since we tend to "overestimate" the average user's statistical literacy and computer skills.

▷ Shiny-apps exploit R's rich and extensively tested (and trusted) set of statistical methods.

▷ The R/shiny configuration is "mature" and meets the demands in terms of reliability (also in a corporate IT environment).

▷ In general, we are faced with a trade-off between computer-intensive (but perhaps more appropriate) methods and simple statistical methods which are easy to compute (and understood by non-statisticians) and thus ensure high responsiveness of the tool. It is the statistician job to decide how much he or she is willing to sacrifice "methodological appropriateness" in favour of computing time and thus ease of use.

▷ If responsiveness of the tool becomes an issue (e.g., unreasonably long delays), we recommend using pre-computed lookup-tables (with stored data and statistics) instead of re-doing heavy computations each time the user modifies the selected items.

References

Chang, W., J. Cheng, J.J. Allaire, Y. Xie, and J. McPherson (2016). *shiny: Web Application Framework for R*. R package version 0.14.2. URL https://CRAN.R-project.org/package=shiny.

DiCiccio, T., J. and B. Efron (1996). Bootstrap Confidence Intervals, *Statistical Science 11*, pp. 189–228.

DiCiccio, T. J. and B. Efron (1992). More accurate confidence intervals in exponential families. *Biometrika 79*, pp. 231–245.

Dowle, M., A. Srinivasan, T. Short, and S. Lianoglou with contributions from R. Saporta and E. Antonyan (2016). data.table: Extension of Data.frame. R package version 1.9.6. URL https://CRAN.R-project.org/package=data.table.

Kendall, M. G. (1975): Rank Correlation Methods, 4th ed., London: Charles Griffin.

R Development Core Team (2016). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna. URL https://www.R-project.org/.

Schenker, N. and F.J. Gentleman (2001). On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals, *The American Statistician* 55, pp. 182–186.

Country Reports

AUSTRALIA

Reporting: Dr. Mark Zhang (ABS) and Dr. Oksana Honchar (ABS)

Predicting survey estimates by state space models using multiple data sources

The Australian Bureau of Statistics (ABS) is embarking on a transformation program, which includes, amongst other things, re-engineering and consolidating collections, using different collection modes for survey data, and using different, but more efficient, sampling frames and estimation methods for official statistics. Whilst this transformation is expected to bring about positive changes to official statistics, there is also a risk that such changes could lead to impacts on some ABS time series. The challenge for the ABS is to develop methodologies to monitor, measure and, where needed, adjust for any such impacts.

The methodology the ABS is proposing to implement for this purpose makes use of the data of related series. A special multiple time series model called a Seemingly Unrelated Time Series Equation (SUTSE) model has been investigated as a basis for predicting a target survey estimate using multiple data sources.

Where related series measure a similar concept to the target survey variable, but are not subject to measurement change, these can assist in understanding the change that occurs on the target survey variable. Under this method, the statistical impact can be assessed by intervention analysis, taking advantage of the cross-correlations and leading properties between the target survey variable and the other related series. The power of this method has been tested by estimating historical supplementary survey effects and the effect of past questionnaire redesign using Australian Labour Force Survey data. This work has also been extended in a number of other directions. A case study involving Labour Force Survey (LFS) unemployment confirmed that a standard bivariate SUTSE model with claimant count data offered improvements in terms of prediction error, detecting outliers and structural changes in the target unemployment estimates.

However, available related data sources may not have appropriate properties for applying a standard SUTSE model to predict survey estimates efficiently. As part of these investigations the ABS developed a strategy to select valuable data sources and adjust the way a SUTSE model is applied to take advantage of SUTSE modelling strength. Another case study considered employment estimates from the LFS, and demonstrated that such a strategy also has the potential to work much better than a univariate structural time series model, by borrowing strength from multiple source data in an efficient way.

CANADA

Reporting: Darren Gray

A framework of sensitivity measures to treat weighted data confidentiality

At many national statistical organizations, linear sensitivity measures are used to quantify the disclosure risk in tabular magnitude data, based on one or more underlying sensitivity rules such as the p-percent, pq (prior/posterior), and nk (dominance) rules. However, these measures are not always well-suited for issues present in survey data such as negative values, respondent waivers and estimation weights. The Precision Threshold and Noise (*PTN*) Framework of sensitivity measures provides a useful alternative to the set of linear sensitivity measures, allowing users to address many of the latter's limitations.

The *PTN* framework builds on the concepts of protection and noise inherent to most sensitivity rules, and introduces the concept of self-noise. These concepts are defined at a microdata level, providing flexibility in scenarios when not all respondent contributions require or provide a proportional degree of protection. (For example, when some respondents sign a confidentiality waiver, or for respondents with different estimation weights.) While not a direct expansion of linear sensitivity, the *PTN* framework does contain natural and intuitive expressions of sensitivity rules that are in common use.

The *PTN* framework was first presented at the Privacy in Statistical Databases conference in Dubrovnik (September 2016). Due to its compatibility with sensitivity-score suppression methods (such as those used by the statistical disclosure control software systems *G-Confid* and τ-*ARGUS*) and in particular its ability to incorporate self-noise, a set of *PTN* measures is currently in development for the treatment of estimation weights at Statistics Canada. This functionality will be released in a new G-Confid package mid-2017. When this new functionality becomes available, it is expected to allow a more accurate assessment of the protective noise offered by respondents according to their weights. The underlying assumption is that the survey weight of each contribution is not known to data users.

References:

Gray, D. (2016). Precision Threshold and Noise: An Alternative Framework of Sensitivity Measures. In *International Conference on Privacy in Statistical Databases* (pp. 15-27). Springer International Publishing.

INDIA

Reporting: Dr. Gayatri Vishwakarma

Clinical Development Services Agency (CDSA), Fariadabad

The Comprehensive National Nutrition Survey, being conducted by UNICEF and the ministry of health and family and welfare, the ministry of statistics and program implementation, not-for-profit entity Population Council of India, private sector firm SRL Diagnostics Ltd and ArcelorMittal SA. CDSA is involved in this project as independent monitoring agency and responsible for lab as well as field monitoring. The survey aims to cover 120,000 children in the 0-19 age group across all Indian states. It is doing piloting and preparation of tools to measure cognitive development, school readiness & learning outcomes among children for comprehensive national nutrition survey (CNNS).

CNNS is the first of its kind in terms of sample size and parameters tested. The collection of samples started in March 2016. The survey aims to cover 120,000 children in the 0-19 age group across all Indian states. The results will likely be released only in 2018, given the scale and complexity of the survey.

The plan of sampling is such that each selected village has 500 to 600 houses, which are divided into social and economic segments of 200 to 250 each. Using statistical methodology, 60 to 65 houses are identified for collecting samples.

CNNS is a multidisciplinary survey that includes biochemical and nutritional samples. It even takes into account cognitive domain, anthropometric, household food security, water sanitation and socioeconomic features.

Comprehensive data is key to addressing malnutrition and so far India has been lacking in this field as most of the surveys have been based on cluster samples limited to height and weight indicators. CNNS includes measuring deficiencies in body mass, micronutrients, vitamins, minerals as well as worm infestations among children.

The survey will help create the right policy interventions to address the root cause of malnourishment for India.

MALAYSIA

Reporting: Siti Asiah Ahmad

National Enterprise-Wide Statistical System

National Enterprise-Wide Statistical System (NEWSS) is an integrated statistical system developed by Department of Statistics Malaysia (DOSM) to systematically manage household and business frame, data collection, processing, analysing and dissemination. Technology used to develop the system are object-oriented Java with the Oracle (RedHat Linux, WebLogic and Oracle DB 11r2) as a database. NEWSS Framework are divided into two components which are Integrated Statistical Survey Framework (ISSF) and Information Support System (ISS).NEWSS which started in 2009, facilitate to standardize, simplify and expedite the data dissemination and also provide a central repository for DOSM to share the statistics within the government agencies. NEWSS also plays an important role as a one-stop centre data of frame for establishment and household.

Currently, NEWSS have 25 modules and six core applications extensively used to support the core business of statistics specifically for establishment and household surveys which include economic survey, monthly manufacturing, labour force, quarterly construction, monthly distributive trade and producer price index. The users of the system include subject matter division, field enumerator, business analyst, respondent and the public. NEWSS has also been utilised for the implementation of two censuses which are Census of Distributive Trade 2014 and Census of Economy 2016 in managing the frame, sampling and operation of the census.

By using the NEWSS, statistician has benefit whereby they are able to consistently monitor and manage the surveys and censuses efficiently and effectively. The system also enables respondents to respond via online and this improve the timelines of data collection as well as cost saving as DOSM embarks on "go green" that is environmental friendly.

In term of dissemination, users are able to access the statistics by DOSM on timely manner and has reduced the red tapes whereby *eStatistik* has been incorporated. *eStatistik* is a medium for user to access the statistics via online. In 2016, there were 275,004 publications downloaded for free and 349,152 transactions via *eStatistik* (*Source portal DOSM).*

For further information, please contact jpbkkp@stats.gov.my.

NEW ZEALAND

Reporting: Felibel Zabala

Small Area Estimates for Disability Measures

Stats NZ released small area estimates for the prevalence of five disability measures by territorial authority (http://www.stats.govt.nz/browse_for_stats/health/disabilities/disability-small-areas-2013.aspx) in April 2017. The estimates were produced under a Bayesian modelling framework, and follow on from work in 2016 that produced cultural well-being measures by kiwi/tribe for New Zealand's indigenous Māori population. The focus for small area work will now change to produce labour force characteristics.

Contact Gareth Minshall at gareth.minshall@stats.govt.nz for more details on the disability estimates.

Upcoming Conferences and Workshops

The list below highlights events that have sessions or main subject related to areas such as survey methods, official statistics, data linkage and confidentiality. For a more wide-ranging list, please check the ISI Calendar of Events at https://www.isi-web.org/index.php/activities/calendar

ISI2017
MARRAKECH
ISI WORLD STATISTICS CONGRESS
16 - 21 July 2017
61st World Statistics Congress – ISI2017

61st World Statistics Congress of the International Statistical Institute
Organized by: The High Commission for Planning – HCP
Where: Marrakech, Morocco
When: July 16 – 21, 2017
Homepage: http://www.isi2017.org

ESRA
7th Conference of the
European Survey Research Association
Lisbon, Portugal, July 17-21, 2017

The 7th Conference of the European Survey Research Association (ESRA)
Organized by: the University of Lisbon, jointly by the School of Economics and Management (ISEG), the Centre for Research in Social Sciences and Management (CSG) and the Institute for Social Sciences (ICS)
Where: Lisbon, Portugal
When: July 17 – 21, 2017
Homepage: http://ec.europa.eu/eurostat/cros/content/ntts-2017

The Joint Statistical Meetings (JSM) 2017
Organized by: Several statistical societies and associations
Where: Baltimore, United States
When: July 29 - August 3, 2017
Homepage: http://ww2.amstat.org/meetings/jsm/2017/

23 de junio de 2017 fecha límite primer Pronto pago
27° SIMPOSIO
INTERNACIONAL DE ESTADÍSTICA
5 IWAS
INTERNATIONAL WORKSHOP ON APPLIED STATISTICS
Agosto 8 - 12 de 2017
Universidad Santo Tomás
Medellín, Colombia

XXVII International Symposium on Statistics
Organized by: National University of Colombia
Where: Medellin, Colombia
When: August 8 - 12, 2017
Homepage: http://simposioestadistica.unal.edu.co/

BALTIC-NORDIC-UKRAINIAN NETWORK ON SURVEY STATISTICS

Baltic-Nordic-Ukrainian Network Workshop on Survey Statistics 2017
Organized by: Baltic-Nordic-Ukrainian Network
Where: Vilnius, Lithuania
When: August 21 - 25, 2017
Homepage: https://wiki.helsinki.fi/display/BNU/Events

European Network for Better Establishment Statistics / European Network for Better Establishment Statistics Home

News and Events

Created and last modified by Boris Lorenc on 02 Nov, 2016

European Establishment Statistics Workshop (EESW) 2017
Organized by: European Network for Better Establishment Statistics (ENBES)
Where: Southampton, UK
When: August 30 - September 1, 2017
Homepage: http://www1.unece.org/stat/platform/display/ENBES/EESW17

Royal Statistical Society 2017 International Conference
Organized by: The Royal Statistical Society
Where: Glasgow, United Kingdom
When: September 4 - 7, 2017
Homepage: http://www.rss.org.uk/conference2017

11th International Conference
on Transport Survey Methods
2017
in the era of big data: facing the challenges
24-29 SEPT
2017
Québec

11th International Conference on Transport Survey Methods
Organized by: Interuniversitary Research Centre on Entreprise Networks, Logistics and Transport (CIRRELT)
Where: Quebec, Canada
When: September 24 - 29, 2017
Homepage: http://www.hksts.org/isctsc.htm

V ESAMP
ESCOLA DE AMOSTRAGEM
E METODOLOGIA DE PESQUISA
IV Workshop Internacional sobre Pesquisas para Avaliação de Políticas Públicas

4th International Workshop on Surveys for Policy Evaluation and the 5th Brazilian School on Sampling and Survey Methodology – ESAMP V
Organized by: Federal University of Mato Grosso and the Brazilian Statistical Association
Where: Mato Grosso, Brazil
When: October 17 - 20, 2017
Homepage: http://www.redeabe.org.br/esamp2017/

New Challenges for Statistical Software
– The Use of R in Official Statistics – | The 5th International Conference

The Use of R in Official Statistics
Organized by: National Institute of Statistics of Romania, Bucharest University of Economic Studies, Universiti Teknologi Mara, Ecological University of Bucharest, The University of Bucharest and Nicolae Titulescu University of Bucharest
Where: Bucharest, Romania
When: November 6 - 7, 2017
Homepage: http://r-project.ro/conference2017/

2018
European Conference
on Quality in Official Statistics
26 - 29 June 2018 - Krakow, Poland

Q2018 is the 9th European Conference on Quality in Official Statistics
Organized by: Q2018
Where: Krakow, Poland
When: 26-29 June 2018
Homepage: https://ec.europa.eu/eurostat/cros/Q2018_en

In Other Journals

Journal of Survey Statistics and Methodology

Volume 5, Number 2 (June 2017)
https://academic.oup.com/jssam/issue

Survey Statistics

An Extension of Kish's Formula for Design Effects to Two- and Three-Stage Designs With Stratification
Sixia Chen; Keith Rust

Model-Assisted Survey Regression Estimation with the Lasso
Kelly S. McConville; F. Jay Breidt; Thomas C. M. Lee; Gretchen G. Moisen

Calibration Weighting for Nonresponse that is Not Missing at Random: Allowing More Calibration than Response-Model Variables
Phillip S. Kott; Dan Liao

Survey Methodology

Explaining Interviewer Effects: A Research Synthesis
Brady T. West; Annelies G. Blom

Modeling the Weekly Data Collection Efficiency of Face-to-Face Surveys: Six Rounds of the European Social Survey
Caroline Vandenplas; Geert Loosveldt

An Item Response Theory Approach to Estimating Survey Mode Effects: Analysis of Data from a Randomized Mode Experiment
Louis T. Mariano; Marc N. Elliott

Effects of Using an Overlapping Dual-Frame Design on Estimates of Health Behaviors: A French General Population Telephone Survey
Jean-Baptiste Richard; Raphaël Andler; Arnaud Gautier; Romain Guignard; Christophe Leon

Survey Methodology

Statistics Canada Statistique Canada Canada

Volume 43, Number 1 (June 2017)
http://www.statcan.gc.ca/pub/12-001-x/12-001-x2017001-eng.htm

The promise and challenge of pushing respondents to the Web in mixed-mode surveys
Dillman, Don A.

A layered perturbation method for the protection of tabular outputs
Jean-Louis Tambay

State space time series modelling of the Dutch Labour Force Survey: Model selection and mean squared errors estimation
Oksana Bollineni-Balabay, Jan van den Brakel and Franz Palm

Bayesian predictive inference of a proportion under a two-fold small area model with heterogeneous correlations
Danhyang Lee, Balgobin Nandram and Dalho Kim

Sample allocation for efficient model-based small area estimation
Mauno Keto and Erkki Pahkinen

A mixed latent class Markov approach for estimating labour market mobility with multiple indicators and retrospective interrogation
Francesca Bassi, Marcel Croon and Davide Vidotto

Variance estimation in multi-phase calibration
Noam Cohen, Dan Ben-Hur and Luisa Burck

Journal of Official Statistics

Volume 33, Issue 2 (June 2017)
https://www.degruyter.com/view/j/jos.2017.33.issue-2/issue-files/jos.2017.33.issue-2.xml

Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ
Reid, Giles / Zabala, Felipa / Holmberg, Anders

Comparing Two Inferential Approaches to Handling Measurement Error in Mixed-Mode Surveys
Buelens, Bart / Van den Brakel, Jan A.

Adjusting for Measurement Error and Nonresponse in Physical Activity Surveys: A Simulation Study
Beyler, Nicholas / Beyler, Amy

Effect of Missing Data on Classification Error in Panel Surveys
Edwards, Susan L. / Berzofsky, Marcus E. / Biemer, Paul P.

Survey Practice

Volume 10, Number 2 (2017)
http://www.surveypractice.org/index.php/SurveyPractice/index

Using Tablet Computers to Implement Surveys in Challenging Environments
Lindsay J Benstead, Kristen Kao, Pierre F Landry, Ellen M Lust, Dhafer Malouche

Is Using a Form Processing Application Cost Saving? A Real-life Example
Julien Danhier

Mitigating Nonresponse Error in Online Surveys Through Research-based Design and Delivery
David Carl Ison

Interviewers in Computer Assisted Telephone Interviews: A Standardization Controversy
Wojciech Jablonski

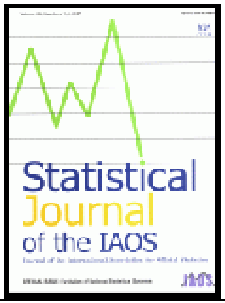
Are Response Rates to Organizational Climate Surveys Influenced by Informing Respondents about the Sampling Strategy?
Taylor H Lewis, Lorraine Latimore

It's About Time: Examining the Effect of Interviewer-Quoted Survey Completion Time Estimates on Survey Efficiency
Christopher D. Ward, Bess Welch, Amie Conley, Philip J. Smith, Stacie Greby

Survey Research Methods

Volume 11, Number 1 (2017)
https://ojs.ub.uni-konstanz.de/srm/index

Table of Contents



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

Volume 33, Number 2 (2017)
http://content.iospress.com/journals/statistical-journal-of-the-iaos/33/2

Editorial
West, Kirsten

Interview with Fritz J. Scheuren: Statistical Mentsch
Condon, Katherine M.

Remembering Hans Rosling (1948 – 2017)
West, Kirsten, Awad, Ola

An international perspective on the undercount of young children in the U.S. Census
O'Hare, William P

Commenting on An international perspective on the undercount of young children in the U.S. Census (DOI: 10.3233/SJI-161008)
Griffin, Deborah H.

An international perspective on the undercount of young children in the U.S. Census
Robinson, J. Gregory

Undercount of young children and young adults in the new French census
Toulemon, Laurent

Commenting on an international perspective on the undercount of young children in the U.S. Census (DOI: 10.3233/SJI-161008)
Dolson, David

The coverage of young children in demographic surveys
Jensen, Eric B., Hogan, Howard R.

An international perspective on the undercount of young children in the U.S. Census: Rejoinder
O'Hare, William P.

Celebrating the establishment, development and evolution of statistical offices worldwide: A tribute to John Koren
Michalopoulou, Catherine; Mimis, Angelos

The concept and commodity of official statistics
Rolland, Asle

Towards a global system of monitoring the implementation of UN fundamental principles in national official statistics
Georgiou, Andreas V.

The 2010 round of population and housing censuses (2005-2014)[1]
Juran, Sabrina; Pistiner, Arona L.

Statistical monitoring systems to inform policy decision-making, and new data sources
Schnorr-Baecker, Susanne

Assessing quality control: Evaluating the quality audit
Nguyen, Justin D.; Hogue, Carma R.

Minding the store: An internal audit program for demographic programs at the U.S. Census Bureau
Levy, Richard; Scott, Jimmie

The assurance of administrative data: A proportionate approach
Babb, Penny

Introducing a framework for process quality in National Statistical Institutes
Brancato, Giovanna; D'Assisi Barbalace, Francesco; Signore, Marina; Simeoni, Giorgia

Explaining political participation: A comparison of real and falsified survey data
Landrock, Uta

Targeted letters: Effects on sample composition and item non-response
Bianchi, Annamaria; Biffignandi, Silvia

Producing multiple tables for small areas with confidentiality protection
Krenzke, Tom[a; *]*; Li, Jianzhu*[a] *; McKenna, Laura*[b]

Innovations on measuring the indigenous population in the 2010 Brazilian Population Census
de Oliveira Martins Pereira, Nilza

Imputation and money income distribution measures
Turek, Joan L.

ICT tools for creating, expanding and exploiting statistical linked Open Data
Kalampokis, Evangelos ; Tambouris, Efthimios; Tarabanis, Konstantinos

Financial Intermediation Services Indirectly Measured (FISIM): The role of reference rate
Das, Abhiman; Jangili, Ramesh

Developing Households' sub-sectors accounts: Pros and cons of the top-down and the bottom-up methods
Coli, Alessandra; Tartamella, Francesca

Comparing random coefficient autoregressive model with and without autocorrelated errors by Bayesian analysis
Araveeporn, Autcha

Data anomaly in mining statistics of India
Chattopadhyay, Molly; Lahiri, Anupam

Contribution of investment in economic growth of major sectors: With focus on Agriculture and Allied sector in Bihar
Sinha, Jitendra Kumar

International Statistical Review
Revue Internationale de Statistique

VOL 85, ISSUE 1 (April 2017)
http://onlinelibrary.wiley.com/doi/10.1111/insr.v85.1/issuetoc

Statistical Scale Space Methods
Lasse Holmström and Leena Pasanen

Discussions on Statistical Scale Space Methods
María Oliveira, Rosa M. Crujeiras and Alberto Rodríguez-Casal
Cheolwoo Park, Yongho Jeon and Kee-Hoon Kang
Fred Godtliebsen, Stein O. Skrøvseth and Susan Wei
Hui-Guo Zhang and Chang-Lin Mei
Subhajit Dutta and Anil K. Ghosh

Rejoinder to Statistical Scale Space Methods
Lasse Holmström and Leena Pasanen

Interview

A Conversation with Roger Koenker
Xuming He

Original Articles

Functional Principal Component Regression and Functional Partial Least-squares Regression: An Overview and a Comparative Study
Manuel Febrero-Bande, Pedro Galeano and Wenceslao González-Manteiga

Accessible Conceptions of Statistical Inference: Pulling Ourselves Up by the Bootstraps
Chris J. Wild, Maxine Pfannkuch, Matt Regan and Ross Parsonage

Multivariate Hill Estimators
Yves Dominicy, Pauliina Ilmonen and David Veredas

Selecting Adaptive Survey Design Strata with Partial R-indicators
Barry Schouten and Natalie Shlomo

Modelling the Ecological Comorbidity of Acute Respiratory Infection, Diarrhoea and Stunting among Children Under the Age of 5 Years in Somalia
Damaris K. Kinyoki, Samuel O. Manda, Grainne M. Moloney, Elijah O. Odundo, James A. Berkley, Abdisalan M. Noor and Ngianga-Bakwin Kandala

Book Reviews:

Statistics in Toxicology Using R L.A. Hothorn Chapman and Hall/CRC
Alice Richardson

Extreme Value Modelling and Risk Analysis Dipak K. Dey and Jun Yan Chapman and Hall/CRC
Lili Zhao

Fundamental Concepts for New Clinical Trialists Scott Evans and Naitee Ting Chapman and Hall/CRC Biostatistics Series
Lili Zhao

Monte Carlo Methods and Stochastic Processes – From Linear to Non-Linear Emmanuel Gobet Chapman and Hall/CRC
Krzysztof Podgórski

Stochastic Volatility Modelling Lorenzo Bergomi Chapman and Hall/CRC, 2016
Krzysztof Podgórski

Time Series Econometrics Klaus Neusser Springer International Publishing, 2016
Tucker S. McElroy

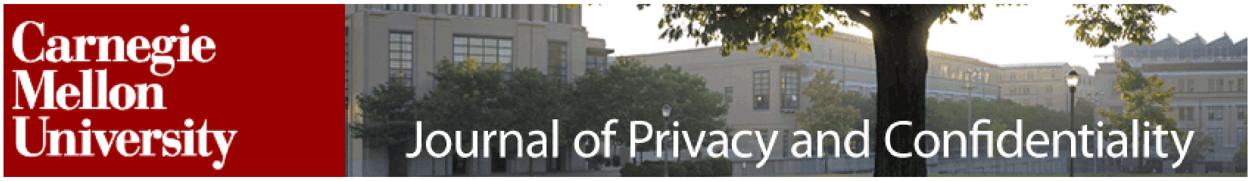
TRANSACTIONS ON
DATA PRIVACY
Foundations and Technologies
http://www.tdp.cat

Volume 10 Issue 1, April 2017
http://www.tdp.cat/issues16/vol10n01.php

Feature Selection for Classification under Anonymity Constraint
Baichuan Zhang, Noman Mohammed, Vachik S. Dave, Mohammad Al Hasan

Enhancing the Utility of Anonymized Data by Improving the Quality of Generalization Hierarchies
Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, Liam Murphy, Christina Thorpe

Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion
Anna Oganian, Josep Domingo-Ferrer



Volume 7 (2015-2017), Issue 3 (2017)
http://repository.cmu.edu/jpc/vol7/iss2/

How Will Statistical Agencies Operate When All Data Are Private?
John M. Abowd

Calibrating Noise to Sensitivity in Private Data Analysis
Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith

On the Meaning and Limits of Empirical Differential Privacy
Anne-Sophie Charest and Yiwei Hou

Practical Data Synthesis for Large Samples
Gillian M. Raab, Beata Nowok, and Chris Dibben

A New Data Collection Technique for Preserving Privacy
Samuel S. Wu Dr, Shigang Chen, Deborah L. Burr, and Long Zhang

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

Vol 180 Issue 3 (June 2017)

http://onlinelibrary.wiley.com/doi/10.1111/rssa.2017.180.issue-3/issuetoc

Book reviews

Happiness Explained: What Human Flourishing Is and How We Can Promote It
Kuldeep Kumar

Managing and Sharing Research Data: a Guide to Good Practice
Carole Sutton

Handbook of Statistical Distributions with Applications
Mark Pilling

Journal of the American Statistical Association

Journal of the American Statistical Association

ASA
AMERICAN STATISTICAL ASSOCIATION

Volume 112, Issue 517 (April 2017)
http://www.tandfonline.com/toc/uasa20/current?nav=tocList

On the Reproducibility of Psychological Science
Valen E. Johnson, Richard D. Payne, Tianying Wang, Alex Asher & Soutrik Mandal

Robust Treatment Comparison Based on Utilities of Semi-Competing Risks in Non-Small-Cell Lung Cancer
Thomas A. Murray, Peter F. Thall, Ying Yuan, Sarah McAvoy & Daniel R. Gomez

A Directional Mixed Effects Model for Compositional Expenditure Data
J. L. Scealy & A. H. Welsh

Bayesian Treed Calibration: An Application to Carbon Capture With AX Sorbent
Bledar A. Konomi, Georgios Karagiannis, Kevin Lai & Guang Lin

Defining Cancer Subtypes With Distinctive Etiologic Profiles: An Application to the Epidemiology of Melanoma
Audrey Mauguen, Emily C. Zabor, Nancy E. Thomas, Marianne Berwick, Venkatraman E. Seshan & Colin B. Begg

The Generalized Higher Criticism for Testing SNP-Set Effects in Genetic Association Studies
Ian Barnett, Rajarshi Mukherjee & Xihong Lin

A Bayesian Race Model for Recognition Memory
Sungmin Kim, Kevin Potter, Peter F. Craigmile, Mario Peruggia & Trisha Van Zandt

Spatiotemporal Modeling of Node Temperatures in Supercomputers
Curtis B. Storlie, Brian J. Reich, William N. Rust, Lawrence O. Ticknor, *Amanda M. Bonnie, Andrew J. Montoya & Sarah E. Michalak*

A Framework for Synthetic Control Methods With High-Dimensional, Micro-Level Data: Evaluating a Neighborhood-Specific Crime Intervention
Michael W. Robbins, Jessica Saunders & Beau Kilmer

Forecasting Generalized Quantiles of Electricity Demand: A Functional Data Approach
Brenda López Cabrera & Franziska Schulz

Discussions

Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing
M. P. Wand
Dustin Tran & David M. Blei
Wanzhu Tu
Philip T. Reiss & Jeff Goldsmith

Residual Weighted Learning for Estimating Individualized Treatment Rules
Xin Zhou, Nicole Mayer-Hamblett, Umer Khan & Michael R. Kosorok

Weighted Statistic in Detecting Faint and Sparse Alternatives for High-Dimensional Covariance Matrices
Qing Yang & Guangming Pan

A Multi-Resolution Approximation for Massive Spatial Datasets
Matthias Katzfuss

Dynamic Multiscale Spatiotemporal Models for Poisson Data
Thaís C. O. Fonseca & Marco A. R. Ferreira

A Dynamic Structure for High-Dimensional Covariance Matrices and Its Application in Portfolio Allocation
Shaojun Guo, John Leigh Box & Wenyang Zhang

Covariance Regression Analysis
Tao Zou, Wei Lan, Hansheng Wang & Chih-Ling Tsai0Altmetric

Partition MCMC for Inference on Acyclic Digraphs
Jack Kuipers & Giusi Moffa

Augmented Particle Filters
Jonghyun Yun, Fan Yang & Yuguo Chen

Conditions for Ignoring the Missing-Data Mechanism in Likelihood Inferences for Parameter Subsets
Roderick J. Little, Donald B. Rubin & Sahar Z. Zangeneh

Randomization Inference and Sensitivity Analysis for Composite Null Hypotheses With Binary Outcomes in Matched Observational Studies
Colin B. Fogarty, Pixu Shi, Mark E. Mikkelsen & Dylan S. Small

Robust Jump Regressions
Jia Li, Viktor Todorov & George Tauchen

Promoting Similarity of Sparsity Structures in Integrative Analysis With Penalization
Yuan Huang, Qingzhao Zhang, Sanguo Zhang, Jian Huang & Shuangge Ma

Semiparametric Modeling and Estimation of the Terminal Behavior of Recurrent Marker Processes Before Failure Events
Kwun Chuen Gary Chan & Mei-Cheng Wang

Geometric Representations of Random Hypergraphs
Simón Lunagómez, Sayan Mukherjee, Robert L. Wolpert & Edoardo M. Airoldi

Nonparametric Estimation of the Leverage Effect: A Trade-Off Between Robustness and Efficiency
Ilze Kalnina & Dacheng Xiu

A New Graph-Based Two-Sample Test for Multivariate and Object Data
Hao Chen & Jerome H. Friedman

A Concave Pairwise Fusion Approach to Subgroup Analysis
Shujie Ma & Jian Huang

Robust Maximum Association Estimators
Andreas Alfons, Christophe Croux & Peter Filzmoser

Cluster-Robust Bootstrap Inference in Quantile Regression Models
Andreas Hagemann

BIOMETRIKA

Volume 104, Issue 2 (June 2017)
https://academic.oup.com/biomet/issue

Data integration with high dimensionality
Xin Gao; Raymond J. Carroll

Automatic tagging with existing and novel tags
Junhui Wang; Xiaotong Shen; Yiwen Sun; Annie Qu

Instrumental variables as bias amplifiers with general outcome and confounding
P. Ding; T.J. Vanderweele; J. M. Robins

An improved and explicit surrogate variable analysis procedure by coefficient adjustment
Seunggeun Lee; Wei Sun; Fred A. Wright; Fei Zou

Ignorability for general longitudinal data
D. M. Farewell; C. Huang; V. Didelez

Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness
A. R. Linero

Bayesian block-diagonal variable selection and model averaging
O. Papaspiliopoulos; D. Rossell

Covariate-assisted spectral clustering
N. Binkiewicz; J. T. Vogelstein; K. Rohe

Sparse Gaussian graphical model estimation via alternating minimization
Onkar Dalal; Bala Rajaratnam

On the number of common factors with high-frequency data
Xin-Bing Kong

Projection-based outlier detection in functional data
Haojie Ren; Nan Chen; Changliang Zou

Testing separability of space-time functional processes
P. Constantinou; P. Kokoszka; M. Reimherr

Multiply robust imputation procedures for the treatment of item nonresponse in surveys
Sixia Chen; David Haziza

Construction of maximin distance Latin squares and related Latin hypercube designs
Qian Xiao; Hongquan Xu

Miscellanea

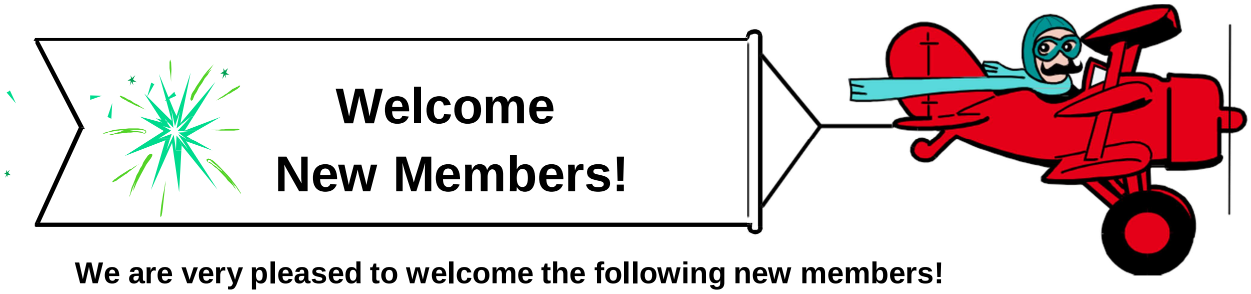
A general rotation method for orthogonal Latin hypercubes
Fasheng Sun; Boxin Tang

Bias-corrected score decomposition for generalized quantiles
W. Ehm; E. Y. Ovcharov

Nonlinear shrinkage estimation of large integrated covariance matrices
Clifford Lam; Phoenix Feng; Charlie Hu

Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression
I. Kosmidis; A. Guolo; C. Varin

Assigning a value to a power likelihood in a general Bayesian model
C. C. Holmes; S. G. Walker



We are very pleased to welcome the following new members!

Title	First name	Surname	Department	Country
Dr.	James	Chipperfield	Australian Bureau of Statistics	Australia
Dr.	Roberto	Olinto Ramos	IBGE	Brazil
Mr.	Anil	Arora	Statistics Canada	Canada
Mr.	Jørgen	Elmeskov	Statistics Denmark	Denmark
Mrs.	Marjo	Bruun	Statistics Finland	Finland
Mr.	Dieter	Sarreither	Statistiches Bundesamt	Germany
Dr.	Faustina	Frempong-Ainguah	Regional Institute For Population Studies	Ghana
Dr.	Girish	Chandra	Division of Forestry Statistic	India
Mrs.	Sigalit	Mazeh	International Rel. & Statistical Coordination	Israel
Prof.	Giorgio	Alleva	ISTAT	Italy
Mr.	Farzan	Madadizadeh	Dept.of Epidemiology and Biostatistics	Republic of Iran, Islamic
Mr.	Hyungsoo	Park	Statistics Korea	Republic of Korea
Dr.	Pieter Cornelis J.	Everaers	European Commission - Eurostat	Luxembourg
Mr.	Ieong Meng	Chao	Direcção dos Serviços de Estatística e Censos	Macao, SAR China
Mr.	Cosme	Vodounou	AFRISTAT	Mali
Mr.	Deepuk	Bahadoor	Statistics Mauritius	Mauritius
Dr.	Julio A.	Santaella Castell	INEGI	Mexico
Dr.	Emilio	Lopez Escobar	Numérika-Medición y Análisis Estad. Avanzado, SC	Mexico
Dr.	Omar	De La Riva Torres		Mexico
Ms.	Liz	MacPherson	Statistics New Zealand	New Zealand
Ms.	Christine Benedicht	Meyer	Statistics Norway	Norway
Mrs.	Alda de Caetano	Carvalho	Inst. Nacional de Estatística (INE)	Portugal
Mr.	Stefan	Lundgren	Statistics Sweden	Sweden
Ms.	Renee	Picanso	National Agriculture Statistics Service	United States
Mr.	Charles	Rothwell	National Center of Health Statistics	United States
Dr.	Graham.	Kalton	WESTAT Inc.	United States
Dr.	Paul	Biemer	Research Triangle Institute	United States
Dr.	Mick P.	Couper	Survey Research Center	United States
Dr.	Stanislav	Kolenikov		United States

IASS Officers and Council Members

Executive Officers

President:	Steve Heeringa (USA)	sheering@isr.umich.edu
President-elect:	Peter Lynn (UK)	Plynn@essex.ac.uk
Vice-Presidents:	Monica Pratesi (Italy)	Monica.pratesi@unipi.it
	Imbi Traat (Estonia)	Imbi.traat@ut.ee
Scientific Secretary:	Denise Silva (Brazil)	denisebritz@gmail.com
Council Members (2013-2017)	Michael Brick (USA)	mikebrick@westat.com
	Daniela Cocchi (Italy)	daniela.cocchi@unibo.it
	Jack Gambino (Canada)	jack.gambino@Canada.ca
	Risto Lehtonen (Finland)	risto.lehtonen@helsinki.fi
	Ralf Münnich (Germany)	muennich@uni-trier.de
	Jean Opsomer (USA)	jopsomer@stat.colostate.edu
(2015-2019)	Hukum Chandra (India)	Hchandra12@gmail.com
	Maria Govanna Ranalli (Italy)	Giovanna.ranalli@stat.unipg.it
	Timo Schmid (Germany)	Timo.schmid@gmx.org
	David Steel (Australia)	dstell@uow.edu.au
	Ineke Stoop (The Netherlands)	i.stoop@scp.nl
	Niko Tzavidis (UK)	n.tzavidis@soton.ac.uk
Chair of the Morocco IASS Program Committee 2017	Marcel Vieira (Brazil)	Marcel.vieira@ice.ufjf.br
Chair of the committee for the Cochran Hansen prize award	Monica Pratesi (Italy)	monica.pratesi@unipi.it

Ex Officio Members

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

Institutional Members

2 International Organisations

AFRISTAT

EUROSTAT

28 Bureaus of Statistics

AUSTRALIA – AUSTRALIAN BUREAU OF STATISTICS

BRAZIL – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE)

CANADA – STATISTICS CANADA

DENMARK –STATISTICS DENMARK

FINLAND – STATISTICS FINLAND

GERMANY –STATISTICHES BUNDESAMT

GHANA – UNIVERSITY OF GHANA

IRAN – ISLAMIC REPULIC OF

ISRAEL- INTERNATIONAL REL. & STATISTICAL COORDINATION

INDIA – INDIAN COUNCIL OF FORESTRY RESEARCH AND EDUCATION PO NEW FOREST

ITALY –INSTITUTO NAZIONALE DI STATISTICSA (ISTAT)

KOREA, REPUBLIC OF – STATISTICS KOREA

LUXEMBOURG – EUROPEAN COMMISSION – EUROSTAT

MACAO, SAR China – DIREÇCAO DOS SERVIÇOS DE ESTATISTICA E CENSOS

MALI – AFRISTAT

MAURITIUS – STATISTICS MAURITIUS

MEXICO – DONATO MIRANDA FONSECA 68col. Adolfo López Mateos

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

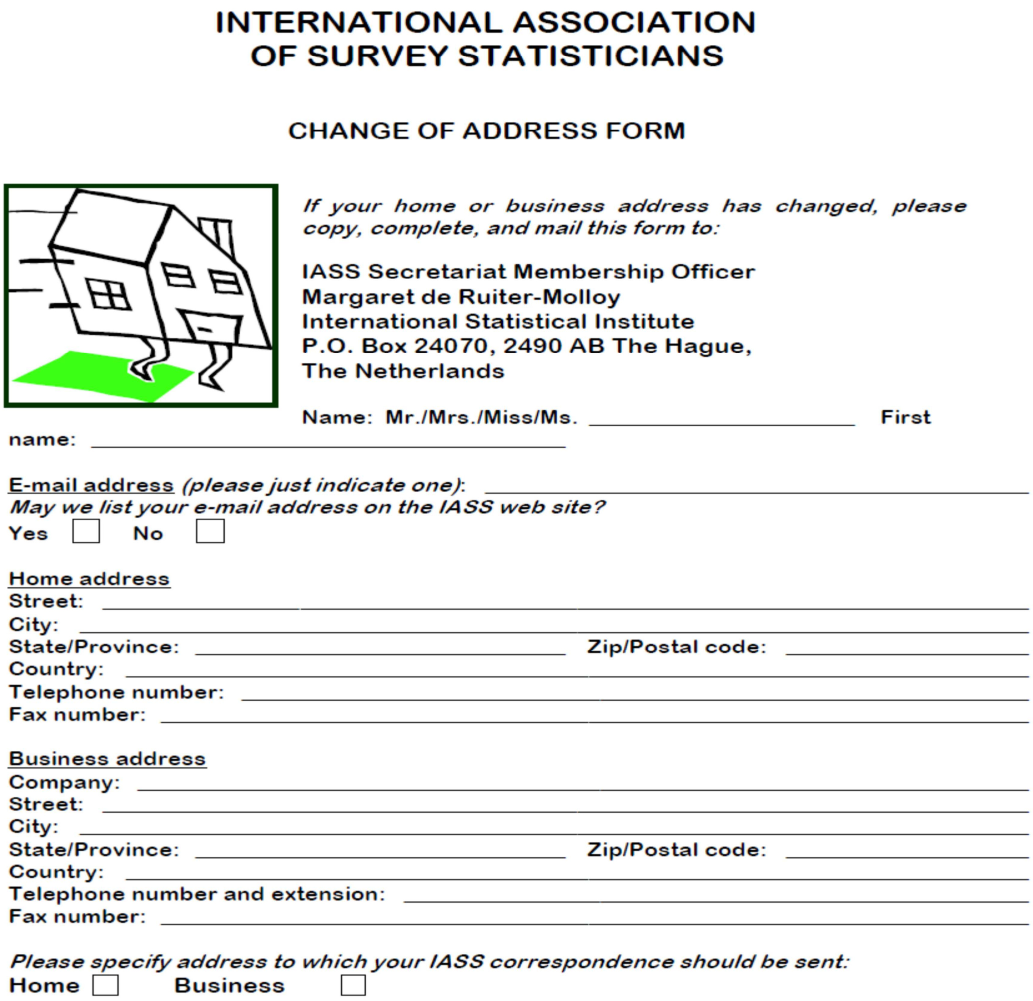
UNITED STATES – NATIONAL AGRICULTURE STATISTICS SERVICE

UNITED STATES – NATIONAL CENTER FOR HEALTH STATISTICS

UNITED STATES – RESEARCH TRIANGLE INSTITUTE

UNITED STATES - SURVEY RESEARCH CENTER

UNITED STATES – WESTAT INC.



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!

the Survey Statistician
The Newsletter of the International Association of Survey Statisticians
No. 49
January 2004
No. 70
July 2014
INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS
International Statistical Institute
ASSOCIATION INTERNATIONALE DES STATISTICIENS D'ENQUETES
Institut International de Statistique

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