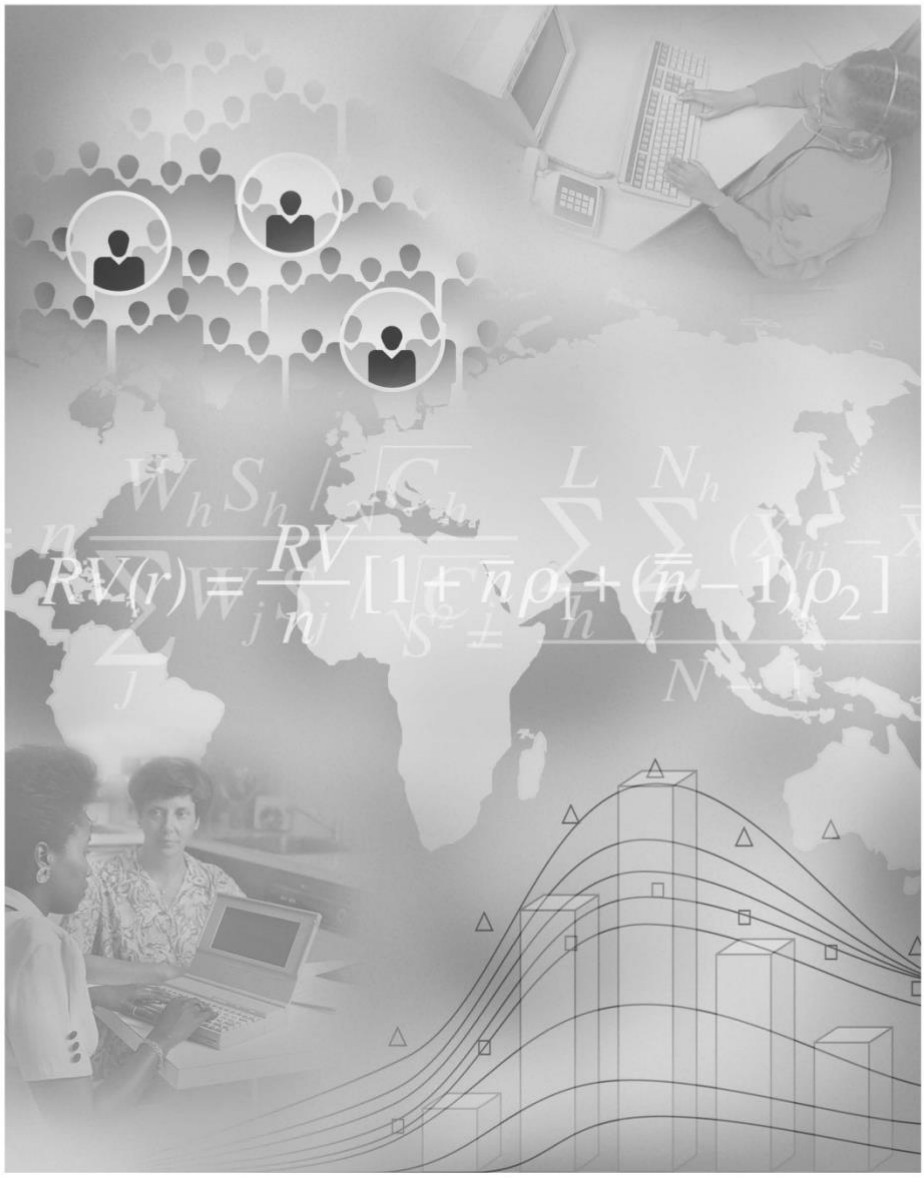
the Survey Statistician

The Newsletter of the International Association of Survey Statisticians

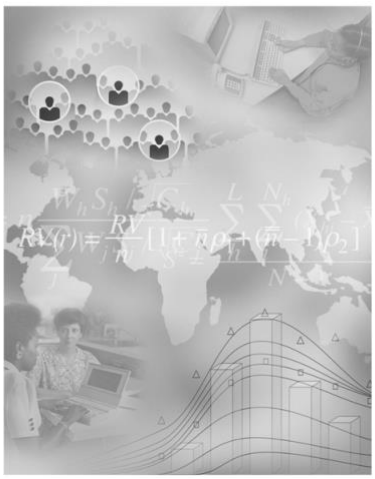
No. 85 **January 2022**



INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

INTERNATIONAL ASSOCIATION OF SURVEY STATISTICIANS

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



The Survey Statistician No. 85, January 2022

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

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

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

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

ISSN 2521-991X

In this Issue

Letter from the Editors

Dear readers,

Happy New Year 2022. We are glad to present the January 2022 issue of TSS, another strong issue.

As usual, we start with the *Letter from the President* by Monica Pratesi. It is followed by the *Report from the Scientific Secretary,* Giovanna Ranalli. We welcome both of them and have already started to enjoy cooperating with them.

In the News and Announcement section, this issue then pays tribute to Jean-Claude Deville who passed away at the end of 2021. The section also contains a report on the 2021 Baltic-Nordic-Ukraine Summer School on Survey Statistics.

In the Ask-the-Experts section, Stefan Bender from the Deutsche Bundesbank & University of Mannheim and Joseph W. Saskhaug from the German Institute for Employment Research & Ludwig Maximilian University of Munich discuss the very much changed context of the use of data sources in business statistics. They explain how this field went from using almost exclusively surveys to incorporating and/or using a variety of administrative and commercial data, as well as text, image and other data from devices and internet.

Then, in the New and Emerging Methods section, Jae-Kwang Kim from Iowa State University, presents an interesting introduction to the ways to integrate data in the context of survey sampling when there are both a probability and a non-probability sample. He considers a number of approaches such as mass imputation, the propensity score method, calibration weighting, and doubly robust estimation methods.

To complete this issue, you will find the country reports (from six countries). As the President explains, we need to confirm and/or renew representatives but in the meantime, we would like to remind all members that this section could benefit from more submissions. This is followed by information on upcoming conferences and workshops and the tables of contents of a number of journals related to survey statistics.

We would like to thank each and everyone of those who devoted their time in organizing and preparing materials for this issue. If you have any information about conferences, events or just ideas you would like to share with other statisticians – please do go ahead and contact any member of the editorial board of the newsletter.

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

Danutė Krapavickaitė (danute.krapavickaite@vilniustech.lt)

Eric Rancourt (eric.rancourt@canada.ca)

Letter from the President

Dear IASS Members,

I hope you are doing well, especially in this unsettling and stress-full time of the coronavirus.

We will celebrate the 50th anniversary of the IASS in 2023. I would like to express my most sincere congratulations to all current and past members and my sincere gratitude to all the Presidents and EC members which brought and will bring the IASS at this important date.

This first formal letter is the occasion to introduce you in the new Executive Committee members and their roles:

a) Vice President for finance: Jairo Arrow;

b) Scientific secretary: M. Giovanna Ranalli (with the support of Nikos Tzavidis);

c) Vice President for liaising with ISI EC and ISI PO plus administrative matters: Natalie Shlomo;

d) Chair of the 2023 Cochran-Hansen prize committee: Nikos Tzavidis;

e) IASS representative on the 2023 WSC scientific programme committee: Natalie Shlomo;

f) IASS representative on the 2023 WSC short course committee: Natalie Shlomo;

g) IASS representative on the ISI publications committee; M. Giovanna Ranalli;

h) IASS Webinars Representatives 2021-2023; Andrea Silva (supported by the EC members as the EC has the collective responsibility of the programme).

IASS today is a prominent and well-respected science organization that brings together many hundreds of scholars from a variety of Institutions, spread all over the world.

The wisdom of our leaders in 1973, when the idea of an IASS was born, and in the following decades, when our association was established, paved the way for a rapid and continuing growth of survey methods in the world.

By the seventies, we have taken a prominent role in shifting the focus of research and policy to understanding and addressing the challenges faced by new data collection tools and methods. During the last decades, survey methods have been progressively evolving, following the development of new tools of data collection: smart devices as meters, mobile phones, GPS systems and several new applications. Methods have been continuously advancing to measuring uncertainty and data quality, integrating Big and small data sources and developing methods to make inferences from the data. At the same time new methods for analyzing large volumes of data, as machine learning and artificial intelligence methods, are emerging. The focus of the IASS has become and will remain engaging with scientific communities involved in data collection and socio-economic data analysis and embracing knowledge systems complementary to survey methods, as applied mathematics, computer science and information technologies. The voice of survey methods is now heard at global and regional conferences, from the Baltic-Nordic-Ukrainian network on survey statistics in Europe, where the IASS is a supporter, to the Workshops organized in Africa and Pakistan, where the IASS is one of only few region-based member associations.

Although our successes are many, it is not the time to rest.

Country IASS representatives need to be confirmed or renewed. Our communication plan using social media needs to be developed. The number of individual members and institution members needs to increase. Young generations of professionals and researchers from developed and developing countries need our support.

In the mandate, the EC's and my priorities will be to

a) engage our members to become even more involved with the IASS both on the Country and Global levels,

b) help our members to grow professionally, through the tools provided by the IASS (Conferences, Workshops, Prizes and Publications) and disseminated through the IASS's network,

c) continue to increase this network with solutions and opportunities through the IASS communication plan and social networks.

We hope that these resources will help you to disseminate and develop survey methodology in your job as well as in your community and also to increase the number of IASS members. I shall devote my following formal letters to tell you of what we have achieved and what we have not.

The new challenges are emerging: we are grappling with an increasing "datafication" of all the aspects of our life: this makes Citizen Science and Citizen Generated Data interesting opportunities to design probability and non-probability surveys, with new "smart" data collection methods, opening new issues in designing training sets of data in artificial intelligence and machine learning, data integration and data analysis. In addition, we face the COVID-19 pandemic. As individuals and as an organization, we are affected by them. COVID-19 postponed some of our supported Conferences and Workshops, which will take place in the next months. COVID-19 transformed the life in our communities and forever altered the ways we conduct research. At the same time, it has given us an opportunity to reflect on what, how and why we pursue science in the survey methods research field.

Many important topics and emerging issues among those mentioned before were and will be treated and discussed in the new-born IASS Webinars series. I invite you all to join the IASS webinars series and act as a promoter of this initiative in your own network of researchers.

This new tool and the Covid 'pause' could provide an opportunity for our community to consider what kind of research in survey methods we would like, to make our partnership more fruitful to serve with accurate evidence and reliable data the scientific community and our stakeholders.

Finally, as we are celebrating the 49th anniversary of the IASS at the computer screens or in socially-distanced settings, we all thank the many colleagues and friends of the ISI PO and of the IASS who invested their passion in making it possible. Thanks also to our current members for support, and rejoice in hoping that the path we are on is the road to working together across disciplines, knowledge systems, borders and generations.

Wishing you a peaceful and serene 2022 with the IASS,

Yours

Monica Pratesi

IASS President

Report from the Scientific Secretary

I have been appointed Scientific Secretary of IASS during the first meeting of the newly elected IASS EC in September. I am very grateful to the members of the EC for their trust, and I am indebted to James Chipperfield for his legacy on this role. As my first duty, I had to choose a topic for the "**New and emerging methods**" section of *The Survey Statistician*. I wanted to give space to methods for data integration as, in my opinion, this is the new framework for survey estimation. In this regard, I didn't have to give it a lot of thought and immediately asked Prof. Jae Kwang Kim (Iowa State University) who has kindly agreed to give *The Survey Statistician* "**A gentle introduction to data integration in survey sampling**". His paper provides a systematic review of data integration techniques for combining, in particular, a probability sample with a non-probability sample when the study variable is observed in the non-probability sample only. In this setting, information bias affects the probability sample, whereas selection bias affects the non-probability sample. Prof. Kim reviews statistical procedures for handling missing data, such as mass imputation, propensity score, calibration weighting, and doubly robust methods to adjust for selection bias in the non-probability sample or adjust for information bias in the probability sample. Please, contact me if you are interested in writing an article for the "New and emerging methods" of future editions of *The Survey Statistician*.

Another activity that has held the EC members busy, and Andrea Diniz da Silva in particular, is continuing with the organization of the **Webinar series** that was inaugurated at the beginning of the Covid Pandemic. The Webinars held after the last Report have covered issues in constructing frames using cost-saving website databases, new sampling approaches to estimate graph related parameters, advances and applications of adaptive survey designs, new challenges for survey methods in the next decade, and robust methods to analyze data coming from sources linked with error. Please, visit the webinar section of our website http://isi-iass.org/home/webinars/` for slides and that of ISI https://www.isi-web.org/events/webinars for upcoming and recorded webinars. At the end of 2021, IASS had reached Webinar number 12 and the EC aims at making it a monthly appointment as this has been a very successful activity that has attracted an audience of up to one hundred attendees. Please, contact Andrea andrea.silva@ibge.gov.br if you have suggestions for topics and/or speakers for the upcoming Webinars.

Writing our **monthly Newsletter** has allowed me to realize how rich the scientific life of survey statisticians is! In the Newsletter we provide news on the life of the IASS, details on Webinars, information on conferences, on the recipients of awards and on call for nominations. Please, feel free to contact me for news and info to be added in the Newsletter by the 15th of each month.

As I am at the beginning of my term, I would like to ask for ideas and suggestions to make my appointment fruitful for the members of the Association.

Maria Giovanna Ranalli

maria.ranalli@unipg.it

IASS Scientific Secretary

News and Announcements

Awards

The Small Area Estimation (SAE) Award Committee selected Prof. Partha Lahiri for the 2020 SAE Award for Outstanding Contribution to Small Area Estimation and Prof. Wayne A. Fuller for the 2021 Award for Outstanding Contribution to Small Area Estimation. Since SAE2020 was cancelled due to the Covid-19 pandemic, both awards were given to the professors in the award ceremony of the SAE2021 conference held virtually from Naples, Italy, during September 20-24, 2021. Congratulations!

Obituary to Jean-Claude Deville

Jean-Claude Deville passed away on November 2021, at the age of 77. A former Inspector General of the French National Institute of Statistics (INSEE), he served as head of the department on statistical methodology. He then arrived at the French National School for Statistics and Information Analysis (ENSAI) in 1998, and stayed there until his retirement in 2010 served as the director of the Laboratory in Survey Statistics in the Center for Research in Economy and Statistics (CREST).

Jean-Claude Deville devoted a large part of his career to statistical research. He made seminal contributions to functional data analysis and factorial analysis. He is also known worldwide for his research in survey sampling, which had a huge impact internationally. His work focused in particular on the theory of calibration estimators, on the generalized share weight method and on balanced sampling. These techniques are nowadays of routine use in statistical offices. Jean-Claude Deville also co-created the famous cube sampling method.

He was elected a member of the International Statistical Institute in 1979 and was involved in the activities of the International Association of Survey Statisticians, of which he was a member of the Board from 1993 to 1997. He also created the conference known as INSEE Statistical Methodology Days, organized since 1991, in which he made more than 20 contributions.

At the head of the Laboratory in Survey Statistics, Jean-Claude Deville supervised numerous PhDs in the field. In 2018, he received the Waksberg Prize, which recognizes prominent statisticians for their innovative work combining theory and practice in the field of survey methodology.

A brilliant and passionate statistician, Jean-Claude Deville was an extremely curious, attentive, and highly educated man. Those who had the privilege of meeting him and interacting with him know how endearing he was. A tribute will be paid to him on the occasion of the next INSEE Statistical Methodology Days (Paris, March 2022).

Guillaume Chauvet

ENSAI, France

Report on the BNU Summer School on Survey Statistics 2021

The Summer School on Survey Statistics 2021 was the 25st in the series of annual scientific and educational events of the Baltic–Nordic–Ukrainian (BNU) Network on Survey Statistics. The event was fully virtual, open for all interested and free from registration fee. The summer school was bilingual and involved sessions in English and Russian. There were three broad main themes of interest: Data integration, Machine Learning, and Small area estimation. Sharing ideas about trends and challenges in both the field of survey sampling and data science was thus made possible. These features together proved successful: a total of 219 registered attendees is by far the highest figure in the series of our events. The main audience consisted of survey statisticians and students from partner universities and national statistical agencies in the Baltic and Nordic countries and Belarus, Ukraine and Poland. There were also a large number of participants from developed and developing countries elsewhere in Europe and also worldwide. Many of them had never before attended the events of the BNU network.

Widely recognized speakers were invited as the keynote lecturers for the series of sessions in English. In her talk, Shu Yang of North Carolina State University, USA, proposed data integration as a new paradigm for survey statistics and presented a systematic review of data integration techniques for combining probability and non-probability samples and for combining probability and big data samples. Many recent data integration methods were covered, including calibration weighting, inverse probability weighting, mass imputation, and doubly robust methods. Piet Daas of Eindhoven University of Technology and Statistics Netherlands, Netherlands, discussed the results of a study on Machine Learning-based classifications of web sites texts in the identification of innovative platform economy and Artificial Intelligence (AI) companies in the Netherlands. Marcin Szymkowiak of Poznan University of Economics and Business and Statistical Office in Poznan, Poland, reviewed the main applications of current SAE methods in official statistics, including labour market, agriculture and business statistics and poverty mapping, and considered possible new developments of SAE, such as the use of big data sources in domain and small area estimation.

Four experts were invited to give lectures on current topics in survey statistics for the series of sessions in Russian: Tetiana Ianevych and Iryna Rozora, both of Taras Shevchenko National University of Kyiv, Ukraine, Tetiana Manzhos of Kyiv National Economic University, Ukraine, and Olga Vasylyk of National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”. Their lectures covered topics on the main estimation methods for sample surveys, calibration estimation for nonresponse bias reduction, problems in the use of Big Data for sample surveys, and estimation for domains and small areas. In particular, the lecture “Main estimation methods for sample surveys” was devoted to basic estimators widely applied in survey sampling: Horvitz-Thompson estimator, nonlinear estimator, regression estimators. Results from a sample survey using StatVillage (hypothetical village in Canada, 1991) were presented. In the lecture on calibration estimation in the presence of nonresponse, the basic concepts, auxiliary information, linear calibration estimators and some other issues were discussed. In the lecture “Estimation for domains and small areas” the following topics were considered: classifications of SAE methods (direct and

indirect, design-based and model-based, and others), classical and new approaches to estimation, and a brief overview of SAE software. Finally, in the lecture on big data and machine learning (ML) different types of ML models (LASSO, SVM, CART etc.) and their usage for responsive/adaptive designs, data processing, and weighting were considered. Two types of ML techniques: supervised and unsupervised learning were presented and examples of ML tasks in a survey research were given. Students from Belarusian and Ukrainian universities were the target audience of the Russian Days program.

The scientific program also included invited talks from partner organizations of the network and a selection of contributed papers on topics in modern survey and official statistics. Abstracts of the presentations are available in "Proceedings of the Summer School on Survey Statistics 2021 of the BNU Network on Survey Statistics", published by Statistics Lithuania (2021). The publication is freely available via the web site of the BNU network https://wiki.helsinki.fi/display/bnu/events.

The summer school was partially funded by the International Association on Survey Statistics (IASS). The support was crucial for organizing the event and made it possible to arrange the lecture sessions in Russian. The event was sponsored by the University of Helsinki, via the organizing and hosting of the virtual Zoom sessions, and the other partner universities and national statistical agencies. Keynote talks and sessions in Russian were recorded and made available to the participants.

The BNU Summer School 2021 was dedicated to the memory of Professor Seppo Laaksonen, who passed away in December 2020. Several generations of students and statisticians have had the opportunity to enjoy his experience and expertise in survey methodology in lectures he has given at the network's numerous educational and scientific events over the years.

Cooperation in education in the field of survey statistics between universities in the Baltic and Nordic countries began in 1992 via the initiative of Professor Gunnar Kulldorff (University of Umeå, Sweden) and has been developed since 1996 as the Baltic-Nordic-Ukrainian Network on Survey Statistics. Today, the network includes partner organizations (universities and national statistical agencies) of eight countries: Belarus, Estonia, Finland, Latvia, Lithuania, Poland, Sweden and Ukraine. More information about the BNU Summer School 2021 and the other activities of the network can be found on the BNU website at https://wiki.helsinki.fi/display/BNU/ .

Maria Valaste, University of Helsinki, Finland

Natalia Bokun, Belarus State Economic University (BSEU), Belarus

Natallia Bandarenka, School of Business of Belarusian State University, Belarus

Olga Vasylyk, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine

Risto Lehtonen, University of Helsinki, Finland

The Survey Statistician, 2022, Vol. 85, 10–18.

Ask the Experts

Data Sources for Business Statistics: What has Changed?

Stefan Bender[1] and Joseph W. Sakshaug[2]

[1] Deutsche Bundesbank & University of Mannheim, stefan.bender@bundesbank.de
[2] German Institute for Employment Research & Ludwig Maximilian University of Munich, joe.sakshaug@iab.de

Abstract[1]

The production of business statistics has been experiencing a shift from a primary reliance on single-source statistics based on survey data to a greater reliance on alternative data sources and multisource statistics. Much of this shift has focused on the potential uses of unstructured data sources originating from digitalization processes. This article provides an overview of the current landscape of data sources for business statistics, highlighting some of their advantages/disadvantages, applications, and opportunities and challenges of linking them.

Keywords: administrative data, Big Data, data linkage, data collection, sample surveys

1 Introduction

Traditionally, data on businesses, establishments, and companies have mostly been used to produce single-source statistics based on surveys in which a coherent and pre-defined set of variables is observed. The advantage of this approach is that units, populations, variables, and timing can be explicitly defined by the researcher or statistician. A substantial part of the production efforts come prior to data collection where an explicit data generating process along the Total Survey Error (TSE) framework can be established (Biemer, 2010). In comparison to administrative data and unstructured business data (discussed later in this article), relatively fewer efforts come after data collection where additional activities such as data quality management and post-processing are performed. The differences in the distribution of pre- and post-data collection effort across different types of data sources (surveys, administrative/commercial data, and unstructured business data) are depicted in Figure 1.

In the last years alternative data sources, such as structured data (e.g. administrative records) and unstructured data (e.g. automated data recording) for businesses, establishments, and companies have received increasing attention and are playing a major role in the production of business

[1] Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem. Acknowledgements: The authors would like to thank Frauke Kreuter (LMU Munich) and Naoual El-Ouche, Maurice Fehr, Elena Triebskorn, Susanne Walter (all Bundesbank).

statistics. The use of these data sources is in a sense secondary because these data are typically collected for purposes other than research or producing business statistics. Because researchers and statisticians are not involved in the data generating process, the effort prior to data collection is low or – if there is some involvement – medium. Most efforts come after data collection because the data must be transformed for statistical or research purposes.

In the case of unstructured data, the distribution of effort is opposite to survey data. Because the data are "organic" or "found", relatively little effort is put into the pre-data generation process. But to transform the data from unstructured to structured or to evaluate the data quality of these unknown and (possibly changing) sources requires a lot of time and effort and specific methodologies to produce accurate estimates for the intended target population.

Bringing these alternative data sources together to supplement more traditional data sources offers new possibilities to increase the information richness of the units being observed. But bringing these different data sources together – for example, with record linkage techniques – into one harmonized data source can be a large effort, because in most cases a common identifier is missing and/or the definition of the units of workplaces, establishments, and companies differ in the data sources.

In this article, we provide an overview of the current landscape of data sources for business statistics, highlighting some of their advantages/disadvantages, applications, and opportunities and challenges of linking them.

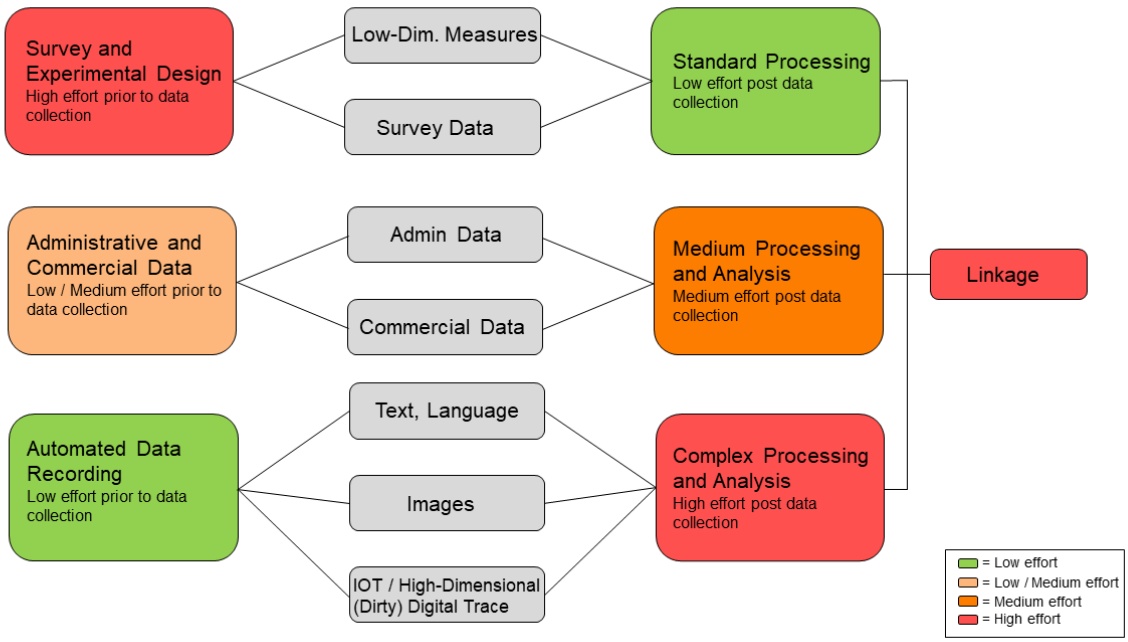


Figure 1. *Distribution of researcher effort for pre- and post-data collection activities for different data sources (adapted and expanded from Stahl et al., 2021, p.6).*

2 Business Survey Data

Business surveys continue to be a dominant source of structured data used to produce official statistics and evaluate and inform economic policies. These data are structured in the sense that the researcher has control over the design of the sample, questionnaire, and data collection procedures, which leads to a standard rectangular dataset of sampled units and variables available for analysis. The amount of pre-data collection effort is high, but the advantage is that every step of the data

generation process is carefully planned and documented and can be deliberately altered to adapt to changing research needs. Further, possible error sources are known in advance according to the TSE framework and the survey can be explicitly designed with those error sources in mind (Biemer, 2010).

Taking a closer look at business surveys, one can immediately see the variations and possibilities that exist for collecting relevant information. In addition to standard cross-sectional and longitudinal designs, there are cross-national business surveys, such as the European Company Survey[2], that allows researchers to perform comparative analyses of business conditions and characteristics. Such data are particularly relevant at the present time as researchers are interested in the impacts of the COVID-19 pandemic on businesses and workplace practices in countries that implemented different containment measures (Jones et al., *in press*). The pandemic has also spurred interest in high-frequency measurements of businesses and how businesses adapt to international crises as they comply with changing regulations. The IAB BeCovid panel survey is one such example of a high-frequency business survey that has collected weekly data from establishments since the early stages of the pandemic (Bellmann et al., 2021).[3]

Although survey data are widely used in the production of business statistics, they are known to be susceptible to errors that can affect their accuracy. For example, nonresponse is a common issue, especially in voluntary business surveys, where response rates have declined over time particularly among larger establishments (König et al*., in press*). Methods to adjust for nonresponse bias, including the use of administrative data (discussed later) and machine learning algorithms, have been the subject of ongoing research (Küfner, Sakshaug, and Zins, 2020). Measurement error and item missing data are also potential issues that affect data quality in business surveys. Given the complex questions asked of businesses and the varying ease with which respondents can access their records or other relevant systems to answer them, there is the potential for misreporting and item nonresponse (Bavdaž, 2010). Although the TSE framework provides an impetus for designing surveys in a way that minimizes the impacts of these error sources, sometimes trade-offs between errors must be made given the survey's budget and research aims.

A key advantage of business surveys is the possibility to embed carefully designed experiments within the data collection. Collecting experimental data is more widespread in household surveys than in business surveys, but recent developments have signaled an increased interest in business survey experimentation (Langeland et al., *in press*). Some experiments are substantive in nature (e.g. vignettes) but also take the form of methodological innovations aimed at reducing survey errors or costs, such as implementing different contact protocols to improve response rates, providing enhanced instructions to complex survey questions in order to reduce item nonresponse, or introducing push-to-web strategies. Sometimes surveys experiment with complete redesigns where multiple changes to the recruitment protocol or questionnaire are implemented simultaneously and compared with the original design on various data quality indicators. However, implementing well-controlled experiments in business surveys can be challenging as production goals are usually prioritized over experimentation, which can lead to unplanned deviations in the implementation of the experiment and possible confounding effects.

3 Administrative and Commercial Business Data

Administrative data typically refers to data generated or collected by governments or other organizations for purposes other than statistics or research. Sources of administrative business data

[2] https://www.eurofound.europa.eu/surveys/european-company-surveys

[3] For the same reason, the Bundesbank has established the Bundesbank Online Panel Firms (Deutsche Bundesbank 2021).

could be business registers or company registrations, records from tax and customs authorities, notifications of social security contributions, reports for fulfilling legal requirements, application forms for loans/credits, and information for subsidies, which were highly relevant during the COVID crisis. Additionally, there is also detailed information from the financial sector available, including investment, trade, financial and capital transactions, financial statements, or insolvency data. For many countries it is possible to bring together administrative data at the employer level with the employee level to have linked employer-employee data. In the field of labor market analysis these linked employer-employee data are one of the main data sources, because they allow researchers to analyze the joint role of worker and firm heterogeneity, both observed and unobserved.

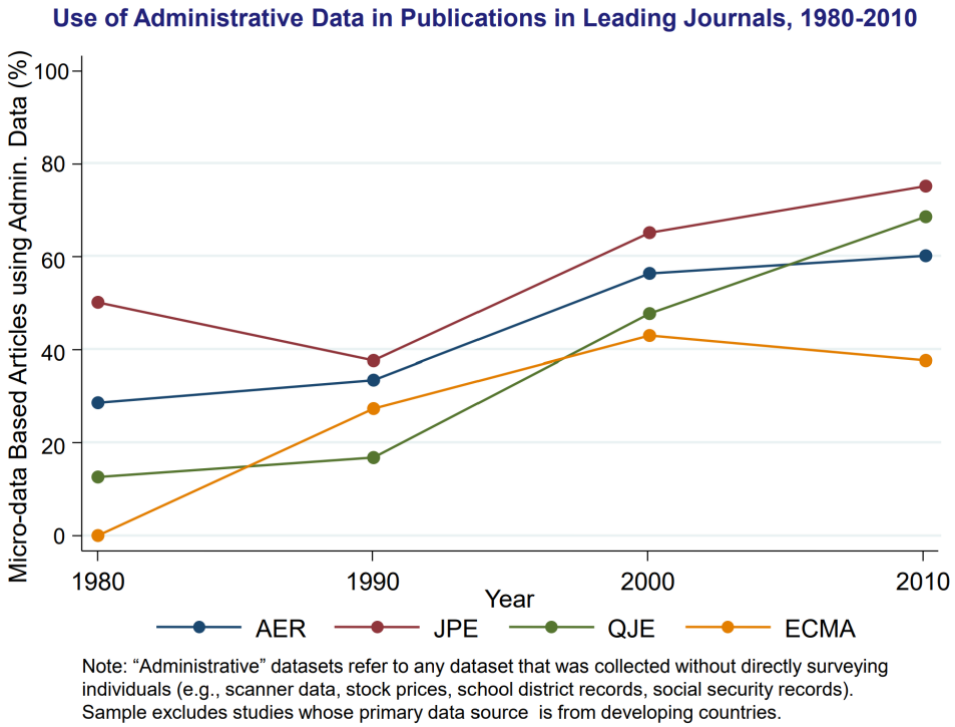


Figure 2. *Articles using administrative data published in leading economic journals between 1980-2010. Source: Chetty (2012)*

Chetty (2012) showed that the use of administrative data in articles published in leading economic journals has increased in recent decades (see Figure 2)[4]. Administrative data have several advantages that have contributed to their popularity in research. As economist and Nobel Laureate David Card and his co-authors (2010) remarked, "Administrative data offer much larger sample sizes and have far fewer problems with attrition, non-response, and measurement error than traditional survey data sources. Administrative data are therefore critical for cutting-edge empirical research and particularly for credible public policy evaluation." Administrative data are often comprised of total populations, but they normally have fewer variables than surveys; though, most variables of interest (e.g. dates, services rendered, status changes) are measured very precisely. They often have a longitudinal structure with a detailed time scale, which allows researchers to follow businesses and their workers over an indefinite time span (and without panel attrition). Given their large quantity, administrative data may also provide more granular information than is possible with surveys and without increasing response burden.

Although administrative data are increasingly used, these data also have some drawbacks. Because most administrative data are not collected for statistics or research, the data generation process is outside the control of the researcher. Therefore, the definitions of some variables might not match 100 percent with the theoretical concept under study, or – in the worst case – the relevant variables

[4] Chetty (2012) also shows a concomitant decrease in the use of secondary survey data in published articles for the same time span and same journals.

are missing. Further, there may also be differences in population coverage, unit types, periodicity, and measurement accuracy compared to survey data. A growing literature investigates the quality issues associated with directly using administrative data in business statistics. An overview of these issues is provided by Van Delden and Lewis (*in press*).

Commercial data is another data source used in business statistics. There are several private companies that compile these data from various sources and offer a variety of data products comprising varying types and amounts of information on business units. Bureau van Dijk[5] and Dun & Bradstreet[6] are two of the major international commercial data providers. There is also an increasing number of local and national commercial data providers, which are closing information gaps by providing business data on sustainability, company networks, and other relevant topics. Federal statistical agencies are increasingly purchasing data from commercial providers as these products promise high data quality. However, commercial data can have similar drawbacks as administrative data in terms of definitional differences, coverage deficiencies, missing data, and measurement error. The quality of these data sources is largely understudied in the literature.

4 Unstructured Business Data

Surveys, administrative data, and – to a lesser extent – commercial data are the backbone of statistical research on businesses. These data are considered as structured data because an underlying data generation model is present, which – in the ideal case – organizes the data in a rectangular format and the relationships between the different rows and columns are known. For surveys, the elements of the data generation model map onto the TSE framework and for administrative data they map onto definitions of units and variables based on regulations or laws. Thus, the data are, in a sense, designed or pre-defined, although they may not necessarily map directly onto the definitions and variables used at a statistical agency or by a researcher.

With the rise of new data science techniques, such as natural language processing, web scraping, text mining, machine learning, advanced visualization techniques, and Artificial Intelligence, so-called unstructured data, such as sensor data, satellite images, scanner data, web sites, data communications, etc. are getting more attention by researchers and statisticians. This attention is also driven by the fact that we are surrounded by unstructured data, and some new research topics require unstructured data as one, or the only, data source.

Some researchers have pointed out that most data are unstructured (e.g. King, 2019). The difference between unstructured and structured data are that unstructured data are not based on an explicit data generation model and the data are not pre-defined. Unstructured data for businesses can be, for example, text information from annual reports, newspaper articles about the business itself, internal records, the management or the location(s) of the business, news/discussions/comments in social media (e.g. Twitter feeds or Facebook), speeches of the higher management, protocols from meetings, or financial or trade information from different sources. In addition to these more text-based sources, pictures can play a significant role, for example, photographs of the company, the company's surroundings, and satellite images. Even marketing videos or videos of CEOs' speeches can be sources for analysis. The use of sensing technology and internet data communication in some industries, including smart farming and transportation, also generates massive amounts of sensor data that can be used for analyzing businesses (Wolfert et al., 2017; Punt and Snijkers, 2019).

In most cases unstructured data must be transformed into structured forms in preparation for analysis. Because the information content is not fixed or determined a priori, different techniques are

[5] https://www.bvdinfo.com/
[6] https://www.dnb.com/

used for different purposes. For example, in text analysis, one can think of the following types of analysis: search for relevant content, clustering, classification, sentiment analysis, synonyms, named entity linkage, general extraction, visualization, summarization, and translation. To transform text into structured data, an analysis pipeline with initial processing, adding linguistic features, converting enriched text to a matrix, and the analysis plan should be established (Klochikhin and Boyd-Graber, 2020).

A hot topic application of unstructured business data is the study of climate change. Businesses play an important role in the discussion of climate change, but there is a lack of high-quality and accessible climate-related data at the business level. The lack of data poses a challenge to policymakers, researchers, statisticians, the private sector, and regulators. Although global progress on improving and making climate data available is underway, in the short- and medium-term such data have to be extracted and collected from mostly unstructured data sources (in addition to some commercial data providers). The German Bundesbank with its Sustainable Finance Data Hub[7], as one example, tries to obtain transition risks, which are typically observed at the business level from unstructured data sources. For example, information on greenhouse gas intensities is published in annual reports, dedicated sustainability reports, on company websites, or are estimated. The information can be reported in the form of tables, pictures, or text.

In addition to climate change, there are other examples of using unstructured data for studying businesses. One example is automatic validation of their economic sector. The economic sector of a business is often (self-)reported in different sources, which leads to different economic sector codes for the same establishments due to misreporting and different measurement schema. A natural question is whether unstructured data, namely, visual information about the company's facilities can be used to validate their economic sector. The Bundesbank is planning to combine information from multiple sources, including structured survey and administrative data with geoinformation, satellite images, and street views to get an indication of the necessity of checking the economic sector of a business (Walter, *in press*).

As with survey and administrative data, these new data sources also carry quality considerations. Representativeness, validity and reliability, coverage issues, and changes in frequency of delivery or data generation processes are just a few such considerations. The various ways in which unstructured data can be prepared for analysis and analyzed also presents a risk of multiple (and possibly conflicting) conclusions being drawn from the same data source. Reproducibility of data preparation and analyses is another important consideration as are data availability, access, sharing, and harmonization. Efforts to adapt the TSE framework to the "Big Data" context are currently underway (Amaya, Biemer, and Kinyon, 2020).

5 Linking Multiple Data Sources

Linking multiple business data sources increases the potential to support statistics, evidence-based policy making, and research. The need for linked business data has increased in recent years, especially for tracking multinationals (large business units) or for describing or analyzing the 2008 financial crisis or the consequences of the COVID-19 pandemic, just to name a few examples. There are a multitude of advantages for combining multiple data sources for businesses, including enhancing the richness of substantive information for a given unit, creating population frames with improved coverage, removing survey questions that are covered by alternative data sources, measurement validation and error adjustment, and improving estimation quality.

[7] Information about the Hub can be found in the presentation of Elena Triebskorn: https://www.bis.org/ifc/events/210709_prog/bundesbank.pdf

A prominent example of linking data sources in the business context is the Longitudinal Employer-Household Dynamics program at the US Census Bureau (Abowd, Haltiwanger, and Lane, 2004), a linkage of various survey and administrative datasets that allow researchers to study labor market dynamics within and across firms, the spatial distribution of employment, and various employment statistics. Linking business registers with trade statistics to compile trade flows by business characteristics is another topic of current research that informs policymakers on the role of businesses in the trade of merchandise, services, and foreign direct investment (Snyder and Jansen, 2015). Statistics Canada has been exploring the integration of administrative data and remote sensing data to supplement or replace survey data, reduce response burden, and implement small area estimation techniques to improve the quality of business statistics (Thomassin, 2018; Duval, Laroche, and Landry. *In press*).

Linking data sources is usually straightforward if there is a unique identifier for each entity in the data sources to be linked. One example of a unique identifier for businesses is the Legal Entity Identifier (LEI), which serves as an international business or entity registration number and allows for the tracing of financial transactions to specific companies or organizations. However, if no unique identifier exists, then identifying the same entities from each of the data sources is more challenging. In this case, the researcher must rely on other indicators that partially identify the entities (e.g. company name, address, economic sector, balance sheets) and link entities that have multiple fields in common. In the context of linking large structured and unstructured business data sources, some of the challenges are linking data sources with very few fields in common, data quality issues (e.g. missing data, misspellings, abbreviations), and duplicate entities. Besides possible linkage errors, other data quality issues can arise when producing multisource statistics, including representation errors and measurement errors. Van Delden et al. (*in press*) provide a framework for conceptualizing these error sources in multisource statistics.

6 Conclusions

The data landscape for business statistics has evolved significantly in recent decades from relying on traditional single-source statistics based on surveys, to greater use of alternative data sources such as administrative/commercial data and unstructured data collected from text, images, video, among others, and multisource statistics based on the combination of these data sources. Each data type is unique in its properties and the amount of pre- and post-data collection processing required to prepare the data for analysis. What remains constant throughout this shift is the importance of understanding the underlying data generation process of each data type, so that researchers are aware of the strengths and limitations of the data when generalizing and drawing conclusions from them. While survey data has established quality frameworks for understanding and quantifying the various error sources that can arise during the data generation process, such frameworks for evaluating the quality of administrative data, commercial data, and unstructured data are only recently starting to emerge. Lastly, the collection of administrative, commercial or unstructured data requires data science skills and methodologies for processing and analyzing these data as well as procedures for accessing, documenting, and archiving these data.[8]

References

Abowd, J.M., Haltiwanger, J., and Lane, J. (2004) Integrated Longitudinal Employer-Employee Data for the United States. *American Economic Review*, **94(2)**, 224–229.

Amaya, A., Biemer, P.P., and Kinyon, D. (2020) Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, **8(1)**, 89–119.

[8] For accessing these types of data see, for example, Bender et al. (in press).

Bavdaž, M. (2010) Sources of Measurement Errors in Business Surveys. *Journal of Official Statistics*, **26(1)**, 25-42.

Bellmann, L., Gleiser, P., Kagerl, C., Kleifgen, E., Koch, T., König, C., Leber, U., Pohlan, L., Roth, D., Schierholz, M., Stegmaier, J., Aminian, A. (2021) *The Impact of the Covid-19 Pandemic: Evidence from a New Establishment Survey*. IAB-Forum, 26th February 2021, https://www.iab-forum.de/en/the-impact-of-the-covid-19-pandemic-evidence-from-a-new-establishment-survey/

Bender, S., Blaschke, J., Hirsch, C. (In Press) Statistical Data Production in a Digitised Age: The Need to Establish Successful Workflows for Micro Data Access In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Biemer, P.P. (2010) Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, **74(5)**, 817-848.

Bundesbank (2021) Assessments and Expectations of Firms in the Pandemic: Findings from the Bundesbank Online Panel Firms. *Deutsche Bundesbank Monthly Report*, **April 2021**, 33-56.

Card, D.E., Chetty, R., Feldstein, M., and Saez, E. (2010) *Expanding Access to Administrative Data for Research in the United States*. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas. http://dx.doi.org/10.2139/ssrn.1888586

Chetty, R. (2012) *Time Trends in the Use of Administrative Data for Empirical Research*. http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf

Duval, M-C., Laroche, R., and Landry, S. (In Press) Integrating Alternative and Administrative Data into the Monthly Business Statistics: Some Applications from Statistics Canada. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Jones, J., Ryan, L., Lanyon, A.J., Apostolou, M., Price, T., König, C., Volkert, M., Sakshaug, J.W., Mead, D., Baird, H., Elliott, D., and McLaren, C.H. (In Press) Producing Official Statistics During the COVID-19 Pandemic. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

King, T. (2019) 80 Percent of Your Data Will Be Unstructured in Five Years. *Data Management Solutions Review*, March 28, 2019. https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

Klochikhin, E., and Boyd-Graber, J. (2020) Text Analysis. In: *Big Data and Social Science, 2nd Edition* (eds. I. Foster, R. Ghani, R. Jarmin, F. Kreuter, and J. Lane), Chapman and Hall/CRC, 193-219.

König, C., Sakshaug, J.W., Stegmaier, J., and Kohaut, S. (In Press) Trends in Establishment Survey Nonresponse Rates and Nonresponse Bias: Evidence from the 2001-2017 IAB Establishment Panel. *Journal of Official Statistics*.

Küfner, B., Sakshaug, J.W., and Zins, S. (2020) *Using Administrative Data and Machine Learning to Address Nonresponse Bias in Establishment Surveys*. Presented at the Big Data Meets Survey Science (BigSurv20) Conference, Virtual, November. https://www.bigsurv20.org/conf20/uploads/15/62/Presentation_BigSurv2020_K_fner.pdf

Langeland, J., Ridolfo, H., McCarthy, J., Ott, K., Kilburg, D., CyBulski, K., Krakowiecki, M., Vittoriano, L., Potts, M., Küfner, B., Sakshaug, J.W., and Zins, S. (In Press) Results from Selected

Experiments in Establishment Surveys. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Punt, T., and Snijkers, G. (2019) *Exploring Sensor-Generated Data in Precision Farming: Towards Official Statistics Using Business Sensor Data*. Heerlen: Statistics Netherlands.

Snyder, N., and Jansen, R. (2015) *Linking Business Registers to Trade Statistics*. Presented at the European Establishment Statistics Workshop 2015 (EESW15), Poznan, Poland, September. https://statswiki.unece.org/download/attachments/123143219/2015Slides_S6_4_Snyder%20Jansen-Linking%20Business%20Registers%20with%20Trade%20Statistics-EESW15.pdf

Stahl, F., Bischl, B., Gehrlein, Kreuter, F., and Tochtermann, K. (2021) *BERD@NFDI in a Nutshell*. https://www.berd-nfdi.de/wp-content/uploads/resources/BERD-NFDI-in-a-nutshell.pdf

Thomassin M. (2018) *The Migration of the Canadian Census of Agriculture to an Integrated Business Program Without Contact with Respondents*. Presented at the Fifth International Workshop on Business Data Collection Methodology, Lisbon, September 2018. https://www.ine.pt/scripts/bdcm/doc/ppt/D2_22_01_BDCMLisbon2018_StatisticsCanada_Mathieu%20Thomassin_PPT.pdf

Van Delden, A., and Lewis, D. (In Press) Methodology for the Use of Administrative Data in Business Statistics. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Van Delden, A., Scholtus, S., De Waal, T., and Csorba, I. (In Press) Methods for Estimating the Quality of Multisource Statistics. In: *Advances in Business Statistics, Methods and Data Collection*, (eds. G. Snijkers, M. Bavdaž, S. Bender, J. Jones, S. MacFeely, J.W. Sakshaug, J. Thompson, and A. Van Delden), John Wiley and Sons.

Wolfert, S., Ge, L., Verbouw, C., Bogaardt, M-J. (2017) Big Data in Smart Farming – A Review. *Agricultural Systems*, **153**, 69-80.

Walter, S (In Press) *A Picture is Worth a Thousand Definitions: Validating Company Data with Satellite Images and Street View*. Technical Report, Deutsche Bundesbank, Research Data and Service Centre.

The Survey Statistician, 2022, Vol. 85, 19–29.

New and Emerging Methods

A gentle introduction to data integration in survey sampling

Jae Kwang Kim

Department of Statistics, Iowa State University, jkim@iastate.edu

Abstract

This article provides a systematic review of data integration techniques for combining a probability sample with a non-probability sample when the study variable is observed in the non-probability sample only. We discuss a wide range of integration methods such as mass imputation, propensity score method, calibration weighting, and doubly robust estimation methods. Finally, we highlight important questions for future research.

Keywords: big data, calibration weighting, doubly robust estimation, mass imputation, propensity score.

1 Introduction

Probability sampling is regarded as the gold-standard in survey statistics for finite population inference. Because probability samples are selected under known sampling designs, they are representative of the target population. Because the selection probability is known, the subsequent inference from a probability sample is often design-based and respects the way in which the data were collected; see Särndal et al. (2003); Cochran (1977); Fuller (2009) for textbook discussions. Kalton (2019) provided a comprehensive overview of the survey sampling research in the last 60 years.

On the other hand, statistical analysis of non-probability survey samples faces many challenges as documented by Baker et al. (2013). Non-probability samples have unknown selection/inclusion mechanisms and typically do not represent the target population. A popular framework in dealing with the biased non-probability samples is to assume that auxiliary variable information on the same population is available from an existing probability survey sample. This framework was first used by Rivers (2007) and followed by a number of other authors including Vavreck and Rivers (2008), Lee and Valliant (2009), Valliant and Dever (2011),

Elliott and Valliant (2017) and Chen et al. (2020), among others. Combining the up-to-date information from a probability sample can be viewed as data integration. Rao (2021) and Yang and Kim (2020) provide comprehensive reviews for data integration for finite population inference.

One can view data integration as a missing data problem, and apply the statistical techniques for handling missing data. Specifically, we consider the following setup for data integration. Let A be a probability sample with observations on auxiliary variable X; let B be the non-probability sample with information on both the study variable Y and the auxiliary variables X. Table 1 presents the general setup of the two sample structure for data integration. As indicated in Table 1, sample B is not representative of the target population.

Table 1: Data Structure for Two Samples

Sample	Type	X	Y	Representative?
A	Probability Sample	✓		Yes
B	Non-probability Sample	✓	✓	No

Under the data structure in Table 1, we wish to develop methods for combining information from two samples. To develop statistical methods for data integration, we may require some assumptions on the outcome model or on the sampling mechanism for sample B.

2 Setup and assumptions

Let $X \in \mathbb{R}^p$ be a vector of auxiliary variables (including an intercept) that are available from two data sources, and let $Y \in \mathbb{R}$ be the study variable of interest. We consider combining a probability sample with X, referred to as sample A, and a non-probability sample with (X, Y), referred to as sample B, to estimate μ_y the population mean of Y. We focus on the case where the study variable Y is observed in sample B only, but the other auxiliary variables are commonly observed in both datasets. The sampling mechanism for sample B is often unknown, and we cannot compute the first-order inclusion probability for Horvitz-Thompson estimation. The naive estimators constructed without adjusting for the sampling process are subject to selection biases. On the other hand, although the probability sample with design weights represents the finite population, it does not contain the study variable. We wish to develop data integration methods that leverage the advantages of both sources.

Let $f(Y \mid X)$ be the conditional distribution of Y given X in the superpopulation model ζ that generates the finite population. Let $\delta_i = 1$ if $i \in B$ and $\delta_i = 0$ otherwise. We make the following assumption.

Assumption 1 *(i) The sampling indicator δ of sample B and the study variable Y are conditionally independent given X; i.e. $P(\delta = 1 \mid X, Y) = P(\delta = 1 \mid X)$; and (ii) $\pi_B(X) \equiv P(\delta = 1 \mid X) > 0$ for all X.*

Assumption 1 (i) and (ii) constitute the strong ignorability condition (Rosenbaum and Rubin; 1983). This assumption holds if the set of covariates contains all predictors for the outcome that affect the possibility of being selected in sample B. Assumption 1 (i) states the ignorability of the selection mechanism to sample B conditional upon the covariates. Under Assumption 1 (i), $E(Y \mid X) = E(Y \mid X, \delta = 1)$ can be estimated based on sample B. Assumption 1 (ii) implies that the support of X in sample B is the same as that in the finite population. Assumption 1

(ii) does not hold if certain units would never be included in the non-probability sample. The plausibility of Assumption 1 (ii) can be checked by comparing the marginal distributions of the auxiliary variables in sample B with those in sample A.

Under the sampling ignorability assumption, there are two main approaches: i) the weighting approach of constructing weights for sample B to improve the representativeness of sample B; ii) the imputation approach of creating mass imputation for sample A using the observations in sample B. There is considerable interest in bridging the findings from a randomized clinical trial to the target population. This problem has been termed as generalizability (Cole and Stuart; 2010; Stuart et al.; 2011, 2015; Keiding and Louis; 2016), external validity (Rothwell; 2005) or transportability (Pearl and Bareinboim; 2011; Rudolph and van der Laan; 2017) in the statistics literature.

3 Mass imputation

In mass imputation, we view the probability sample as having 100% missing values for the study variable. We can then use the non-probability sample as training data to develop an imputation model and construct a synthetic dataset for the probability sample. Mass imputation was originally developed in the context of two-phase sampling (Breidt et al.; 1996; Kim and Rao; 2012) to create synthetic data for the probability sample. Rivers (2007), Kim et al. (2021), and Chen et al. (2021) develop mass imputation for a probability sample using observations from a non-probability sample. Even though the observations in the non-probability sample are not necessarily representative of the target population, the relationships among variables in the non-probability sample can be used to develop a predictive model for mass imputation. Thus, the non-probability sample can be used as training data for developing a model for mass imputation.

We use $\mathbf{x}$ and y to denote the realized value of X and Y in the sample, respectively. In a parametric approach, let $m(\mathbf{x};\boldsymbol{\beta})$ be the posited model for $m(\mathbf{x}) = E(Y \mid \mathbf{x})$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter. Under Assumption 1, a consistent estimator of $\boldsymbol{\beta}$ can be obtained by fitting the model to sample B. Thus, we can estimate $\boldsymbol{\beta}$ by finding the minimizer of

$$Q(\boldsymbol{\beta}) = \sum_{i \in B} \left\{ y_i - m(\mathbf{x}_i;\boldsymbol{\beta}) \right\}^2 / v(\mathbf{x}_i;\boldsymbol{\beta}) = 0$$

for some $v(\mathbf{x};\boldsymbol{\beta}) = V(Y \mid \mathbf{x};\boldsymbol{\beta})$. Thus, we use the observations in sample B to obtain $\widehat{\boldsymbol{\beta}}$ and construct $\widehat{y}_i = m(\mathbf{x}_i;\widehat{\boldsymbol{\beta}})$ for all $i \in A$.

Using $\widehat{y}_i = m(\mathbf{x}_i;\widehat{\boldsymbol{\beta}})$ for all $i \in A$, we can construct

$$\widehat{\mu}_{\mathrm{I}} = N^{-1} \sum_{i \in A} d_{A,i} \widehat{y}_i$$

as the mass imputation estimator of $\mu = N^{-1} \sum_{i=1}^{N} y_i$, where $d_{A,i}$ is the design weight of unit i for sample A. The justification for $\widehat{\mu}_{\mathrm{I}}$ relies on correct specification of $m(\mathbf{x};\boldsymbol{\beta})$ and the consistency of $\widehat{\boldsymbol{\beta}}$. For variance estimation, either linearization method or bootstrap method can be used. See Kim et al. (2021) for more details.

Instead of using parametric mass imputation with a parametric model, we can develop non-

parametric mass imputation using nonparametric models. Rivers (2007) first proposed using nearest neighbor imputation for mass imputation and its asymptotic theory is rigorously discussed by Yang et al. (2021).

4 Propensity Score Method

Under Assumption 1, we can further build a model for $P(\delta = 1 \mid \mathbf{x})$ and use it to construct the propensity score weights for sample B. Suppose that $\pi(\mathbf{x}) = P(\delta = 1 \mid \mathbf{x})$ has a parametric form such that $\pi(\mathbf{x}) = \pi(\mathbf{x}; \phi)$ for some ϕ. The population log-likelihood function for ϕ can be written as

$$l(\phi) = \sum_{i=1}^{N} \left[\delta_i \log \pi(\mathbf{x}_i; \phi) + (1 - \delta_i) \log\{1 - \pi(\mathbf{x}_i; \phi)\}\right].$$

Thus, the (population-based) maximum likelihood estimator of ϕ can be obtained by solving

$$S_p(\phi) \equiv \sum_{i=1}^{N} \left\{\frac{\delta_i}{\pi(\mathbf{x}_i; \phi)} - \frac{1 - \delta_i}{1 - \pi(\mathbf{x}_i; \phi)}\right\} \dot{\pi}(\mathbf{x}_i; \phi) = 0,$$

which is equivalent to solving

$$\sum_{i=1}^{N} \delta_i h(\mathbf{x}_i; \phi) = \sum_{i=1}^{N} \pi(\mathbf{x}_i; \phi) h(\mathbf{x}_i; \phi) \tag{1}$$

for ϕ, where

$$h(\mathbf{x}_i; \phi) = \frac{\dot{\pi}(\mathbf{x}_i; \phi)}{\pi(\mathbf{x}_i; \phi)\{1 - \pi(\mathbf{x}_i; \phi)\}}$$

and $\dot{\pi}(\mathbf{x}; \phi) = \partial \pi(\mathbf{x}; \phi)/\partial \phi$. The left side of (1) can be constructed from sample B. Thus, we have only to estimate the right side of (1). Using the sampling weights, we can use

$$\sum_{i=1}^{N} \delta_i h(\mathbf{x}_i; \phi) = \sum_{i \in A} d_{A,i} \pi(\mathbf{x}_i; \phi) h(\mathbf{x}_i; \phi), \tag{2}$$

which does not require identification of the elements in both samples. Chen et al. (2020) first proposed estimation using (2) for propensity score method for voluntary samples. The final propensity score (PS) estimator for μ is

$$\hat{\mu}_{PS} = \frac{\sum_{i \in B} \hat{\pi}_i^{-1} y_i}{\sum_{i \in B} \hat{\pi}_i^{-1}}, \tag{3}$$

where $\hat{\pi}_i = \pi(\mathbf{x}_i; \hat{\phi})$. If $n_B = |B|$ is small compared with N, then the estimated probability $\hat{\pi}(\mathbf{x}_i)$ can take small values, and the resulting PS estimator in (3) can be unstable.

Elliott and Valliant (2017) proposed a different approach of propensity score method for data integration. Note that

$$P(\delta = 1 \mid \mathbf{x}) \propto P(I_A = 1 \mid \mathbf{x}) \cdot \frac{f(\mathbf{x} \mid \delta = 1)}{f(\mathbf{x} \mid I_A = 1)},$$

where I_A is the sample inclusion indicator function for sample A. Thus,

$$\frac{1}{P(\delta = 1 \mid \mathbf{x})} \propto \{P(I_A = 1 \mid \mathbf{x})\}^{-1} \cdot \frac{f(\mathbf{x} \mid I_A = 1)}{f(\mathbf{x} \mid \delta = 1)} := \tilde{w}(\mathbf{x}) \cdot R(\mathbf{x}).$$

Elliott and Valliant (2017) proposed estimating two terms separately. To estimate the first term $\tilde{w}(\mathbf{x}_i)$, using

$$E(w_i \mid \mathbf{x}_i, I_{A,i} = 1) = \frac{1}{P(I_{A,i} = 1 \mid \mathbf{x}_i)},$$

one can apply regression of w_i on $\mathbf{x}_i$ from sample A. To estimate the second term, Elliott and Valliant (2017) proposed using

$$R(\mathbf{x}) \equiv \frac{f(\mathbf{x} \mid I_A = 1)}{f(\mathbf{x} \mid \delta = 1)} \propto \frac{P(I_A = 1 \mid \mathbf{x}, I_A + \delta \geq 1)}{P(\delta = 1 \mid \mathbf{x}, I_A + \delta \geq 1)}.$$

One can apply a suitable classification method from the combined sample to estimate $R(x)$. The final pseudo weight for sample B is then

$$\hat{w}_i = \tilde{w}_i \hat{R}(\mathbf{x}_i).$$

Rafei et al. (2020) uses Bayesian Additive Regression Trees (BART) to estimate the two components in the pseudo weights for voluntary big data sample.

5 Calibration weighting

The second weighting strategy is calibration weighting, or benchmarking weighting (Deville and Särndal; 1992; Kott; 2006; Breidt and Opsomer; 2017). This technique can be used to calibrate auxiliary information in the non-probability sample with that in the probability sample, so that after calibration the non-probability sample is similar to the target population (Lee and Valliant; 2009).

Instead of estimating the propensity score model and inverting the propensity score to correct for the selection bias of the non-probability sample, the calibration strategy estimates the weights directly. Toward this end, we assign a weight $\omega_{B,i}$ to each unit i in the sample B so that

$$\sum_{i \in B} \omega_{B,i} \mathbf{x}_i = \sum_{i \in A} d_{A,i} \mathbf{x}_i, \tag{4}$$

where $\sum_{i \in A} d_{A,i} \mathbf{x}_i$ is a design-weighted estimate of the population total of X from the probability sample. Constraint (4) is referred to as the covariate balancing constraint (Imai and Ratkovic; 2014), and weights $\mathcal{Q}_B = \{\omega_{B,i} : i \in B\}$ satisfying (4) are the calibration weights. The balancing constraint calibrates the covariate distribution of the non-probability sample to the target population in terms of X. Instead of calibrating each X, one can use model calibration (Wu and Sitter; 2001). In this approach, one can posit a parametric model for $E(Y \mid \mathbf{x}) = m(\mathbf{x}; \boldsymbol{\beta})$ and estimate the unknown parameter $\boldsymbol{\beta}$ from sample B. The model-based calibration specifies the constraints for $\mathcal{Q}_B$ as

$$\sum_{i \in B} \omega_{B,i} m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}) = \sum_{i \in A} d_{A,i} m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}). \tag{5}$$

Suppose that the finite population follows the following superpopulation model:

$$y_i = m(\mathbf{x}_i) + e_i \tag{6}$$

with $E(e_i \mid \mathbf{x}_i) = 0$ and $V(e_i \mid \mathbf{x}_i) = \sigma^2$. If we can express $m(\mathbf{x}) = \sum_{k=1}^{L} \beta_k b_k(\mathbf{x})$ for some $\beta_k, k = 1, 2, \cdots, L$, that is $m(\mathbf{x}) \in \text{span}\{b_1(\mathbf{x}), \cdots, b_L(\mathbf{x})\}$, then we may use

$$\sum_{i \in B} \omega_{B,i} \left[b_1(\mathbf{x}_i), \cdots, b_L(\mathbf{x}_i)\right] = \sum_{i \in A} d_{A,i} \left[b_1(\mathbf{x}_i), \cdots, b_L(\mathbf{x}_i)\right] \tag{7}$$

in the calibration estimation. As long as $m(\mathbf{x}) \in \text{span}\{b_1(\mathbf{x}), \cdots, b_L(\mathbf{x})\}$ holds, the calibration weights in (7) satisfy (5) without estimating $\boldsymbol{\beta}$. The dimension L may increase with the sample size. In this case, some regularization method can be used to choose L. For example, Montanari and Ranalli (2005) used neural network models and Breidt et al. (2005) used penalized Spline models for nonparametric calibration estimation.

Writing $\hat{\mu}_w = N^{-1} \sum_{i \in B} \omega_{B,i} y_i$, we can express

$$\begin{aligned} \hat{\mu}_w - \mu &= N^{-1} \left\{ \sum_{i \in B} \omega_{B,i} m(\mathbf{x}_i) - \sum_{i=1}^{N} m(\mathbf{x}_i) \right\} + N^{-1} \left\{ \sum_{i \in B} \omega_{B,i} e_i - \sum_{i=1}^{N} e_i \right\} \\ &:= C + D. \end{aligned}$$

Since $E(D) = 0$ by model (6), we may require $E(C) = 0$ to get unbiased estimation. A sufficient condition for $E(C) = 0$ under model (6) is the model calibration condition in (5) or (7). To find the optimal calibration estimator that minimizes variance of $\hat{\mu}_w$ in the class of unbiased estimators under model (6), we have only to minimize $E(D^2)$ subject to the calibration constraints. Note that

$$\begin{aligned} E(D^2) &= \text{var} \left\{ N^{-1} \sum_{i=1}^{N} (\delta_i \omega_{B,i} - 1) e_i \right\} \\ &= N^{-2} \sum_{i=1}^{N} (\delta_i \omega_{B,i} - 1)^2 \sigma^2 = \sigma^2 N^{-2} \sum_{i \in B} (\omega_{B,i} - 1)^2 + \text{constant}. \end{aligned}$$

Thus, we can formulate the calibration weighting problem as finding the minimizer of $Q_0(\boldsymbol{\omega}_B) = \sum_{i \in B} (\omega_{B,i} - 1)^2$ subject to (4) or (7) with $\boldsymbol{\omega}_B = \{\omega_{B,i}; i \in B\}$. However, using $Q_0(\boldsymbol{\omega}_B)$ as the objective function for the calibration problem can lead to negative calibration weights.

To avoid negative calibration weights, following Hainmueller (2012), we may consider the entropy divergence

$$Q(\boldsymbol{\omega}_B) = \sum_{i \in B} \omega_{B,i} \log(\omega_{B,i}) \tag{8}$$

as the objective function for optimization. Thus, we find the minimizer of $Q(\boldsymbol{\omega}_B)$ subject to $\omega_{B,i} \geq 0,$ for all $i \in B$; $\sum_{i \in B} \omega_{B,i} = N$, and the balancing constraint (4) or (7). This optimization problem can be solved using convex optimization with a Lagrange multiplier. Other objective functions can also be considered. By introducing Lagrange multiplier $\boldsymbol{\lambda}$, the objective function becomes

$$L(\boldsymbol{\omega}_B, \boldsymbol{\lambda}) = \sum_{i \in B} \omega_{B,i} \log \omega_{B,i} - \boldsymbol{\lambda}' \left\{ \sum_{i \in B} \omega_{B,i} \mathbf{x}_i - \sum_{i \in A} d_{A,i} \mathbf{x}_i \right\}. \tag{9}$$

Thus, by minimizing (9), the estimated weights are

$$\omega_{B,i} = \omega_B(\mathbf{x}_i; \widehat{\boldsymbol{\lambda}}) = N \frac{\exp\left(\widehat{\boldsymbol{\lambda}}' \mathbf{x}_i\right)}{\sum_{i \in B} \exp\left(\widehat{\boldsymbol{\lambda}}' \mathbf{x}_i\right)},$$

where $\widehat{\boldsymbol{\lambda}}$ solves

$$U(\boldsymbol{\lambda}) \equiv \sum_{i \in B} \exp\left(\boldsymbol{\lambda}' \mathbf{x}_i\right) \left\{ \mathbf{x}_i - N^{-1} \sum_{i \in A} d_{A,i} \mathbf{x}_i \right\} = 0. \tag{10}$$

Finally, the calibration weighting estimator is

$$\widehat{\mu}_{\text{cal}} = \frac{1}{N} \sum_{i \in B} \omega_{B,i} y_i. \tag{11}$$

Variance estimation of $\widehat{\mu}_{\text{cal}}$ can be obtained by the standard M-estimation theory by treating $\boldsymbol{\lambda}$ as the nuisance parameter and (10) as the corresponding estimating equation.

Chan et al. (2016) generalize the calibration idea further to develop a general calibration weighting method that satisfies the covariate balancing property with increasing dimensions of the control variables for $m(\mathbf{x})$. Zhao (2019) developed a unified approach of covariate balancing method using Tailored loss functions. The regularization techniques using penalty terms in the loss function can be incorporated into the framework. The covariate balancing condition, or calibration condition, in (4), can be relaxed using soft calibration (Rao and Singh; 1997; Guggemos and Tille; 2010). Wong and Chan (2018) used the theory of reproducing Kernel Hilbert space to develop a uniform approximate balance for covariate functions.

6 Doubly robust estimation

To improve the robustness against model misspecification, one can consider combining the weighting and imputation approaches (Kim and Haziza; 2014). The doubly robust (DR) estimator employs both the propensity score and the outcome models, which is given by

$$\widehat{\mu}_{\text{dr}} = \widehat{\mu}_{\text{dr}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\delta_i}{\pi_B(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}})} \{ y_i - m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}) \} + I_{A,i} d_{A,i} m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}) \right]. \tag{12}$$

The estimator $\widehat{\mu}_{\text{dr}}$ is doubly robust in the sense that it is consistent if either the propensity score model or the outcome model is correctly specified, not necessarily both. Moreover, it is locally efficient if both models are correctly specified (Bang and Robins; 2005; Cao et al.; 2009). Let $\widehat{\mu}_{\text{HT}} = N^{-1} \sum_{i \in A} d_{A,i} y_i$ be the Horvitz–Thompson estimator that could be used if y_i were observed in sample A. Note that

$$\widehat{\mu}_{\text{dr}} - \widehat{\mu}_{\text{HT}} = -\frac{1}{N} \sum_{i=1}^{N} \left\{ I_{A,i} d_{A,i} - \delta_i \{ \pi_B(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}) \}^{-1} \right\} \widehat{e}_i,$$

where $\widehat{e}_i = y_i - m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}})$. To show the double robustness of $\widehat{\mu}_{\text{dr}}$, we consider two scenarios. In

the first scenario, if $\pi_B(\mathbf{x};\boldsymbol{\alpha})$ is correctly specified, then

$$E\left(\widehat{\mu}_{\mathrm{dr}} - \widehat{\mu}_{\mathrm{HT}} \mid \mathcal{F}_N\right) \cong -\sum_{i\in A} d_{A,i}\widehat{e}_i + \sum_{i\in U}\widehat{e}_i$$

which is design-unbiased for zero. In the second scenario, if $m(\mathbf{x};\boldsymbol{\beta})$ is correctly specified, then $E(\widehat{e}_i) \cong 0$. In both cases, $\widehat{\mu}_{\mathrm{dr}} - \widehat{\mu}_{\mathrm{HT}}$ is unbiased for zero and therefore $\widehat{\mu}_{\mathrm{dr}}$ is unbiased for μ_y. Asymptotic expansion of the DR estimator is simplified if the model parameters satisfy the orthogonality condition of Randles (1982). That is, if

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\hat{\mu}_{\mathrm{dr}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathbf{0} \;\text{ and }\; \frac{\partial}{\partial\boldsymbol{\beta}}\hat{\mu}_{\mathrm{dr}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathbf{0} \tag{13}$$

at $(\boldsymbol{\alpha},\boldsymbol{\beta}) = (\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$, then we can safely ignore the uncertainty of estimating $(\boldsymbol{\alpha},\boldsymbol{\beta})$ in the final DR estimation. We can impose (13) in constructing the estimating equation for model parameters.

Yang et al. (2019) extended DR estimation to the high dimensional covariate problem. If both the outcome model and the propensity score model are nonparametric, then the estimator of the form (12) is no longer doubly robust. In this case, estimation using sample splitting can be considered. See Chernozhukov et al. (2017) for details.

7 Discussion

Data integration is a new field of study with a wide range of prospective research subjects. We have considered the situation of merging data from two samples, one from probability sampling and the other from non-probability sampling, where the probability sample lacks the study variable of interest. As a result, information bias affects the probability sample, whereas selection bias affects the non-probability sample. We can adjust for selection bias in the non-probability sample or adjust for information bias in the probability sample using statistical procedures for handling missing data. The majority of data integration methods are based on the unverifiable assumption that the sampling mechanism for the non-probability sample is non-informative. Suppose the non-probability sample is big data. In that case, we can develop the dual frame estimator approach as in Kim and Tam (2021), and the non-informativeness assumption of the sampling mechanism is unnecessary.

Even when the non-informativeness assumption (Assumption 1) is true, the proposed data integration methods employ explicit assumptions for the outcome regression model or sample selection model. Modest model misspecification does not necessarily lead to biased point estimation, but may increase the variance. In this case, the proposed variance estimators based on the assumed model may underestimate the true variance of the data integration estimators. Achieving robustness and assessing uncertainty under modest model misspecification is an important future research topic.

If the sampling mechanism is informative, imputation techniques can be developed under the strong model assumptions for the sampling mechanism (Morikawa and Kim; 2020). As in the non-informative sampling case, the informative sampling assumptions are unverifiable. Thus, sensitivity analysis is recommended to evaluate the robustness of the study conclusions to unverifiable assumptions. Or, if budget is allowed, a follow-up subsampling can be used to build a realistic model for the informative sampling mechanism. Developing tools for data integration under informative sampling is another important research topic.

Acknowledgements

The article is a concisely updated version of the review paper of Yang and Kim (2020). The author is grateful to professors Wayne Fuller and Maria Giovanna Ranalli for their very constructive comments. The research was partially supported by by the National Science Foundation grant (MMS-1733572) and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling, *Journal of Survey Statistics and Methodology* **1**: 90–143.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models, *Biometrics* **61**: 962–973.

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

Breidt, F. J., McVey, A. and Fuller, W. A. (1996). Two-phase estimation by imputation, *Journal of the Indian Society of Agricultural Statistics* **49**: 79–90.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques, *Statistical Science* **32**: 190–205.

Cao, W., Tsiatis, A. A. and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data, *Biometrika* **96**: 723–734.

Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting, *Journal of the Royal Statistical Society, Series B* **78**: 673–700.

Chen, S., Yang, S. and Kim, J. K. (2021). Nonparametric mass imputation for data integration, *Journal of Survey Statistics and Methodology* . Accepted for publication.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples, *Journal of the American Statistical Association* **115**: 2011–2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects, *The American Economic Review* **107**: 261–265.

Cochran, W. G. (1977). *Sampling Techniques*, 3 edn, New York: John Wiley & Sons, Inc.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial, *American Journal of Epidemiology* **172**: 107–115.

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

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples, *Statistical Science* **32**: 249–264.

Fuller, W. A. (2009). *Sampling Statistics*, Wiley, Hoboken, NJ.

Guggemos, F. and Tille, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models, *Journal of statistical planning and inference* **140**: 3199–3212.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis* **20**: 25–46.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score, *Journal of the Royal Statistical Society, Series B* **76**: 243–263.

Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective, *International Statistical Review* **87**: S10–S30.

Keiding, N. and Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys, *Journal of the Royal Statistical Society, Series A* **179**: 319–376.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling, *Statistica Sinica* **24**(1): 375–394.

Kim, J. K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation, *Journal of the Royal Statistical Society, Series A* **184**: 941–963.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach, *Biometrika* **99**: 85–100.

Kim, J. K. and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference, *International Statistical Review* **89**: 382–401.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology* **32**: 133–142.

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

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling, *Journal of the American Statistical Association* **100**: 1429–1442.

Morikawa, K. and Kim, J. K. (2020). Semiparametric optimal estimation with nonignorable nonresponse data, *Annals of Statistics* **49**: 2991–3014.

Pearl, J. and Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach, *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, IEEE, pp. 540–547.

Rafei, A., Flannagan, C. A. C. and Elliott, M. R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees, *Journal of Survey Statistics and Methodology* **8**: 148–180.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters, *Annals of Statistics* **10**: 462–74.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources, *Sankhya B* **83**: 242–272.

Rao, J. N. K. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling, *ASA Proceedings of the Section on Survey Research Methods*, pp. 57–85.

Rivers, D. (2007). Sampling for web surveys, *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA).

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**(1): 41–55.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?", *The Lancet* **365**: 82–93.

Rudolph, K. E. and van der Laan, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**: 1509–1525.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Stuart, E. A., Bradshaw, C. P. and Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations, *Prevention Science* **16**: 475–485.

Stuart, E. A., Cole, S. R., Bradshaw, C. P. and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials, *Journal of the Royal Statistical Society, Series A* **174**: 369–386.

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

Vavreck, L. and Rivers, D. (2008). The 2006 cooperative congressional election study, *Journal of Elections, Public Opinion and Parties* **18**: 355–366.

Wong, R. K. W. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies, *Biometrika* **105**: 199–213.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association* **96**: 185–193.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review, *Japanese Journal of Statistics and Data Science* **3**: 625–650.

Yang, S., Kim, J. K. and Hwang, Y. (2021). Integration of survey data and big observational data for finite population inference using mass imputation, *Survey Methodology* **47**: 29–58.

Yang, S., Kim, J. K. and Song, R. (2019). Doubly robust inference when combining probability and non-probability samples with high-dimensional data, *Journal of the Royal Statistical Society, Series B* **82**: 445–465.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions, *Annals of Statistics* **47**: 965–993.

The Survey Statistician, 2022, Vol. 85, 30-32.

Book and Software Review

Big Data Meets Survey Science. A Collection of Innovative Methods

Alina Matei

Institute of Statistics, University of Neuchâtel, Switzerland. E-mail : alina.matei@unine.ch

Wiley published in 2021 the book entitled Big Data Meets Survey Science. A Collection of Innovative Methods, edited by Craig A. Hill, Paul P. Biemer, Trent D. Buskirk, Lilli Japec, Antje Kirchner, Stas Kolenikov, and Lars E. Lyberg. The book includes selected papers presented at the first conference on Big Data Meets Survey Science (BigSurv18), hosted by the University Pompeu Fabra's Research and Expertise Centre for Survey Methodology in Barcelona, Spain in 2018, and conducted under the auspices of the European Survey Research Association.

Featuring a broad range of topics, the book includes 24 chapters organized in six sections, and offers a wide range of tools, methods, and approaches that illustrate how Big Data sources and methods are used in survey and social sciences to improve official statistics and estimates. Finding answers to the two following questions is essential for survey statisticians/official statisticians:

1) How are Big Data defined?
2) What does '*survey science'* mean?

First, the editors provide in the book introduction the following definition: 'In contrast to censuses or surveys that generate designed and sample data, we view Big Data as nonsampled data that are organic or found in sources for which the primary purpose is not statistical inference *per se*. In particular, in this book we use the term of Big Data to refer to a collection of datasets so large, complex, and rapidly changing that they become difficult to process using extant database management tools or traditional data processing applications.' Second, it seems that a formal definition of 'survey science' is not given in the book. One understands that it represents a mixture between survey methodology and data science.

The information provided in the book is huge, reason for which I focus my discussion on some parts related to official statistics and survey estimation, mostly presented in Section 3 ('Big Data in Official Statistics'). Compared to the definition advocated above, census and administrative data are also seen as Big Data in Section 3, when referring to large populations. It is difficult to define what is 'large'. Chapter 11 (Tam et al., 2021), for instance, provides an application in Subsection 11.6, where the population size is 1,000,000. Does this represent Big Data? Numerous existing surveys around have been dealing with such or even larger amount of data.

Following Holt (2007), Chapter 9 (Japec and Lyberg, 2021) advocates and illustrates several possible issues for official statistics using Big Data: 'wider, deeper, quicker, better, and cheaper'. The aspect 'cheaper' is illustrated, for example, by the estimation of the Consumer Price Index using scanner data from retail stores, first used by Statistics Sweden, and employed now by several countries. Chapter 9 also discusses the very important aspect of the Big Data quality. 'Selectivity, in that Big Data subpopulations often do not coincide with target populations studied in official statistics' represents one of the most important challenges for National Statistical Offices.

The future will tell us if 'wider, deeper, quicker, better, and cheaper' will represent characteristics of Big Data statistics. Definitely, Big Data provide nowadays a 'cheaper' solution in some cases. Nevertheless, this may not always be the case in the future. Big Data statistics are quicker to obtain, yet they may not be wider or deeper if they do not represent, for instance, the whole population under study, or if a carefully designed study is not in place. Last, but not least, the aspect 'better' will depend upon all the other mentioned characteristics.

Chapter 10 (Braaksma et al., 2021) offers a very good description on how Big Data are used at Statistics Netherlands, and how they may be used in general in official statistics. It furthermore describes innovative experiments useful to develop other methods to deal with Big Data. This chapter also presents two strategies to handle Big Data in official statistics: 1) use Big Data as an indicator of the society, assuming however some imperfections due to the way these data are collected, and 2) use Big Data into a model or model-assisted approach and conduct similar analyses as for classical survey data. The authors underline the high degree of changes in Big Data, and a lack of information about the data-generating process. They also advocate the link between multiple sources, and provide the following example among others: individual social media information may be combined to some information known from registers for instance. Nevertheless, many users keep personal information private, to protect their privacy on social media. In this context, it is quite difficult to successfully use such a data combination.

Chapter 11 (Tam et al., 2021) offers a more technical content. It is dedicated to data combination between a nonprobability sample or Big Data set and a probability sample. The method is based on calibration, assuming that different totals are known. Several limitations of the method are underlined by the authors. The method is, however, promising, and opens the way for new methodological developments.

It is important to mention the use of machine learning in clustering and prediction models. Chapter 1 (Buskirk and Kirchner, 2021) provides an informative review on the use of machine learning methods (MLMs) in surveys. The use of algorithms is not new in survey statistics. Methods based on clustering algorithms are used for a while, for example, to create imputation classes; see for instance Haziza and Beaumont (2007). 'Compared to traditional statistical methods, MLMs are more prone to overfitting the data, that is, to detecting patterns that might not generalize to other data', underline the authors of Chapter 1. Overfitting is not, however, an MLM issue, and traditional statistical criteria and methods may lead to overfitted models (especially when cross-validation techniques are not used during the model selection process). Recently, several papers in survey estimation have used model-assisted estimators, that attenuate the potential impact of overfitting produced by MLMs (McConville et al., 2017; Mehdi Dagdoug et al., 2022).

I welcome the attempt of the cited chapters' authors to also underline the drawbacks of Big Data sources and methods. However, my impression in reading the book is that a gap between survey statisticians and Big Data defenders is still present. More research must be done from the methodological point of view to accommodate Big Data in our 'routine' as survey statisticians. In any case, Big Data should be used together with conventional statistical sources and methods whenever they can bring new insights.

In conclusion, the book includes a large diversity of topics, making it informative for a broad audience including survey and social science researchers, and survey statisticians/official statisticians. Without any doubt, the book represents an important contribution to survey science. The Big Data debate still continues, but I hope that the book will help to diminish the mentioned gap between survey statisticians and Big Data defenders.

References

Braaksma, B., Zeelenberg, K., and de Broe, S. (2021). Big Data in Official Statistics: A Perspective from Statistics Netherlands. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 303-338. Wiley.

Buskirk, T. D. and Kirchner, A. (2021). Why Machines Matter for Survey and Social Science Researchers: Exploring Applications of Machine Learning Methods for Design, Data Collection,

and Analysis. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 9-62. Wiley.

Haziza, D. and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. International Statistical Review, 75(1):25-43.

Holt, D. T. (2007). The Official Statistics Olympic Challenge: Wider, Deeper, Quicker, Better, Cheaper. *The American Statistician*, 61(1):1-8.

Japec, L. and Lyberg, L. (2021). Big Data Initiatives in Official Statistics. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 275-302. Wiley.

McConville, K. S., Breidt, F. J., Lee, T. C. M., and Moisen, G. G. (2017). Model-Assisted Survey Regression Estimation with the Lasso. *Journal of Survey Statistics and Methodology*, 5(2):131-158.

Dagdoug, M., Goga, C., and Haziza, D. (2022). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, to appear.

Tam, S. M., Kim, J.-K., Ang, L., and Pham, H. (2021). Mining the New Oil for Official Statistics. In Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., and Lyberg, L. E., editors, *Big Data Meets Survey Science: A Collection of Innovative Methods*, pages 339-357. Wiley.

Country Reports

ARGENTINA

Reporting: **Verónica Beritich**

INDEC starts the National Survey on Time Use

The National Institute of Statistics and Censuses (INDEC) reports that from October until December 2021, the National Time Use Survey (ENUT) will be carried out. This statistical survey will enable the lives of people of various ages to be characterized, as well to understand the time they allocate to the activities performed inside and outside their home.

The ENUT 2021 target is to meet the new demands for information from the population and expanding the statistical map to other dimensions of daily life and people's well-being. Once data are obtained, an exhaustive analysis will be carried out to know the balance between people's life and work and to know the amount of time they dedicate to reading, studying, caring for other people, doing housework, and participating in recreational and cultural activities among other topics.

This survey will interview 28,520 selected dwellings from urban areas throughout the country. It will inquire about paid work, domestic tasks and caring for other members of the households, and personal activities. In addition, it will be possible to know the contribution to GDP of unpaid work, and to monitor the commitments assumed by Argentina in relation to the Sustainable Development Goals of the United Nations 2030 Agenda.

The operation will be accomplished in person by means of a questionnaire in two blocks of questions, one about households and the other about people. For the first time, an activity diary will be introduced which will include the tasks that people had inside and outside their homes during the 24 hours of the day prior to the date of the interview. Recording multitasking will be permitted with up to three simultaneous activities in each of the 10-minute segments presented in the diary.

The data provided by those who participate in the survey are strictly confidential and protected by statistical secrecy, in accordance with the provisions of Law 17,622 and Decree 3,110 / 70.

The institutional video of the operation is available at the following link: https://youtu.be/0OHHbzEF73I

General information can be found at www.indec.gob.ar.

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

BRAZIL

Reporting: **Dr. Andrea Diniz da Silva**

Regional Hub for Big Data in Brazil

Since April 2021, the Brazilian Institute of Geography and Statistics (IBGE) is hosting the United Nations Global Platform Regional Hub for Big Data. The Regional Hub supports projects in the use of Big Data and data science for official statistics and SDG indicators in Latin America and the

Caribbean region. Cooperation, training, research, and conferences are the four workstreams of the Regional Hub to leverage sharing of knowledge on newly developed methods, algorithms and tools. The Regional Hub is a milestone for the region and can improve statistical production enhancing the use of Big Data, in complement to the surveys and census, for official and experiential statistics. Preliminary results of a consultation on the use of Big Data for official statistics with the national statistical offices revealed that in several countries in the region experimental statistics as well as studies are already in course, nevertheless such a practice is not yet imbedded into the regular production processes. Ongoing activities can be followed at https://hub.ibge.gov.br.

CANADA

Reporting: **Steve Matthews**

Statistics Canada is pushing the envelope to provide more timely information on the Canadian economy

Statistics Canada delivers high quality statistics on various components of the Canadian economy to enable evidence-based decision making by our data users. High quality information historically comes at a cost in timeliness – it takes time to apply the many steps required to produce traditional statistical estimates, ranging from sampling and data collection, through to analysis and dissemination. In recent years, the agency has been exploring methods to publish more timely economic information, and the COVID-19 pandemic increased the urgency of this initiative. Beginning in the spring of 2020, *flash estimation* methods were used to publish early indicators of key economic measures including monthly gross domestic product, and monthly retail sales which received increasing attention from data users. What is a flash estimate? At Statistics Canada the term *flash estimate* refers to an indicator that is available earlier than the official statistic, and is produced by applying traditional methodological approaches to a partial information set (e.g. survey responses received early in data collection). This approach has provided early indicators that, despite being available up to one month in advance, have predicted the economic indicators very closely. Unfortunately, this approach is not effective for all indicators and further gains in timeliness are increasingly difficult.

With this in mind, methods based on statistical models, called *nowcasting* methods, have been studied to further advance the timeliness of information, and harness the predictive power of the ever-growing body of available data sources. What is a nowcast? At Statistics Canada, the term nowcast refers to an estimate of an indicator made available soon after the reference month, produced by building and applying statistical models to predict the indicator of interest. Statistical agencies are well positioned to apply nowcasting; the approach based on statistical models makes it possible to include information beyond the confidential data available internally (e.g. survey responses or administrative sources). Gains in accuracy and timeliness are also possible by introducing data from social media, big data sources, and other information that may be publicly available. Based on this data, statistical modelling techniques using time series, machine learning, or classical statistics can be applied to nowcast a given indicator of interest. Ideally, many data sources for the reference period that we aim to predict are already available, which makes the nowcast more robust to unexpected shocks and distinguishes nowcasting from classical forecasting. In comparison to flash estimation, nowcasting increases the potential for timeliness gains without imposing additional burden on respondents to provide timely data.

Nowcasting methods have recently been applied to a number of Statistics Canada's indicators to evaluate their potential and have shown promise. Current work is focussed on Canada's monthly Gross Domestic Product – a key indicator used heavily in economic policy, and represents a challenge for nowcasting as it encompasses all sectors of the Canadian economy. Providing early

indicators along with the existing GDP release (two month lag) would provide users with much more up-to-date economic information. In particular we aim to produce estimates based on flash estimation or nowcasting methods extending as close as possible to real-time, with quality sufficient to meet user needs. This ongoing work is also expected to evolve to potentially target more granular industrial and geographic domains, and higher-frequency economic indicators which could completely change the information that we offer on the current economic situation in Canada.

JAPAN

Reporting: **Dr. Ryozo Yoshino**

Recent developments in the Japanese data archives, and newly released results of the Japanese National Character Survey

The Japan Society for the Promotion of Science (JSPS) and the National Institute of Informatics (NII) have released the "Japanese Data Catalogue for the Humanities and Social Sciences" (JDCat), a system for searching research data in the humanities and social sciences. JDCat is a cross-disciplinary search system for research data currently maintained by five research institutes selected by the JSPS through an open call for proposals as part of the JSPS's "Program for Building a Data Infrastructure for the Humanities and Social Sciences." The five research institutions are Hitotsubashi University's Institute of Economic Research, the University of Tokyo's Institute of Social Science's Center for Social Research and Data Archives, Keio University's Panel Data Research Center, Osaka University of Commerce's JGSS Research Center, and the University of Tokyo's Institute of Archives and History. The current release targets research data in the social sciences, but it is planned to add research data in the humanities around October this year. (For more information on JDCat please see
https://www.jsps.go.jp/english/edi/data/JDCat_NII_20210716_en .pdf).

In addition, the Institute of Statistical Mathematics has released the outline and basic tabulation of the survey results of the 14th "Japanese National Character Survey" (2018), which has been conducted every five years for about 70 years since 1953: https://www.ism.ac.jp/survey/index_ks14.html (in Japanese). The results of the 1st to 13th surveys (1953 to 2013) are available at the following website: https://www.ism.ac.jp/kokuminsei/en/index_e.html (in English). For an overview of the history and data analysis of this survey and related international comparative surveys, please refer to the recently published "Cultural Manifold Analysis on National Character" [Yoshino, 2021] and its references (https://link.springer.com/book/10.1007/978-981-16-1673-0). For a series of related international comparative studies, please refer to the following: https://www.ism.ac.jp/~yoshino/index_e.html/.

NEW ZEALAND

Reporting: **Dr. Hannes Diener**

An experimental Administrative Population Census

In what will sound very familiar to many national statistics offices, Stats NZ is looking into moving towards an administrative-first census approach supported by surveys. Stats NZ's census transformation programme (CT) has been investigating alternative census models since 2012. This August, we have released the first iteration of the experimental Administrative Population Census

(APC). The APC is an instrument for us to engage with our customers and treaty partners and invite their feedback on the strengths and weaknesses of moving to an administrative-first census.

Much like a full field enumeration census, the goal of the APC is to provide fine-grained population, social, and economic statistics for small areas and communities. Unlike a fully enumerated census that happens in the field, the enumeration happens in linked administrative data. The APC is built using Stats NZ's integrated data infrastructure (IDI). The IDI provides researchers with de-identified and linked unit-record information. The IDI data-holdings includes tax, border movements, birth and death registrations, health service provider data, educational enrolments, and more. The production of the APC also built on years of CT research and the practical experience of admin enumeration used in the 2018 Census to compensate for a low field response rate. The APC is an annual time series from 2006–2020. It includes an admin NZ resident population and a selection of demographic and identity variables: age, sex, geography, ethnicity, Māori descent, birthplace, and years since arrival in NZ. While analysis and development are ongoing, advantages of an administrative-first census over a full field enumeration one are discernible:

- it can be produced more frequently and reduce respondent burden;
- it can provide more accurate responses such as determining income from tax data versus self-reported income band;
- It can be a longitudinal time series. This allows for cohort analysis. For example, it is possible to follow annual population flows between lower geographics which usually cancel out on an aggregate level.

We are still working our way through to some of the more difficult methodological questions, and there is still a lot more work we need to do before we are confident that we could make a smooth transition to an admin first census. One area that we will be focusing on is to figure out how to collect census information which is not covered in administrative data sources and integrate this data into an administrative census. Most likely this information will come from a regular large-scale (attribute) survey, similar in scale to the U.S. Census Bureau's American Community Survey.

Engagement has been a main element of the APC. A successful transition to an administrative-first census relies on genuine engagement with our customers, stakeholders, and our Te Tiriti o Waitangi (Treaty of Waitangi) partners to ensure we get their buy-in, meet their needs, and build their trust and confidence. The iterative design of the APC includes an active engagement plan to seek early feedback, and continually incorporate it into the following releases. We are excited about the possibilities the APC has to offer and are looking forward to adding more puzzle pieces in the upcoming years.

For more details see Stats NZ (2021), Experimental administrative population census: Data sources and methods. For more information on APC, please contact Hannes.Diener@stats.govt.nz.

UNITED STATES

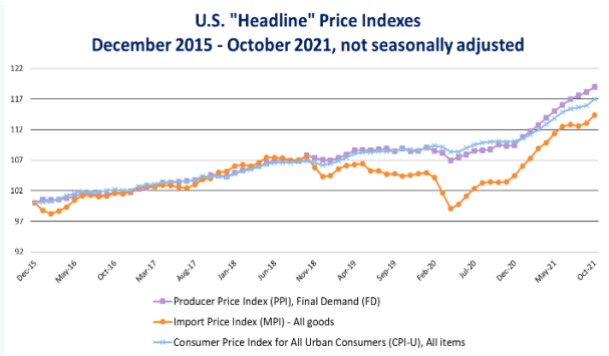
Reporting: **Jeffrey Hill**

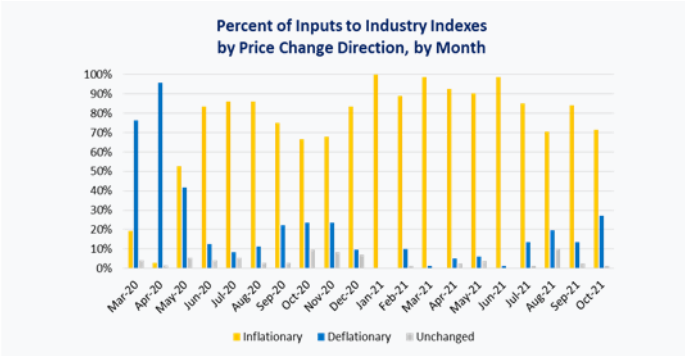
U.S. Input to Industry price indexes reflect inflationary pressures facing businesses

The U.S. Bureau of Labor Statistics (BLS) now publishes a satellite inputs to industry data series. These indexes measure price change for the net inputs consumed by most 3-digit North American Industry Classification System (NAICS) industry groups, excluding capital investment and labor. To construct an overall input to industry index, the BLS first calculates two separate indexes, one measuring price change for domestically produced inputs and the other measuring price change for imported inputs. BLS uses its Producer Price Index (PPI) commodity series to construct the domestic portion of the overall index and its Import Price Indexes (MPIs) to construct the imported portion.

The two indexes are then aggregated to an overall price index that measures price change for inputs to the industry sector regardless of their country of origin.

While the most straightforward use of the net inputs to industry price indexes is to measure changes in industry input costs over time, they also provide data users with an opportunity to analyze price transmission between BLS input and output price indexes for industry groups. During this post-2020 recession recovery period, these indexes demonstrate that inflationary pressures facing businesses are continuing as 2021 comes to a close.





The “headline,” or most aggregate MPI, representing all imported goods, and the headline PPI for Final Demand, representing prices received by domestic producers for products sold to end users, fell sharply at the start of the pandemic (February-April 2020) and then began increasing in May 2020 such that by 2021 prices were above pre-pandemic levels. Prices for consumer goods, tracked by the overall CPI for all urban consumers, followed suit.

One explanation for the consumer price change is that producers passed fluctuations in their input costs on to consumers. At the start of the pandemic when many prices were falling, the inputs to industry series reflects that more industry groups experienced lower input costs, with nearly 96% of them seeing falling input prices in April 2020. For example, prices for petroleum products (used as inputs by many industries) declined substantially during the beginning of the pandemic. As prices for imports and prices received by domestic producers increased beginning in May 2020, the series reflects that many industry groups faced higher input costs, with 100% of them seeing increases in costs for January 2021. Examples of rising input costs during this period include lumber purchased by construction industries and furniture manufacturers, organic chemicals purchased by plastics and rubber manufacturers, and wheat purchased by food manufacturers. Industries experiencing decreased costs in the autumn of 2021 were those that faced large price increases earlier and for which prices then significantly fell, as was the case with lumber.

The new satellite series are not official statistics, but the BLS welcomes feedback from data users at Satellite_Series_Feedback@bls.gov as it considers publishing the series as an official data product. More information is available on the BLS website at https://www.bls.gov/ppi/input-indexes/home.htm.

Upcoming IASS-Supported Conferences in 2022

Latest Developments in the Theory and Practice of Sample Surveys and Censuses will be held on 12th March 2022 and followed by a workshop **Utilization of Remote Sensing in Sample Surveys and Censuses** held on 13th March, 2022. Organised by Pak Institute of Statistical Training And Research (PISTAR) Website: http://pistar.org/

Other Conferences on survey statistics and related areas

SAE 2022 – The 2022 **Small Area Estimation** international conference will take place at the University of Maryland, College Park, USA campus during May 23-27, 2022. https://sae2022.org/ In addition to traditional topics in SAE, the conference will cover a few emerging topics in survey and official statistics (e.g., nonprobability sampling, probabilistic record linkage, data fusion, etc.) In principle, this will be an in-person conference following the University of Maryland, College Park, guidelines. However, in view of the on-going pandemic, international participants can join the conference virtually.

ITACOSM2022 – The 7th ITAlian Conference on Survey Methodology

The Conference “Survey methods for Data Integration and New Data Sources” will be hosted by the Department of Political Sciences of the University of Perugia (Italy), 8-10 June, 2022. A short course on “Survey Data Integration” will be held in Assisi (Palazzo Bernabei, Italy) on June 7th by Jae-Kwang Kim. https://meetings3.sis-statistica.org/index.php/ITACOSM2022/ITACOSM2022

Q2022 – the European Conference on Quality in Official Statistics

The conference will be held on 8 to 10 June 2022 in Vilnius, Lithuania https://q2022.stat.gov.lt/lt/ The event will focus on the institutional challenges of quality management, quality assurance in the emerging data ecosystem. This conference should also draw the attention of governmental bodies to the importance of high-quality, timely and more detailed statistics, and increase awareness among other stakeholders of the challenges faced by producers of official statistics, especially in times of crisis.

In addition, one-day training courses on quality management, the role of statistics in the era of Big Data and in a future society, innovation and modernisation practices, as well as data journalism and data visualisation.

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

Writing manuscripts for Official Statistics journals: Guidelines for practitioners and researchers

Under the auspices of the ISI, the *Statistical Journal of the IAOS* (IAOS), *The Survey Statistician* (IASS), *Journal of Official Statistics* (JOS, Statistics Sweden), *Survey Methodology* (SMJ, Statistics Canada), IOS Press and Wiley are organizing a workshop comprising three separate webinars of two hours each. The workshop will be held online February 8, 10, 15, 2022. https://www.isi-web.org/events/node-1221.

The objective of this workshop is to prepare Official Statisticians and researchers to draft and submit manuscripts to Official Statistics journals. The workshop focuses on manuscript drafting and structuring skills as well as on anticipating the knowledge level and expectations of the audiences and on organizing and preparing a manuscript for submission to a journal in the field of Official Statistics.

In Other Journals

Journal of Survey Statistics and Methodology

Volume 9, Issue 3, June 2021

https://academic.oup.com/jssam/issue/9/3

Survey Methodology

Telephone Sample Surveys: Dearly Beloved or Nearly Departed? Trends in Survey Errors in the Era of Declining Response Rates
David Dutwin, Trent D Buskirk

Telephone Sample Surveys: Dearly Beloved or Nearly Departed? Trends in Survey Errors in the Era of Declining Response Rates
David Dutwin, Trent D Buskirk

Transitions from Telephone Surveys to Self-Administered and Mixed-Mode Surveys: AAPOR Task Force Report
Kristen Olson, Jolene D Smyth, Rachel Horwitz, Scott Keeter, Virginia Lesser, Stephanie Marken, Nancy A Mathiowetz, Jaki S McCarthy, Eileen O'Brien, Jean D Opsomer, Darby Steiger, David Sterrett, Jennifer Su, Z Tuba Suzer-Gurtekin, Chintan Turakhia, James Wagner

An Experimental Evaluation of an Online Interview Scheduler: Effects on Fieldwork Outcomes
Katherine McGonagle, Narayan Sastry

The Relationship Between Interviewer-Respondent Rapport and Data Quality
Hanyu Sun, Frederick G Conrad, Frauke Kreuter

Fit for Purpose in Action: Design, Implementation, and Evaluation of the National Internet Flu Survey
Jill A Dever, Ashley Amaya, Anup Srivastav, Peng-Jun Lu, Jessica Roycroft, Marshica Stanley, M Christopher Stringer, Michael G Bostwick, Stacie M Greby, Tammy A Santibanez, Walter W Williams

Exploring Scale Direction Effects and Response Behavior across PC and Smartphone Surveys
Dagmar Krebs, Jan Karem Höhne

An Evaluation of Mixture Confirmatory Factor Analysis for Detecting Social Desirability Bias
Alexandru Cernat, Caroline Vandenplas

Survey Statistics

Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data
Jörg Drechsler, Jingchen Hu

Combining Multiple Imputation and Hidden Markov Modeling to Obtain Consistent Estimates of Employment Status
Laura Boeschoten, Danila Filipponi, Roberta Varriale

Effects of Outcome and Response Models on Single-Step Calibration Estimators
Daifeng Han, Richard Valliant

Combining Information from Multiple Data Sources to Assess Population Health
Trivellore Raghunathan, Kaushik Ghosh, Allison Rosen, Paul Imbriano, Susan Stewart, Irina Bondarenko, Kassandra Messer, Patricia Berglund, James Shaffer, David Cutler

Oversampling of Minority Populations Through Dual-Frame Surveys
Sixia Chen, Alexander Stubblefield, Julie A Stoner

Volume 9, Issue 4, September 2021

https://academic.oup.com/jssam/issue/9/4

Survey Methodology

Comparing Methods for Assessing Reliability
Roger Tourangeau, Hanyu Sun, Ting Yan

Assessing Response Quality by Using Multivariate Control Charts for Numerical and Categorical Response Quality Indicators
Jiayun Jin, Geert Loosveldt

The Interviewer Contribution to Variability in Response Times in Face-to-Face Interview Surveys
Patrick Sturgis, Olga Maslovskaya, Gabriele Durrant, Ian Brunton-Smith

Disentangling Interviewer and Area Effects in Large-Scale Educational Assessments using Cross-Classified Multilevel Item Response Models
Theresa Rohm, Claus H Carstensen, Luise Fischer, Timo Gnambs

Moving from Face-to-Face to a Web Panel: Impacts on Measurement Quality
Alexandru Cernat, Melanie Revilla

Viewing Participation in Censuses and Surveys through the Lens of Lifestyle Segments
Mary H Mulry, Nancy Bates, Matthew Virgile

Survey Statistics

Finding a Flexible Hot-Deck Imputation Method for Multinomial Data
Rebecca Andridge, Laura Bechtel, Katherine Jenny Thompson

Multiply Robust Bootstrap Variance Estimation in the Presence of Singly Imputed Survey Data
Sixia Chen, David Haziza, Zeinab Mashreghi

Bayes-Raking: Bayesian Finite Population Inference with Known Margins
Yajuan Si, Peigen Zhou

Multivariate Logistic-Assisted Estimators of Totals from Clustered Survey Samples
Timothy L Kennel, Richard Valliant

Sample Bias Related to Household Role
Marcin Hitczenko

Corrigendum

CORRIGENDUM TO: Methods for Exploratory Assessment of Consent-To-Link in a Household Survey (JSSM, 2019, 7(1),118-155)
Daniel Yang, Scott Fricker, John Eltinge

Volume 9, Issue 5, April 2021

https://academic.oup.com/jssam/issue/9/5

Survey Methodology

Survey Costs: Where are We and What is the Way Forward?
Kristen Olson, James Wagner, Raeda Anderson

Using Time Series Models to Understand Survey Costs
James Wagner, Heidi Guyer, Chrissy Evanchek

Survey Reliability: Models, Methods, and Findings
Roger Tourangeau

Using Placeholder Text in Narrative Open-Ended Questions in Web Surveys
Tanja Kunz, Franziska Quoß, Tobias Gummer

Machine Learning for Occupation Coding—A Comparison Study
Malte Schierholz, Matthias Schonlau

Survey Statistics

Dealing with Inaccurate Measures of Size in Two-Stage Probability Proportional to Size Sample Designs: Applications in African Household Surveys
Graham Kalton, Ismael Flores Cervantes, Carlos Arieira, Mike Kwanisai, Elizabeth Radin, Suzue Saito, Anindya K De, Stephen McCracken, Paul Stupp

A Note on Chromy's Sampling Procedure
Guillaume Chauvet

Application of Probability-Based Link-Tracing and Nonprobability Approaches to Sampling Out-of-School Youth in Developing Countries
Tom Krenzke, Leyla Mohadjer

Boosted Kernel Weighting – Using Statistical Learning to Improve Inference from Nonprobability Samples
Christoph Kern, Yan Li, Lingxiao Wang

Applications

Blending Probability and Nonprobability Samples with Applications to a Survey of Military Caregivers
Michael W Robbins, Bonnie Ghosh-Dastidar, Rajeev Ramchand

Total Error and Variability Measures for the Quarterly Workforce Indicators and LEHD Origin-Destination Employment Statistics in Onthemap
Kevin L Mckinney, Andrew S Green, Lars Vilhuber, John M Abowd

Information Entropy and Scale Development
Daniel Friesner, Carl Bozman, Matthew McPherson, Faith Valente, Anqing Zhang

---

Survey Methodology
Statistics Canada Statistique Canada Canada

Survey Methodology, December 2021, vol. 47, no.2

https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2021002-eng.htm

Waksberg Invited Paper Series

Multiple-frame surveys for a multiple-data-source world
Sharon L. Lohr

Regular Papers

Replication variance estimation after sample-based calibration
Jean D. Opsomer and Andreea L. Erciulescu

Two local diagnostics to evaluate the efficiency of the empirical best predictor under the Fay-Herriot model
Éric Lesage, Jean-François Beaumont and Cynthia Bocci

Estimating the false negatives due to blocking in record linkage
Abel Dasylva and Arthur Goussanou

With-replacement bootstrap variance estimation for household surveys Principles, examples and implementation
Pascal Bessonneau, Gwennaëlle Brilhaut, Guillaume Chauvet and Cédric Garcia

Short notes

An alternative jackknife variance estimator when calibrating weights to adjust for unit nonresponse in a complex survey
Phillip S. Kott and Dan Liao

Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling
Yong You

Assessing the coverage of confidence intervals under nonresponse. A case study on income mean and quantiles in some municipalities from the 2015 Mexican Intercensal Survey
Omar De La Riva Torres, Gonzalo Pérez-de-la-Cruz and Guillermina Eslava-Gómez

Volume 37 (2021): Issue 3 (September 2021)

Special Issue on Population Statistics for the 21st Century

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

Preface
Jakub Bijak, Johan Bryant, Elżbieta Gołata and Steve Smallwood

Letter to the Editors
Giampaolo Lanzieri

Fertility Projections in a European Context: A Survey of Current Practices among Statistical Agencies
Rebecca Folkman Gleditsch, Astri Syse and Michael J. Thomas

Modelling Frontier Mortality Using Bayesian Generalised Additive Models
Jason Hilton, Erengul Dodd, Jonathan J. Forster and Peter W.F. Smith

Probabilistic Projection of Subnational Life Expectancy
Hana Ševčíková and Adrian E. Raftery

Spatio-Temporal Patterns in Portuguese Regional Fertility Rates: A Bayesian Approach for Spatial Clustering of Curves
Zhen Zhang, Arnab Bhattacharjee, João Marques and Tapabrata Maiti

Optimal Sampling for the Population Coverage Survey of the New Italian Register Based Census
Paolo Righi, Piero Demetrio Falorsi, Stefano Daddi, Epifania Fiorello, Pierpaolo Massoli and Marco Dionisio Terribili

A General Framework for Multiple-Recapture Estimation that Incorporates Linkage Error Correction
Daan Zult, Peter-Paul de Wolf, Bart F. M. Bakker and Peter van der Heijden

Exploratory Assessment of the Census of Pakistan Using Demographic Analysis
Asif Wazir and Anne Goujon

A Simulation Study of Diagnostics for Selection Bias
Philip S. Boonstra, Roderick J.A. Little, Brady T. West, Rebecca R. Andridge and Fernanda Alvarado-Leiton

Fay-Herriot Model-Based Prediction Alternatives for Estimating Households with Emigrated Members
Jairo Fúquene-Patiño, César Cristancho, Mariana Ospina and Domingo Morales Gonzalez

Volume 37 (2021): Issue 4 (December 2021)

https://sciendo.com/issue/JOS/37/4

Freedom of Information and Personal Confidentiality in Spatial COVID-19 Data
Michael Beenstock and Daniel Felsenstein

Response Burden and Data Quality in Business Surveys
Marco Bottone, Lucia Modugno and Andrea Neri

Evaluating the Utility of Linked Administrative Data for Nonresponse Bias Adjustment in a Piggyback Longitudinal Survey
Tobias J.M. Büttner, Joseph W. Sakshaug and Basha Vicari

Combining Cluster Sampling and Link-Tracing Sampling to Estimate Totals and Means of Hidden Populations in Presence of Heterogeneous Probabilities of Links
Martín Humberto Félix-Medina

Comparing the Response Burden between Paper and Web Modes in Establishment Surveys
Georg-Christoph Haas, Stephanie Eckman and Ruben Bach

Trends in Establishment Survey Nonresponse Rates and Nonresponse Bias: Evidence from the 2001-2017 IAB Establishment Panel
Corinna König, Joseph W. Sakshaug, Jens Stegmaier and Susanne Kohaut

Robust Estimation of the Theil Index and the Gini Coeffient for Small Areas
Stefano Marchetti and Nikos Tzavidis

Occupation Coding During the Interview in a Web-First Sequential Mixed-Mode Survey
Darina N. Peycheva, Joseph W. Sakshaug and Lisa Calderwood

Nowcasting Register Labour Force Participation Rates in Municipal Districts Using Survey Data
Jan van den Brakel and John Michiels

The Robin Hood Index Adjusted for Negatives and Equivalised Incomes
Marion van den Brakel and Reinder Lok

Estimation of Domain Means from Business Surveys in the Presence of Stratum Jumpers and Nonresponse
Mengxuan Xu, Victoria Landsman and Barry I. Graubard

Book Review
Alina Matei

Survey Practice

Vol. 14, Issue 1, 2021

https://www.surveypractice.org/issue/2728

Articles

Web and paper survey mode patterns and preferences, Health & Employment Survey, World Trade Center Health Registry
Kacie Seil, Shengchao Yu, Robert Brackbill, Lennon Turner

What to Do With All Those Open-Ended Responses? Data Visualization Techniques for Survey Researchers
Jessie Rouder, Olivia Saucier, Rachel Kinder, Matt Jans

Impact of demographic survey questions on response rate and measurement: A randomized experiment
Jeanette Y. Ziegenfuss, Casey A. Easterday, Jennifer M. Dinh, Meghan M. JaKa, Thomas E. Kottke, Marna Canterbury

Validating the Sixteen-Item Transportation Security Index in a Nationally Representative Sample: A Confirmatory Factor Analysis
Alexandra K. Murphy, Alix Gould-Werth, Jamie Griffin

Survey Research Methods

Journal of the European Survey Research Association

Vol 15 No 2 (2021)

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

Establishing a baseline: bringing innovation to the evaluation of cross-national probability-based online panels
Gianmaria Bottoni, Rory Fitzgerald

Dependent interviewing: a remedy or a curse for measurement error in surveys?
Paulina Pankowska, Bart F. M. Bakker , Daniel L. Oberski, Dimitris Pavlopoulos

Enhancing the Demand for Labour survey by including skills from online job advertisements using model-assisted calibration
Maciej Eryk Beręsewicz, Greta Białkowska, Krzysztof Marcinkowski, Magdalena Maślak, Piotr Opiela, Pawlukiewicz Katarzyna, Robert Pater

Directional Pattern based Clustering for Quantitative Survey Data: Method and Application
Roopam Sadh, Rajeev Kumar

Evaluation of Estimated Survey Duration Equations Using a Health Risk Assessment
Brittany Carter, James Bennett, Elric Sims

More Clarification, Less Item Nonresponse in Establishment Surveys? A Split-Ballot Experiment
Benjamin Küfner, Joseph W. Sakshaug, Stefan Zins

Vol 15 No 3 (2021)

https://ojs.ub.uni-konstanz.de/srm/issue/view/227

Reducing Respondent Burden with Efficient Survey Invitation Design
Hafsteinn Einarsson, Alexandru Cernat, Natalie Shlomo

Measurement quality of 67 common social sciences questions across countries and languages based on 28 Multitrait-Multimethod experiments implemented in the European Social Survey
Carlos Poses, Melanie Revilla, Marc Asensio, Hannah Schwarz, Wiebke Weber

Assessing consent for and response to health survey components in an era of falling response rates: National Health and Nutrition Examination Survey, 2011-2018
Geraldine McQuillan, Deanna Kruszon-Moran, Hua Di, Denise Schaar, Susan Lukacs, Tala Fakhouri, Eric Tolliver, Ryne Paulose-Ram

Personal Values Strongly Predict Study Dropout
Johannes Beller, Siegfried Geyer

Understanding the patterns of mode switching in longitudinal studies
Alexandru Cernat, Joseph W. Sakshaug

The Response Entropy Index: Comparative Assessment of Performance and Cultural Bias across Indices of Careless Responding
John Tawa

Other Journals

- **Statistical Journal of the IAOS**
  - https://content.iospress.com/journals/statistical-journal-of-the-iaos/
- **International Statistical Review**
  - https://onlinelibrary.wiley.com/journal/17515823
- **Transactions on Data Privacy**
  - http://www.tdp.cat/
- **Journal of the Royal Statistical Society, Series A (Statistics in Society)**
  - https://rss.onlinelibrary.wiley.com/journal/1467985x
- **Journal of the American Statistical Association**
  - https://amstat.tandfonline.com/uasa20
- **Statistics in Transition**
  - https://sit.stat.gov.pl

Welcome New Members!

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

PROF.			
DR.	Annamaria	Bianchi	Italy
MS	Margaret Beryl-Ann	Clarkson	Grenada
DR.	John Lamont	Eltinge	United States
MR.	Antoine	Simonpietri	France
MS	Katherine Jenny	Thompson	United States
DR.	Arnout	Van Delden	The Netherlands

IASS Executive Committee Members

Executive officers (2022 – 2024)

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

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

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

Institutional Members

International organisations:

- Eurostat (European Statistical Office)

National statistical offices:

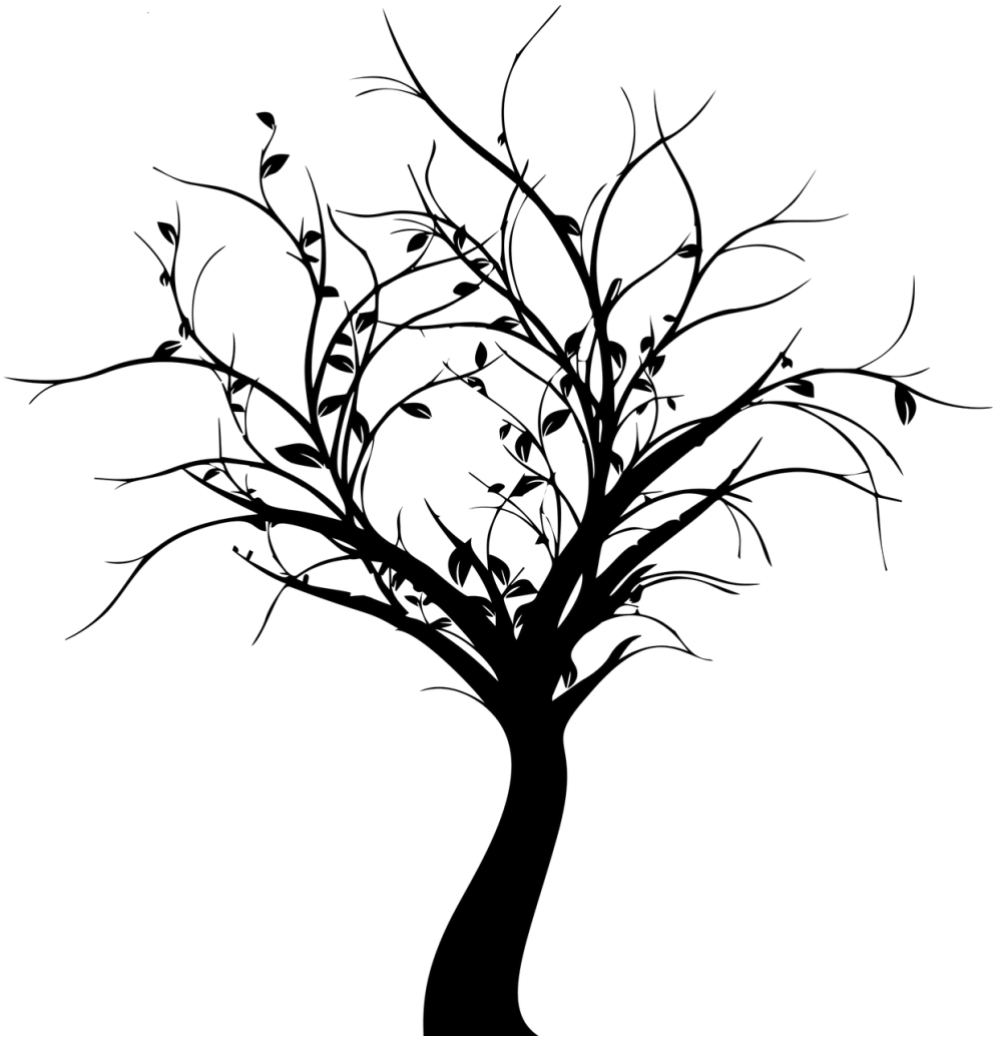
- Australian Bureau of Statistics, Australia
- Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil
- Statistics Canada, Canada
- Statistics Denmark, Denmark
- Statistics Finland, Finland
- Statistisches Bundesamt (Destatis), Germany
- Israel Central Bureau of Statistics, Israel
- Istituto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

Private companies:

- Westat, United States

Save a tree!

Read *the Survey Statistician* online!



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