Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Survey Methodology
43-2

Release date: December 21, 2017

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2017

•

Volume 43

•

Number 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (statcan.smj-rte.statcan@canada.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 43, Number 2, December 2017

Contents

Special paper

Regular papers

Short note

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 145-160
Statistics Canada, Catalogue No. 12-001-X

Sample survey theory and methods: Past, present, and future directions

J.N.K. Rao and Wayne A. Fuller[1]

Abstract

We discuss developments in sample survey theory and methods covering the past 100 years. Neyman's 1934 landmark paper laid the theoretical foundations for the probability sampling approach to inference from survey samples. Classical sampling books by Cochran, Deming, Hansen, Hurwitz and Madow, Sukhatme, and Yates, which appeared in the early 1950s, expanded and elaborated the theory of probability sampling, emphasizing unbiasedness, model free features, and designs that minimize variance for a fixed cost. During the period 1960-1970, theoretical foundations of inference from survey data received attention, with the model-dependent approach generating considerable discussion. Introduction of general purpose statistical software led to the use of such software with survey data, which led to the design of methods specifically for complex survey data. At the same time, weighting methods, such as regression estimation and calibration, became practical and design consistency replaced unbiasedness as the requirement for standard estimators. A bit later, computer-intensive resampling methods also became practical for large scale survey samples. Improved computer power led to more sophisticated imputation for missing data, use of more auxiliary data, some treatment of measurement errors in estimation, and more complex estimation procedures. A notable use of models was in the expanded use of small area estimation. Future directions in research and methods will be influenced by budgets, response rates, timeliness, improved data collection devices, and availability of auxiliary data, some of which will come from "Big Data". Survey taking will be impacted by changing cultural behavior and by a changing physical-technical environment.

Key Words: Data collection; History of survey sampling; Probability sampling; Survey inference.

1 Introduction

This paper was prepared at the invitation of Dr. Danny Pfeffermann, 2015 President of the International Association of Survey Statisticians, who provided the ambitious title. The paper was presented at the meetings of the International Statistical Institute in Rio de Janeiro, Brazil in 2015.

The title defines an area too large for us to address in a single paper. Furthermore, there are a number of review papers that address the topics of the title, including Kish (1995), Bellhouse (2000), Rao (2005), Bethlehem (2009), Brick (2011), Groves (2011), and Brewer (2013). Our discussion draws on those papers, but we do not attempt completeness. We provide a brief appraisal of the three topics and project a number of current situations into the future. Our aim is to stimulate further discussion, especially on the future directions. Beyond the discussion of controversies related to purposive sampling, we will concentrate on probability-based sampling. Because survey sampling is an applied field, some of the problems encountered and methods employed in practice will be addressed. Our discussion is most relevant for large general purpose samples, the surveys where we have the most experience. Likewise, our knowledge of applications is concentrated in Canada and the United States.

The paper is organized as follow. Section 2 presents the early landmark contributions from 1920-1960. Inferential issues are covered in Section 3. The paper concludes with a discussion on the future in Section 4.

1. J.N.K. Rao is Distinguished Research Professor, Carleton University, Ottawa, Canada, K15-5B6. E-mail: jrao@math.carleton.ca; Wayne A. Fuller is Distinguished Professor Emeritus, Iowa State University, Ames, IA, USA. 50011. E-mail: waf@iastate.edu.

2 Early landmark contributions: 1920-1960

Kiaer (1897) is perhaps the first to promote sampling (or what was then called the representative method) over complete enumeration (census), although the oldest reference can be traced back to 1000 BC. In the representative method, the objective is for the sample to mirror the parent finite population and this may be achieved either by balanced sampling on known auxiliary totals, through purposive selection or by random sampling leading to equal inclusion probabilities. By the 1920s the representative method was widely used. The International Statistical Institute (ISI) played a vital role by creating an expert committee to report on this method. Bowley's (1926) contribution to the ISI report includes his fundamental work on stratified random sampling with proportional allocation, leading to equal inclusion probabilities. Bowley (1936) states that the "first application of this principle" of inferring the population from the sample was the 1912 study in Reading. Bowley specified the sampling procedure for that study as a systematic sample from a list of houses. Bowley called the systematic procedure a "pure method of sampling" and stated, "This is literally the method of stratified sampling". Bowley gives a number of examples where systematic sampling was used after 1912. Bowley (1936) emphasized the importance of a complete frame and equal probabilities of selection. But it was Neyman (1934) who laid the foundations of probability sampling (or design-based approach). He demonstrated that stratified random sampling is preferable to balanced (representative) sampling as it was used then. He also introduced the concept of efficiency and optimal sample allocation, now called Neyman allocation, that minimizes the total size of the sample for a specified precision by relaxing Bowley's condition of equal inclusion probabilities. In fact, Tchuprow (1923) derived the Neyman allocation ten years earlier, in a paper discovered after the Neyman paper appeared. Neyman (1934) also showed that for large enough samples one could obtain confidence intervals on the population mean of a variable of interest such that the frequency of errors in the confidence statement in repeated sampling does not exceed the limit prescribed in advance, "whatever the unknown properties of the population". In recent years, balanced sampling, originally advocated by Gini and Galvani, has been refined to incorporate the nice features of both probability sampling and balanced sampling on known auxiliary totals (Deville and Tillé, 2004). The new balanced sampling method is now used in Europe, especially in France, to select samples for establishment surveys. A second method of probability controlled selection is rejective sampling, introduced by Hájek (1964) as a method for controlling the sample size in Poisson sampling. Fuller (2009a) extended the procedure to restrict acceptable samples to the set where estimates of the means of auxiliary variables are close to the population mean.

The 1930s witnessed a rapid growth in demand for socio-economic information, and the advantages of probability sampling in terms of greater scope, reduced cost, and greater speed relative to censuses, were soon recognized worldwide. This led to an increase in number and type of surveys based on probability sampling and covering large populations. Neyman's probability sampling (or design-based approach) was almost universally accepted and it became a standard tool for empirical research in social sciences and official statistics. It was also recognized that the precision of an estimator is determined largely by the sample size and not by the sampling fraction. The 1940's saw a number of studies on the properties of systematic sampling for different populations. See Madow and Madow (1944), Cochran (1946), and Yates (1948). Cochran (1977, Chapter 8) is an excellent discussion of systematic sampling, a discussion that makes

clear why only model-based estimators of variance are possible. Also see Bellhouse (1988). In the early development of sampling theory, focus was on estimating totals and means and associated sampling errors. Non-sampling errors such as nonresponse, coverage errors, and measurement errors, were largely ignored in theoretical research.

We now list a few important post-Neyman theoretical developments in the design-based approach. Mahalanobis used multi-stage sampling designs for crop surveys in India as early as 1937. His classic 1944 paper (Mahalanobis, 1944) rigorously formulated cost and variance functions for the efficient design of surveys. He was instrumental in creating the National Sample Survey of India, the largest multi-subject continuing survey with full-time staff using personal interviews for socio-economic surveys and physical measurements for crop surveys. Sukhatme, who studied under Neyman, also made pioneering contributions to the design and analysis of large scale agricultural surveys in India, using stratified multi-stage sampling. Classic text books on sampling by Cochran (1953), Deming (1950), Hansen, Hurwitz and Madow (1953), Sukhatme (1954) and Yates (1949) benefited students as well as practitioners.

Survey statisticians at the U.S. Census Bureau, under the leadership of Morris Hansen, made fundamental contributions to sample survey theory and methodology, during the period 1940-1960. This period is regarded as the golden era of the Census Bureau. Hansen and Hurwitz (1943) developed the basic theory of stratified two-stage cluster sampling with one cluster (or primary sampling unit) within each stratum drawn with probability proportional to size (PPS) and then subsampled at a rate to ensure a self-weighting sample (equal overall probabilities of selection). Unequal probability selection of clusters can lead to significant variance reduction by controlling the variability arising from unequal cluster sizes. Another major contribution from the U.S. Census Bureau is the use of rotation sampling with partial replacement of households to handle response burden in surveys repeated over time, such as the monthly U.S. Current Population Survey for measuring unemployment rates. Hansen, Hurwitz, Nisselson and Steinberg (1955) developed simple but efficient composite estimators under rotation sampling. Rotation sampling and composite estimation are widely used in large-scale continuing surveys.

Prior to the 1950s, the primary focus was on estimating population totals and means. Woodruff (1952) of the U.S. Census Bureau developed a unified approach for constructing confidence intervals for quantiles (in particular, the median), applicable to general sampling designs. The procedure remains a cornerstone for quantile estimation (Francisco and Fuller, 1991).

After the consolidation of the basic design-based sampling theory, Hansen, Hurwitz, Marks and Mauldin (1951) and others paid attention to measurement or response errors in survey data. Under additive measurement error models with minimal model assumptions on the observed responses treated as random variables, total variance of an estimator can be decomposed into sampling variance, simple response variance and correlated response variance (CRV) due to interviewers.

Mahalanobis (1946) had developed the method of interpenetrating subsamples for assessing both sampling and interviewer errors. By assigning the subsamples at random to interviewers, both the total variance and the interviewer component can be estimated. The interviewer component can dominate total variance when the number of interviewers is small. To remove the CRV component due to interviewers, self-enumeration by mail was introduced in the 1960 U.S. Census.

Nonresponse in surveys was also addressed in early survey sampling development. Hansen and Hurwitz (1946) proposed two-phase sampling in which the sample is contacted by mail in the first phase and a subsample of nonrespondents is then subjected to personal interview, assuming complete response or negligible nonresponse at the second phase. This method was used recently in Canada when the compulsory long form sample census was replaced by a voluntary National Household Survey. After the change of Government in 2015, the Prime Minister of Canada reinstated the long form census. Two phase sampling is retained but to a lesser extent. The Hansen-Hurwitz two phase sampling method has also been used in other surveys including the American Community Survey.

Attention was also given to inferences for unplanned subpopulations (called domains) such as age-sex groups within a state. Hartley (1959) and Durbin (1958) developed a unified theory for domain estimation applicable to general designs and yet requiring only existing formulae for population totals and means.

Most of the survey sampling theory in the early period was developed by official statisticians while academic researchers, especially in USA, paid little attention to survey sampling. An exception was Iowa State University, where faculty played a leading role from the early stages under the leadership of Cochran, Jessen and Hartley. Another institution making early contribution to survey practice and research is the Survey Research Center at the University of Michigan established in 1947, with Leslie Kish as one of its first members.

In the 1950s formal theoretical frameworks for design-based inference on totals and means were proposed by regarding the sample data as a set of sample labels together with the associated variables of interest. Horvitz and Thompson (1952) derived the well-known estimator with weight inversely proportional to the inclusion probability. Narain (1951) also proposed this estimator. Godambe (1955) developed a general class of linear estimators by letting the sample weight of a unit depend on the label as well as on the labels of the other units in the sample. He then showed that the best linear unbiased estimator does not exist in this general class even under simple random sampling.

3 Inferential issues: 1950 -

3.1 Theoretical foundations

Attempts were made to integrate sample survey theory with mainstream statistical inference via the likelihood function. Godambe (1966) showed that the likelihood function from the full sample data including labels, regarding the vector of unknown population values as the parameter, provides no information on the non-sampled values and hence on the population total or mean. This uninformative feature of the likelihood function is due to the inclusion of labels in the data which makes the sample unique. An alternative design-based route ignores some aspects of the sample data to make the sample non-unique and thus arrive at informative likelihood functions (Hartley and Rao, 1968; Royall, 1968). This non-parametric likelihood approach is similar to the currently popular empirical likelihood (EL) approach in mainstream statistical inference (Owen, 1988). The EL approach has been applied to sampling problems in recent years to estimate not only totals and means but also more complex parameters. So the integration efforts with main stream statistics was partially successful.

The model-dependent approach provides an alternative route to inference from survey data. The approach requires that the population structure obeys a specified super-population model. The distribution induced by the assumed model provides the basis for inferences (Brewer, 1963 and Royall, 1970). Such conditional (conditional on the sample) inferences can be appealing. However, the resulting estimators may be design inconsistent and, as such, they can perform poorly in large samples under model misspecification (Hansen, Madow and Tepping, 1983).

A hybrid approach, called the model-assisted approach, attempts to combine the desirable features of the design-based and model-dependent methods, see Cassel, Särndal and Wretman (1976). The approach typically includes the use of data external to the collected data, called auxiliary data. Procedures using auxiliary data include regression estimation, ratio estimation, and raking, methods with estimators linear in the variable of interest. Estimators using auxiliary information, particularly regression, were recognized very early as powerful estimators (Cochran, 1953). Computing power made regression estimation practical in the 1970's, but to be acceptable in large scale surveys the regression weights need to be nonnegative. An early definition of nonnegative weights is Huang and Fuller (1978). Deville and Särndal (1992) gave a general method of constructing weights for design consistent estimators. Model assisted methods entertain only design consistent estimators of the total that are also model unbiased under a working model. This approach is useful for large samples and it leads to valid design-based inferences in large samples, regardless of the validity of the working model. However, efficiency of the estimators does depend on the degree to which the working model approximates the true population structure. The most popular form of model-assisted estimators are known as generalized regression estimators (GREGs) and are implemented in survey software packages.

Theoretical results for probability-based sampling emphasize the first two moments of the sample statistics. Central limit theorems have been used to provide justification for normality-based confidence intervals. An early central limit theorem for simple random samples is that of Madow (1948). Hájek (1960) gave a central limit theorem for simple random sampling and a theorem for rejective sampling in Hájek (1964). Bickel and Freedman (1984) gave a central limit theorem for stratified random sampling. Recent literature considers both sequences of fixed finite populations and sequences of finite populations that are samples from a superpopulation (Fuller, 2009b; Section 1.3.2).

Variance estimation was very costly, nearly prohibitive, in the 1930's and 1940's, and remains expensive today. Replication was adopted as an efficient variance estimation method from the beginning. As we noted, an early replication form was introduced by Mahalanobis (1939, 1946) called "interpenetrating" samples by him, and called "random groups" by later authors. The method of random groups based on half samples, was used by the U.S. Census Bureau in the 1950's and 1960's. McCarthy (1966, 1969) developed and described balanced half-sample variance estimation. Also see Kish and Frankel (1970). Wolter (2007) contains an extensive discussion on balanced half samples. Also see Dippo, Fay and Morgenstein (1984), Kish and Frankel (1974), Krewski and Rao (1981), and Rao and Shao (1999). The jackknife and bootstrap are the current versions of early replication procedures. Wolter (2007, Chapter 4) credits Durbin (1959) with the first use of the jackknife in finite population estimation. The use of the bootstrap in the classical setting dates from Efron (1979) but application to unequal probability samples and finite populations is not immediate. Among the large number of papers on jackknife and bootstrap for survey samples are McCarthy

and Snowden (1985), an early with-replacement sampling version, and Rao and Wu (1988), a modified bootstrap based on "rescaling" for survey samples. Sitter (1992) discussed several topics including suggestions for obtaining integer sample sizes. Antal and Tillé (2011) gave bootstrap methods appropriate for a wide range of designs, including Poisson sampling. Beaumont and Patak (2012) gave general bootstrap procedures.

3.2 Analytic use of survey data

As we have remarked, the early work on probability sampling emphasized totals and means and many estimation procedures were developed for official statistics. However, from the beginning, survey samples were used by social scientists to answer subject matter questions with relevance beyond the finite population sampled. Deming and Stephan (1940) and Deming (1953) gave explicit consideration to the difference between "enumerative" and "analytic" use of survey and census data, also see Hartley (1959). The analytic estimates are sometimes called estimates for a superpopulation. Early analysts often treated survey sample data as a simple random sample and constructed estimates on that basis. The potential for bias that arises from ignoring the design led to estimation theory for analytic estimates. One component is comprised of tests for the effect of weights on estimates, see DuMouchel and Duncan (1983), Fuller (1984), and Korn and Graubard (1995). A second component has been the development of design based theory for complicated statistics. See Fuller (1975), Rao and Scott (1981, 1984), and Binder and Roberts (2003). The third approach builds the sampling design into the model (Skinner, 1994 and Pfeffermann and Sverchkov, 1999). A number of computer packages (SAS, SUDAAN, R, STATA) are now available for probability-based statistics and standard errors. Many of the algorithms date from the work at Iowa State University (Hidiroglou, Fuller and Hickman, 1976).

3.3 Missing data

Almost all samples (and experiments) have missing and incorrect data. Missing data in survey sampling are placed in two categories; unit-missing and item-missing, where, as the name implies, a missing unit means that all items in the response record are missing. An indicator of the importance of missing data in survey research is the monograph set edited by Madow, Nisselson and Olkin (1983). One method of handling missing data is to report the nature and number of missing items and tabulate the remaining items. This was common in the early years, but the implied assumption of exchangeability in such a procedure was often not reasonable. An early method of correcting for unit nonresponse was to use a substitute respondent, often interviewing someone "close" to the nonrespondent. A common modification at the analysis stage was, and remains, post stratification. (Deming, 1953; Thomsen, 1973; Kalton, 1983 and Jagers, 1986). In the missing data literature, post strata are often called cells. Regression estimators are direct extensions of cell estimators and are an important method of correcting for missing data (Fuller and An, 1998). Weighting methods for handling unit nonresponse are reviewed in Brick and Montaquila (2009).

Various forms of imputation for item nonresponse have been used over time, with imputation performed by clerks prior to use of computers. An early formal model-based and computer-based imputation was the hot deck imputation procedure used by the U.S. Census Bureau in the 1947 Current Population Survey, see

the description in Andridge and Little (2009). Improved computing power and theoretical advances (Little, 1982; Kalton and Kish, 1984; Rubin, 1974, 1976, 1987; Little and Rubin, 1987; Kim and Fuller, 2004) have made imputation a standard part of estimation for survey samples and an active area of research. Recent books are Kim and Shao (2013) and Little and Rubin (2014).

3.4 Small area estimation

The increased use of models for small-domain estimates is the result of the pairing of two factors. The first is the demand for estimates for small domains (e.g. geographic areas) in policy formulation, fund allocation and regional planning. The second is the large standard errors for many of the design-based domain estimators. Schaible (1996) and Purcell and Kish (1979) gave early examples of small area estimation, also see Gonzalez (1973) and Steinberg (1979). The U.S. Census Bureau used model-based methods for small area estimation as early as 1947 (Hansen et al., 1953; Vol. I, pages 483-486). More recently, linear mixed models involving both fixed and random effects have become important. Early uses of mixed models for small area estimation are Fay and Herriot (1979) and Battese, Harter and Fuller (1988). Some sets of small area estimates can be viewed as a reallocation of the domain estimates, retaining the direct design-consistent estimate of the grand total. Bayesian methods, in particular hierarchical Bayes, are increasingly being used because of the ability to handle complex models; see Rao and Molina (2015, Chapter 10). On the basis of growing demand, there has been a large increase in literature and the field now boasts regular meetings and a book (Rao, 2003) with a recent second edition, (Rao and Molina, 2015).

3.5 Survey practice

Sample design and estimation topics that we have discussed are critical parts of a survey operation, but represent a small fraction of the total. The quality of the final product is determined by frame materials, collection instrument, data collection, editing, processing, and presentation of results. Many error sources are difficult to measure, but those designing surveys make implicit cost estimates when they allocate resources to different parts of the survey operation. Groves and Lyberg (2010) is a review of attempts to enumerate the components of survey quality and to bring them under a single umbrella. They credit Deming (1944) for an early description of error sources in sample surveys and describe the contributions of Dalenius (1974), Anderson, Kasper and Frankel (1979), Groves (1989), Biemer and Lyerg (2003), among others. Groves and Herringa (2006) proposed tools for actively controlling survey errors and costs that can lead to responsive designs for household surveys. In particular, para data (measurements related to the process of collecting survey data) can be used to monitor field work, to make intervention decisions during data collection and to deal with measurement error, nonresponse and coverage errors (Kreuter, 2013).

4 The future

We can project a number of current situations into the future. Budgets will be tight and requests for products will expand. There will be demand for forecasts, and for improved access by users. There will be requests for statistics to be produced more rapidly and, naturally, with no compromise in quality. There will be pressure to bring estimates from different sources into agreement.

We expect faster computing to influence all aspects of the field. More complex edit and imputation algorithms will be developed. The time from collection to publication will be shortened. More complex analyses will be performed on survey data. Record linkage procedures will be improved. Data will be made available in different forms. Searchable databases where the user provides queries will become more common. The use of auxiliary data of all kinds, and in particular administrative data, will increase. Administrative data will be used both as auxiliary data and as the direct estimates for certain items. Citro (2014) gives examples of items where administrative data can be used to replace answers to questions in a questionnaire. Uses of auxiliary data where matching to collected data is imperfect will be a research area.

Modern communication methods and social media have resulted in vast quantities of data, much generated with short term and poorly identified purpose. The term "Big Data" is not well defined, but most would agree that social media data are a part of Big Data. The AAPOR report on Big Data (2015) is an excellent analysis of the potential and the challenges associated with Big Data. Tam and Clarke (2015) and Pfeffermann (2015) discuss the issues from the perspective of a governmental statistical organization. As part of modern society, social media are of interest to social scientists in their own right. Therefore, indexes and summaries of these data are, and will be, produced. An example is the University of Michigan Social Media Job Loss Index. Sampling has a large role to play in the creation of products from these data.

A challenge is transforming some types of Big Data into a form useful as auxiliary data. One example is the Porter, Holan, Wikle and Cressie (2014) use of Google trends of Spanish words as functional covariates to estimate state proportions of people speaking Spanish using American Community Survey estimates as dependent variables in small area models.

One of the often quoted advantages of samples relative to censuses is cost. The cost structure has changed with increased computing power and seems destined to continue to change. In the United States, the National Land Cover Database is a census of land cover (Han, Yang, Di and Mueller, 2012). Classification procedures are expected to improve so that use of such data as auxiliary data will increase. Data collection agencies will invest more in constructing improved auxiliary data files at the population level so that some data now collected on a sample basis will be collected at a population level. The same types of data development will continue for population and business statistics.

Of necessity, our discussion has little on collection. The way in which data collection procedures have been modified with changing technology is perhaps more obvious than the link between technology and theory. For the links to theory see Bellhouse (2000). Computer-assisted data collection is the evolving standard. The use of geo-location technology can be expected to increase. It is safe to forecast the increased use of remote sensing and remote data collection devices. For example, it would be easy to incorporate physical data collected by something like the Apple Watch or Fitbit into a health study. Larger and less attractive monitoring devices are currently in use in physical activity surveys (van Remoortel, Giavedoni, Raste, Burtin, Louvaris, Gimeno-Santos, Langer, Glendenning, Hopkinson, Vogiatzis, Peterson, Wilson, Mann, Rabinovich, Puhan, Troosters and PROactive consortium, 2012).

The recent experience is that phone and personal interview data collection is becoming more and more difficult. Respondents are facing expanded organized data collection activities. The ubiquitous questionnaire on satisfaction for everything from medical services to tooth paste surely must impact an

individual's willingness to respond. It seems reasonable to forecast increased difficulty in obtaining cooperation for traditional methods of data collection. Associated with that trend will be increased study of the nature of non-respondents and of non-response. Likewise efforts will be made to adapt data collection to the changing methods of communication.

Nonprobability samples have been a part of survey activity throughout the post-Neyman period. In particular, quota sampling is commonly used in marketing research and other areas for cost reasons (Sudman, 1966; 1976). Moser and Stuart (1953) and Stephan and McCarthy (1958) made early comparisons between quota sampling and probability sampling. Cochran (1977, page 136) says "The quota method seems likely to produce samples that are biased on characteristics such as income, education and occupation, although it often agrees with the probability samples on questions of opinion and attitude". Use of procedures such as post stratification and regression estimation in nonprobability samples has continued at pace with use in probability samples. The changing nature of human communication offers opportunities for both model-based and probability-based procedures. Because of cost structures, new methods such as web-based procedures will often be used first in nonprobability settings and for nongovernmental purposes.

As matching procedures improve and as demand for detailed data increases, disclosure limitation procedures and associated research will receive increased attention.

Survey sampling is an application discipline, functioning in the current social, geographic, cultural, and technological world. To forecast how our field will be impacted by social and cultural changes, even in the short run, is a challenge. Will the fact that one must assume that almost all of one's public activity and a great deal of one's private activity has potential of being recorded lead to a more relaxed attitude in responding to questions? Will improved monitoring devices make respondents more willing to permit their physical activities be monitored? Or will all of the incidental monitoring lead to a reaction against organized data collection? Will increased availability of results based on collected data have a positive or negative effect on data collection efforts? What is the impact of various Social Media?

This discussion makes clear that factors external to our discipline will determine our future activities. We will be required to adapt in data collection, data processing, and data presentation-dissemination.

Acknowledgements

We thank Graham Kalton for comments and suggestions that led to improvements in the original draft. We thank the four discussants, Graham Kalton, Sharon Lohr, Danny Pfeffermann and Chris Skinner, for their supplements on the history, insightful observations on the present, and comments on the future of survey sampling. We elected not to prepare a rejoinder because we found much to appreciate and little basis for disagreement.

References

AAPOR Big Data Task Force (2015). *AAPOR Report on Big Data*. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/images/BigDataTaskForceReport_FINAL_2_12_15_b.pdf.

Anderson, R., Kasper, J. and Frankel, M. (1979). *Total Survey Error: Applications to Improve Health Surveys*. San Francisco, CA: Jossey-Bass.

Andridge, R.H., and Little, R.J. (2009). The use of sample weights in hot deck imputation. *Journal of Official Statistics,* 25, 21-36.

Antal, E., and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association,* 106, 534-543.

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

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review/Revue Internationale de Statistique,* 80, 127-148.

Bellhouse, D.R. (1988). Systematic sampling. In *Handbook of Statistics,* (Eds., P.R. Kreshnaiah and C.R. Rao), Elsevier, 6, 125-145.

Bellhouse, D.R. (2000). Survey sampling theory over the twentieth century and its relation to computing technology. *Survey Methodology,* 26, 1, 11-20. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2000001/article/5174-eng.pdf.

Bethlehem, J. (2009). The rise of survey sampling. Discussion Paper (09015), Statistics Netherlands, The Hague.

Bickel, P.J., and Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12, 470-482.

Biemer, P.P., and Lyberg, L. (2003). *Introduction to Survey Quality*. New York: John Wiley & Sons, Inc.

Binder, D.A., and Roberts, G.A. (2003). Design-based and model-based methods for estimating model parameters. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner), Wiley, Chichester, UK, 29-48.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute,* 22(1), 1-62.

Bowley, A.L. (1936). The application of sampling to economics and sociological problems. *Journal of the American Statistical Association,* 31, 474-480.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics,* 5, 93-105.

Brewer, K.R.W. (2013). Three controversies in the history of survey sampling. *Survey Methodology,* 39, 2, 249-262. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2013002/article/11883-eng.pdf.

Brick, M.J. (2011). The future of survey sampling. *Public Opinion Quarterly,* 75, 872-888.

Brick, M.J., and Montaquila, J.M. (2009). Non response and weights. In *Handbook of Statistics,* (Eds., D. Pfeffermann and C.R. Rao), Elsevier, Amsterdam, 29A, 163-185.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40, 2, 137-161. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2014002/article/14128-eng.pdf.

Cochran, W.G. (1953). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

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

Dalenius, T. (1974). Ends and Means of Total Survey Design. Report in "Errors in Surveys", Stockholm University.

Deming, E. (1944). On errors in surveys. *American Sociological Review*, 9, 359-369.

Deming, E. (1950). *Some Theory of Sampling*. New York: John Wiley & Sons, Inc.

Deming, W.E. (1953). On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response. *Journal of the American Statistical Association,* 48, 743-772.

Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 4, 427-444.

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

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Dippo, C.S., Fay, R.E. and Morgenstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 489-494.

DuMouchel, W.H., and Duncan, G.J. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association,* 78, 535-543.

Durbin, J. (1958). Sampling theory for estimates based on fewer individuals than the number selected. *Bulletin of the International Statistical Institute,* 36, 113-119.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika*, 46, 477-480.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics,* 7, 1-26.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places. An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 366, 269-277.

Francisco, C.A., and Fuller, W.A. (1991). Estimation of quantiles with survey data. *Annals of Statistics,* 19, 454-469.

Fuller, W.A. (1975). Regression analysis sample survey. *Sankhyā, Series C,* 37, 117-132.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology,* 10, 1, 97-118. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1984001/article/14352-eng.pdf.

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 1-12.

Fuller, W.A. (2009b). *Sampling Statistics*. New York: John Wiley & Sons, Inc.

Fuller, W.A., and An, A.B. (1998). Regression adjustments for nonresponse. *Journal of Indian Society of Agricultural Statistics,* 51, 331-342.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B,* 17, 269-278.

Godambe, V.P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society, Series B,* 28, 310-328.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section of the American Statistical Association,* 33-36.

Groves, R.M. (1989). *Survey Errors and Survey Costs.* New York: John Wiley & Sons, Inc.

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly,* 73, 861-871.

Groves, R.M., and Heeringa, S.G. (2006). Responsive designs for household surveys: Tolls for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A,* 169, 439-457.

Groves, R.M., and Lyberg, L. (2010). Total survey error: Past, present and future. *Public Opinion Quarterly,* 74, 849-879.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy,* 5, 361-374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics,* 35, 1491-1523.

Han, W., Yang, Z., Di, L. and Mueller, R. (2012). CropScape: A Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture,* 84, 111-123.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics,* 14, 333-362.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association,* 41, 517-529.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory,* Vols. I and II, New York: John Wiley & Sons, Inc.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association,* 78, 776-793.

Hansen, M.H., Hurwitz, W.N., Marks, E.S. and Mauldin, W.P. (1951). Response errors in surveys. *Journal of the American Statistical Association,* 46, 147-190.

Hansen, M.H., Hurwitz, W.N., Nisselson, H. and Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association,* 50, 701-719.

Hartley, H.O. (1959). Analytical studies of survey data. In Volume in Honor of Corrado Gini, Instituto di Statistica, Rome, 1-32.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Hidiroglou, M.A., Fuller, W.A. and Hickman, R.D. (1976). *SUPER CARP* Statistical Laboratory, Survey Section, Iowa State University, Ames, IA.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association,* 47, 663-685.

Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 300-305.

Jagers, P. (1986). Post-stratification against bias in sampling. *International Statistical Review/Revue Internationale de Statistique*, 54, 159-167.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Survey Research Center, University of Michigan, Ann Arbor, Michigan.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics,* A13, 1919-1939.

Kiaer, A. (1897). The representative method of statistical surveys (1976 English translation of the original Norwegian), Oslo, Central Bureau of Statistics of Norway.

Kim, J.K., and Fuller, W.A. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

Kim, J.K., and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press, Boca Raton, FL.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2, 813-830.

Kish, L., and Frankel, M.R. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association,* 65, 1071-1094.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Korn, E.L., and Graubard, B.I. (1995). Analysis of large health surveys: Accounting for the sampling designs. *Journal of the Royal Statistical Society, Series A*, 158, 263-295.

Kreuter, F. (2013). *Improving Surveys with Paradata.* Hoboken: Wiley.

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

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A., and Rubin, D.B. (1987, 2014). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc., (Second Edition 2014).

Madow, W.G. (1948). On the limiting distribution of estimates based on samples from finite universes. *Annals of Mathematical Statistics*, 19, 535-545.

Madow, W.G., and Madow, L.H. (1944). On the theory of systematic sampling, I. *The Annals of Mathematical Statistics*, 15, 1-24.

Madow, W.G., Nisselson, H. and Olkin, I. (Eds.) (1983). *Incomplete Data in Sample Surveys*, 1, 2, 3, New York: Academic Press.

Mahalanobis, P.C. (1939). A sample survey of the acreage under jute in Bengal. *Sankhyā*, 4, 511-531.

Mahalanobis, P.C. (1944). On large-scale sample surveys. *Philosophical Transactions of the Royal Society of London, Series B*, 231, 329-451.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 325-378.

McCarthy, P.J. (1966). Replication: An approach to the analysis of data from complex surveys. *Vital and Health Statisitcs*, Series 2, No. 14, National Center for Health Statistics, Public Health Service, Washington DC.

McCarthy, P.J. (1969). Pseudoreplication: Further evaluation and application of the balanced half-sample technique. *Vital and Health Statistics*, Series 2, No. 31, National Center for Health Statistics, Public Health Service, Washington, DC.

McCarthy, P.J. (1969). Pseudoreplication: Half-samples. *International Statistical Review/Revue Internationale de Statistique*, 37, 239-264.

McCarthy, P.J., and Snowden, L.B. (1985). The bootstrap and finite population sampling. *Vital Health Statistics*, 2-95, *Public Health Service Publication*, 85-1369, U.S. Government Printing Office, Washington, D.C.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3, 169-174.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Owen, A.B. (1988). Empirical likelihood ration confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *Journal of Survey Statistics and Methodology*, 3, 425-483.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā, Series B,* 61, 166-186.

Porter, A.T., Holan, S.H., Wikle, C.K. and Cressie, N. (2014). Spatial Fay-Herriot model for small area estimation with functional covariates. *Spatial Statistics,* 10, 27-42.

Purcell, N., and Kish, L. (1979). Estimation for small domains. *Biometrics,* 35, 365-384.

Rao, J.N.K. (2003). *Small Area Estimation.* Hoboken: Wiley.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation: Second Edition.* Hoboken: Wiley.

Rao, J.N.K., and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association,* 76, 221-230.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics,* 12, 46-60.

Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika,* 86, 403-415.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association,* 83, 231-241.

Royall, R.M. (1968). An old approach to finite population sampling. *Journal of the American Statistical Association,* 63, 1269-1279.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* 57, 377-387.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association,* 69, 467-474.

Rubin, D.B. (1976). Inference and missing data. *Biometrika,* 63, 581-590.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

Schaible, W.L. (Ed.) (1996). *Indirect Estimators in U.S. Federal Programs.* New York: Springer.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association,* 87, 755-765.

Skinner, C.J. (1994). Sample models and weights. In *Proceedings of the Survey Research Methods Section,* American Statistical Association, 133-142.

Steinberg, J. (Ed.) (1979). *Synthetic Estimates for Small Areas: Statistical Workshop Papers and Discussion.* NIDA Research Monograph No. 24. U.S. Government Printing Office, Washington, D.C., U.S.A.

Stephan, F., and McCarthy, P.J. (1958). *Sampling Opinions*. New York: John Wiley & Sons, Inc.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, 749-791.

Sudman, S. (1976). *Applied Sampling.* New York: Academic Press.

Sukhatme, P.V. (1954). *Sampling Theory of Surveys with Applications*. Iowa State College Press, Ames.

Tam, S.-M., and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review/Revue Internationale de Statistique*, 83, 436-448.

Tchuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2, 461-493, 646-683.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift*, 11, 278-283.

van Remoortel, H., Giavedoni, S., Raste, Y., Burtin, C., Louvaris, Z., Gimeno-Santos, E., Langer, D., Glendenning, A., Hopkinson, N.S., Vogiatzis, I., Peterson, B.T., Wilson, F., Mann, B., Rabinovich, R., Puhan, M.A., Troosters, T. and PROactive consortium (2012). Validity of activity monitors in health and chronic disease: A systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 9.

Wolter, K.M. (2007). *Introduction to Variance Estimation.* New York: Springer-Verlag.

Woodruff, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47, 635-636.

Yates, F. (1948). Systematic sampling. *Philosophocal Transaction of the Royal Society of London, Series A*, A241, 345-377.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. Griffin, London.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 161-165
Statistics Canada, Catalogue No. 12-001-X

Comments on the Rao and Fuller (2017) paper

Danny Pfeffermann[1]

Abstract

This note by Danny Pfeffermann presents a discussion of the paper "Sample survey theory and methods: Past, present, and future directions" where J.N.K. Rao and Wayne A. Fuller share their views regarding the developments in sample survey theory and methods covering the past 100 years.

Key Words: Data collection; History of survey sampling; Probability sampling; Survey inference.

I happily take the guilt of inviting J.N.K. Rao and Wayne Fuller to present a paper at a special session sponsored by the International Association of Survey Statisticians (IASS) during the International Statistical Institute (ISI) meeting in Rio de Janeiro in 2015. Credit goes however to Professor Vijay Nair, the ISI president at the time, who initiated this kind of invited session for all the sections of the ISI. And so, as soon as I became aware of the possibility to organize this kind of session, I immediately thought of Rao and Fuller, the two uncrowned kings of modern sample survey theory and methods as my natural candidates to present a paper on the past, present and future directions of sample survey theory and methods, and fortunately to all of us, they immediately agreed. Quite honestly, I didn't expect such an immediate agreement and while inviting them, I already prepared a list of arguments to convince them to agree, but as we all know, kings have responsibility for their natives. What I didn't know at the time is that two years later I shall be invited to discuss a paper based on their presentation, an invitation which I gratefully agreed to.

In the discussion that follows I shall concentrate mostly on the third part of the paper, the future of sample surveys, a topic that occupies me more and more, since changing my career four years ago to become the National Statistician and Director of the Central Bureau of Statistics in Israel (ICBS). Naturally, my discussion is based on my experience in Israel, but I have good reasons to believe that it represents to a great part what is occurring in other countries as well.

In the section titled "The Future", Rao and Fuller (Hereafter, RF, like my other king, Roger Federer) provide a long list of items, which they predict will dominate the future of sample surveys. I totally agree with that list with one reservation. Many of the items in the list already dominate present sample surveys. Budgets are already tight and requests for products expand constantly, not only within countries but also by international organizations like the United Nations, the OECD, Eurostat, the IMF and the World Bank. The statistics requested by these organizations are often similar and sometimes even overlap, but are required in different forms and at different times, covering different time periods, thus adding extra burden to the work of National Statistical Offices (NSO). The use of administrative data already occupies the work of NSOs all over the world. In Scandinavian countries the population censuses are based solely on administrative data and many other European countries and Israel as well, invest a lot of resources in establishing reliable

1. Danny Pfeffermann, National Statistician and Director of Central Bureau of Statistics, Israel, Hebrew University of Jerusalem, Israel; Southampton Statistical Sciences Research Institute, UK. E-mail: MsDanny@cbs.gov.il.

administrative databases that will replace their present censuses in the next generation of censuses, to be carried out around 2030. The present censuses in these countries already heavily use administrative data, but still require samples to correct for under- and over coverage. There is an important legal issue about the use of administrative data, having to do with getting access to them. Public and private organizations simply refuse to transfer the data. In Israel, the National Statistician is authorized by law to obtain any set of data which he or she thinks is important for the production of official statistics, but some organizations maintain that they are committed to their customers not to transfer any private data. I know that other countries face a similar conflict. Will we be able to foster a culture of data sharing in the future? Looks like a formidable challenge right now.

As mentioned in the paper, phone and personal interview data collection is becoming more and more difficult, resulting in increased rates of nonresponse. The situation is even worse in business surveys because very often the same (mostly big) establishments are requested to participate annually in many surveys (sometimes seven or more). This is also our experience in Israel although unlike in other countries, we are still able to maintain reasonable response rates (around 70%), because most of our surveys are mandatory. There are several directions to deal with this problem. The first is to develop new sampling procedures to ease the response burden. Several procedures based on the use of permanent random numbers are already in use, but more sophisticated methods have to be established, although the big or unique establishments and firms will hardly if ever benefit from them. This calls for better imputation methods that use past data and data observed for other sampled units, although imputation for the relatively large units could be practically impossible. An ideal way to ease the response burden would be to get the micro data as-is from the businesses along with the relevant metadata, and then process them at the NSO. This of course will require cooperation of the businesses and more data science competence at their offices than what is currently the case. Is there a realistic chance for such cooperation? At my age I am allowed to be sceptic but thinking of it, with proper data protection arrangements, why not?

In Section 3.3, RF discuss several imputation methods, listing key references. As far as I can tell, a common assumption underlying all these methods is the availability of external or internal data (past and other data observed for the same sampled units), that fully explain the missing data. However, from my experience at the ICBS, this is often not the case and in many surveys the nonresponse is what is known as not missing at random (NMAR nonresponse). Handling NMAR nonresponse is a very difficult problem for the simple reason that the target variables of interest are not observed for the nonrespondents, requiring making assumptions about the nonresponse mechanism in the form of statistical models, with limited possibilities to test them. This discussion is not supposed to highlight my own research with my colleagues, but I do like to mention an approach proposed in Pfeffermann and Sikov (2011), Feder and Pfeffermann (2015) and Pfeffermann (2017) for handling NMAR nonresponse, which allows testing the response model, at least partly. The idea behind this approach is to assume a model for the population data (parametric or nonparametric), and a parametric model for the response mechanism, and then test if the implied model holds for the observed data, using conventional model testing procedures. See also Riddles, Kim and Im (2016) for an important identifiability condition for the model holding for the observed data, and Sverchkov

and Pfeffermann (2017) for application of this condition in the context of small area estimation under NMAR nonresponse.

Two other directions to deal with nonresponse that will require further extensive research in the coming years is the use of adaptive survey designs (ASD) and the allowance for alternative data collection modes. The general idea underlying ASD is to use auxiliary information available from registry data and/or interviewer observations, in order to tailor the survey design so as to optimize response rates and consequently reduce nonresponse bias. See the recent book by Schouten, Peytchev and Wagner (2017) and the numerous references therein for details and illustrations. RF discuss briefly data collection methods and their effects on inference. We are all familiar with the traditional and more modern modes of data collection; personal interview, phone (cell phone) interview, mail (email) and nowadays by the internet, the cheapest mode of data collection and hence the preferred one, although it still requires sending out a web address and password to the sampled units. In order to increase response rates, survey organizations tend to assign different response modes to different sampled units or more generally, let the sampled units choose their preferred mode. Such surveys are called mixed-mode surveys. Another variant of mixed-mode surveys and the one often applied in practice is where the different modes of response are offered sequentially to those who do not respond with a previous mode. However, an inferential problem with these procedures is the possibility of mode effects, resulting from a selection effect (the effect of differences between the characteristics of respondents preferring to respond with different modes and consequently, possible differences in the values of study variables), and a measurement effect (the effect of responding differently by the same sample member, depending on the mode of response). In Pfeffermann (2015) I review briefly available methods to deal with this problem and propose another approach, but much more research is required on this topic. Notice that a mode (measurement) effect may exist even if only a single mode of response is offered.

One other aspect of data collection that is missing in the paper but is very common in practice, for example, in labour-force surveys, health surveys and even in modern censuses, is the use of proxy surveys, whereby one member of the household provides information about all the members of the household. There is a possible ethical problem with this kind of survey in that the not interviewed household members are not asked for their consent that information about them is supplied by the member who is interviewed. From a statistical point of view, the use of proxy surveys potentially increases the possibility for (correlated) measurement errors. Does the household member interviewed know the health status of all other household members or whether they were seeking work in the week prior to the interview? And if he or she is wrong about one member of the household, will they not be wrong about other members? Has this problem been researched systematically in the literature and solutions have been proposed? To make things even worse, what about the interplay between the measurement errors and nonresponse?

Last, but probably the main issue when discussing the future of sample surveys is the possible use of big data as an alternative to the use of traditional surveys based on probability samples. In recent years, I found myself making many presentations on this issue and my main thoughts are already summarised in Pfeffermann (2015). In what follows I shall repeat some of them, wearing the hat of a National Statistician, concerned about the production of official statistics.

RF state correctly that the term big data is not well defined but mention social media data as an example of such data. I totally agree that this is a good example, which, however, also illustrates the problems with handling this kind of data. It is generally diverse, unstructured, appears irregularly and may even cease to exist. (Notwithstanding, not all types of big data are like this and other types could be considered as just big versions of what is usually referred to as administrative data). RF add that data from social networks are of interest to social scientists. Again, I agree, but should NSO's publish estimates of social indexes obtained from social networks? Maybe yes, but which population will they represent? RF also make the point that big data should serve as auxiliary data. They don't go into detail but I presume that they think of using them as covariates in model assisted or model dependent inference, similarly to the use of what is known as administrative data. This is obvious and many examples have been published in the literature, although it is not as straightforward as it may seem because of all the computational hurdles that will need to be overcome before the data is ready for use, including record linkage, if big and survey micro data are to be matched. Clearly, proper models need to be fitted and tested.

Where I see the real challenge, however, is in the possible use of big data as substitute for traditional sample surveys. There is no question that it is much more efficient and cheaper to get sale prices (and quantities) electronically, directly from stores, for computing the CPI, instead of sending surveyors to collect prices. But price indexes are often sought for sub-groups of the population as defined by age, origin, etc., and the big data obtained electronically does not generally include demographic information. Will we be able to use credit card information to link buyers to purchases? Will credit card companies provide the required information? How will this happen? And what about coverage problems of the big data? Do opinions expressed in social networks represent the opinions held by the general public? Can job advertisements on the internet replace business surveys inquiring about vacant jobs? At a conference to celebrate Jon Rao's 80th birthday earlier this year, Professor Jae Kim considered three possible procedures to correct for possible coverage bias of big data. At the ISI meeting in Marrakesh I proposed a fourth procedure. All four procedures seem reasonable but clearly, much more theoretical and applied research is needed before any of the procedures can be recommended for actual use.

One of the procedures proposed by Kim requires linking the big data with an appropriate sample, so as to estimate the probability of being included in the big data. This forms a very neat example of combining big data with survey data. Di Zio, Zhang and De Waal (2017) discuss another use of traditional sampling, namely, to "assist" model building and validation. An important question in this respect is how to incorporate sampling errors with the model errors in subsequent inference. Should one evaluate the big data estimator only from a model-based point of view, when the model building makes use of sample data which is subjected to sampling errors? Warning: when combining big data with survey data, one should check carefully the definition of variables which appear in both data sets even if they seem to measure the same phenomenon. Unemployment in a big data set may be defined very differently from the International Labour Organization (ILO) definition adopted in labour force surveys. Another aspect in the possible combination of big data with survey data and even more so in the sole use of big data, is the development of new sampling algorithms to be applied to the big data. Sampling of big data reduces storage space, it helps in protecting privacy and disclosure, and it produces manageable data sets on which algorithms can run to fit models and

produce estimates. But random sampling from big, versatile dynamic data is obviously different from sampling finite populations, requiring new sampling algorithms. Will finite network sampling of big data replace traditional finite population sampling?

To conclude this brief discussion, my own opinion is that traditional sample surveys will continue to be vital in the foreseeable future. However, quoting from Marker (2017), "the existence of big data has changed the expectation of timeliness and NSOs will need to figure out how to carry out surveys and censuses quicker, or users will rely on available big data without understanding what they are losing."

References

Di Zio, M., Zhang, L.C. and De Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, 17-26.

Feder, M., and Pfeffermann, D. (2015). Statistical inference under non-ignorable sampling and nonresponse-an empirical likelihood approach. Southampton Statistical Sciences Research Institute, University of Southampton, UK. http://eprints.soton.ac.uk/id/eprint/378245.

Marker, D. (2017). How have National Statistical Institutes improved quality in the last 25 years? *Statistical Journal of the IAOS*, 1, 1-11.

Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics. *The Journal of Survey Statistics and Methodology* (*JSSAM*), 3, 425-483.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples. A unified approach. *Calcutta Statistical Association (CSA) Bulletin*, 69, 35-63.

Pfeffermann, D., and Sikov, A. (2011). Estimation and imputation under non-ignorable non-response with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Riddles, K.M., Kim, J.K. and Im, J. (2016). A propensity-scoreadjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*. Chapman&Hall/CRC Statistics in the Social and Behavioral Sciences.

Sverchkov, M., and Pfeffermann, D. (2017). Small area estimation under informative sampling and not missing at random nonresponse. (Under Review).

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 167-172
Statistics Canada, Catalogue No. 12-001-X

Comments on the Rao and Fuller (2017) paper

Graham Kalton[1]

Abstract

This note by Graham Kalton presents a discussion of the paper "Sample survey theory and methods: Past, present, and future directions" where J.N.K. Rao and Wayne A. Fuller share their views regarding the developments in sample survey theory and methods covering the past 100 years.

Key Words: Data collection; History of survey sampling; Probability sampling; Survey inference.

Jon Rao and Wayne Fuller's brief paper is wide-ranging in its coverage. Reading it stimulated me to reflect on my experience of the history of survey research over the past half-century or more, mostly working as an applied survey researcher on social surveys in the United Kingdom and the United States. Overall, amazing advances have taken place in all aspects of survey research during my working life, including in both survey sampling and data collection methods. Jon and Wayne have, of course, been major contributors to those advances. For my discussion, I have chosen two broad areas of the changes that have taken place.

Changing role of models in survey sampling inference

At the outset of my career, Neyman's design-based mode of inference was dominant, but its dominance has been decreasing over time, particularly more recently. The attraction of design-based inference is that the consistency of estimators of population parameters based on a probability sample from a finite population is not dependent on models, unlike model-dependent estimation where the inference depends on the validity of the model assumptions. From the early days, probability sampling and design-based inference were model-assisted, for instance with sample allocation in stratified sampling and regression estimation. However, while misspecification of these working models affects the precision of the survey estimators, the estimators' consistency remains intact.

The conditions needed for pure design-based inference are that every unit in the target population has a known non-zero selection probability (or at least a known *relative* selection probability) and that valid survey data are obtained for every sampled unit. In social surveys these conditions are almost never fully met in practice because of inevitable missing data arising from noncoverage and nonresponse (both unit and item nonresponse). Models are essential for addressing missing data, whether that be through weighting and imputation methods or by ignoring the problem–implicitly employing a model that treats the missing data as missing completely at random (MCAR).

As I recollect the situation in the early days of my career, there was little recognition of the use of models in producing survey estimates and, indeed, there was a strong opposition to a dependence on models in survey inference. Many researchers fiercely resisted imputation when it began to become more common around the 1980s on the grounds that it involves "fabricating data"; instead, analysts would employ complete

1. Graham Kalton is a Senior Vice President at Westat, 1600 Research Blvd., Rockville, MD, 20850, U.S.A. E-mail: grahamkalton@westat.com.

case analysis, implicitly making the MCAR assumption. In those early days, nonresponse rates were low, so that the dependence on models assumptions was not so significant; when simple nonresponse weighting adjustments were employed, they received little attention. With the growing nonresponse rates in recent years, the situation has changed markedly: survey estimates are now highly model-dependent. As a result, there has been a considerable amount of research on methods for modeling missing data, as noted by Rao and Fuller.

Another way that models arise in survey practice is by the use of nonprobability sampling methods or sampling methods that do not strictly adhere to the requirement of known selection probabilities. Cost considerations play an important role in sample implementation, where they may lead to a choice of a sample design with unknown selection probabilities that are then approximated based on model assumptions. One well-known example is quota sampling, a nonprobability sampling method that has been widely used in market research studies, and that was used in a number of early social research studies (see Stephan and McCarthy, 1958). Sudman (1966) describes a quota allocation scheme for interviewers that employed quota controls to create cells within which persons were assumed to have the same selection probabilities. Within a given area, the interviewer is then free to select anyone subject to the condition that the resultant sample satisfies the quota controls. Another example is random route (random walk) sampling that avoids the cost of listing the dwellings in a sampled area; with this method, the interviewer is instructed to start at a specified location and to follow a given route. Bauer (2016) describes how this method fails to provide the equal probability sample that is assumed. Listing costs are also avoided with the World Health Organization's Expanded Programme on Immunization (EPI) rapid assessment surveys that are designed to estimate the immunization coverage of children in a given area. Within a sampled cluster (e.g., a village), the interviewer starts with a "randomly selected" household and then proceeds to the next closest household and so on in seriatim until the specified sample size is reached (often seven eligible children). As well as making the assumption that the children are sampled at random within the sampled cluster, the method also makes the flawed assumption that the clusters are sampled with probabilities exactly proportional to the number of eligible children in the cluster at the time of the survey. Bennett (1993) and others have suggested modifications to avoid the biases occurring from the assumed model of equal selection probabilities with this very widely used EPI sampling methodology.

In recent years, there has been a major growth in the demand for surveys to study rare populations, some of which are defined in terms of sensitive characteristics (including illegal behaviors). See, for example, Tourangeau, Edwards, Johnson, Wolter and Bates (2014). Nonprobability sampling methods are needed in situations where probability sampling is deemed infeasible. However, they lack the security of design-based inference. Widely used nonprobability sample designs for surveying difficult-to-sample rare populations include snowball sampling, respondent-driven sampling, location (venue-based) sampling, and web surveys.

Web surveys have the attractions of obtaining survey responses inexpensively and almost instantaneously. Web surveys come in many different forms, including self-selected web surveys, volunteer panels of internet users, and internet panels based on probability samples (Couper, 2000). The sample sizes of self-selected web surveys and volunteer panels are often very large, but the key concern is the potential biases in the survey estimates. As the infamous 1936 *Literacy Digest* poll indicates, large samples are not a

protection against bias in the survey estimates. That poll was a mail survey sent to around 10 million individuals selected mainly from telephone directories and car registration lists, and about 2 million responded. It predicted a landslide victory for Alf Landon in the 1936 U.S. Presidential Election whereas Franklin Roosevelt won by a large margin (see Converse, 1987, pages 456-457 for references that attempted to explain the failure of this poll). What remains to be seen is whether modern methods of weighting adjustments applied to large-scale nonprobability web data collections can overcome the *Literary Digest* poll problems and, more critically, under what conditions one can safely rely on the quality of the model-dependent estimates produced. Even when a web panel is recruited using a probability sample design, the security of design-based inference is severely challenged by the generally extremely low overall response rate.

After years of opposition, model-dependent small area estimation methods are nowadays widely accepted, as noted by Rao and Fuller who have both made major contributions to the literature on this topic. This acceptance has come about because the great demand for small area estimates by policy makers and others cannot be met by design-based methods with affordable sample sizes. Although small area estimation starts from data collected in a probability sample, it then "borrows strength" from models that make use of administrative data, past censuses, and other data available at the small area level. The small area models are carefully constructed and evaluated to the extent possible, but nevertheless, the resulting small area estimates are model-dependent.

In sum, complete reliance on design-based inference is not realistic these days for a variety of reasons. Greater attention should be given for ways to communicate the uncertainty about the estimates produced from hybrid data containing both design-based and model-dependent components, taking into account plausible levels of model misspecification.

Developments in computing capability in the past decades

Developments in computing capability in the past decades have had a major influence on all aspects of survey research. When I started my career, survey analysis was carried out with punch cards on counter-sorters and other such equipment. Tabulation was virtually the only form of analysis. Standard errors that reflected complex sample designs were seldom computed; instead, simple rules of thumb were applied to modify the simple random sampling standard errors. In his text *Sample Design in Business Research*, Deming (1960) advocated that samples be designed in 10 replicates to facilitate variance estimation, and furthermore, he proposed that the standard error of an estimate be obtained by the simple calculation of dividing the difference between the largest and smallest replicate value by 10. In *Survey Sampling*, Kish (1965) emphasized the simplicity of variance calculations based on a paired sample design in which two primary sampling units are selected in each stratum, and he laid out the way to perform these calculations by hand. Now variance estimates for simple and complex statistics based on complex sample designs are readily computed in one of a number of software packages using such techniques as balanced repeated replication, jackknife repeated replication, the bootstrap, and the linearization approach. Moreover, with the replication methods, recomputing even complex weighting adjustments for each replicate is

straightforward, thus enabling the variance estimates to incorporate the variability associated with these adjustments.

The impact of computers on survey statistics is not restricted to variance estimation. They also enable more complex designs to be applied, and they have led to a great increase in complex methods of analysis (as discussed by Rao and Fuller). Consider the case of deep stratification as one example of a more complex design. Goodman and Kish (1950) describe a method of deep stratification known as controlled selection that could be carried out by simple computations. The more recent balanced sampling method of cube sampling and the related method of rejective sampling referenced by Rao and Fuller are far more complex to apply.

Computers have also had major effects on other aspects of the survey process. Fifty years ago, survey data were collected by means of paper-and-pencil interviews (PAPI) or by means of mail questionnaires. The PAPI method has largely been replaced by computer-assisted data collection (Couper, Baker, Bethlehem, Clark, Martin, Nicholls and O'Reilly, 1998). Computer-assisted personal interviews (CAPI) are conducted using laptops or, now more commonly, tablet computers. Some of the data–particularly sensitive data–may be collected by audio computer-assisted self-interviews (audio-CASI). With a CAPI data collection, all or parts of an interview may be recorded (computer-assisted recorded interviews–CARI); CARI can be useful for pretesting and for checking on interviewer performance throughout the data collection period. Computers can also collect the GPS locations of interviews, thus providing a check on interviewer fabrication and providing data for a variety of location-based analyses. More recently, web data collection has emerged as an attractive cost-efficient mode of data collection, but in an era when response rates are falling, it often has to be supplemented by face-to-face interviews or some other method. Mixed-mode surveys are becoming increasingly popular, and their use seems likely to increase in the future (with due attention to possible response differences for some questions by mode). See Dillman (2017) for a review of issues associated with pushing respondents initially to the web in mixed-mode surveys.

Conclusion

The history of survey research is one of rapidly increasing and more complex demands for survey data. This demand has led to the release of public-use files (PUFs) for individual analyses, and to concerns about protecting the respondents' confidentiality. Similar concerns arise with the release of many tabular and other analyses. These concerns are being addressed by the ongoing developments of methods for statistical disclosure control for PUFs (e.g., see Hundepool, Domingo-Ferrer, Franconi, Giessing, Schulte Nordholt, Spicer and de Wolf, 2012), to the provision of restricted-use files and to the establishment of statistical enclaves where analysts can go to perform analyses under supervision. In addition, survey data archives have emerged as places to store and administer survey datasets.

The trend of rapidly increasing demand for survey data in the past few decades is likely to continue, making the future for survey research seems rosy. However, as Paul Valery remarked, "The trouble with our times is that the future is not what it used to be", a remark that seems particularly apposite for survey research at this time. Some see Big Data and administrative records as serious competitors to surveys, but I

am not so convinced. Both may serve some needs, but the multivariate nature of surveys and often the need to collect some items that can be obtained only from respondents (e.g., opinions, level of adult literacy, household expenditures, diabetes) mean that surveys will continue to have a major role to play. Administrative data may produce estimates for some official statistics (especially economic statistics), particularly with the merging of sets of administrative files where permitted. However, I see a main use of administrative records in official social surveys as a supplement that may reduce burden by replacing the survey items with record data, and that can provide longitudinal data for both the time before and the time after the survey data collection. Of course, the quality of the record data and the record linkages needs to be assessed. As I see it, the biggest threat to survey research lies in the decreasing willingness of the public to answer surveys. To date, no good solutions have been found for this threat.

References

Bauer, J.J. (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4, 263-287.

Bennett, S. (1993). Cluster sampling to assess immunization: A critical appraisal. *Bulletin of the International Statistical Institute*, 49th Session, 55(2), 21-35.

Converse, J.M. (1987). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.

Couper, M.P. (2000). Web surveys. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. and O'Reilly, J.M. (Eds.) (1998). *Computer Assisted Survey Information Collection.* New York: John Wiley & Sons, Inc.

Deming, W.E. (1960). *Sample Design in Business Research*. New York: John Wiley & Sons, Inc.

Dillman, D.A. (2017). The promise and challenge of pushing respondents to the Web in mixed-mode surveys. *Survey Methodology*, 43, 1, 3-30. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14836-eng.pdf.

Goodman, R., and Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K. and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. Chichester, UK: Wiley.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons, Inc.

Stephan, F.F., and McCarthy, P.J. (1958). *Sampling Opinions*. New York: John Wiley & Sons, Inc.

Sudman, S. (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61, 749-771.

Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M. and Bates, N. (Eds.) (2014). *Hard-to-survey populations*. Cambridge, UK: Cambridge University Press.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 173-178
Statistics Canada, Catalogue No. 12-001-X

Comments on the Rao and Fuller (2017) paper

Sharon L. Lohr[1]

Abstract

This note by Sharon L. Lohr presents a discussion of the paper "Sample survey theory and methods: Past, present, and future directions" where J.N.K. Rao and Wayne A. Fuller share their views regarding the developments in sample survey theory and methods covering the past 100 years.

Key Words: Data collection; History of survey sampling; Probability sampling; Survey inference.

Rao and Fuller deserve thanks for their succinct review of a field to which they both have contributed so much. It is no small feat to summarize the history of probability sampling and outline future directions in 16 pages!

It is always hazardous to predict the future. But reviewing the history of survey sampling allows us to see how the pioneers of the field dealt with the challenges of their day, and how those challenges and their solutions relate to today's issues.

To begin, let us compare the advantages and disadvantages of probability sampling today with the advantages and disadvantages that were perceived in the middle of the twentieth century, when probability samples were starting to have widespread use. The following lists are derived from Parten (1950, Chapter 4); the early sampling books by Deming (1950) and Hansen, Hurwitz and Madow (1953a) have similar descriptions. Parten's advantages of probability samples, relative to taking a census of the population or taking a convenience sample, are of four types:

A1. Estimates can be obtained faster from a sample than from a census. Fewer interviews are needed and, in the 1950s, data processing and tabulation could be done faster for a small data set than for a large one. Parten wrote: "This time-saving advantage is especially important in studies of our modern dynamic society. Conditions change so rapidly that unless short-cut methods are devised for measuring social situations, the measurement is out of date before the survey or poll is completed." (Parten, 1950, page 109).

A2. Estimates from a sample are less expensive than a census because fewer interviews are needed. This translates into lower costs for field staff and training.

A3. The survey can be tailored to the estimates of interest. The sampler can be more careful in the data collection, asking exactly the questions wanted and taking steps to minimize bias from nonresponse and other sources. A census, by contrast, may have few questions and limited opportunity for follow-up.

A4. Probability sampling allows the sampler to design the survey to achieve a desired precision and later report the achieved precision, without relying on model assumptions. Deming (1950, page 10) emphasized that not only can the sampling errors be calculated from probability

1. Sharon L. Lohr, Arizona State University, Tempe, AZ. E-mail: sharon.lohr@asu.edu.

> samples, but that "the biases of selection, nonresponse, and estimation are virtually eliminated or contained within known limits." Hansen et al. (1953a, page 10) stated: "With probability sampling methods one can get away completely from dependence upon judgment for determining precision. Under these circumstances, and with reasonably large samples, the precision of the results from the sample can be measured from the sample itself."

Parten also reviewed the disadvantages of taking a sample rather than a census:

D1. It is difficult to do sampling well, and to obtain representative samples. Mistakes in following the sampling protocol may introduce errors into the estimates, and results can be misleading if a sample is designed or analyzed badly. Additionally, a shortage of experienced survey statisticians makes it difficult for the would-be survey-taker to obtain technical assistance.

D2. The small size of the sample limits the information that can be obtained. Rare subpopulations have few observations in a sample. Additionally, the number of cross-tabulations is limited because there are too few cases in some subclassifications of interest.

How do Parten's advantages and disadvantages of probability samples hold up today? The disadvantages still exist. In particular, the demand for more detailed, more up-to-date, and more comprehensive information is increasing every year (D2). But while surveys still have the advantage (A3) that they can be tailored to answer the questions of interest, advantages (A1) and (A2) have diminished. In the 1950s, it was often expensive to collect *any* type of data. Even data from a small convenience sample could require expensive-to-collect interviews or labor-intensive transcription of paper records. But today, huge convenience samples can often be obtained with much less cost, while probability samples such as the American Community Survey or the National Crime Victimization Survey are increasingly expensive as response rates continue to decline. The convenience sample information may also be available faster than data from a high-quality probability sample, which requires months to weight the data, compute estimates, and perform quality checks.

The advantage (A4) of being able to plan for a desired precision has also diminished. Most large surveys still use design-based methods to report the precision of the survey. But the design-based margin of error that is reported generally includes only the sampling error and has the implicit assumption that the survey weighting has removed the nonresponse bias from the estimates. As response rates decrease, there is increasing reliance on judgment, through the use of model assumptions, to determine precision.

The landscape for data collection is thus different than it was in the 1930s, 1940s, and 1950s when probability sampling techniques were developed and implemented. Then, probability sampling in the United States answered an urgent need for faster and cheaper information about agriculture, business activity, manufacturing output, characteristics of the labor force, and other social and economic indicators. The pioneers of survey sampling methods revolutionized data collection during this period. Duncan and Shelton (1978) argued that this revolution was made possible by parallel developments in statistical theory, national income and product accounts, computing capacity, and organization of the statistical system.

Although the available data sources, infrastructure, technology, and methods have changed, the main problem facing us today is the same as in 1950: How can we best collect and make inferences from data to inform policy and research questions? If the current framework of probability samples did not exist, and we were tasked with constructing a system of collecting data, what would we do? For many problems, we would want to build a data collection system that is modular and can adapt to new sources of data and new technology for data collection. Much of the methodology that Rao and Fuller reviewed would be useful for this system, but new infrastructure and methodology–and perhaps another revolution–are needed.

As an example, consider the U.S. National Automotive Sampling System (NHTSA, 2017a). The system has two component surveys. The first component is a stratified multistage probability sample of 50,000 to 60,000 police accident reports (PARs) from the universe of approximately 6 million annual PARs, where PARs from serious crashes are sampled with higher probabilities than PARs from crashes involving only minor property damage. Data elements from the sampled PARs are coded into the electronic database; no information external to the PAR is obtained. The second survey is a smaller probability sample of about 5,000 PARs with much more labor-intensive data collection, where specially trained crash investigators visit the crash scene, inspect the vehicle(s) involved in the crash, get permission to access medical records, interview witnesses, and obtain other detailed information about the crash. The data from these two surveys are used to investigate time trends in vehicle crashes and effects of vehicle features on traffic safety (see, for example, NHTSA, 2017b), and are used in thousands of research papers.

But suppose we were asked to design this data collection system afresh. I want to stress that these suggestions are my fantasy, and have no connection to any plans for the surveys, which are constrained by current practical considerations and budgetary limitations to have a multistage sampling structure. If the current system did not exist, would it not be desirable to design the first survey to take a census of PARs instead of a sample? This task would not necessarily be easy. Hetzel (1997) described the long and laborious process by which the United States established the Vital Statistics System, requiring cooperation from state and local government agencies, uniform data collection procedures, and intensive research to validate the accuracy and coverage of records. Obtaining a census of PARs would similarly take huge initial investments to develop infrastructure and to secure cooperation from states and police jurisdictions. After that investment, however, the data collection would be established and PARs could be transmitted electronically as they were collected or updated.

The advantages of having a census of PARs instead of a sample would be numerous. One advantage would be that statistics would be available much more quickly, since one would not need to wait until the end of the data collection year to weight and publish the survey data: statistics could be updated as the data came in. The largest advantage of a census, however, would be the extra information on subpopulations. This would allow better monitoring of the data to detect potential safety hazards. In a sample of size 50,000, a particular make/model/year of vehicle may be represented by only a handful of observations (if any); the sample size would be much larger in the census. In some surveys, as Rao and Fuller point out, small area estimation methods can be used to model results for subpopulations with small sample sizes. But for crash data, often a subpopulation is of interest because it is suspected to be an outlier–it is suspected there are more crashes for a certain car make or vehicle feature than would be predicted by a model. These outliers

cannot be detected from a small area model. The only way to obtain information on potentially outlying subpopulations is to collect more data on them.

With a census of PARs, though, where do survey research methods come in? There would inevitably be missing data that would need to be investigated and modeled, and a two-phase design might be used to obtain information from nonresponding states or police jurisdictions. But the main survey design problem would be for two aspects: first, sampling could be used to audit the census of accident reports, and second, sampling would be needed for the labor-intensive crash investigation part of the system. The census of PARs would provide a rich sampling frame for the crash investigation system and other investigations. That rich frame information could be exploited in the sample design, possibly by using balanced sampling or a sampling design that can be dynamically adapted to data needs and to the continuously updated frame.

Of course, even a census of PARs might be outdated or insufficient for data needs in the future. As more vehicles are equipped with cameras and sensors, or as self-driving vehicles and surveillance systems become more prevalent, a sample or census of PARs may be replaced or supplemented with passively collected data. Increased use of large-scale passive data sources raises serious issues about privacy and data ownership, requiring much debate and research, and these issues are beyond the scope of this discussion. But, beyond the societal questions about the ethics of data collection, what new statistical methodology is needed to deal with the revolution in data availability?

I see three major areas of interconnected research needed for the short-term future, and these are related to the research problems that faced Parten, Deming, and Hansen in the middle of the last century.

- Better measures of uncertainty for survey estimates. When Hansen and Hurwitz (1949, page 365) wrote about the superiority of probability samples over judgment samples, they emphasized that the assumption-free nature of inference from probability samples depends on achieving high response rates: “In the Census Bureau it is usually assumed that if the required information is obtained from more than 95 per cent of the designated households one is entitled to feel fairly secure in assuming that the sample was taken in conformance with sampling theory, even though assumptions may be necessary for the remaining 5 per cent. It has been found that for some purposes trouble arises even when making assumptions for only 5 per cent.” Deming (1950, page 13) also used 95 percent as the lower bound for the validity of inference from probability samples: “A sample that is 95 or 98 percent a probability-sample and the other 5 or 2 percent a judgment-selection or judgment-adjustment for refusals, for people not at home, etc., may still be an excellent sample, although it is important to investigate the remaining 5 or even 2 percent as soon as possible.”

  Over the years, the 95 percent threshold for using probability sampling methods for inference has drifted downward, to the point that now the same weighting methods are used for a sample with a response rate of 10 percent as for one with a response rate of 95 percent. As response rates have decreased, increasingly strong model assumptions have been made about response mechanisms and the undercovered and nonresponding populations, but uncertainty about these assumptions is generally not reflected in the reported confidence intervals: these are still primarily based on the sampling error. Lohr and Brick (2017) ascribed recent polling failures to the systems used to

derive confidence or posterior prediction intervals, and argued that new statistical methods are needed for reporting interval estimates that better reflect uncertainty about the estimates. As Parten (1950, page 403) said: "It is not unusual to find very refined statistical techniques used to measure random errors in data which are so biased that all the corrective devices known would not enable the surveyor to determine what the correct results should be."

- Combining multiple sources of data. Methods such as record linkage, multiple frame surveys, and hierarchical models can facilitate combining data, and can also be used to evaluate data quality from different sources. Lohr and Raghunathan (2017) reviewed statistical methods for combining data from different sources, and argued that using multiple sources of information could also help with the problem of evaluating and incorporating bias errors into uncertainty estimates, although the statistical methods reviewed do not solve the problem of how to obtain estimates for subpopulations that may be missing from all sources.

  While probability sampling methods have served society well for the last 70 years, they may play a more limited role in the future, perhaps being used in some data collections to validate and check other data sources instead of serving as the primary data sources themselves. Hansen, Hurwitz and Pritzker (1953b) viewed the census post-enumeration surveys, in which intensive efforts were made to determine the most accurate information possible from an area sample, in this light. They also wrote about the need to weigh the costs of obtaining accurate statistics against the option of obtaining larger sample sizes or more variables. Hansen, Madow and Tepping (1983) argued that it is precisely at times of societal change that the unbiasedness guaranteed by probability sampling is essential, and targeted use of high-quality probability samples can provide assessments of the coverage and accuracy of information from other data sources.

- Finally, there is a need for a renewed focus on design, which dominated much of the literature between 1920 and 1960. A modular data collection system, relying on different data sources for different information needs, could adapt to changing societal needs and new technology for data collection. Designs are needed that are robust to errors in individual sources and allow assessment of those errors, and are also robust to potential changes in the data sources.

As Rao and Fuller pointed out, the statistical research community has repeatedly met the information needs of society through new innovations. The challenges of dealing with new data sources and missing data are great, but so were the problems faced in the past that led to the development of probability sampling, small area estimation, replication variance estimation, and imputation theory. The next revolution in sampling may be just around the corner.

Acknowledgements

Part of this discussion was adapted from the author's 2016 JPSM Distinguished Lecture "The Essential Survey Statistician," available at https://www.jpsmclasses.umd.edu/ Mediasite/Catalog/catalogs/default.

References

Deming, W.E. (1950). *Some Theory of Sampling.* New York: Dover.

Duncan, J.W., and Shelton, W.C. (1978). *Revolution in United States Government Statistics 1926-1976.* Washington, D.C.: U.S. Department of Commerce.

Hansen, M.H., and Hurwitz, W.N. (1949). Dependable samples for market surveys. *Journal of Marketing*, 14, 363-372.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953a). *Sample Survey Methods and Theory. Volume I: Methods and Applications.* New York: John Wiley & Sons, Inc.

Hansen, M.H., Hurwitz, W.N. and Pritzker, L. (1953b). The accuracy of census results. *American Sociological Review*, 18, 416-423.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 384, 776-793.

Hetzel, A.M. (1997). *U.S. Vital Statistics System: Major Activities and Developments, 1950-95*. Hyattsville, MD: National Center for Health Statistics. Available from https://www.cdc.gov/nchs/data/misc/usvss.pdf, last visited May 5, 2017.

Lohr, S.L., and Brick, J.M. (2017). Roosevelt predicted to win: Revisiting the *Literary Digest* poll of 1936. *Statistics, Politics, and Policy*, 8, 65-84.

Lohr, S.L., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

National Highway Transportation Safety Administration (NHTSA, 2017a). *National Automotive Sampling System (NASS)*. Available from https://www.nhtsa.gov/research-data/national-automotive-sampling-system-nass, last visited May 5, 2017.

National Highway Transportation Safety Administration (NHTSA, 2017b). *Traffic Safety Facts, 2015*. Available from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384, last visited May 17, 2017.

Parten, M. (1950). *Surveys, Polls, and Samples.* New York: Harper & Brothers.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 179-181
Statistics Canada, Catalogue No. 12-001-X

Comments on the Rao and Fuller (2017) paper

Chris Skinner[1]

Abstract

This note by Chris Skinner presents a discussion of the paper "Sample survey theory and methods: Past, present, and future directions" where J.N.K. Rao and Wayne A. Fuller share their views regarding the developments in sample survey theory and methods covering the past 100 years.

Key Words: Data collection; History of survey sampling; Probability sampling; Survey inference.

This paper provides an outstanding account of sample survey theory and methods, distilling in a concise and elegant way a huge amount of wisdom about both the theory and practice of the field. I shall not comment on the past and present but, following the invitation I received, will present some thoughts on the future. I am fully in agreement with the final part of the paper regarding the future and see my thoughts as overlapping.

This discussion will emphasise a National Statistics Institute (NSI) perspective, and I expect (and hope) that NSIs will play a key role in driving methodological developments, even though the statistical environment will change, e.g., with an increased range of bodies which supply data to NSIs and/or produce statistical outputs themselves.

Inferential targets: I expect the same kinds of descriptive finite population targets (viewed methodologically) to remain of core interest. The importance of analytic needs will also continue but how these will be met will depend on how data access arrangements evolve, in the context e.g., of concerns about confidentiality, and the impact of developing data science practice, such as greater emphasis on predictive modelling.

Sample surveys and other data sources: The nature and extent of relevant data sources will be a critical area of development. I do not believe that surveys will disappear – there will always be a huge number of variables of interest which require primary data collection. But I do expect the sample survey to be increasingly an integrated part of a wider set of data sources including census, administrative data and 'big data' sources (e.g., Lohr and Raghunathan, 2017; Zhang, 2012). The methodological challenge will be how to integrate such a range of sources effectively. The different sources may have multiple owners and access arrangements will have an important bearing on how sources can be integrated. I also do not believe that sampling will disappear – it will be needed not just for primary data collection but also for supplementary surveys (see below) and for managing big data sources.

Supplementary surveys: the need for supplementary survey samples to check validity or to improve inference is likely to grow. 'Reference surveys' may augment non-probability samples (Elliot and Valliant, 2017); coverage sample surveys may be needed to check for both under- or over-coverage, e.g., in

1. Chris Skinner, London School of Economics and Political Science. E-mail: C.J.Skinner@lse.ac.uk.

administrative data sources, and to correct for such errors (Zhang, 2015); surveys linked at the unit level may be needed to check for measurement error in data sources.

Non-response and sampling: Unit nonresponse will become ever more problematic and it will invariably be necessary for inference to take account of both nonresponse error and sampling error. The key challenge will be to avoid (reduce) selection bias. The use of randomisation in sampling to achieve this goal and to justify certain modelling assumptions may become at least as prominent a property of probability sampling as its use for design-based inference. It may become sensible to consider protocols for sampling and managing nonresponse in a more integrated way and research into such options might allow for sampling protocols which include non-probability features, providing the goal of reducing selection bias remains central. Flexible multi-mode options for response seem likely to be natural candidate options to consider. The nature of auxiliary data sources and estimation considerations must also, of course, be considered carefully when evaluating options for sampling and nonresponse management in a combined way.

Estimation methods and theory: estimation methods will evolve to exploit new kinds of statistical relationships within and between data sources, both to control for potential selection bias effects and to gain efficiency. Many estimation problems may be formulated in terms of outcome variables Y which can only be obtained on selective samples and predictor variables X for which large databases approximating 100% coverage can be achieved. Constructing such databases may be a key goal in both business and social statistics contexts in NSIs. In the latter case, this goal may be aligned to population census developments, involving for example administrative data sources (Skinner, 2017). In such settings, a broad approach to estimation may combine population-level X distributions with conditional distributions of Y given X from the selective sample sources under assumptions similar to missing at random. The importance of temporal considerations, such as the benefits of borrowing strength over time, seems likely to increase and may exploit opportunities afforded by administrative sources which are typically longitudinal. Existing methods of calibration and Fay-Herriot model-based small area estimation will continue to be used for sources linked in an aggregate way. Linkage at the level of the individual or GPS-based unit (e.g., building or address) may open up further methods (e.g., Lohr and Raghunathan, 2017). Survey sampling theory, including model-based prediction methods and small area estimation methods, will continue to play a key role. Missing data theory provides a natural framework for handling integrated data sources and I would expect further confluence between sampling and missing data theory. The treatment of linkage errors and measurement errors, e.g., arising from measurement differences between sources and modes of data collection, will also be important.

Quality assessment and accuracy estimation: in the face of likely continuing pressures on budgets, it will be essential that the importance of high standards of quality is promoted and recognised among users of statistical outputs if high quality sample surveys are not to be replaced by cheap, untrustworthy alternatives. This could benefit from a strengthening of the quality assessment role of national bodies set up to oversee and enhance public confidence in statistical outputs, especially if the number and diversity of suppliers of such outputs is to increase. More specifically, accuracy assessment will be critical. Traditional variance estimation methods can play a role and may be extended to capture wider sources of variation, e.g., by extending the definition of replicates in replication methods. But, with the increasing use of model-based inference, accuracy assessment will need also to embrace the assessment of the impact of departures from

assumptions on estimation methods. Approaches, such as model checking, diagnostics and sensitivity analysis will likely grow in importance.

References

Elliot, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science,* 32, 249-264.

Lohr, S.L., and Raghunathan, T.E. (2017). Combining survey data and other data sources. *Statistical Science,* 32, 293-312.

Skinner, C.J. (2018). Issues and challenges in census taking. *Annual Review of Statistics and its Application*, Volume 5.

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

Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, 31, 381-396.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 183-210
Statistics Canada, Catalogue No. 12-001-X

Social media as a data source for official statistics; the Dutch Consumer Confidence Index

Jan van den Brakel, Emily Söhler, Piet Daas and Bart Buelens[1]

Abstract

In this paper the question is addressed how alternative data sources, such as administrative and social media data, can be used in the production of official statistics. Since most surveys at national statistical institutes are conducted repeatedly over time, a multivariate structural time series modelling approach is proposed to model the series observed by a repeated surveys with related series obtained from such alternative data sources. Generally, this improves the precision of the direct survey estimates by using sample information observed in preceding periods and information from related auxiliary series. This model also makes it possible to utilize the higher frequency of the social media to produce more precise estimates for the sample survey in real time at the moment that statistics for the social media become available but the sample data are not yet available. The concept of cointegration is applied to address the question to which extent the alternative series represent the same phenomena as the series observed with the repeated survey. The methodology is applied to the Dutch Consumer Confidence Survey and a sentiment index derived from social media.

Key Words: Big data; Design-based inference; Model-based inference; Nowcasting; Structural time series modelling; Cointegration.

1 Introduction

National statistical institutes traditionally use probability sampling in combination with design-based or model-assisted inference for the production of official statistics. The concept of random probability sampling has been developed mainly on the basis of the work of Bowley (1926), Neyman (1934) and Hansen and Hurwitz (1943). See for example Cochran (1977) or Särndal, Swensson and Wretman (1992) for an extensive introduction in sampling theory. This is a widely accepted approach, since it is based on a sound mathematical theory that shows how under the right combination of a random sample design and estimator, valid statistical inference can be made about large finite populations based on relative small samples. In addition, the amount of uncertainty by relying on small samples can be quantified through the variance of the estimators.

There is persistent pressure on national statistical institutes to reduce administration costs and response burden. In addition, declining response rates stimulate the search for alternative sources of statistical information. This could be accomplished by using administrative data like tax registers, or other large data sets – so called big data – that are generated as a by-product of processes not directly related to statistical production purposes. Examples of these include time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook and internet search behaviour from Google Trends. A common problem with this type of data sources is that the process that generates the data is unknown and likely selective with respect to the intended target population. A challenging problem in this context is to use this data for the production of official statistics that are representative of the target population. There is no randomized sampling design that facilitates the generalization of conclusions and

1. Jan van den Brakel, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands and Maastricht University School of Business and Economics, Department of Quantitative Economics, The Netherlands. E-mail: ja.vandenbrakel@cbs.nl; Emily Söhler, Student Econometrics, Maastricht University; Piet Daas and Bart Buelens, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands.

results obtained with the available data to an intended larger target population. Hence, extracting statistically relevant information from these sources is a challenging task (Daas and Puts, 2014a).

Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013) address the problem of using non-probability samples and mention the possibility of applying design-based inference procedures to correct for selection bias. Buelens, Burger and van den Brakel (2015) explore the possibility of using statistical machine learning algorithms to correct for selection bias. Instead of replacing survey data for administrative data or big data, these sources can also be used to improve the accuracy of survey data in model-based inference procedures. Marchetti, Giusti, Pratesi, Salvati, Giannotti, Perdreschi, Rinzivillo, Pappalardo and Gabrielli (2015) and Blumenstock, Cadamuro and On (2015) used big data as a source of auxiliary information for cross-sectional small area estimation models.

Many surveys conducted by national statistical institutes are conducted repeatedly. In this paper, a multivariate structural time series modelling approach is applied to combine the series obtained by a repeated survey with series from alternative data sources. This serves several purposes. First, a model based estimation procedure based on a time series model increases the precision of the direct estimates by using the temporal correlation between the direct estimates in the separate editions of the survey. The use of time series modelling with the aim of improving the precision of survey data has been considered by many authors dating back to Blight and Scott (1973). Second, extending the time series model with an auxiliary series allows to model the correlation between the unobserved components of the structural time series models, e.g., trend and seasonal components. Harvey and Chung (2000) propose a time series model for the Labour Force Survey in the UK extended with a series of claimant counts. If such a model detects strong positive correlations between these components, then this might further increase the precision of the time series estimates for the sample survey. Indicators derived from social media are generally available at a higher frequency than related series obtained with periodic surveys. This allows to use this time series modelling approach to make early predictions for the survey outcomes in real time at the moment that the outcomes for the social media are available, but the survey data not yet. In this case, the social media are used as a form of nowcasting. Third, the concept of cointegration in the context of multivariate state space models can be used to evaluate to which extent both series are identical. If the trend components of two observed series are cointegrated, then both series are driven by one underlying common trend. It can be argued that if an auxiliary series is cointegrated with the series of the survey, they represent the same underlying stochastic process. This could be used as an argument to motivate that a statistic measured with a big data source is representative for an intended target population. This is, however, more an empirical argument and not as strong as the theory underlying probability sampling, that proves that random sampling in combination with an (approximately) design-unbiased estimator results in representative statistics.

The Dutch Consumer Confidence Survey (CCS) is a monthly survey based on approximately 1,000 respondents with the purpose of measuring the sentiment of the Dutch population about the economic climate by means of the so-called Consumer Confidence Index (CCI). Daas and Puts (2014b) developed a sentiment index, independently of the CCS, that is derived from social media platforms that was found to mimic the CCI very well. This index is referred to as the Social Media Index (SMI). In this paper, the aforementioned multivariate structural time series modelling approach is applied to both series in an attempt

to improve the precision of the CCI. It is also illustrated how the SMI in this time series model can be used to make early predictions or nowcasts of the CCI.

In Section 2, the survey design of the CCS and the estimation procedure for the CCI is described. The approach followed by Daas and Puts (2014b) to construct a sentiment index from social media platforms is also described. In Section 3, a structural time series model for the CCI series and SMI series is proposed. Results obtained with the model are presented in Section 4. The paper concludes with a discussion in Section 5.

2 Data

2.1 Dutch Consumer Confidence Survey

The Consumer Confidence Index (CCI) is based on a monthly survey, called the Consumer Confidence Survey (CCS), and measures the opinion of households residing in the Netherlands about the economic climate in general and their own financial situation. The CCS is a continuous survey. Each month a self-weighted sample of approximately 2,500 households is drawn by stratified two-stage sampling from a sample frame derived from the Dutch Municipal Register. Households for which a known telephone number is available are contacted by an interviewer who completes the questionnaire by computer assisted telephone interviewing during the first ten working days of the month. On average a net sample of about 1,000 responding households is obtained, which comes down to a response rate of about 40%. A major part of the nonresponse are households for which no known telephone number of a land-line connection is available. The response among households for which a known telephone number is available is about 60%.

The CCI is based on five questions that can be answered positively, neutral or negatively. The questions refer to the economic or financial situation in the last 12 month or the respondents expectations in the future 12 months. Let $P_{1,t}^{q}$, $P_{2,t}^{q}$, and $P_{3,t}^{q}$, denote the percentage of respondents that answered question $q = 1, \ldots, 5,$ in month t positively, neutral or negatively, respectively. Now the CCI is defined as the difference between the percentage of positive and negative respondents, averaged over the five questions:

$$I_t = \frac{1}{Q} \sum_{q=1}^{Q} \left(P_{1,t}^{q} - P_{3,t}^{q} \right). \tag{2.1}$$

Since the sample is self-weighted, and no auxiliary information is used in the estimation procedure, the percentages are estimated with the sample mean, i.e.,

$$P_{j,t}^{q} = \frac{100}{n_t} \sum_{i=1}^{n} \delta_{i,j,t}^{q}, \tag{2.2}$$

for question $q = 1, \ldots, 5,$ and answer category $j = 1, 2, 3.$ In (2.2) n_t is the net sample size in month $t,$ and $\delta_{i,j,t}^{q}$ is a dummy indicator that is equal to one if respondent i chose category j to question $q.$ Assuming simple random sampling without replacement for the households, it can be proved that the variance of (2.1) can be estimated by

$$\begin{aligned}\mathrm{Var}\left(I_t\right) &= \frac{1}{Q^2}\sum_{q=1}^{Q}\left[\mathrm{Var}\left(P_{1,t}^{q}\right)+\mathrm{Var}\left(P_{3,t}^{q}\right)\right]-\frac{2}{Q^2}\sum_{q=1}^{Q}\sum_{q'=1}^{Q}\mathrm{Cov}\left(P_{1,t}^{q},P_{3,t}^{q'}\right)\\ &+\frac{1}{Q^2}\sum_{q=1}^{Q}\sum_{q'\neq q}^{Q}\left[\mathrm{Cov}\left(P_{1,t}^{q},P_{1,t}^{q'}\right)+\mathrm{Cov}\left(P_{3,t}^{q},P_{3,t}^{q'}\right)\right],\end{aligned} \tag{2.3}$$

with

$$\mathrm{Var}\left(P_{j,t}^{q}\right)=\frac{1}{n_t}P_{j,t}^{q}\left(100-P_{j,t}^{q}\right),\qquad \mathrm{Cov}\left(P_{j,t}^{q},P_{j',t}^{q'}\right)=\frac{1}{n_t}\left(P_{jj',t}^{qq'}-P_{j,t}^{q}P_{j',t}^{q'}\right),$$

$$\mathrm{Cov}\left(P_{j,t}^{q},P_{j',t}^{q'}\right)=\frac{1}{n_t}\left(P_{jj',t}^{qq'}-P_{j,t}^{q}P_{j',t}^{q'}\right),\quad \mathrm{Cov}\left(P_{j,t}^{q},P_{j',t}^{q}\right)=-\frac{1}{n_t}P_{j,t}^{q}P_{j',t}^{q},$$

$$P_{jj',t}^{qq'}=\frac{100}{n_t}\sum_{i=1}^{n}\delta_{i,j,t}^{q}\,\delta_{i,j',t}^{q'}.$$

Figure 2.1 shows the CCI with a 95% confidence interval calculated using the approach described in this section, observed during the period December 2000 through March 2015. In October 2013, the official publication of the CCI is missing.

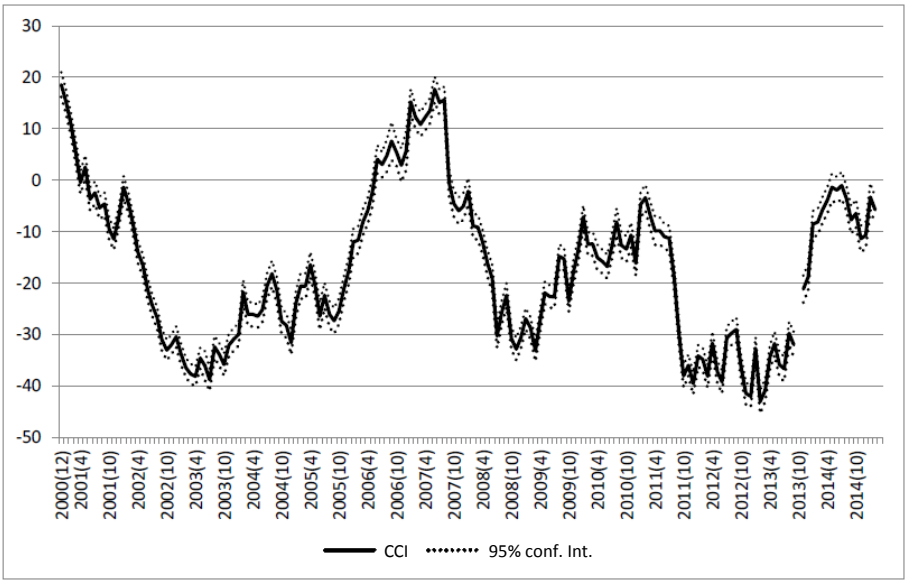

Figure 2.1 Consumer confidence index (CCI) with a 95% confidence interval.

2.2 Social media sentiment

In an attempt to reduce administration costs and response burden, Daas and Puts (2014b) developed a sentiment index from social media sources that could be used as an alternative indicator for the CCI. They used messages posted on the most popular social media platforms in the Netherlands, written in the Dutch language. These messages are classified as containing positive, neutral, or negative messages using a variant of sentence-level based classification (Pang and Lee, 2008). An index is calculated by taking the difference between the percentage of positive and negative messages.

Combinations of all Facebook and Twitter messages with and without certain filters on phrases were compared with the CCI. The combination of all publicly available Facebook messages together with filtered Twitter messages containing personal pronouns had the highest correlation with the CCI. The Twitter messages had to be filtered due to the fact that a lot of Twitter messages are not very informative. See Daas and Puts (2014b) for further details. In their research Daas and Puts (2014b) also found that major changes in the behaviour of the public on social media, such as those caused by huge events and changes in the number of messages posted on each platform, have a disturbing effect on the series. The final indicator proposed is the average of the sentiment in the Facebook and Twitter messages during each period.

In Figure 2.2, the Social Media Index (SMI) is compared with the CCI for the period June 2010 until March 2015. Both series are clearly on a different level but show a more or less similar evolution. During the presented period, the CCI is always negative, while the SMI is always positive. The size or amplitude of the movements of the CCI are also considerably larger compared to the SMI. Many factors are responsible for this difference since the CCI is based on a survey where data collection is conducted by telephone and the SMI is based on classifying messages on Twitter and Facebook. The interesting question is to which extent the evolution of both series is similar.

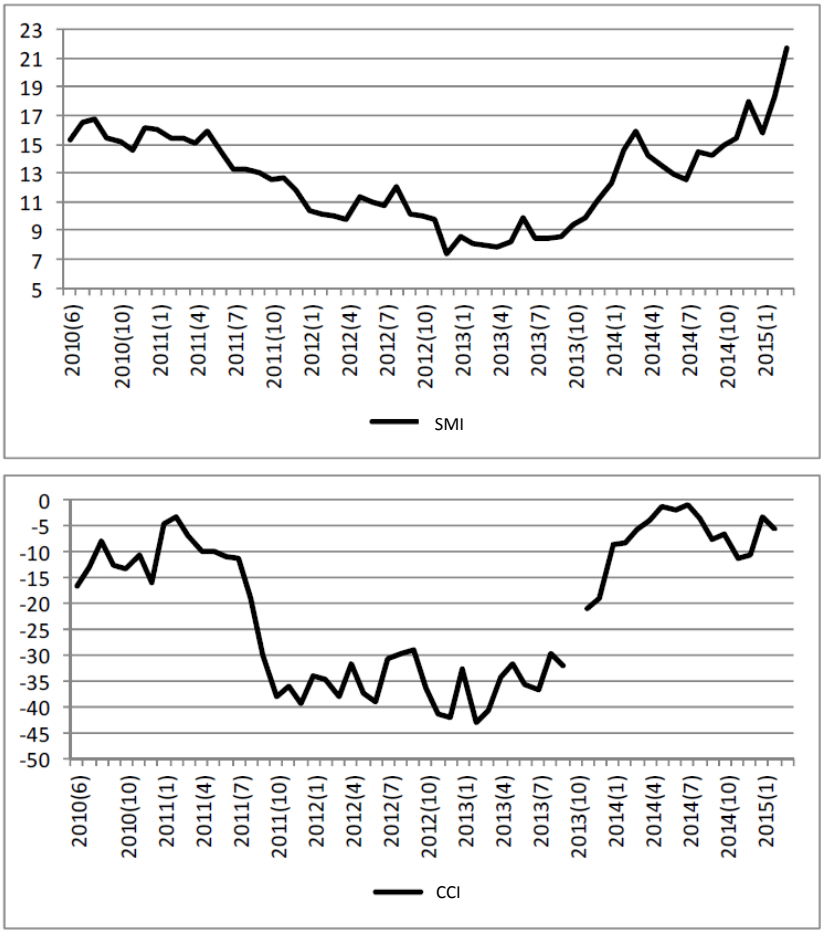

Figure 2.2 Comparison of the Social media index (SMI, upper panel) with the Consumer confidence index (CCI, lower panel).

2.3 Quality aspects of the CCI and the SMI

The accuracy of statistics are measured with its variance and bias. For simplicity we only distinguish between selection bias and measurement bias. The variance of survey sample statistics, like the CCI, depends on the sample size and will typically constitute a substantial part of the uncertainty of sample statistics. In big data sources the concept of sampling variance is meaningless since the data generating process is not a probability sample from a finite target population. Variance components of the model used to describe the assumed data generating process could be used as an accuracy measure instead. The model-based variance of statistics obtained with time series models applied to series obtained from internet or social media will always be positive depending of the volatility of the series, which predominantly depends on the frequency of the observed series and the dynamics of the stochastic process instead of the volume of the data.

The selection bias of sample survey statistics is approximately zero under complete response. In practice however, selection bias arises due to selective nonresponse, under coverage of the sample frame and to which extend with the field work strategy the target population is successfully reached. In the case of the CCI, only the population with a known telephone number of a land-line connection is reached and the response among this subpopulation is about 60%. The selection bias of big data sources is generally unknown. In this paper, we apply the concept of cointegration to evaluate to which extend the SMI measures the same concept as the CCI. Note, however, that in the case of cointegration, the SMI might reflect a similar nonresponse and coverage selection bias as the CCI. Baker et al. (2013) pointed out that there are similarities between selection bias in probability samples and the non-probabilistic approach followed with data sources like social media.

The measurement bias in sample statistics typically depends on the extend that the conceptual variables to be measured, are implemented in the questionnaire, but also on data collection mode and the quality of the interviewers. Problems with measurement bias in surveys arises, since measurements of the variables of interest are indirect in that respondents are asked to report about their behaviour, introducing all kind of measurement errors. In the case of the CCI the question can be raised to which extend respondents are capable to express their long-term confidence in the economy and to which extend it is influenced by short-term emotions. These problems do not arise with big data if they contain direct measurements of people behaviour. With an index derived from social media like the SMI the question can be raised to which extend it measures a similar concept as the CCI. In Subsection 2.2, it was already mentioned that major changes in the behaviour of the public on social media have a disturbing effect on the series. Particularly at the end of the series, a sudden change in behaviour on social media will be very hard to distinguish from a real turning point. For example, a Google-trend series on search related to vacancies might track an official series on unemployment. It does measure unemployment, however, search behaviour before the start of the financial crisis in 2009 might be completely different compared to the period directly after the financial crisis, invalidating the concept intended to be measured.

3 Structural time series modelling of the CCI and the SMI

In this section, univariate and bivariate structural time series models for the CCI and SMI are developed. With a structural time series model, a series is decomposed in a trend component, seasonal component, other

cyclic components, regression component and an irregular component. For each component a stochastic model is assumed. This allows the trend, seasonal, and cyclic component but also the regression coefficients to be time dependent. If necessary autoregressive-moving-average (ARMA) components can be added to capture the autocorrelation in the series beyond these structural components. See Harvey (1989) or Durbin and Koopman (2012) for details about structural time series modelling.

The question addressed in this paper is to which extent the SMI follows a similar pattern as the CCI such that the SMI can be used in the estimation procedure of the CCI or, in the most extreme case, even can replace the CCI. This question is addressed by developing a bivariate structural time series model for the CCI and the SMI and modeling the correlation between the disturbance terms of the different components of the structural time series model for both series. The concept of cointegration is used to investigate to which extent the unobserved components of both series are driven by common factors. If e.g., the trends of both series are driven by one underlying common trend an argument can be made that the SMI represents similar evolution of sentiment feelings compared to the CCI. Alternatively, the SMI can be used as an auxiliary series in a model based estimation procedure for the CCI or in a nowcasting procedure to obtain more precise real time estimates.

3.1 Univariate model CCI series

As a first step, a univariate time series model for the CCI series is proposed. With the design-based approach described in Section 2.1, the sample information observed in each separate month is used to obtain an estimate for the CCI in that month. A drawback of this approach is that information observed in preceding periods is not used to obtain more accurate estimates for the CCI. In survey methodology, time series models are frequently applied to develop estimates for periodic surveys. Blight and Scott (1973) and Scott and Smith (1974) proposed to regard the unknown population parameters as a realization of a stochastic process that can be described with a time series model. This introduces relationships between the estimated population parameters at different time points in the case of non-overlapping as well as overlapping samples. The explicit modelling of this relationship between these survey estimates with a time series model can be used to combine sample information observed in the past to improve the precision of estimates obtained with periodic surveys. Some key references to authors that applied the time series approach to repeated survey data to improve the efficiency of survey estimates are Scott, Smith and Jones (1977), Tam (1987), Binder and Dick (1989, 1990), Bell and Hillmer (1990), Tiller (1992), Rao and Yu (1994), Pfeffermann and Burck (1990), Pfeffermann (1991), Pfeffermann and Rubin-Bleuer (1993), Pfeffermann, Feder and Signorelli (1998), Pfeffermann and Tiller (2006), Harvey and Chung (2000), Feder (2001), Lind (2005) and van den Brakel and Krieg (2009, 2015).

Developing a time series model for survey estimates observed with a periodic survey starts with a model, which states that the survey estimate can be decomposed in the value of the population variable and a sampling error:

$$I_t = \theta_t + e_t, \tag{3.1}$$

where θ_t denote the real CCI in month t under a complete enumeration of the target population and e_t the sampling error.

The CCI is observed at a monthly frequency. Therefore, as a first step, the series of the finite population parameter can be decomposed in a stochastic trend, seasonal component to model systematic deviations from the trend within a year, and a white noise component for the remaining unexplained variation. These considerations lead to the following model for the series of the finite population parameter:

$$\theta_t = L_t + S_t + \xi_t, \tag{3.2}$$

where L_t denotes a stochastic trend, S_t a stochastic seasonal component and ξ_t the unexplained variation of the finite population parameter. Inserting (3.2) into measurement model (3.1) gives

$$I_t = L_t + S_t + \xi_t + e_t. \tag{3.3}$$

In a cross-sectional survey it is difficult to separate the sampling error from the white noise of the population parameter. Therefore, both components are combined in one disturbance term

$$\upsilon_t = \xi_t + e_t. \tag{3.4}$$

It is assumed that $E(\upsilon_t) = 0$ and $\text{Var}(\upsilon_t) = \sigma_\upsilon^2$. To allow for nonhomogeneous variance in the sampling errors, Binder and Dick (1990) proposed a measurement error where the disturbance terms υ_t are proportional to the standard errors of I_t, i.e.,

$$\upsilon_t = \sqrt{\text{Var}(I_t)}\ \tilde{\upsilon}_t, \tag{3.5}$$

with $E(\tilde{\upsilon}_t) = 0$, $\text{Var}(\tilde{\upsilon}_t) = \sigma_\upsilon^2$, and where $\text{Var}(I_t)$ is defined by (2.3) and is used as a priori information in the time series model. Such a model would be useful if the sampling error dominates the white noise in the population parameter. Initial analyses indicate that in this application the variance of the population white noise is substantial, invalidating (3.5) for this application. In addition, the variance of the sampling error in this application is constant over time. Therefore, it is decided to combine the sampling error with the population white noise and assume a constant variance over time. The question how to account for sampling variance is also an issue in seasonal adjustment variances (Pfeffermann and Sverchkov, 2014). Bell (2005) studied the contribution of the sampling variance in the variance of the estimation error of seasonally adjusted series and in the nonseasonal component. In the case of (rotating) panels, the sampling error can be separated from the population white noise. In cross-sectional repeated surveys, it is difficult to identify the separate components and therefore both terms are combined in one disturbance term that captures both the sampling variance and the unexplained variation of the population parameter.

An extensive model selection showed that a smooth trend model is the most appropriate model to capture the trend and the economic cycle in the CCI series. The smooth trend model is defined as (Durbin and Koopman, 2012):

$$L_t = L_{t-1} + R_t,$$

$$R_t = R_{t-1} + \eta_t, \quad \mathrm{E}(\eta_t) = 0, \tag{3.6}$$

$$\mathrm{Cov}(\eta_t, \eta_{t'}) = \begin{cases} \sigma_\eta^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}.$$

Adding a random component for the level in (3.6) improves the log-likelihood with five units but results in an overfit of the data in a sense that the smoothed signal almost exactly follows the observed series with a very small measurement error variance. A local level model (random level without a slope) improves the log-likelihood with three units but also intends to overfit the data.

The seasonal component is modelled with a trigonometric model, which is defined as (Durbin and Koopman, 2012):

$$S_t = \sum_{j=1}^{6} \gamma_{jt}, \tag{3.7}$$

with

$$\begin{aligned} \gamma_{jt} &= \gamma_{jt-1} \cos(\lambda_j) + \tilde{\gamma}_{jt-1} \sin(\lambda_j) + \omega_{jt}, \\ \tilde{\gamma}_{jt} &= -\gamma_{jt-1} \sin(\lambda_j) + \tilde{\gamma}_{jt-1} \cos(\lambda_j) + \tilde{\omega}_{jt}. \end{aligned}$$

Here λ_j denotes the frequency of the different cycles in radians and is defined as

$$\lambda_j = \frac{2\pi j}{12}, \quad \text{for} \quad j = 1, \ldots, 6.$$

For the disturbance terms, it is assumed that

$$E(\omega_{jt}) = 0, \quad E(\tilde{\omega}_{jt}) = 0,$$

$$\mathrm{Cov}(\omega_t, \omega_{t'}) = \begin{cases} \sigma_\omega^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}.$$

For reasons of parsimony, the same variance structure is assumed with the same hyperparameter for $\tilde{\omega}_{jt}$. Furthermore, it is assumed that ω_t and $\tilde{\omega}_t$ are uncorrelated.

After including the stochastic trend component (3.6) and seasonal component (3.7), no additional cycle components are required. The model selection procedure indicated that two level interventions are needed to model sudden jumps in the series. The first one is due to the financial crisis in September 2008, and the second one is due to the economic downturn in September of 2011. Finally, an outlier is required for September 2007. Adding these three components increases the log-likelihood with 15 units. These considerations lead to the following model for the observed CCI series

$$I_t = L_t + S_t + \beta^{07}\delta_t^{07} + \beta^{08}\delta_t^{08} + \beta^{11}\delta_t^{11} + \upsilon_t, \tag{3.8}$$

with

$$\delta_t^{07} = \begin{cases} 1 & \text{if} \quad t = 2007(9) \\ 0 & \text{if} \quad t \neq 2007(9) \end{cases}, \qquad \delta_t^{08} = \begin{cases} 1 & \text{if} \quad t \geq 2008(9) \\ 0 & \text{if} \quad t < 2008(9) \end{cases}, \qquad \delta_t^{11} = \begin{cases} 1 & \text{if} \quad t \geq 2011(9) \\ 0 & \text{if} \quad t < 2011(9) \end{cases},$$

and β^x the corresponding regression coefficients.

Finally, autoregressive (AR) and moving average (MA) components can be added to the structural time series model (3.8). In this application, there were no indications that such components are required, since there are no clear signs of remaining serial correlation in the standardized innovations. Adding an AR(1) or an MA(1) to (3.8) increases the log-likelihood with 5 and 4.5 units respectively. Adding second-order AR or MA models does not further improve the log-likelihood. Adding an ARMA(1,1) also does not further increases the log-likelihood. An AR(1) or MA(1) slightly improves the correlogram but also increases the standard error of the filtered smoothed signals. Therefore, model (3.8) was finally selected for the CCI series.

State space models assume that the disturbance terms are normally and independently distributed. These assumptions translate into the assumption that the innovations are normally and independently distributed. Table A.1 in the appendix contains an overview of goodness of fit statistics applied to the standardized innovations. The values for skewness, kurtosis and the Bowman-Shenton test do not indicate deviations from normality of the standardized innovations. The values for the Ljung-Box test and Durban-Watson test do not indicate serial correlations in the standardized innovations. This is also confirmed by a correlogram (not shown). In conclusion, these diagnostics indicate that (3.8) fits the series of the CCI reasonably well.

3.2 Bivariate model CCI and SMI series

The next step is to combine the univariate model for the CCI with the series for the SMI. Before combining CCI and SMI in a bivariate model, a univariate model for the SMI is developed with the purpose to better understand the behaviour of this series. A model selection procedure, similar to the one conducted for the CCI series in Subsection 3.1, indicated that the observed series for the SMI can be modelled with a smooth trend model and a white noise component for the unexplained variation. No significant seasonal component or business cycle is established. There are no signs for outliers or level shifts. AR(1) and MA(1) components are not included since there is no serial correlation in the standardized innovations. These considerations led to a bivariate model for the CCI and SMI where the CCI contains a trend and a seasonal component and the SMI a trend component.

Tables A.2 and A.3 in the appendix contain an overview of goodness of fit statistics for the standardized innovations of the CCI and SMI respectively. There are no indications that the standardized innovations of both series deviate from a normal distributions. The null hypothesis of no serial correlation in the standardized innovations could not be rejected. The correlogram of the innovations for the SMI, however, show a non-significant seasonal pattern (not shown). The innovations of the SMI, also contain heteroscedasticity.

The disturbance terms of the trend of both series are correlated. Since the series for the SMI is available from June 2010, the model for the CCI also contains the last intervention for September 2011, but not the

outlier in September 2007 and the intervention in September 2008. As a result the following bivariate model is obtained:

$$\begin{pmatrix} I_t \\ X_t \end{pmatrix} = \begin{pmatrix} L_t^I \\ L_t^X \end{pmatrix} + \begin{pmatrix} S_t^I \\ 0 \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ 0 \end{pmatrix} + \begin{pmatrix} \upsilon_t^I \\ \upsilon_t^X \end{pmatrix}, \tag{3.9}$$

with L_t^I and L_t^X the smooth trend model as defined in (3.6) with covariance structure

$$\begin{aligned} \text{Cov}\left(\eta_t^I, \eta_{t'}^I\right) &= \begin{cases} \sigma_{\eta I}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \\ \text{Cov}\left(\eta_t^X, \eta_{t'}^X\right) &= \begin{cases} \sigma_{\eta X}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \\ \text{Cov}\left(\eta_t^I, \eta_{t'}^X\right) &= \begin{cases} \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}. \end{aligned}$$

In the last expression ρ_η denotes the correlation between the slope disturbances of the CCI and SMI. Furthermore, S_t^I is the seasonal effect defined by (3.7) and δ_t^{11} the intervention for September 2011 with β^{11} the corresponding regression coefficient. Finally, υ_t^I and υ_t^X are the disturbance terms for the CCI and SMI series and are defined as:

$$\begin{aligned} E(\upsilon_t^I) &= E(\upsilon_t^X) = 0, \\ \text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^I\right) &= \begin{cases} \sigma_{\upsilon I}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \\ \text{Cov}\left(\upsilon_t^X, \upsilon_{t'}^X\right) &= \begin{cases} \sigma_{\upsilon X}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \\ \text{Cov}\left(\upsilon_t^I, \upsilon_{t'}^X\right) &= 0 \text{ for all } t \text{ and } t'. \end{aligned}$$

If the model detects a strong correlation between the trends of the CCI and the SMI, then the trends of both series will develop into the same direction more or less simultaneously. In this case, the additional information from the SMI series will result in an increased precision of the estimates of the CCI figures. In the case of strong correlation between the disturbances of the trends, i.e., if $\rho_\eta \to 1$, the trends are said to be cointegrated. In that case, there is one underlying common trend that drives the evolution of the trends of the two observed series. To see this, it is noted that the covariance matrix of the slope disturbances is implemented as a singular value decomposition:

$$\text{cov}\begin{pmatrix} \eta_t^I \\ \eta_t^X \end{pmatrix} = \begin{pmatrix} \sigma_{\eta I}^2 & \sigma_{\eta I}\sigma_{\eta X}\rho_\eta \\ \sigma_{\eta I}\sigma_{\eta X}\rho_\eta & \sigma_{\eta X}^2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}. \tag{3.10}$$

Instead of estimating $\sigma_{\eta I}^2$, $\sigma_{\eta X}^2$, and ρ_η, parameters d_1, d_2, and a are estimated. If $d_2 \to 0$, it follows that $\rho_\eta \to 1$. In that case, the covariance matrix of the slope disturbances is of reduced rank and

both trends are driven by one common trend. This implies that the slope disturbances of both series simultaneously move up or down and that the slope disturbances of the SMI can be perfectly predicted from slope disturbances of the CCI by $\eta_t^X = a\eta_t^I$. Furthermore, the slope for the SMI series can be expressed as a linear combination of the slope for the CCI series as $R_t^X = aR_t^I + \overline{R}$. Similarly, the trend for the SMI series can be expressed as a linear combination of the trend for the CCI series as $L_t^X = aL_t^I + \overline{L} + \overline{R}t$. Note that $\overline{R}$ and $\overline{L}$ are constants that are derived from the estimated states at the last two time periods of the series.

Cointegration increases the precision of the estimated trend and signal of the CCI series, allows for formulating more parsimonious models, but could also be seen as an argument to replace the CCI series by the SMI series since both series are driven by and represent the same common trend. For a more detailed discussion about cointegration in the context of state space modelling, see Koopman, Harvey, Shephard and Doornik (2009, Sections 6.4 and 9.1).

3.3 Estimation of structural time series models

The general way to analyse a structural time series model is to express it in the so-called state space representation and apply the Kalman filter to obtain optimal estimates for the state variables, see e.g., Durbin and Koopman (2012). The software for the analysis and estimation of the time series models is developed in Ox in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman, Shephard and Doornik (2008).

All state variables are non-stationary and initialised with a diffuse prior, i.e., the expectation of the initial states are equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. In Ssfpack 3.0, an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997). Maximum likelihood (ML) estimates for the hyperparameters, i.e., the variance components of the stochastic processes for the state variables are obtained using a numerical optimization procedure (Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, Doornik, 2009). To avoid negative variance estimates, the log-transformed variances are estimated. More technical details about the analysis of state space models can be found in Harvey (1989) or Durbin and Koopman (2012).

Under the assumption of normally distributed disturbance terms, the Kalman filter can be applied to obtain optimal estimates for the state variables, see e.g., Durbin and Koopman (2012). The Kalman filter assumes that the variance and covariance terms are known in advance and are often referred to as hyperparameters. In practise, these hyperparameters are not known and are therefore substituted with their ML estimates. Estimates for state variables for period t based on the information available up to and including period t are referred to as the *filtered estimates*. They are obtained with the Kalman filter where the ML estimates for the hyperparameters are based on the complete time series. The filtered estimates of past state vectors can be updated, if new data become available. This procedure is referred to as smoothing and results in *smoothed estimates* that are based on the complete time series.

Standard errors of the Kalman filter estimates do not reflect the additional uncertainty of using the ML estimates for the unknown hyperparameters. Therefore, the estimates of the standard errors are too optimistic.

4 Results

4.1 Univariate model CCI series

The univariate analysis is based on model (3.8) from Section 3.1 applied to the series of the CCI obtained from December 2000 until March 2015. In Table 4.1, the ML estimates for the hyperparameters of the model are specified.

Table 4.1
Maximum Likelihood estimates hyperparameters univariate model CCI

Standard deviation	ML estimate
Trend (σ_η)	1.18
Seasonal (σ_ω)	0.0025
Measurement equation (σ_υ)	2.46

The average of the standard errors of the direct estimates for the CCI equals 1.21. The standard deviation of the disturbance terms of the measurement equation equals 2.46, as follows from Table 4.1. This illustrates that the population white noise dominates the variance of the measurement disturbance terms as mentioned by the choice of the variance structure for (3.4) in Section 3.1.

In the upper panel of Figure 4.1, the smoothed trend plus interventions are compared with the direct estimates for the CCI. In the lower panel of Figure 4.1, the smoothed signal, defined as trend plus interventions plus seasonals, are compared with the direct estimates for the CCI. In the series of the smoothed trend and interventions, the seasonal effect, the white noise of the population parameter and the sampling error are removed from the original series. It follows from Figure 4.1 that with the time series model a more stable estimate for the CCI can be obtained. The filtered trend plus interventions is compared with the smoothed estimates in Figure 4.2. This filtered series approximates what would be obtained in the production of official statistics if no revisions would be published. It follows that even in this case a considerable part of the high-frequency variation and seasonal fluctuations can be removed. Both figures illustrate that the Kalman filter provides plausible smoothed but also filtered imputations for the missing observation in October 2013.

Figure 4.3 shows the smoothed seasonal pattern of the CCI series. Since the seasonal effects are almost time invariant, the effects are displayed for the 12 months of one year only. There are clear significant negative effects in October, November and December and clear positive effects in January and August. The intention of the CCI is to measure a long-term confidence of respondents, since all questions refer to the respondents financial and economic situation over the last 12 month or the expectations for the future 12 months. The clear significant seasonal pattern, however, indicates that answers given by the respondents are clearly driven by a much shorter emotion, which is, among other things, subject to seasonal fluctuations.

In Figure 4.4, the standard error of the direct estimates for the CCI are compared with the standard errors of the filtered and smoothed trend plus interventions. The spikes in the standard error of the filtered and smoothed estimates are the result of the intervention variables and the missing observation in 2013. If at a certain point in time an intervention variable is activated, a new regression coefficient has to be estimated. This results in additional uncertainty in the model estimates, and shows up as a sudden peak in the standard

error of the filtered and smoothed trend. In 2013, one observation is missing, which also results in additional uncertainty since the state space model produces a prediction for this missing value.

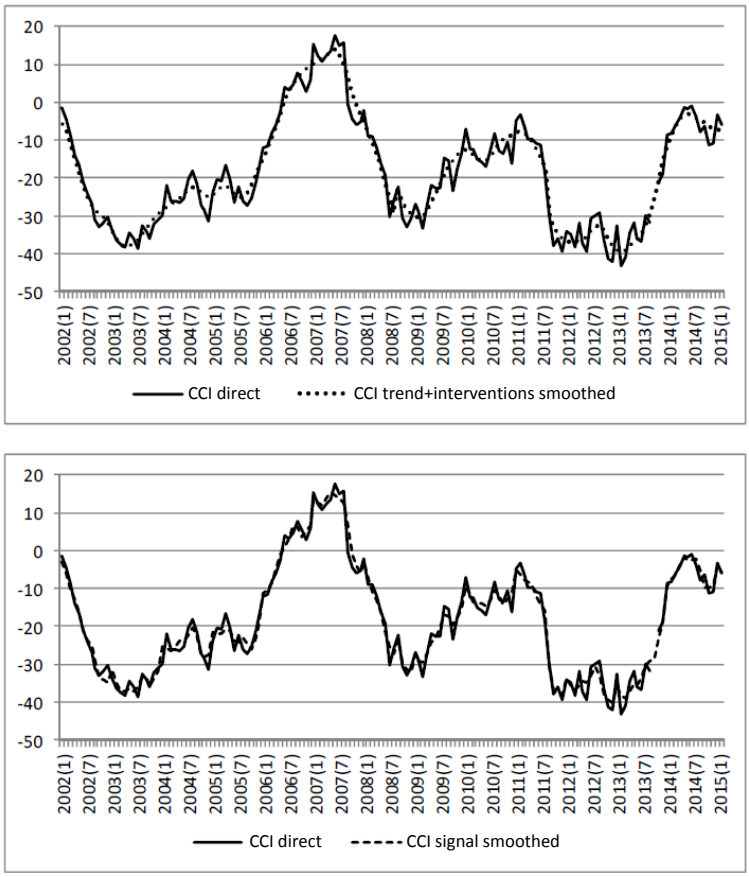

Figure 4.1 Smoothed trend plus interventions compared with direct estimates CCI (upper panel) and smoothed signal (trend plus intervention plus seasonal) compared with direct estimates CCI (lower panel).

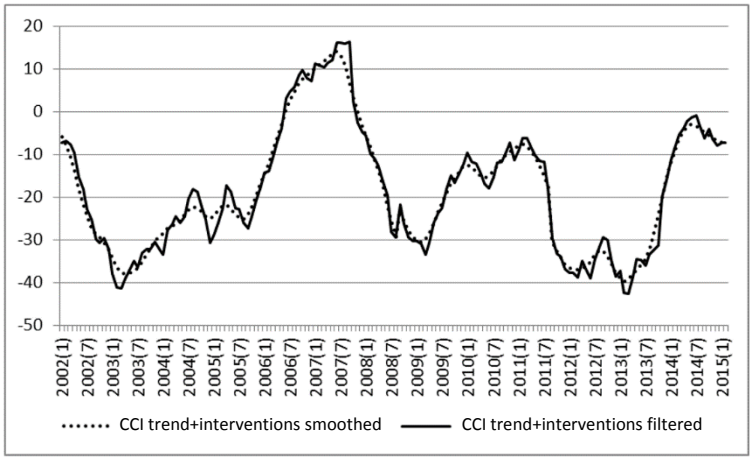

Figure 4.2 Filtered trend plus interventions compared with smoothed trend plus interventions CCI.

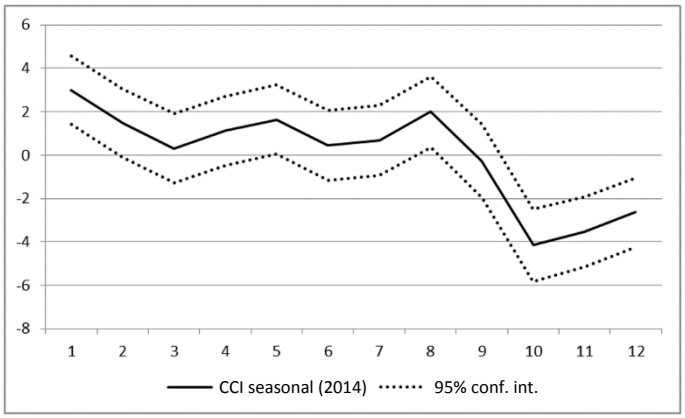

Figure 4.3 Smoothed seasonal pattern CCI for 2014.

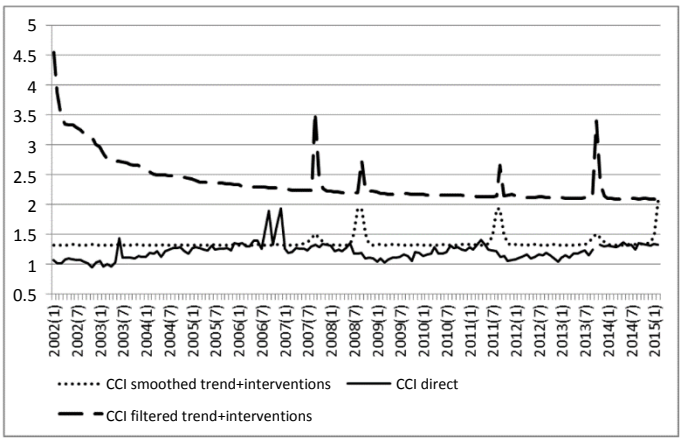

Figure 4.4 Standard error smoothed and filtered trend plus interventions compared with direct estimates CCI.

The standard errors of the smoothed estimates are slightly larger than the standard errors of the direct estimates. The standard errors of the filtered estimates are considerably larger than the standard errors of the direct estimates. This is a remarkable result. Filtered and smoothed estimates based on the time series model are based on a considerably larger set of information since sample information from preceding periods (in the case of filtered estimates) or the entire series (in the case of smoothed estimates) are used to obtain an optimal estimate for the monthly CCI. The direct estimates, on the other hand, are based on the observed sample in that particular month only. Most applications where structural time series models are applied as a form of small area estimation, result in substantive reductions of the standard error compared to the direct estimates, see e.g., van den Brakel and Krieg (2009, 2015) and Bollineni-Balabay, van den Brakel and Palm (2015, 2017).

The reason that in this application a time series modelling approach results in standard errors for filtered and smoothed times series model estimates that are larger than the standard errors of the direct estimates is a result of a large white noise component in the real population value of the CCI. Recall from Section 3.1 that the disturbance term of (3.8) contains two components; the sampling error and the unexplained high-frequency variation of the real population value, as expressed by (3.4). Recall from Table 4.1 that σ_υ is

equal to 2.46 and is twice as large as the average value of the standard errors of the direct estimates. This is a strong indication that the variance of the white noise component in the true population variable is of the same order as the variance of the sampling error. The direct estimator for the CCI derived in Section 2 considers the CCI in each particular month as a fixed but unknown variable. The variance of the direct estimator only measures the uncertainty since a small sample instead of the entire population is observed to estimate the CCI. It does not measure the high-frequency variation of the population value over time. This explains why the time series modelling approach does not result in a reduction of the standard error of the estimated CCI.

Although the gain in precision of level estimates obtained with the time series model is limited, the estimates for the trend are more stable as follows from Figures 4.1 and 4.2. A time series model will therefore still be useful to filter a more stable long term trend from the high-frequency variation in the population parameter and the sampling error. Because the state variables of the trend component of subsequent periods will have a strong positive correlation, more gain from the time series modelling approach can be expected by focussing on month-to-month changes, see e.g., Harvey and Chung (2000). Filtered estimates for the month-to-month change of the CCI are defined as

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\,\delta^{08}_{t} - \beta^{08}_{t-1|t}\,\delta^{08}_{t-1} + \beta^{11}_{t|t}\,\delta^{11}_{t} - \beta^{11}_{t-1|t}\,\delta^{11}_{t-1}, \tag{4.1}$$

where the notation $\Theta_{t|t'}$ stands for the estimate for state variable Θ for period t given the data observed until period t'. The outlier in 2007(9) is, naturally, removed from the signal. Furthermore, the regression coefficients are time invariant. Therefore, $\beta^{x}_{t|t} = \beta^{x}_{t-1|t}$ for $x = 08$ and 11. Since $t = 2008(9)$ and $t = 2011(9)$ are the months that δ^{08}_{t} and δ^{11}_{t} change form value, expression (4.1) can be simplified to

$$\Delta_{t|t} = L_{t|t} - L_{t-1|t} + \beta^{08}_{t|t}\,\tilde{\delta}^{08}_{t} + \beta^{11}_{t|t}\,\tilde{\delta}^{11}_{t}, \tag{4.2}$$

with $\tilde{\delta}^{08}_{t} = 1$ if $t = 2008(9)$ and $\tilde{\delta}^{08}_{t} = 0$ for all other periods and $\tilde{\delta}^{11}_{t} = 1$ if $t = 2011(9)$ and $\tilde{\delta}^{11}_{t} = 0$ for all other periods. Smoothed estimates for the month-to-month change of the CCI are defined as

$$\Delta_{t|T} = L_{t|T} - L_{t-1|T} + \beta^{08}_{t|T}\,\tilde{\delta}^{08}_{t} + \beta^{11}_{t|T}\,\tilde{\delta}^{11}_{t}. \tag{4.3}$$

To compare the month-to-month changes based on (4.2) and (4.3) with the direct estimates, the smoothed seasonal effects in (3.8) are subtracted from the direct estimates. The standard errors of the direct estimates are not corrected for this adjustment.

Figure 4.5 compares the direct estimates for the month-to-month change with the smoothed estimates (upper panel) and the filtered estimates (middle panel) obtained with the time series model. The lower panel compares the standard errors of the smoothed, filtered and direct estimates. The filtered and in particular the smoothed estimates for month-to-month change have a more stable pattern compared to the direct estimates. This is also reflected by the standard errors. The strong positive correlations of the states of the trend component between subsequent periods results in standard errors for filtered and smoothed estimates of the month-to-month change that are clearly smaller compared to the direct estimator. Exceptions are the two periods where a level intervention is required. Introducing a level shift results for a short period in an increased level of uncertainty.

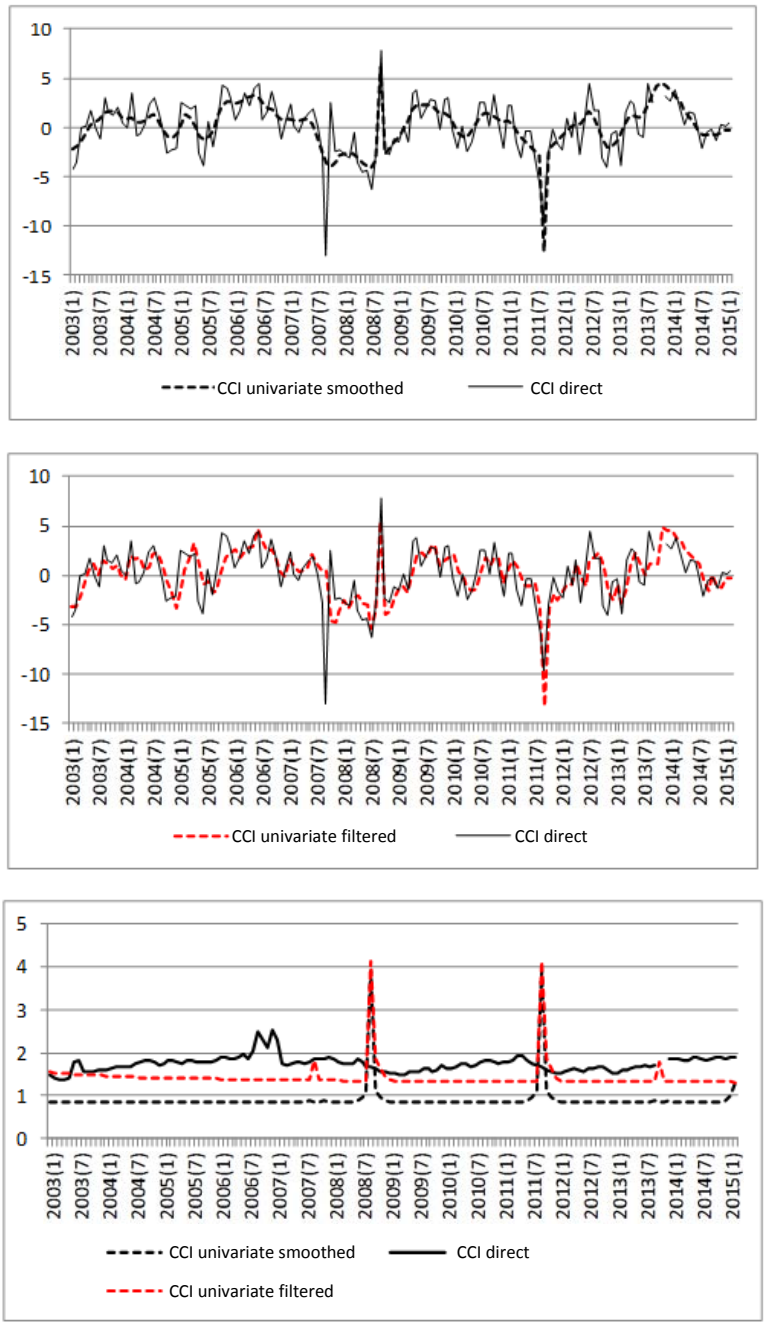

Figure 4.5 Comparison month-to-month change univariate model and direct estimates. Upper panel: smoothed estimates, middle panel: filtered estimates, lower panel standard errors.

The reduction in standard error is measured as the Mean Relative Difference in Standard Error (MRDSE), and is for filtered estimates defined as $\text{MRDSE} = 100/(T - t') * \sum_{t=t'}^{T} \left[\text{se}\left(\hat{\Delta}_t\right) - \text{se}\left(\Delta_{t|t}\right)\right] \Big/ \text{se}\left(\hat{\Delta}_t\right)$, with $\text{se}\left(\hat{\Delta}_t\right)$ the standard error for the direct estimate for the month-to-month change. The MRDSE for smoothed estimates is obtained by replacing $\text{se}\left(\Delta_{t|t}\right)$ for $\text{se}\left(\Delta_{t|T}\right)$. During the period observed from 2003(1), the MRDSE for smoothed estimates equals 47% and for the filtered estimates 17%.

4.2 Bivariate model for CCI and SMI series

In this section, the bivariate model (3.9) proposed in Section 3.2 is applied to the series of the CCI and SMI, which are available from June 2010 until March 2015. Note that the time series components for the CCI are re-estimated using the shorter series. Maximum likelihood estimates for the hyperparameters are specified in Table 4.2. The model detects a strong positive correlation of about 0.92 between the slope disturbances of the CCI and the SMI. There is, however, no indication that both trends are cointegrated and share one common trend. A likelihood ratio test is applied to further investigate the significance of the correlation between the slope disturbances in the bivariate model. If the correlation parameter is set to zero, the log likelihood drops from -229.9 to -233.9. The p – value of the corresponding likelihood ratio test equals 0.0047, indicating that the correlation between the trends of both series is clearly significantly different from zero and should not be removed from the bivariate model. If the correlation parameter is set equal to one (by choosing d_2 in (3.10) equal to zero), the log likelihood drops from -229.9 to -242.1. The p – value of the corresponding likelihood ratio test with one degree of freedom equals zero, indicating that the trends are not cointegrated.

Table 4.2
Maximum Likelihood estimates hyperparameters bivariate model CCI and SMI

Standard deviation	ML estimate
Trend CCI $(\sigma_{\eta I})$	1.25
Seasonal CCI (σ_{ω})	7.5E-6
Trend SMI $(\sigma_{\eta X})$	0.25
Measurement equation CCI $(\sigma_{\upsilon I})$	2.68
Measurement equation SMI $(\sigma_{\upsilon X})$	0.84
Correlation trend CCI and SMI (ρ_{η})	0.92

Figure 4.6 compares the smoothed estimates for the slope of the CCI (x-axis) and SMI (y-axis) under the model without correlation, the model with an ML estimate for the correlation $(\rho_{\eta} = 0.92)$ and the common trend model with $\rho_{\eta} = 1.0$. The model with uncorrelated slopes shows a clearly positive correlation between the slopes if both series are estimated independently (left panel Figure 4.6). This is picked up by the model that allows for correlation (mid panel Figure 4.6). There is however a clear deviation between the slopes of both series, which can be seen if the cross-plot of the model with a correlation estimated with ML (mid panel Figure 4.6) is compared with the cross-plot of a common factor model (right-panel Figure 4.6).

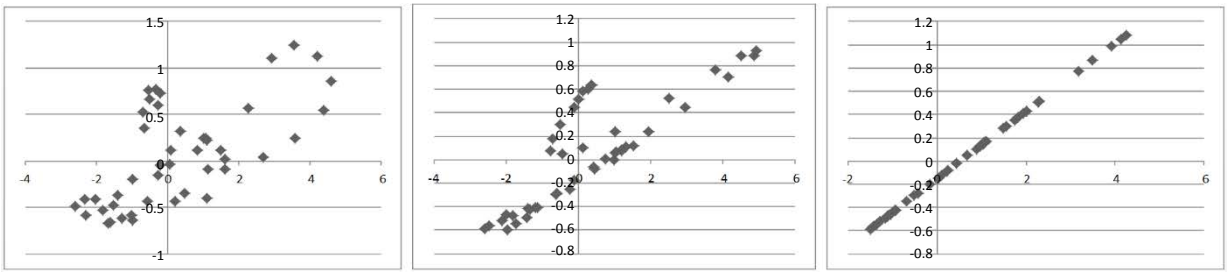

Figure 4.6 Cross-plot smoothed slopes CCI (x-axis) and SMI (y-axis) for a model without correlation (left panel), correlation estimated with ML (mid panel) and correlation set equal to one (right panel).

Figure 4.7 compares the observed SMI series with the smoothed trend obtained under the bivariate model. Figure 4.8 compares the direct estimates for the CCI series with the smoothed trend plus intervention under the univariate model and the bivariate model. As follows from Figure 4.8, the level and evolution of the smoothed estimates for the CCI series are almost identical under the univariate and bivariate models.

Figure 4.9 compares the standard errors of the direct estimates for the CCI series with the smoothed trend plus intervention under the univariate model and the bivariate model. For a fair comparison, the results for the univariate model and bivariate model are based on series of equal length. Therefore, the univariate model is re-estimated with the series from June 2010 until March 2015. As follows from Figure 4.9, the standard error under the bivariate model is slightly smaller compared to the standard error under the univariate model if both models are applied to series of equal length, as expected given the strong and significant positive correlation between the trend disturbance terms of both series. If, however, the univariate model is applied to the series available from December 2000, then the standard errors for the smoothed estimates under the univariate model are slightly smaller compared to the bivariate model as follows from Figure 4.10.

In conclusion, it follows that the bivariate model detects a strong correlation between the CCI and SMI series. Using the SMI series as an auxiliary series slightly improves the precision of the model based estimates for the CCI. Since the series of the CCI is nine years longer than the SMI series, the increased precision obtained with the auxiliary series is compensated in the univariate model with the additional information in the CCI series available before 2010.

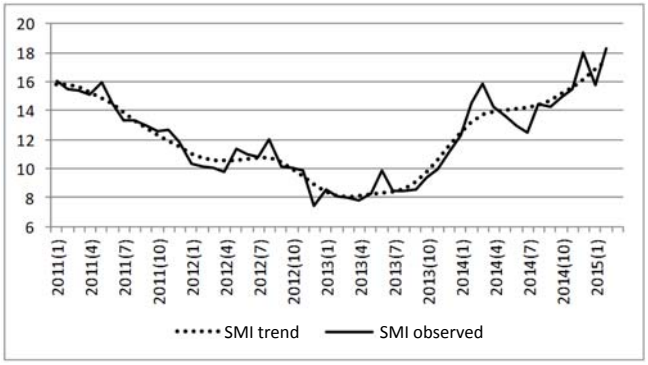

Figure 4.7 Observed series and smoothed trend SMI.

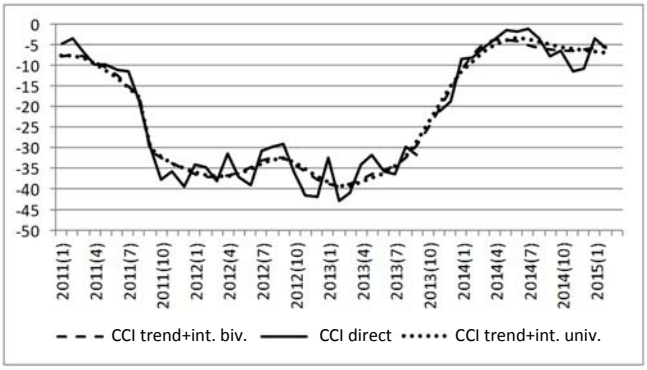

Figure 4.8 CCI comparison of the direct estimates and smoothed trend plus intervention under the bivariate and univariate models for CCI.

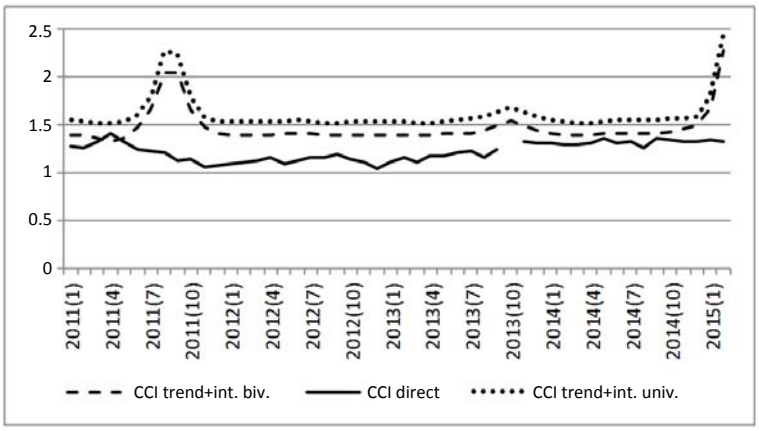

Figure 4.9 CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate models for CCI if both models are applied to a series of equal length (June 2010-March 2015).

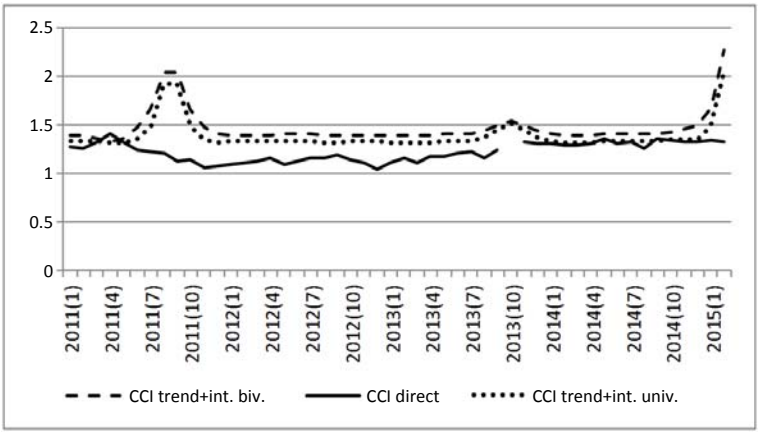

Figure 4.10 CCI comparison of standard errors direct estimates and smoothed trend plus intervention under the bivariate and univariate models for CCI if the univariate model is applied to the complete CCI series (December 2000).

The upper panel of Figure 4.11 compares the direct estimates for the month-to-month change with the smoothed estimates obtained with the univariate and bivariate time series models (both based on the series observed from June 2010). The lower panel compares the standard errors of these estimates. During the period observed from 2011(1), the MRDSE for smoothed estimates under the univariate model equals 39% and under the bivariate model 43%. The MRDSE for filtered estimates under the univariate model equals 7% and under the bivariate model 14%. As in the case of the univariate model, the time series modelling approach results in more stable and more precise estimates for the month-to-month change. The use of the SMI series slightly improves the precision of the month-to-month changes compared to the univariate model.

Once the direct estimate for the CCI for month t becomes available, the additional value of the SMI series is limited to improve a time series estimate for the CCI for month t. A drawback of sample surveys, however, is that they generally are less timely compared to social media sources. The additional value of the SMI becomes more clear when the higher frequency of this series is used to produce early predictions or nowcasts for the CCI with the bivariate state space model. If during month t or directly at the end of month t a first early prediction for the CCI is required, the univariate model can only produce a one-step-ahead prediction. As soon as during month t or at the end of month t results for the SMI series become available, the bivariate model exploits the strong correlation between the series to make a more precise prediction for the CCI, already before the direct estimate for month t becomes available.

To illustrate the additional value of the SMI in a nowcast procedure for the CCI, we compare in the upper panel of Figure 4.12, the one-step-ahead predictions for the trend plus intervention of the CCI series obtained with the univariate model with the estimate obtained with the bivariate model if the SMI for month t is available but the direct estimate of the CCI is still missing. The smoothed estimates for the trend plus intervention of the CCI obtained with the univariate model are included as a benchmark. In the lower panel the standard errors of these three estimates are compared.

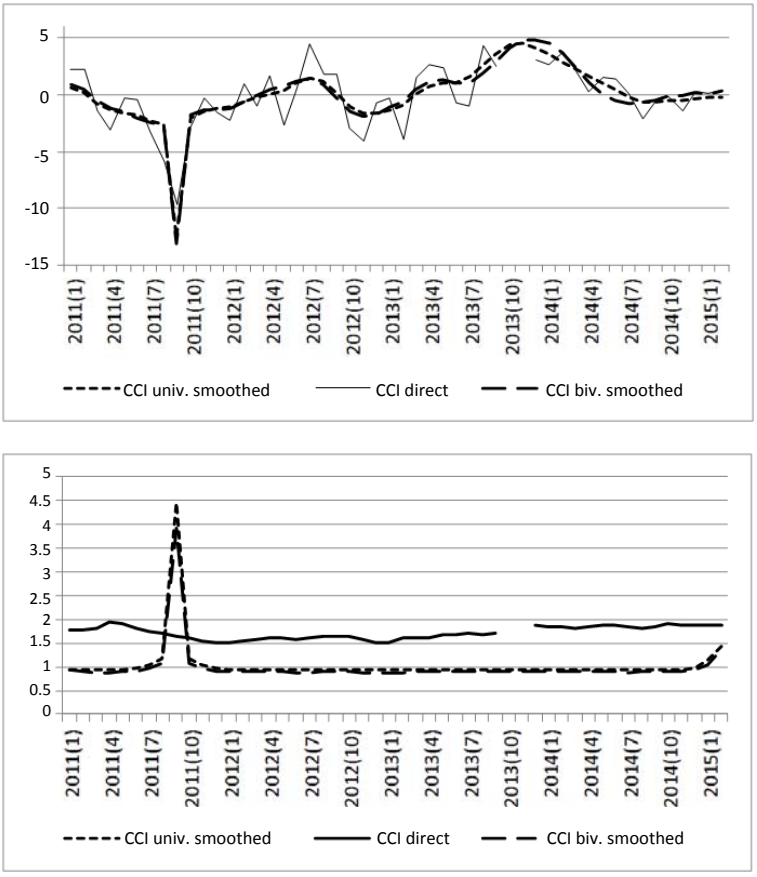

Figure 4.11 Comparison month-to-month change bivariate model, univariate model and direct estimates. Upper panel: smoothed estimates, lower panel standard errors.

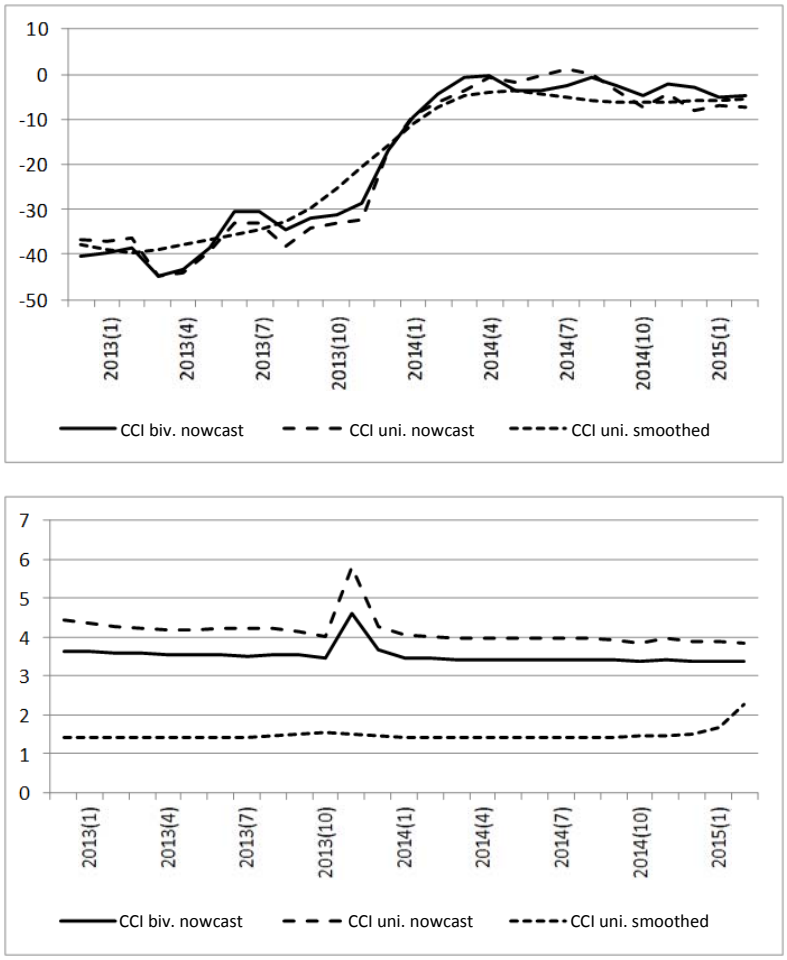

Figure 4.12 Comparison estimates for trend plus intervention CCI series; one-step-ahead prediction univariate model (CCI uni. nowcast), bivariate model if the SMI for month *t* is available but the direct estimate of the CCI is missing (CCI biv. nowcast) and smoothed estimates with the univariate model (CCI uni. smoothed). Upper panel compares point estimates. Lower panel compares standard errors.

If the smoothed estimates obtained with the univariate model are used as a benchmark, the Mean Absolute Relative Difference (MARD) between nowcasts and smoothed estimates is used as a measure for the size of the revision and is defined as $\text{MARD} = 100/(T - t') * \sum_{t=t'}^{T} \left|\theta_{t|T} - \theta_{t|t-1}\right| \Big/ \left|\theta_{t|T}\right|$, where $\theta_t = L_t + \beta^{11}\delta_t^{11}$ denotes the trend plus intervention of the CCI series. Based on the months observed from $t = 2013(1)$ the MARD for nowcasts obtained with the univariate model equals 35% and for the bivariate model 31%. This shows that the size of the revisions is a bit smaller and thus more stable with nowcasts for the CCI with the bivariate model. The difference in precision between the nowcasts obtained with the univariate model and the bivariate model are measured with the MRDSE and is in this case defined as $\text{MRDSE} = 100/(T - t') * \sum_{t=t'}^{T} \left[\text{se}\left(\theta_{t|t-1}^{\text{uni}}\right) - \text{se}\left(\theta_{t|t-1}^{\text{biv}}\right)\right] \Big/ \text{se}\left(\theta_{t|t-1}^{\text{biv}}\right)$. Based on the months observed from $t = 2013(1)$ the difference in precision of both nowcasts based on this MRDSE equals 17%. Figure 4.12 as well as the MARD and the MRDSE illustrate that the SMI improves the stability and precision of nowcasts for the CCI.

5 Discussion

For decades, national statistical institutes relied on probability sampling in the production of official statistics. This approach is based on a sound theory to draw valid statistical inference for large finite target populations based on relatively small random samples. Over the last decades, more and more alternative data sources, such as administrative and big data, have become available and the question is raised how to use these data sources in the production of official statistics. An important question is how results obtained with these sources can be generalized to an intended finite target population. Since the data generating process is generally unknown, it is not obvious how to draw valid inference with such data sources.

In this paper, the question is addressed how administrative and big data sources can be used in the production of official statistics. In the most extreme approach, survey data are replaced by related alternative data sources, running the risk of introducing e.g., selection bias. Since most surveys are conducted repeatedly, a time series modelling approach is proposed to investigate to which extent related alternative data sources reflect a similar evolution compared to the series obtained with a repeated survey. With a multivariate state space model, the correlation between the underlying unobserved components of both series can be modelled. In the case that components of the time series model are cointegrated, there are strong indications that both data sources are driven by the same underlying factor. This could be used as an argument that an alternative source can replace existing surveys since they reflect the same evolution of a process, generally at a different level.

The theory underlying probability sampling for finite population inference is stronger than reliance on the concept of cointegration. Series obtained from social media or Google Trends are selected by maximizing the correlation with the series from the sample survey and does not necessarily measure the same concept as the survey. There is no guarantee that this correlation is based on true causality and that the correlation will remain to exist in the future. Sampling theory, in contrast, provides a rigid mathematical theory showing that under a correct sampling strategy, i.e., the right combination of a probability sample with an approximately design-unbiased estimator, results in valid statistical inference for intended target populations.

Even in the case of cointegrated series, an extensive model evaluation, e.g., by some form of cross validation, will be required to assure that the alternative data source is a valid replacement. See in this context also Eichler (2013) for a discussion about the use of Granger causality for causal inference in multiple time series data. Instead of replacing a periodic survey for related data sources, they can be used in a multivariate time series modelling approach as an auxiliary series to improve the precision of the direct estimates or period-to-period change of the direct estimates obtained with a periodic survey. Another important benefit with big data sources is to use the higher frequency of these data sources to make more precise early predictions or nowcasts if in real time the survey estimate is not yet available but the covariate is already available. The time series model applied in this paper, initially proposed by Harvey and Chung (2000), is a generic approach for a model-based estimation procedure for periodic surveys. There are of course also issues with survey sampling. For example, continuously declining response rates and data

collection modes that does not reach the intended target population result in selection bias either. In this case, cointegration with a related series derived from social media might be indication that there are similarities between the selection bias in the non-probabilistic big data sources and the non-response selection and coverage bias in a survey sample as pointed out by Baker et al. (2013).

In the application to the CCI, the time series modelling approach does not decrease the variance of the direct estimator if it is used for making level estimates. The reason is that the standard error of the time series model reflects the sampling error and the white noise of the population parameter. The standard error of the direct estimator only reflects the sampling error. In the case of the CCI, the variance component of the white noise of the population parameter is as large as the variance of the sampling error. The state space approach is still useful for producing official figures of the CCI, since it filters a more stable trend of the respondents opinion about the economic climate from the observed series of direct estimates. The situation, however, becomes different if the time series model is used to estimate month-to-month change. The stable trend estimates are the result of a strong positive correlation between the trend estimates between subsequent periods. As a result the standard errors of month-to-month change obtained with the time series model are clearly smaller than those of the direct estimates. Standard errors of smoothed month-to-month changes are about 47% smaller than those of the direct estimates. Standard errors of the filtered estimates are about 17% smaller than the standard errors of the direct estimates.

Using the SMI as an auxiliary series in a bivariate state space model slightly reduces the standard error of the model estimates of the CCI. However, since the available series of the SMI is relative short, the reduction obtained with this auxiliary series does not outweigh the loss of information in the CCI series that is observed in the period before the SMI became available. However, since both series reflect a similar evolution and social media is rapidly available, the SMI proved to be useful as an auxiliary series in the bivariate model to produce more reliable nowcasts for the CCI in real time at the moment that the SMI becomes available but the CCI is not available yet. In this application the SMI reduces the standard errors of the CCI in a nowcasting procedure with about 17%.

The question can be raised whether the SMI in its current operationalization measures the same concept as the CCI attempts and how the full potentials of social media or other big data sources can be used to measure consumer confidence better than the current CCI and SMI. Instead of constructing a social media index by taking the difference between positive and negative classified messages, an SMI could be constructed by looking at the concepts of the questions used for the CCI. If for example consumer confidence is measured by the amount of purchases of expensive goods during the last 12 months, or with the tendency of households to buy expensive goods, social media indices should be constructed that measure internet search for such goods (cars, houses, white goods, etc.) as well as actual purchases of such goods during the previous months. The strong advantage of this approach is that now actual behaviour of households is measured directly, while a survey measures it indirectly inducing more measurement error. This might eventually result in cointegrated series that measure similar concepts and further improves or even replaces the CCI.

Acknowledgements

The authors are grateful to the Associate Editor and the reviewers for careful reading of a former draft of this paper and providing constructive comments, which significantly improved the content of this paper. The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

Appendix A

Model diagnostics

Table A.1
Univariate model (3.8) for CCI 172-24 obs

Diagnostic	Value	$p-$value	95% conf. int.	
			L	U
Log-likelihood	-464			
Mean std. innovations	0.0152			
Variance std. innovations	1.0851			
Skewness std. innovations	0.0276			
Kurtosis std. innovations	2.8901			
Bowman-Shenton test[1] on normality in the std. innovations	0.0926	0.955		
Ljung-Box test[2] on serial correlation in std. innovations	24.108	0.287		
Durban-Watson test[3] on serial correlation of std. innovations $(T = 148)$	2.082		1.68	2.32
$F-$test[4] on heteroscedasticity of std. innovations $(df_{\text{num}} = df_{\text{denom}} = 60)$	0.913		0.60	1.67

Table A.2
Bivariate model (3.9) for CCI 57-24 obs

Diagnostic	Value	$p-$value	95% conf. int.	
			L	U
Log-likelihood	-230			
Mean std. innovations	-0.0872			
Variance std. innovations	0.9777			
Skewness std. innovations	0.0982			
Kurtosis std. innovations	2.545			
Bowman-Shenton test[1] on normality in the std. innovations	0.3382	0.844		
Ljung-Box test[2] on serial correlation in std. innovations	18.060	0.645		
Durban-Watson test[3] on serial correlation of std. innovations $(T = 33)$	2.133		1.32	2.68
$F-$test[4] on heteroscedasticity of std. innovations $(df_{\text{num}} = df_{\text{denom}} = 15)$	0.783		0.35	2.86

1) Bowman-Shenton statistic: χ^2_2 distribution.
2) Ljung-Box test statistic for serial correlation in the first 24 lags: χ^2_{21} distribution.
3) Durban-Watson test statistic approximated with $N(2, 4/T)$.
4) $F-$statistic: $F^{df_{\text{num}}}_{df_{\text{denom}}}$ distribution.

Table A.3
Bivariate model (3.9) for SMI 57 -12 obs

Diagnostic	Value	p – value	95% conf. int.	
			L	U
Log-likelihood	-230			
Mean std. innovations	0.0954			
Variance std. innovations	1.0437			
Skewness std. innovations	-0.1311			
Kurtosis std. innovations	2.5331			
Bowman-Shenton test[1] on normality in the std. innovations	0.5377	0.764		
Ljung-Box test[2] on serial correlation in std. innovations	24.208	0.283		
Durban-Watson test[3] on serial correlation of std. innovations $(T = 45)$	2.028		1.42	2.58
$F-$test[4] on heteroscedasticity of std. innovations $(df_{\text{num}} = df_{\text{denom}} = 20)$	0.329		0.41	2.46

1) Bowman-Shenton statistic: χ^2_2 distribution.
2) Ljung-Box test statistic for serial correlation in the first 24 lags: χ^2_{21} distribution.
3) Durban-Watson test statistic approximated with $N(2, 4/T)$.
4) $F-$statistic: $F^{df_{\text{num}}}_{df_{\text{denom}}}$ distribution.

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, first published online September 26, 2013, doi:10.1093/jssam/smt008.

Bell, W.R. (2005). Some considerations of seasonal adjustment variances. Census Bureau. Paper available at https://www.census.gov/ts/papers/jsm2005wrb.pdf.

Bell, W.R., and Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, 16, 2, 195-215. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14535-eng.pdf.

Binder, D.A., and Dick, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 1, 29-45. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1989001/article/14579-eng.pdf.

Binder, D.A., and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 2, 239-253. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14533-eng.pdf.

Blight, B.J.N., and Scott, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-66.

Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 1073-1076.

Bollineni-Balabay, O., van den Brakel, J.A. and Palm, F. (2015). Multivariate state-space approach to variance reduction in series with level and variance breaks due to sampling redesigns. Accepted for publication in *Journal of the Royal Statistical Society, Series A*.

Bollineni-Balabay, O., van den Brakel, J.A. and Palm, F. (2017). State space time series modelling of the Dutch Labour Force Survey: Model selection and mean squared errors estimation. *Survey Methodology*, 43, 1, 41-67. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14819-eng.pdf.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin de l'Institut International de Statistique*, 22, Supplement to Book 1, 6-62.

Buelens, B., Burger, J. and van den Brakel, J.A. (2015). Predictive inference for non-probability samples: A simulation study. Discussion paper 2015-13, Statistics Netherlands, Heerlen.

Cochran, W. (1977). *Sampling Theory*. New York: John Wiley & Sons, Inc.

Daas, P., and Puts, M. (2014a). Big data as a source of statistical information. *The Survey Statistician*, 69, 22-31.

Daas, P., and Puts, M. (2014b). Social media sentiment and consumer confidence. European Central Bank Statistics paper series No. 5, Frankfurt Germany.

Doornik, J.A. (2009). An Object-oriented Matrix Programming Language Ox 6. London: Timberlake Consultants Press.

Durbin, J., and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods, Second Edition*. Oxford: Oxford University Press.

Eichler, M. (2013). Causal inference with multiple time series: Principles and problems. *Philosophical transactions of the Royal Statistical Society A*, 371, issue 1997.

Feder, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Statistica Neerlandica*, 55, 182-199.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14, 333-362.

Harvey, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge: Cambridge University Press.

Harvey, A.C., and Chung, C.H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-339.

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.

Koopman, S.J., Shephard, N. and Doornik, J.A. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*, London: Timberlake Consultants Press.

Koopman, S.J., Harvey, A., Shephard, N. and Doornik, J.A. (2009). *STAMP 8.2*, London: Timberlake Consultants Press.

Lind, J.T. (2005). Repeated surveys and the Kalman filter. *Econometrics Journal*, 8, 418-427.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Perdreschi, D., Rinzivillo, S., Pappalardo, L. and Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31, 263-281.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 1-135.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Rubin-Bleuer, S. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 2, 149-163. Paper available at http://www.statcan.gc.ca/pub/12-001-x/1993002/article/14458-eng.pdf.

Pfeffermann, D., and Sverchkov, M. (2014). Estimation of mean squared error of X-11-ARIMA and other estimators of time series components. *Journal of Official Statistics*, 30, 811-838.

Pfeffermann, D., and Tiller, R. (2006). Small area estimation with state space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, 339-348.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Scott, A.J., and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of* the *American Statistical Association*, 69, 674-678.

Scott, A.J., Smith, T.M.F. and Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review/Revue Internationale de Statistique*, 45, 13-28.

Tam, S.-M. (1987). Analysis of repeated surveys using a dynamic linear model. *International Statistical Review/Revue Internationale de Statistique*, 55, 1, 63-73.

Tiller, R.B. (1992). Time series modelling of sample survey data from the U.S. current population survey. *Journal of Official Statistics*, 8, 149-166.

van den Brakel, J.A., and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 35, 2, 177-190. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2009002/article/11040-eng.pdf.

van den Brakel, J.A., and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41, 2, 267-296. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14231-eng.pdf.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 211-234
Statistics Canada, Catalogue No. 12-001-X

Decomposition of gender wage inequalities through calibration: Application to the Swiss structure of earnings survey

Mihaela-Catalina Anastasiade and Yves Tillé[1]

Abstract

This paper proposes a new approach to decompose the wage difference between men and women that is based on a calibration procedure. This approach generalizes two current decomposition methods that are re-expressed using survey weights. The first one is the Blinder-Oaxaca method and the second one is a reweighting method proposed by DiNardo, Fortin and Lemieux. The new approach provides a weighting system that enables us to estimate such parameters of interest like quantiles. An application to data from the Swiss Structure of Earnings Survey shows the interest of this method.

Key Words: Blinder-Oaxaca; Gender wage discrimination; Quantiles; Reweighting; Wages.

1 Introduction

Wage discrimination can be based on different criteria, such as gender, race or religion. Gender wage discrimination occurs when a man and a woman receive different remuneration for a job that requires the same qualifications or which implies identical productivity (see, for instance Neumark, 1988; Gardeazabal and Ugidos, 2005). Since a quantification of discrimination is required in order to assess its magnitude, the topic has awakened the interest of statisticians. The original technique proposed by Blinder (1973) and Oaxaca (1973) estimates how much of the difference between the average wages of men and the average wages of women is due to discrimination. However, in general, there is an uneven allocation of women and men among jobs (see, for instance Bielby and Baron, 1986). If the members of one of these two groups, usually women, are concentrated in lower paying jobs, the difference in average wages might not be of great relevance. So instead of analysing the discrimination level in average wages, it might be interesting to see if discrimination occurs uniformly in all types of jobs. A detailed reference of the different statistical papers devoted to the estimation of discrimination can be found in Fortin, Lemieux, and Firpo (2011).

While there are many decomposition methods available in the literature, only two of them will be discussed in this paper. These two methods are not presented in their original forms, but by taking into account survey weights. They are the Blinder-Oaxaca method (hereafter, BO) and the semi-parametric method developed by DiNardo, Fortin, and Lemieux (1996) (hereafter, DFL). Originally, the BO method analysed the difference between the average wages of men and the average wages of women. However, it does not allow for an analysis of the wage differences for other parameters, such as quantiles. The original DFL method addresses this issue. Its starting point is a logistic model where, for each observation, the probability of being a man or a woman is modelled as a function of the observed characteristics. The ratio of these probabilities is used to construct a reweighting factor. Its aim is to approach the distribution of the characteristics of women to the distribution of characteristics of men. By having similar distributions of the

1. Mihaela-Catalina Anastasiade and Yves Tillé, Institute of Statistics, University of Neuchâtel, 51 Avenue de Bellevaux, 2000 Neuchâtel, Switzerland. E-mail: mihaela.anastasiade@unine.ch; yves.tille@unine.ch.

characteristics, an estimation of the discrimination level at parameters other than the mean is achievable. However, the reweighting factor may have a large variance in cases where one or more characteristics are good predictors of the gender. Moreover, the reweighted distribution of characteristics of women may not match the distribution of characteristics of men. We address the problems related to the two methods through a calibration approach. The idea behind calibration is the same as that of the DFL method. It consists of approaching the distribution of characteristics of women to that of men, in order to estimate the discrimination level along the entire wage distributions.

The paper is structured as follows: after the definition of the notation in Section 2, the BO decomposition is re-expressed with the use of survey data in Section 3. Sampling weights are taken into account in order to correct for the difference between the sample and the population of interest. Therefore, the decomposition will be termed "weighted BO". The key concept of women's counterfactual wage distribution is also presented. It is defined as the wage distribution of women if they had the same characteristics as men. Next, we discuss the use of the counterfactual wage distribution in the wage difference decomposition. In Section 4, the DFL method is developed, again using survey weights. Since the original method does not include survey weights, it will be termed "weighted DFL". Next in Section 5, a new approach to compute the counterfactual wage distribution is proposed, using the calibration method (Deville and Särndal, 1992). The use of two particular cases of calibration are discussed. These are the linear calibration and the raking-ratio calibration. The first case yields the same result as the weighted BO method for average wages. The second case has a similar approach to the weighted DFL method, but without assuming a logistic model. In other words, the proposed technique can be regarded as a generalization of the two methods discussed above. Section 6 includes an overview of the dataset used as well as descriptive statistics on the observed wages. A brief description of the model used and the results obtained using the discussed methods are presented. Finally, Section 7 summarizes the conclusions and in Appendix B, the computation of the variance of the counterfactual wage is shown.

2 Problem and notation

The question of interest is the estimation of the wage differences between women and men, more specifically, how much of this difference is attributable to discrimination. Assume there is a finite population U of size N that can be divided into two subpopulations, women and men, denoted by $U_h, h \in \{F, M\}$, of size N_h. Additionally, a random sample S is drawn from U, which contains both women and men. Sample S is selected by means of a sampling design $p(s) = \Pr(S = s)$ for any $s \subset U$, where

$$p(s) \geq 0 \quad \text{and} \quad \sum_{s \subset U} p(s) = 1.$$

Sample S can be split into two subsamples, $S_h, h \in \{F, M\}$, women and men, such that $S = \cup S_h$. The variable of interest, denoted by y, is in this case the logarithm of the wage. The totals of the variable of interest in the two subpopulations are given by

$$Y_h = \sum_{k \in U_h} y_k, h \in \{F, M\},$$

where y_k is the logarithm of the wage of the k^{th} individual. Since not all units in the subpopulations are observed, the totals can be estimated by

$$\hat{Y}_h = \sum_{k \in S_h} d_k y_k, h \in \{F, M\},$$

where d_k is a sampling weight assigned to the k^{th} unit of the sample. Sampling weights are obtained after several statistical treatments (for example, adjustment for non-response).

The population means of the logarithms of the wages are given by

$$\bar{Y}_h = \frac{1}{N_h} \sum_{k \in U_h} y_k, h \in \{F, M\},$$

and can be estimated by

$$\hat{\bar{Y}}_h = \frac{\sum_{k \in S_h} d_k y_k}{\sum_{k \in S_h} d_k}, h \in \{F, M\}.$$

Moreover, assume that for each k^{th} individual in either of the two subsamples, there is a vector of p auxiliary variables denoted by

$$\mathbf{x}_k = \left(x_{k1}, \ldots, x_{kj}, \ldots, x_{kp}\right)^\top \in \mathbb{R}^p.$$

This vector is supposed to be known for each unit selected in the sample. The auxiliary variables contain some characteristics of the individual, for instance the age, the education level or the seniority level. They can be quantitative or qualitative variables, thus x_{kj} can be a categorical variable or a quantity. Also assume that the first auxiliary variable is a constant, i.e., $x_{k1} = 1,$ for all $k \in U$.

The totals of these auxiliary variables at the subpopulation level are given by

$$\mathbf{X}_h = \sum_{k \in U_h} x_k, h \in \{F, M\}.$$

Using the weights d_k defined above, these two totals can be estimated by

$$\hat{\mathbf{X}}_h = \sum_{k \in S_h} d_k \mathbf{x}_k, h \in \{F, M\}.$$

Vectors of average values can be analogously estimated. The average values at the subpopulation levels are given by

$$\bar{\mathbf{X}}_h = \frac{1}{N_h} \sum_{k \in U_h} \mathbf{x}_k, h \in \{F, M\},$$

and estimated by

$$\hat{\bar{\mathbf{X}}}_h = \frac{\sum_{k \in S_h} d_k \mathbf{x}_k}{\sum_{k \in S_h} d_k}, h \in \{F, M\}. \quad (2.1)$$

3 The weighted BO decomposition

3.1 The decomposition

Using the setup in Section 2, the findings of Blinder (1973) and Oaxaca (1973) are summarized in the context of sampling theory, namely by using sampling weights. Assume that in each sample, a linear relationship is suitable between the p characteristics that are available and the logarithm of the wage. A regression is done separately in each subpopulation U_h, $h = \{M, F\}$. At the subpopulation level, the values of the regression coefficients are given by

$$\boldsymbol{\beta}_h = \left(\sum_{k \in U_h} \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U_h} \mathbf{x}_k y_k .$$

They can be estimated from the sample by

$$\hat{\boldsymbol{\beta}}_h = \left(\sum_{k \in S_h} d_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_h} d_k \mathbf{x}_k y_k, \tag{3.1}$$

where d_k are the sampling weights. The regression coefficients $\hat{\boldsymbol{\beta}}_h$ are called the group wage structure or the returns on characteristics and they represent the contribution of each characteristic to the wage.

Result 1 *A sufficient condition to obtain the following equalities*

$$\bar{Y}_h = \bar{\mathbf{X}}_h^\top \boldsymbol{\beta}_h \quad \textit{and} \quad \hat{\bar{Y}}_h = \hat{\bar{\mathbf{X}}}_h^\top \hat{\boldsymbol{\beta}}_h$$

is that there exists a vector $\boldsymbol{\varsigma} \in \mathbb{R}^p$, *such that* $\boldsymbol{\varsigma}^\top \mathbf{x}_k = 1$, *for all* $k \in U_h$.

Since it is assumed that $x_{k1} = 1$ for all $k \in U$, with $\boldsymbol{\varsigma}^\top = (1\ 0 \ldots 0)$, the equality is always fulfilled. The proof of Result 1 can be found in Appendix A. Putting together the result above, equations (2.1) and (3.1), the average difference between the wages of two groups can be written as

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right). \tag{3.2}$$

The difference between average wages of the groups contains two elements: an explained part, also called the *composition effect* $\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F \right)^\top \hat{\boldsymbol{\beta}}_F$ and an unexplained part, or the *structure effect* $\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F \right)$. The former encompasses differences in characteristics between the two groups. The latter is the difference in the returns on characteristics between the two groups, the part that is not attributable to objective factors (Oaxaca, 1973; Blinder, 1973). It is obtained using characteristics as a proxy for productivity. The estimation of the *structure effect* is the central element of this paper. Equation (3.2) has the same elements as the one proposed by Oaxaca (1973) and Blinder (1973). The methodology applied to obtain the estimated average values and coefficients differs from the traditional regression technique. The BO method uses the estimated regression coefficients obtained through ordinary least squares (OLS) and the vectors of average values of the observed explanatory variables. The proposed approach takes into

account the survey weights. However, the weighted BO method is the same as the original BO method if the sampling weights are all equal to 1.

3.2 A note on the structure effect

The two elements in equation 3.2 have different names across the literature. The first one, whose denomination we retained as *composition effect* is also termed *endowments effect*. The second one, which we call *structure effect* is also found in the literature as *unexplained residual, price effect, sex effect, calculated effect* or *unequal treatment* (Weichselbaumer and Winter-Ebmer, 2006). Using the BO method, the structure effect is an estimation of the discrimination level. However, discrimination is an intricate phenomenon that might not be always fully observed. Unobserved variables, selection bias or some mechanisms on the labour market can help to increase the explained part of the wage difference. Moreover, Weichselbaumer and Winter-Ebmer (2005) note two potential issues regarding the chosen model. First, if the characteristics chosen in the linear model are themselves subject to discrimination, then the resulting structure effect will be over-estimated. Second, if the characteristics are not a proper measure of the productivity, then again, the structure effect might be under- or over-estimated. Weichselbaumer and Winter-Ebmer (2006) warn about the legitimacy of the characteristics as productivity indicators, since "wages may also be determined by bargaining power, compensating differentials or efficiency wages". However, for simplicity, in what follows, we will assume that there are no such issues and that the estimated structure effect is the result of discrimination on the labor market. Moreover, we do not examine sample selection bias or other mechanisms underlying the distribution of men and women in certain jobs.

3.3 The counterfactual wage distribution

In general, the counterfactual wage distribution is an artificial distribution obtained by using the characteristics of a group to estimate the wages of another group (see, for instance Bourguignon, Ferreira, and Leite, 2002). Examples of counterfactual distributions are found in DiNardo et al. (1996) or DiNardo (2002). The term $\hat{\bar{\mathbf{X}}}_M\hat{\boldsymbol{\beta}}_F$ that appears in equation (3.2) is called the women's counterfactual average wage. It is interpreted as the estimated average wage of women if they had the same average characteristics as men and if their return on characteristics remained unchanged. Women's counterfactual wage distribution is obtained by using the characteristics of men $(\mathbf{X}_M)$ and the wage structure of women $(\boldsymbol{\beta}_F)$. In terms of interpretation, it is the wage distribution of women, if they had the same characteristics as men.

Using Result 1 from the previous section, women's counterfactual mean wage equals

$$\bar{Y}_{F|M} = \bar{\mathbf{X}}_M^\top\boldsymbol{\beta}_F,$$

and is estimated from the sample by

$$\hat{\bar{Y}}_{F|M} = \hat{\bar{\mathbf{X}}}_M^\top\hat{\boldsymbol{\beta}}_F,$$

where $\hat{\bar{\mathbf{X}}}_M$ are estimated in equation (2.1) and $\hat{\boldsymbol{\beta}}_F$ are the coefficients estimated by means of equation (3.1). With this notation, the BO decomposition given in (3.2) is re-expressed as

$$\Delta = \hat{\bar{Y}}_M - \hat{\bar{Y}}_F = \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F + \hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right). \quad (3.3)$$

3.4 Using the counterfactual distribution to estimate the composition and the structure effects

Building the counterfactual average wage allows for the estimation of the two effects that make up the wage difference at the average levels. From equation (3.3), the composition effect is equal to

$$\left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_F\right)^\top \hat{\boldsymbol{\beta}}_F = \left(\hat{\bar{Y}}_{F|M} - \hat{\bar{Y}}_F\right).$$

The composition effect can be interpreted as the difference between what women would earn on average if they had the characteristics of men and what they actually earn. Thus, it reflects the inequality due to the differences in characteristics. The structure effect in equation (3.3) is equal to

$$\hat{\bar{\mathbf{X}}}_M^\top \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F\right) = \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}\right).$$

The structure effect is the difference between the actual average wage of men and what women would earn if they had the average characteristics of men and their wage own structure. The equations above express the composition and structure effects at the average levels, since this is the limitation of the BO method. The next section presents a method that allows for the construction of the entire counterfactual distribution. This in turn results in the ability of estimating the composition and structure effects along the entire wage distribution.

4 The weighted DFL method

4.1 The method

The method proposed by DiNardo et al. (1996) uses a reweighting function by which women's distribution of characteristics is rendered similar to men's distribution of characteristics. The reweighted distribution is the women's counterfactual distribution of characteristics. The DFL method is presented through the use of survey weights in order to take the sampling design into account.

The reweighting function is equal to

$$\psi(\mathbf{x}_k) = \frac{\Pr(D_{Mk} = 1 | \mathbf{x}_k) / \Pr(D_{Mk} = 1)}{\Pr(D_{Mk} = 0 | \mathbf{x}_k) / \Pr(D_{Mk} = 0)},$$

where $D_{Mk} = 1$ if individual k is a man and $D_{Mk} = 0$ otherwise and $\mathbf{x}_k$ is the vector of observed characteristics for individual k. Obviously, $\Pr(D_{Mk} = 1 | \mathbf{x}_k)$ and $\Pr(D_{Mk} = 0 | \mathbf{x}_k)$ must be estimated. For this type of estimation, DiNardo et al. (1996) suggested the use of a logit or a probit model. Using the information from the sample,

$$\hat{\psi}(\mathbf{x}_k) = \frac{\widehat{\Pr}(D_{Mk} = 1 | \mathbf{x}_k) / \widehat{\Pr}(D_{Mk} = 1)}{\widehat{\Pr}(D_{Mk} = 0 | \mathbf{x}_k) / \widehat{\Pr}(D_{Mk} = 0)}. \tag{4.1}$$

Using the reweighting factor, women's counterfactual wage mean is estimated by

$$\hat{\bar{Y}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}, \tag{4.2}$$

and women's counterfactual means of characteristics by

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)}. \tag{4.3}$$

The estimated reweighting factor defined in equation (4.1) will be equal to

$$\hat{\psi}(\mathbf{x}_k) = \hat{a} \ \exp\left(\mathbf{x}_k^\top \hat{\boldsymbol{\gamma}}\right),$$

where $\hat{\boldsymbol{\gamma}}$ is the estimation of $\boldsymbol{\gamma}$ from the sample using empirical likelihood and $\hat{a}$ is the ratio of estimated proportions of women and men. It is given by:

$$\hat{a} = \frac{\widehat{\Pr}(D_{Mk} = 0)}{\widehat{\Pr}(D_{Mk} = 1)} = \frac{\sum_{k \in S_F} d_k}{\sum_{k \in S_M} d_k}.$$

Since the DFL method is presented taking the survey weights into account, the reweighting factor ψ_k will be multiplied by $d_k, k \in S_F$. This resulting factor will be termed "weighted DFL factor". Women's estimated counterfactual wage mean can be re-expressed as

$$\hat{\bar{Y}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) y_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right)}. \tag{4.4}$$

Women's counterfactual means of characteristics are estimated as

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}} = \frac{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k}{\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k)} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}^\top \hat{\boldsymbol{\gamma}}\right)}.$$

Through the use of the reweighting factor, the counterfactual coefficients in the women's sample are given by

$$\boldsymbol{\beta}_F^{\text{DFL}} = \left(\sum_{k \in U_F} \psi(\mathbf{x}_k) \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in U_F} \psi(\mathbf{x}_k) \mathbf{x}_k y_k,$$

and estimated by

$$\hat{\boldsymbol{\beta}}_F^{\text{DFL}} = \left(\sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \sum_{k \in S_F} d_k \hat{\psi}(\mathbf{x}_k) \mathbf{x}_k y_k. \tag{4.5}$$

The coefficients above have to be computed, because under the same condition as in Result 1, women's counterfactual wage mean defined in (4.2) is given by

$$\hat{\bar{Y}}_{F|M}^{\text{DFL}} = \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}}.$$

The BO decomposition formula can now be expressed as

$$\begin{aligned} \hat{\bar{Y}}_M - \hat{\bar{Y}}_F &= \left(\hat{\bar{Y}}_{F|M}^{\text{DFL}} - \hat{\bar{Y}}_F\right) + \left(\hat{\bar{Y}}_M - \hat{\bar{Y}}_{F|M}^{\text{DFL}}\right) \\ &= \left(\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}} - \hat{\bar{\mathbf{X}}}_F^{\top} \hat{\boldsymbol{\beta}}_F\right) + \left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}}\right), \end{aligned} \tag{4.6}$$

where $\hat{\boldsymbol{\beta}}_M$ and $\hat{\boldsymbol{\beta}}_F$ are defined in (3.1). The first term of equation (4.6) is the composition effect and the second one the structure effect.

4.2 Further decomposition of the structure effect

As Fortin et al. (2011) note, the purpose of the DFL reweighting factor is to render the distribution of women's characteristics identical to that of men. This implies that the means of the auxiliary variables in the two groups should be equal. However, with the DFL method, it is not the case. Indeed,

$$\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}} \neq \hat{\bar{\mathbf{X}}}_M \tag{4.7}$$

(see, for instance, Fortin et al. 2011; Donzé, 2013). The reweighting factor thus fails to match the two distributions perfectly.

The structure effect in equation (4.6) can be further divided in the following elements

$$\left(\hat{\bar{\mathbf{X}}}_M^{\top} \hat{\boldsymbol{\beta}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}\top} \hat{\boldsymbol{\beta}}_F^{\text{DFL}}\right) = \hat{\bar{\mathbf{X}}}_M^{\top} \left(\hat{\boldsymbol{\beta}}_M - \hat{\boldsymbol{\beta}}_F^{\text{DFL}}\right) + \left(\hat{\bar{\mathbf{X}}}_M - \hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}}\right) \hat{\boldsymbol{\beta}}_F^{\text{DFL}}, \tag{4.8}$$

where $\hat{\bar{\mathbf{X}}}_{F|M}^{\text{DFL}}$ and $\hat{\boldsymbol{\beta}}_F^{\text{DFL}}$ are defined in equations (4.3) and (4.5), respectively (Fortin et al. 2011). The first element of the right-hand side of equation (4.8) is the *pure effect* and the second the *residual effect* or the *total reweighting error* (Fortin et al. 2011). The pure effect is the actual unexplained part of the wage difference. The residual effect contains the misfit of the model, in other words, what the reweighting factor fails to match between men's and women's distribution of characteristics. This method allows for the construction of a counterfactual wage distribution. This in turn allows for the comparison between this new distribution and the observed wage distributions of women and men. The drawback of the method is that it may happen that at least one characteristic is a good predictor of the gender (for instance, the economic sector). This implies that $\Pr(D_{Mk} = 1 | \mathbf{x}_k)$ may get close to 1 and that the reweighting factor will take take on a large value (Fortin et al. 2011). This obviously leads to a large variance of the factor. This will be shown in Section 8.

5 The calibration approach

5.1 The calibration method

The calibration method was introduced by Deville and Särndal (1992). The idea behind the technique is to make use of the information known at the population level on some auxiliary variables to estimate a

function of a variable of interest. Usually, the auxiliary variables and the variable of interest are correlated. The resulting estimates are consistent and efficient.

Assuming that the sampling weights d_k are available and that the totals of auxiliary information at the population level given by

$$\mathbf{X} = \sum_{k \in U} \mathbf{x}_k ,$$

are known, new weights $w_k , k \in S$ should be constructed, such that the following constraint (or calibration equation) is respected

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k . \tag{5.1}$$

The weights are determined by solving in $\boldsymbol{\lambda}$ the calibration equations that become

$$\sum_{k \in S} w_k \mathbf{x}_k = \sum_{k \in S} d_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k ,$$

where $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right)$ is the calibration function. The resulting calibration estimation of Y is

$$\hat{Y} = \sum_{k \in S} d_k y_k F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right). \tag{5.2}$$

In what follows, we will use the linear case, where the pseudo-distance function is the chi-square distance and the calibration function is given by $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = 1 + \mathbf{x}_k^\top \boldsymbol{\lambda}.$ In the second case, we will use the raking-ratio, which uses the Entropy pseudo-distance and where the calibration function is given by $F_k \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right) = \exp \left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right).$

5.2 Calibration of women's characteristics on the men's characteristics

Suppose that for all the units of the sample, there is a given sampling weight d_k. In the current context, the auxiliary variables that are used in the calibration process are some selected characteristics measured for every individual. The aim is to 'divert' the calibration technique in order to compute a weighting system that adjusts the totals of the auxiliary variables of women on the totals of men. The variable of interest is the logarithm of the wage.

In the women sample, new weights w_k close to d_k are computed, such that $\sum_{k \in S_F} G \left(w_k , d_k \right)$ is minimized. The following calibration equation is satisfied

$$\sum_{k \in S_F} w_k \mathbf{x}_k = \hat{\tilde{\mathbf{X}}}_M , \tag{5.3}$$

where the vector $\hat{\tilde{\mathbf{X}}}_M$ stores the totals of men's characteristics adjusted on the total of the weights of the women over the total of the weights of the men.

$$\hat{\tilde{\mathbf{X}}}_M = \frac{\sum_{k \in S_F} d_k}{\sum_{k \in S_M} d_k} \sum_{k \in S_M} d_k \mathbf{x}_k .$$

Dividing the calibration equation (5.3) by $\sum_{k\in S_F} d_k$ yields

$$\frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} d_k} = \frac{\hat{\hat{\mathbf{X}}}_M}{\sum_{k\in S_F} d_k} = \hat{\hat{\mathbf{X}}}_M . \tag{5.4}$$

So with the new weights w_k, the new women's means of characteristics are equal to those of men. Another interesting equality is

$$\sum_{k\in S_F} w_k = \sum_{k\in S_F} d_k , \tag{5.5}$$

which holds because $x_{k1} = 1, k \in S_M$ and calibration is performed on it. If

$$\hat{\hat{\tilde{\mathbf{X}}}}_M = \frac{\sum_{k\in S_F} w_k \mathbf{x}_k}{\sum_{k\in S_F} w_k},$$

by putting together equations (5.4) and (5.5), this means that

$$\hat{\hat{\tilde{\mathbf{X}}}}_M = \hat{\hat{\mathbf{X}}}_M . \tag{5.6}$$

Women's counterfactual wage mean estimator is thus

$$\hat{\hat{\bar{Y}}}_{F|M} = \frac{\sum_{k\in S_F} w_k y_k}{\sum_{k\in S_F} d_k} = \frac{\sum_{k\in S_F} w_k y_k}{\sum_{k\in S_F} w_k} .$$

5.3 Linear calibration

Result 2 *Women's counterfactual wage mean obtained using linear calibration is equal to the counterfactual wage mean obtained using the weighted BO method, i.e.,* $\hat{\hat{\bar{Y}}}_{F|M} = \hat{\hat{\mathbf{X}}}^\top \hat{\boldsymbol{\beta}}_F$.

Proof

In order to determine the vector $\boldsymbol{\lambda}$ in the case when the chi-squared pseudo-distance is used, the following equation must be solved

$$\begin{aligned}\hat{\hat{\mathbf{X}}}_M &= \sum_{k\in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k\in S_F} d_k \mathbf{x}_k \left(1 + \mathbf{x}_k^\top \boldsymbol{\lambda}\right) \\ &= \sum_{k\in S_F} d_k \mathbf{x}_k + \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right) \boldsymbol{\lambda}.\end{aligned}$$

Thus,

$$\boldsymbol{\lambda} = \left(\sum_{k\in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top\right)^{-1} \left(\hat{\hat{\mathbf{X}}}_M - \sum_{k\in S_F} d_k \mathbf{x}_k\right) = \mathbf{T}^{-1}\left(\hat{\hat{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right),$$

where

$$\mathbf{T} = \sum_{k \in S_F} d_k \mathbf{x}_k \mathbf{x}_k^\top .$$

Thus

$$w_k = d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = d_k \left\{1 + \mathbf{x}_k^\top \mathbf{T}^{-1}\left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)\right\}.$$

Using the result from the previous equation, the numerator of expression (5.2) becomes

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k \\ &= \sum_{k \in S_F} d_k y_k + \left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k , \end{aligned} \tag{5.7}$$

where $\hat{Y}_{F|M}^{\mathrm{LC}}$ denotes the total of the logarithm of the wage in the women sample, when the total is constructed using the chi-squared pseudo-distance. Let

$$\hat{\boldsymbol{\beta}}_F = \mathbf{T}^{-1} \sum_{k \in S_F} d_k \mathbf{x}_k y_k .$$

Vector $\hat{\boldsymbol{\beta}}_F$ has already been defined in the same way in equation (3.1) for the weighted BO method. Equation (5.7) is rewritten as

$$\begin{aligned} \hat{Y}_{F|M}^{\mathrm{LC}} &= \sum_{k \in S_F} d_k y_k + \left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{Y}_F + \left(\hat{\tilde{\mathbf{X}}}_M - \hat{\mathbf{X}}_F\right)^\top \hat{\boldsymbol{\beta}}_F \\ &= \hat{\tilde{\mathbf{X}}}_M^\top \hat{\boldsymbol{\beta}}_F , \end{aligned} \tag{5.8}$$

because under the condition of Result 1, $\hat{\mathbf{X}}_F^\top \hat{\boldsymbol{\beta}}_F = \hat{Y}_F$. By dividing (5.8) by $\sum_{k \in S_F} w_k$, Result 2 is obtained.

Using the chi-squared pseudo-distance, the resulting weights have no bounds. This means that the calibration weights might be negative. Even though this calibration instance yields the same results as the BO method for average wages, we advocate for the use of an instance that gives nonnegative weights.

5.4 Raking-ratio calibration

The second instance of calibration uses the entropy pseudo-distance. It is also known as “raking-ratio” calibration. Using the entropy pseudo-distance, equation (5.3) becomes

$$\hat{\tilde{\mathbf{X}}}_M = \sum_{k \in S_F} d_k \mathbf{x}_k F\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) = \sum_{k \in S_F} d_k \mathbf{x}_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right). \tag{5.9}$$

This resulting system of equations cannot be solved analytically. However, the value of $\boldsymbol{\lambda}$ can be found through the Newton-Raphson algorithm.

The equation (5.2) can be now written as

$$\hat{Y}_{F|M}^{\text{RRC}} = \sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k,$$

where $\hat{Y}_{F|M}^{\text{RRC}}$ denotes the total of the logarithm of the wage in the women sample, when the total is constructed using the raking-ratio calibration. The counterfactual wage mean of women is written as

$$\hat{\bar{Y}}_{F|M}^{\text{RRC}} = \frac{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) y_k}{\sum_{k \in S_F} d_k \exp\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right)}.$$

The equation above is very similar to equation (4.4). The only difference lies in the estimation of the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$. The vector $\boldsymbol{\lambda}$ contains the Lagrangian multipliers solving equation 5.9 under constraint (5.1), while the vector $\boldsymbol{\gamma}$ is found through maximum likelihood.

After computing the calibration weights w_k defined in (5.3) and by using the information in equation (5.6), it results that

$$\hat{\bar{\mathbf{X}}}_M = \hat{\bar{\mathbf{X}}}_{F|M}^{\text{RRC}} = \frac{\sum_{k \in S_F} \mathbf{x}_k w_k}{\sum_{k \in S_F} w_k},$$

which ensures that the residual part of the structure effect defined in equation (4.8) will equal 0. This is a solution to the problem shown in Section 4.3. This instance of calibration also remedies the issue of the negative weights that may arise when using the chi-squared pseudo-distance.

6 Application to the Swiss Structure of Earnings Survey

6.1 Data description

The dataset used contains information collected in 2008 by the Swiss Federal Statistical Office from a survey called Survey on Earnings Structure. A questionnaire was sent to public and private organizations from the secondary and tertiary sectors to collect information on particular aspects. These aspects include the size of the organization, employment contract types and employee remuneration within the organization. The questionnaire was filled in by an authorized member of the organization and not by employees. This enhances data reliability and makes it less prone to approximations. The analyses that follow were restricted to the private sector. The valid observations that were included were the individuals with no missing values, who worked more than one hour per week and whose difference between the age and the work experience was greater than or equal to 15 (according to the Swiss employment laws, this represents the legal minimum

age to be eligible to work). Thus, 29,048 cases were excluded from the original dataset. The final dataset contains 647,139 men and 435,507 women. The sampling weights are also provided in the dataset by the Swiss Federal Statistical Office, therefore no treatment or computation of these weights were done in this application.

In the next tables, the values expressed in Swiss francs are given in parentheses. However, the figures are plotted using the logarithms of the wages. The values are obtained taking the survey weights into consideration.

Table 6.1 contains the median and wage averages for the entire sample and for women and men.

Table 6.1
Wage mean and median computed for the entire dataset, women and men, in Swiss francs

	Mean	Median
Entire dataset	6,977	5,905
Women	5,843	5,220
Men	7,725	6,346

Both the wage mean and the median values of men are above the values in the entire dataset, while those of women are below. Table 6.2 shows the distribution of women and men in low and high paying jobs. The weighted quantiles of the wage of the entire dataset are computed on the first row. The following two lines show the cumulative proportions of women and men who earn less than the value of the quantile.

Table 6.2
Weighted quantiles of the logarithm of the wage and proportions of women and men who earn less than the value that represents a particular quantile of the wage computed for the entire dataset (values in Swiss francs are given in parantheses)

	Quantile										
	1%	10%	20%	30%	40%	50%	60%	70%	80%	90%	99%
Logarithm of wage	7.89 (2,683)	8.27 (3,897)	8.39 (4,412)	8.50 (4,905)	8.59 (5,400)	8.68 (5,905)	8.78 (6,488)	8.89 (7,233)	9.03 (8,380)	9.27 (10,667)	10.09 (24,202)
Cumulative proportion of women	0.02	0.17	0.32	0.43	0.53	0.63	0.72	0.81	0.89	0.96	1
Cumulative proportion of men	0.006	0.06	0.12	0.21	0.31	0.42	0.52	0.63	0.74	0.86	0.99

While 43% of women have a wage of under CHF 4,905 (as opposed to only 21% of men), there are only 11% of women who earn between CHF 8,380 and CHF 24,202 (compared to 25% of men). Moreover, 63% of women earn below the median value of the wage of the entire dataset, compared to only 42% of men. The potential generating mechanisms of this allocation should be investigated. Nevertheless, it is not the purpose of this paper. For a closer insight into the distribution of the wages in each sample, Table 6.3 displays the weighted quantiles of the logarithms of the wages of women and men, as well as the difference between them. A surprising value of the difference between the wages is observed at the quantile of order 1%. It is expected that these jobs fall into the type of jobs that do not require extensive qualifications or high

education levels. While only 0.6% of men occupy such positions (see Table 6.3), they earn more than the 2% of women who have similar jobs. Figure 6.1 shows the data presented in Table 6.3 below in a graphical form.

Table 6.3
Wages of women and men and the difference between wages of men and women, in terms of logarithms (values in Swiss francs are given in parantheses)

	Quantile										
	1%	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**99%**
Women	7.80	8.19	8.30	8.38	8.47	8.56	8.66	8.76	8.88	9.06	9.67
	(2,432)	(3,602)	(4,005)	(4,344)	(4,756)	(5,220)	(5,743)	(6,353)	(7,154)	(8,577)	(15,761)
Men	8.01	8.36	8.49	8.58	8.67	8.76	8.86	8.98	9.14	9.38	10.26
	(3,000)	(4,259)	(4,850)	(5,344)	(5,820)	(6,346)	(7,012)	(7,908)	(9,291)	(11,905)	(28,571)
Difference	0.21	0.17	0.19	0.21	0.20	0.20	0.20	0.22	0.26	0.33	0.59
	(568)	(657)	(845)	(1,000)	(1,064)	(1,126)	(1,269)	(1,555)	(2,137)	(3,328)	(12,810)

The distance between the two sets of points increases toward the higher-level quantiles, which means that the differences between the wages become higher. It has to be established how much of these differences are not attributable to differing characteristics of women and men. As a final graphical evidence of wage inequalities, Figure 6.2 shows the distributions of the logarithm of the wages of women and men.

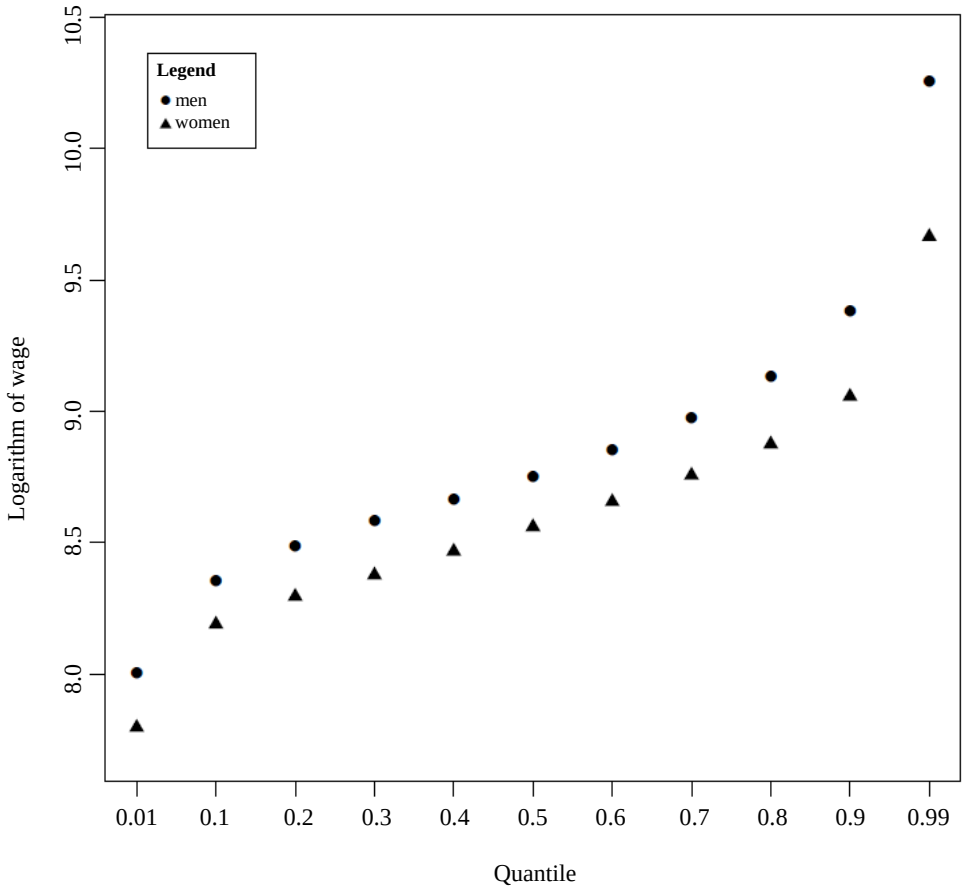

Figure 6.1 Weighted quantiles of the logarithm of the wages of women and men.

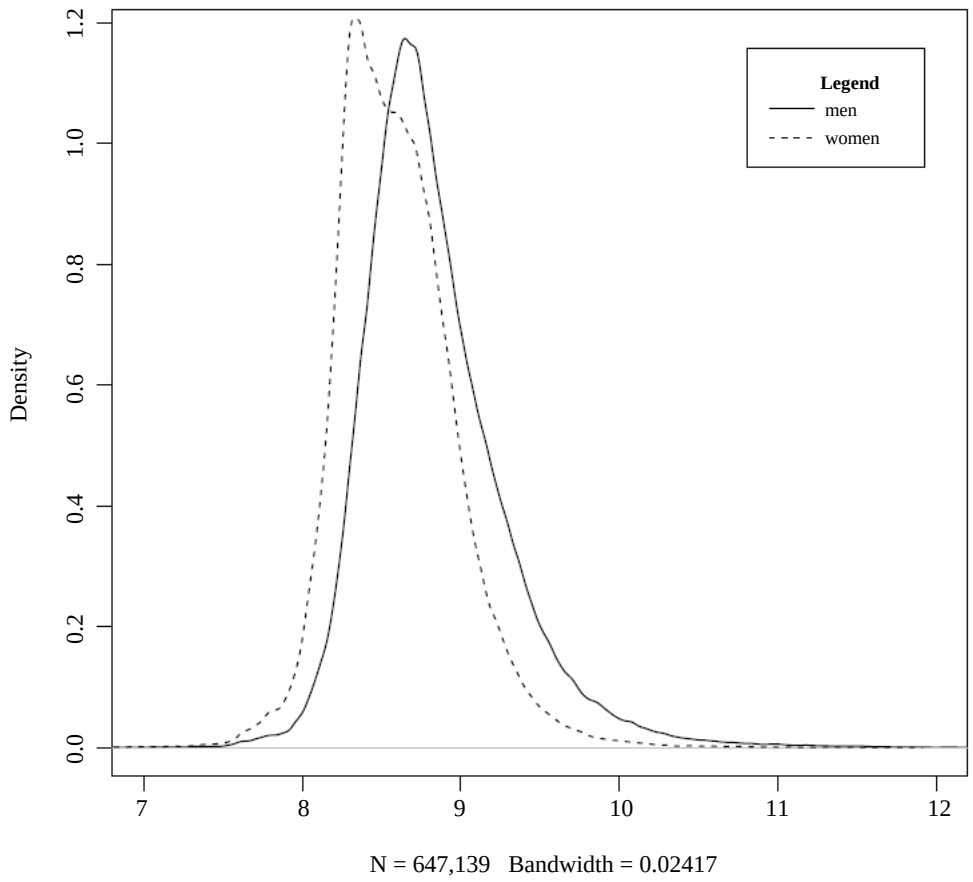

Figure 6.2 Estimated densities of the logarithms of wages of women and men.

6.2 The model

The regression model includes eight explanatory variables:

- education level : nominal variable with 9 categories indicating the highest educational degree attained;
- number of years of service in the current position (proxy for work experience);
- qualification requirements : ordinal variable with 4 levels indicating the level of qualification required for the position;
- region of the institution: nominal variable with 7 categories;
- economic sector: nominal variable with 10 categories;
- degree of occupation - the occupation rate of the employee (if the value is 1, then the employee works full-time);
- age: the actual age;
- the square of the age: the square of the age is also included, because it has been observed that the wage increases until a certain age and decreases afterwards (see, for instance Williams, 2010).

The model was selected from a number of models with several variables using the minimum AIC criterion. The dependent variable is the logarithm of the standardized wage. By standardized wage of an individual, we mean the wage computed for that individual if they worked full-time. This variable is provided by the Swiss Federal Statistical Office in the dataset, therefore no computation was done by the authors.

6.3 Weights and counterfactual distributions

This section only includes results in terms of logarithms. When using the BO method, the difference between average wages of men and women is 0.23, out of which only 0.09 represent the explained part and 0.14 the unexplained part. The results obtained through the methods presented above are compared. The calibration method through the chi-squared pseudo-distance is denoted as "linear", the calibration through the Kullback-Leibler divergence as "raking-ratio" and the method proposed by DiNardo et al. (1996) adjusted to take the survey weights into consideration as "DFL". First, Table 6.4 shows the minimum and the maximum values of the weights, as well as the standard deviations, obtained using the linear calibration, the raking-ratio calibration and the weighted DFL method.

Table 6.4
Weights minimum, maximum and standard deviation

Method	Minimum	Maximum	Standard deviation
Linear	-39.06	319.8	4.97
Raking-ratio	0.0011	904.7	6.79
Weighted DFL	0.0022	804.4	6.16

The linear case yields the same results as the weighted BO method. However, as seen in Table 6.4, this particular case yields negative weights. There were 69,553 such weights (14.59%). The raking-ratio alternative always yields positive weights, however, the standard deviation of the weights is higher. The weighted DFL factor has a smaller standard deviation than the weights obtained by the raking-ratio calibration method. There are 1,319 cases where the conditional probability of being a man is larger than 0.98. Originally, the DFL factor is multiplied by the ratio between the sum of sampling weights of women and the sum of sampling weights of men. Since $\hat{a}$ is smaller than one, the reweighting factor will shrink. If on the other hand, $\hat{a}$ is larger than one (for instance, for sectors such as the public sector), the reweighting factor might be larger. Table 6.5 shows the structure effect estimated at the average levels of the wages. The two calibration approaches yield equal structure and composition effects. Using the DFL reweighting factor, results in a slightly lower structure effect and a higher composition effect than the other two methods.

Table 6.5
Estimated composition and structure effects in the difference in mean averages

Method	Composition effect	Structure effect	Total
Linear	0.09	0.14	0.23
Raking-ratio	0.09	0.14	0.23
Weighted DFL	0.10	0.13	0.23

Given that negative weights are obtained in the first case of calibration, the corresponding estimated density can not be graphically represented. Only women's counterfactual wage distributions constructed using the raking-ratio and the DFL reweighting factor are constructed. They are presented in Figure 6.3.

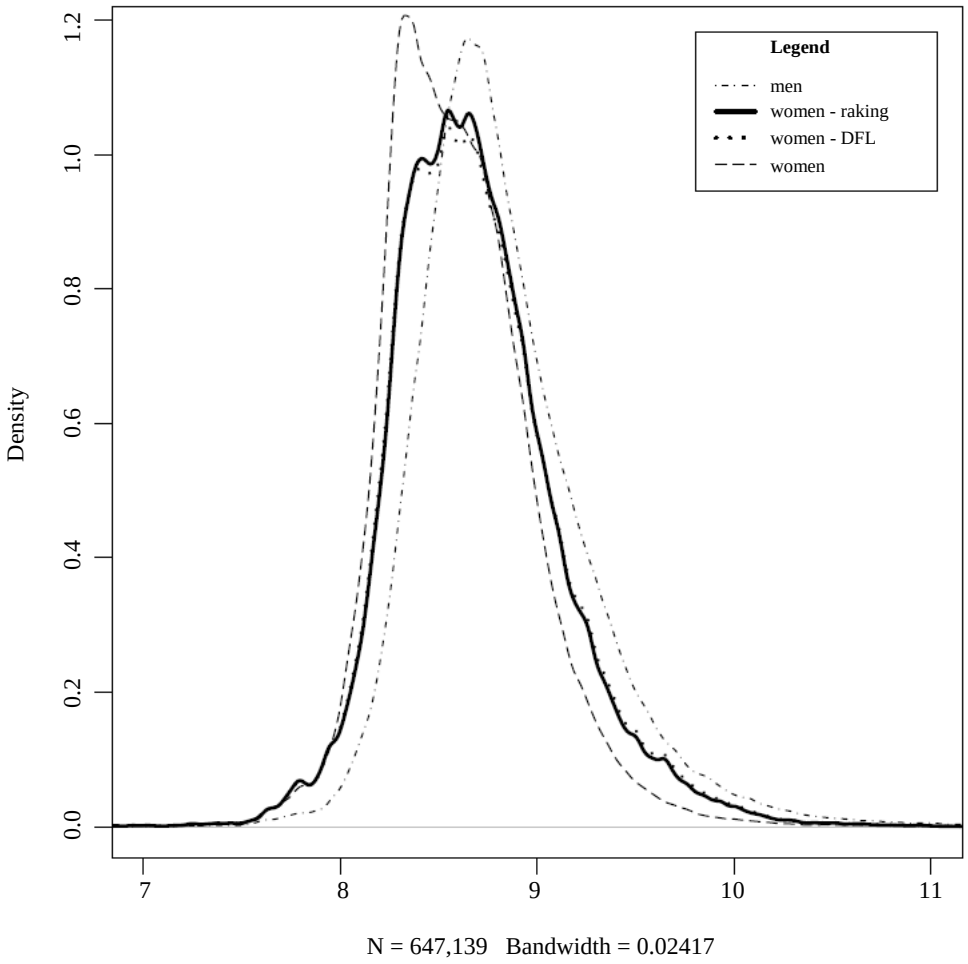

Figure 6.3 Estimated densities of the logarithm of the wage of women and men and the counterfactual distributions of the logarithm of the wage of women constructed using the raking-ratio and the reweighted DFL factor, respectively.

Figure 6.3 shows that the two counterfactual wage distributions are very close to each other around the tails. However, toward the middle, the two methods do not yield the same results. As previously mentioned, using DFL reweighting and calibration methods allow the estimation the composition and structure effects not only at the average levels, but also along the entire distribution. Table 6.6 displays the estimated structure and composition effects of the wage differences between men and women computed using the three methods at some selected quantiles.

Table 6.6
Estimated composition and structure effects of the wage difference at selected quantiles

Quantile	Method	Composition effect (%)	Structure effect (%)	Total
1%	Linear	0.01 (3%)	0.20 (97%)	0.21
	Raking	-0.01 (-3.5%)	0.22 (103.5%)	0.21
	Weighted DFL	-0.01 (-3.4%)	0.22 (103.4%)	0.21
10%	Linear	0.05 (28.8%)	0.12 (71.2%)	0.17
	Raking	0.04 (22.4%)	0.13 (77.6%)	0.17
	Weighted DFL	0.03 (19.4%)	0.14 (80.6%)	0.17
20%	Linear	0.07 (34.2%)	0.13 (65.8%)	0.20
	Raking	0.06 (29.7%)	0.13 (70.3%)	0.19
	Weighted DFL	0.05 (28.2%)	0.14 (71.8%)	0.19
50%	Linear	0.09 (46.3%)	0.10 (53.7%)	0.19
	Raking	0.09 (44.7%)	0.11 (55.3%)	0.20
	Weighted DFL	0.09 (45.7%)	0.11 (54.3%)	0.20
80%	Linear	0.11 (43.9%)	0.15 (56.1%)	0.26
	Raking	0.12 (46.5%)	0.14 (53.5%)	0.26
	Weighted DFL	0.13 (50.8%)	0.13 (49.2%)	0.26
90%	Linear	0.15 (46.0%)	0.18 (54.0%)	0.33
	Raking	0.17 (51.6%)	0.16 (48.4%)	0.33
	Weighted DFL	0.19 (58.0%)	0.14 (42.0%)	0.33
99%	Linear	0.24 (40.0%)	0.36 (60.0%)	0.60
	Raking	0.27 (45.3%)	0.33 (54.7%)	0.60
	Weighted DFL	0.29 (49.4%)	0.30 (50.6%)	0.59

The proportion of the structure effect of the entire wage difference between men and women decreases as the order of the quantile increases. This means that for jobs with higher salaries, more of the wage differences can be explained by differences in group characteristics than for jobs with lower salaries. The raking-ratio and the DFL reweighting factor yield similar results up to the quantile of order 90%. The composition effect at the first percentile is estimated to be negative, meaning that at this point, the differences in wages are due solely to discrimination.

Figure 6.4 shows the weighted quantiles of the logarithms of the wage of men, those of women and contrast the counterfactual distributions obtained through the raking-ratio calibration and the DFL reweighting factor. Because the linear calibration yielded negative weights, the same graph is not reproduced for it.

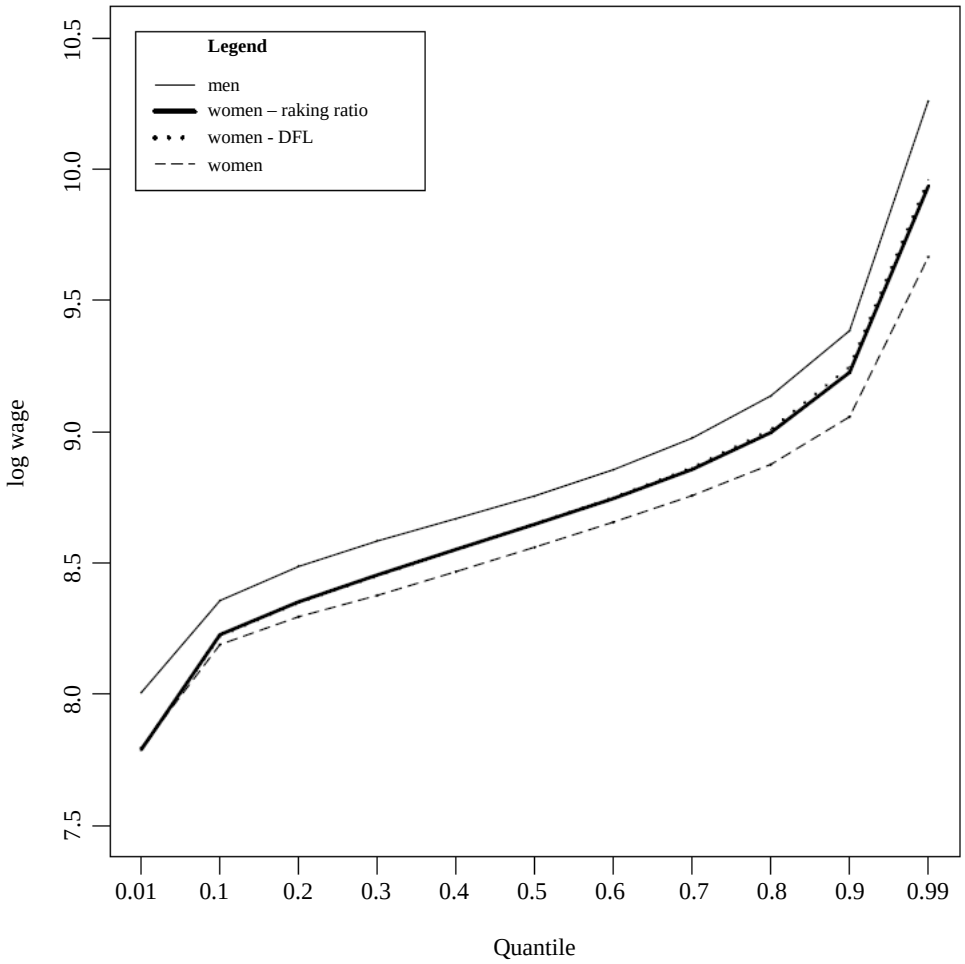

Figure 6.4 Weighted quantiles of the logarithms of the wage of women and men and the weighted quantiles of the counterfactual distribution of the logarithm of the wage of women constructed using the raking-ratio calibration and the weighted DFL factor.

6.4 Further decomposition of the structure effect

A logistic model for the probability of being a man yields estimated values between 0.002 and 0.99. For the variables “years in the current position”, “age” and “square of the age” the difference between the average values of men and the reweighted averages of women computed using the reweighting factor are the largest. In equation (4.8), the structure effect is composed of the pure effect and the residual effect. Using the DFL reweighting factor, the residual effect equals -0.00474. In contrast, by using either one of the calibration techniques, in both cases, it equals 0. Moreover, the calibration approach allows overriding the

computation of the counterfactual regression coefficients. This is because the technique ensures the equality between the means $\hat{\bar{\mathbf{X}}}_M$ and $\hat{\bar{\mathbf{X}}}_{F|M}$. Calibration thus represents a generalization of the DFL reweighting factor technique, because it allows for a more precise estimation of the structure effect, since the resulting value only includes the pure part.

7 Conclusion

The phenomenon of discrimination has multiple facets and there are many mechanisms that can generate it. However, this paper only examines its estimation from a methodological point of view. The two calibration cases taken into consideration represent a generalization of two existing decomposition methods, the technique of Blinder (1973) and Oaxaca (1973) and the semi-parametric method of DiNardo et al. (1996), both expressed using sampling weights. The original methods can also be obtained, if all the sampling weights are considered to be equal to 1. The linear case yields the same result as the BO method. However, since the resulting weights are unbounded, negative values might be observed. Just as the DFL method, the calibration approach allows for the decomposition of wage differences at other points other than the mean, such as quantiles. However, the raking-ratio calibration is an improvement of the DFL method, in that the estimation of the structure effect will always include a residual effect equal to 0. Therefore, the structure effect will only be composed of the pure effect. Decomposing wage differences along quantiles enables the conclusion that in low-paying jobs, the inequalities are due solely to discrimination. In this article, the emphasis was placed on the generalization of two well-established decomposition methods through the calibration approach.

Acknowledgements

The authors are grateful to the Swiss federal statistical office for the financial support and the LOHN department for providing the data. However, the opinions expressed in this paper do not necessarily reflect those of the Swiss federal statistical office.

Appendix A

Proof of Result 1

$$\begin{aligned}
\hat{\mathbf{X}}_h\hat{\boldsymbol{\beta}}_h &= \hat{\mathbf{X}}_h\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\
&= \sum_{j\in S_h} d_j\mathbf{x}_j^\top\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\
&= \left(\sum_{j\in S_h}\boldsymbol{\varsigma}^\top d_j\mathbf{x}_j\mathbf{x}_j^\top\right)\left(\sum_{k\in S_h} d_k\mathbf{x}_k\mathbf{x}_k^\top\right)^{-1}\sum_{l\in S_h} d_l\mathbf{x}_l y_l \\
&= \sum_{l\in S_h}\boldsymbol{\varsigma}^\top d_l\mathbf{x}_l y_l = \sum_{l\in S_h} d_l y_l = \hat{Y}_h.
\end{aligned}$$

By dividing this equation by $\sum_{k\in S_h} d_k$, Result 1 is obtained.

Appendix B

B.1 Linearization of the means

In order to compute the variance of the average means and of the counterfactual means we have used the linearization method proposed by Graf (2011). The author proposes to compute the partial derivative of the estimator with respect to the sample indicator. This derivative provides the linearized variable that can be plugged in the variance estimator. The average means are defined by:

$$\hat{\bar{Y}}_F = \frac{\sum_{k \in S_F} d_k y_k}{\sum_{k \in S_F} d_k},$$

and

$$\hat{\bar{Y}}_M = \frac{\sum_{l \in S_M} d_l y_l}{\sum_{l \in S_M} d_l}.$$

For the two average wages, we obtain the linearized variables:

$$\frac{\partial \hat{\bar{Y}}_F}{\partial I_j} = \begin{cases} \dfrac{d_j \left(y_j - \hat{\bar{Y}}_F \right)}{\sum_{k \in S_F} d_k} & j \in S_F \\ 0 & j \in S_M \end{cases},$$

and

$$\frac{\partial \hat{\bar{Y}}_M}{\partial I_j} = \begin{cases} \dfrac{d_j \left(y_j - \hat{\bar{Y}}_M \right)}{\sum_{l \in S_M} d_l} & j \in S_M \\ 0 & j \in S_F \end{cases}.$$

B.2 Linearization of the counterfactual

In order to compute the counterfactual mean, we compute the weights v_k defined by the system

$$\mathbf{A} = \sum_{k \in S_F} v_k d_k \mathbf{x}_k = \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \sum_{l \in S_M} d_l \mathbf{x}_l = \hat{\bar{\mathbf{X}}}_M \sum_{k \in S_F} d_k,$$

with

$$v_k = F\left(\mathbf{x}_k^\top \boldsymbol{\lambda} \right).$$

For the linearized variables, we have to consider two cases:

- If $j \in S_F$

$$\frac{\partial \mathbf{A}}{\partial I_j} = v_j d_j \mathbf{x}_j + \left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \hat{\bar{\mathbf{X}}}_M .$$

Thus

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = -\mathbf{T}^{-1} d_j \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right),$$

where

$$\mathbf{T} = \sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top .$$

- If $j \in S_M$

$$\frac{\partial \mathbf{A}}{\partial I_j} = \left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k \mathbf{x}_k^\top\right] \frac{\partial \boldsymbol{\lambda}}{\partial I_j} = d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l} .$$

Thus

$$\frac{\partial \boldsymbol{\lambda}}{\partial I_j} = \mathbf{T}^{-1} d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right) \frac{\sum_{k\in S_F} d_k}{\sum_{l\in S_M} d_l} .$$

Since we have supposed that there exists a vector $\boldsymbol{\gamma}$ such that $\boldsymbol{\gamma}^\top \mathbf{x}_k = 1$ for all $k \in U$, then, we have

$$\boldsymbol{\gamma}^\top \mathbf{A} = \sum_{k\in S_F} v_k d_k = \sum_{k\in S_F} d_k .$$

Now consider

$$\hat{\bar{Y}}_{F|M} = \frac{\sum_{k\in S_F} v_k d_k y_k}{\sum_{k\in S_F} v_k d_k} = \frac{\sum_{k\in S_F} v_k d_k y_k}{\sum_{k\in S_F} d_k} .$$

Again, two cases must be considered:

- If $j \in S_F$

$$\begin{aligned}
\frac{\hat{\bar{Y}}_{F|M}}{\partial I_j} &= \frac{d_j \left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) + \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j} \left[\sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k\right]}{\sum_{k\in S_F} d_k} \\
&= \frac{d_j \left(v_j y_j - \hat{\bar{Y}}_{F|M}\right) - d_j \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{T}^{-1} \sum_{k\in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k\in S_F} d_k} \\
&= \frac{d_j \left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F\right]}{\sum_{k\in S_F} d_k},
\end{aligned}$$

where

$$\mathbf{B}_F = \mathbf{T}^{-1} \sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k .$$

- If $j \in S_M$

$$\begin{aligned}
\frac{\partial \hat{\bar{\mathbf{Y}}}_{F|M}}{\partial I_j} &= \frac{\partial \boldsymbol{\lambda}^\top}{\partial I_j} \frac{\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k \in S_F} d_k} \\
&= d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{\sum_{k \in S_F} d_k}{\sum_{l \in S_M} d_l} \mathbf{T}^{-1} \frac{\sum_{k \in S_F} F'\left(\mathbf{x}_k^\top \boldsymbol{\lambda}\right) d_k \mathbf{x}_k y_k}{\sum_{k \in S_F} d_k} \\
&= d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \frac{1}{\sum_{l \in S_M} d_l} \mathbf{B}_F .
\end{aligned}$$

Thus the linearized variable is

$$z_k = \begin{cases} \dfrac{d_j \left[v_j y_j - \hat{\bar{Y}}_{F|M} - \left(v_j \mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F \right]}{\sum_{k \in S_F} d_k} & \text{if} \quad j \in S_F \\ \dfrac{d_j \left(\mathbf{x}_j - \hat{\bar{\mathbf{X}}}_M\right)^\top \mathbf{B}_F}{\sum_{l \in S_M} d_l} & \text{if} \quad j \in S_M . \end{cases}$$

The linearized variable must only be plugged in the variance estimator corresponding to the sampling design. Note that the variance of the counterfactual depends on the variance computed for the sample of men for the part that is explained by the regression and the variance computed for the sample of women for the part that remains unexplained.

References

Bielby, W.T., and Baron, J.N. (1986). Men and women at work: Sex segregation and statistical discrimination. *American Journal of Sociology*, 759-799.

Blinder, A.S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4), 436-455.

Bourguignon, F., Ferreira, F.H. and Leite, P.G. (2002). Beyond Oaxaca-Blinder: Accounting for differences in household income distributions across countries. *Inequality and Economic Development in Brazil*, 105.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. Discussion paper, University of Michigan.

DiNardo, J., Fortin, N.M. and Lemieux, T. (1996). Labor market Institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5), 1001-44.

Donzé, L. (2013). Erreurs de spécification dans la décomposition de l'inégalité salariale. In *International Conference Ars Conjectandi*, 1713-2013.

Fortin, N., Lemieux, T. and Firpo, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics*, (Eds., O. Ashenfelter and D. Card), 4, 1-102. Elsevier.

Gardeazabal, J., and Ugidos, A. (2005). Gender wage discrimination at quantiles. *Journal of Population Economics*, 18(1), 165-179.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis: Theory and Applications*, (Eds., V. Pawlowsky-Glahn and A. Buccianti), 114-127. Wiley, Chichester.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage Discrimination. *Journal of Human Resources*, 23(3), 279-295.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3), 693-709.

Weichselbaumer, D., and Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, 19(3), 479-511.

Weichselbaumer, D., and Winter-Ebmer, R. (2006). Rhetoric in economic research: The case of gender wage differentials. *Industrial Relations: A Journal of Economy and Society*, 45(3), 416-436.

Williams, C. (2010). Economic Well-being. Women in Canada: A Gender-based Statistical Report, Statistics Canada.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 235-240
Statistics Canada, Catalogue No. 12-001-X

A note on Wilson coverage intervals for proportions estimated from complex samples

Phillip S. Kott[1]

Abstract

This note discusses the theoretical foundations for the extension of the Wilson two-sided coverage interval to an estimated proportion computed from complex survey data. The interval is shown to be asymptotically equivalent to an interval derived from a logistic transformation. A mildly better version is discussed, but users may prefer constructing a one-sided interval already in the literature.

Key Words: Effective sample size; Confidence interval; Logistic transformation.

1 Introduction

Brown, Cai and Dasgupta (2001) show that a method proposed by Wilson (1927) can produce reasonably well-behaved two-sided coverage intervals for a proportion under simple random sampling *with* replacement. Section 2 of this note discusses the theoretical foundations for extending this interval-construction method to estimated proportions computed from a complex survey. Section 3 shows that such a Wilson-type interval can be asymptotically equivalent to an interval derived from a logistic transformation. Section 4 offers some concluding remarks.

The term "coverage interval" is used here in place of the more common "confidence interval" because a 95% Wilson coverage interval does not attempt to cover the true proportion at least 95% of the time no matter what that proportion is. Instead, it merely tries to cover the true proportion 95% of the time for reasonable values of the true proportion. For some values it overcovers, for others it undercovers as shown in Brown et al. (2001). By limiting its applicability to two-sided coverage intervals, the Wilson methodology is (mostly) able to ignore the asymmetry of the distribution of an estimated proportion.

2 The extension

It is not hard to generalize Wilson coverage intervals (also called "score intervals") to complex survey data. See, for example, Kott and Carr (1997). As with the Wilson itself, one simply solves this equation for the true proportion P:

$$\frac{(p-P)^2}{\left[\dfrac{P(1-P)}{n^*}\right]} \leq z^2_{1-\alpha/2}, \tag{2.1}$$

1. Phillip S. Kott, RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. E-mail: pkott@rti.org.

where p is a consistent estimator for P under probability-sampling theory, and $z_{1-\alpha/2}$ is the Normal $z-$ score for $(1-\alpha/2)$ given the goal is to produce a $(1-\alpha)\,\%$ coverage interval (α is often set at 0.05). The missing piece to equation (2.1) is n^*, the so-called "effective sample size", which in the standard Wilson formulation is the sample size n. In our more general context, $n^* = p\,(1-p)/\mathrm{var}\,(p)$, where $\mathrm{var}\,(p)$ is a consistent estimator for the variance of p, $\mathrm{Var}\,(p)$.

In order to calculate n^*, we need both $p\,(1-p)$, and $\mathrm{var}\,(p)$ to be positive. In addition, let us assume that $1/n^* = O_P\left(1/n^a\right)$ for some positive $a \le 1$, $p - P = O_P\left(1/\sqrt{n^*}\right)$, $0 < \mathrm{Var}\,(p) = O\left(1/n^*\right)$, and $\mathrm{var}\,(p)/\mathrm{Var}\,(p)$ is $1 + O_P\left(1/\sqrt{n^*}\right)$. Note that the last three are always true under simple random sampling with replacement so long as $P\,(1-P) \ge B > 0$.

Dropping $O_P\left(1/[n^*]^{3/2}\right)$ terms, but allowing $p\,(1-p)$ to be small (effectively $o_p(1)$), one can derive this Wilson-like interval for P from equation (2.1):

$$p + \frac{1-2p}{n^*}\frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2}\left(\frac{p\,(1-p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4\,(n^*)^2}\right)^{1/2} \le P$$

$$\le p + \frac{1-2p}{n^*}\frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2}\left(\frac{p\,(1-p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4\,(n^*)^2}\right)^{1/2}. \qquad (2.2)$$

We can call this the "complex-sampling Wilson coverage interval". WesVar (2007) computes a variant of this interval that does not drop $O_P\left(1/[n^*]^{3/2}\right)$ terms. It is dropped here because other terms of that size will be dropped later in this note.

If it is reasonable to drop $O_P\left(1/[n^*]^{3/2}\right)$ terms in deriving equation (2.2), one can also safely ignore the difference between $1/n$ and $1/(n-1)$. Under simple random sampling *without* replacement, $n^* = n/(1-f)$ (or $(n-1)/(1-f)$) where f is the sampling fraction. When f is very small, the distinction between with and without replacement sampling can be ignored.

Observe that under simple random sampling with replacement, the denominator of the pivotal appearing on the left-hand side of equation (2.1) has no variance at all. By contrast, the denominator in the traditional Wald pivotal, $\mathrm{var}\,(p) = p\,(1-p)/(n-1)$, can have considerable variance, especially when p or $1-p$ is small. That is why Wilson intervals have superior performance under simple random sampling, whether with or without replacement.

That superiority carries over to complex sampling (see, for example, Kott, Andersson and Nerman, 2001), where the pivotal's denominator is

$$\frac{P\,(1-P)}{n^*} = \mathrm{var}\,(p)\frac{P\,(1-P)}{p\,(1-p)} = \mathrm{var}\,(p)\left[1 - \frac{(p-P) - (p^2 - P^2)}{p\,(1-p)}\right]$$

$$= \mathrm{var}\,(p)\left[1 - \frac{(p-P) - (p-P)\,(p+P)}{p\,(1-p)}\right]$$

$$= \mathrm{var}\,(p) - \frac{1-2P}{n^*}(p-P) + \mathrm{O}_P\left(1/[n^*]^2\right),$$

which is likely to have less variance than $\text{var}(p)$ in most applications. For an intuition into why this is so, observe that a putative variance estimator of the form $\text{var}_1(p) = \text{var}(p) - b(p - P)$ is minimized when $b = \text{Cov}[\text{var}(p), p]/\text{Var}(p)$. Under simple random sampling, whether with or without replacement, b is exactly $(1-2P)/n^*$.

Although the minimizing b is not exactly equal to $(1-2P)/n^*$, under more complex sampling designs, the optimal b is likely to be closer to $(1-2P)/n^*$ than to 0. It is thus not surprising that the variance of $\text{var}(p) - [(1-2P)/n^*](p - P)$ will usually be less than the variance of $\text{var}(p)$. Nevertheless, a slight improvement on the complex-sampling Wilson coverage interval can be made by replacing n^* in equation (2.2) by

$$\tilde{n} = [(1-2p)\,\text{var}(p)]/\text{cov}[\text{var}(p), p]$$

when $\text{cov}[\text{var}(p), p]$, a consistent estimator for $\text{Cov}[\text{var}(p), p]$, exists (see Kott et al., 2001).

As with the standard Wilson, the center of the complex-sample Wilson interval in equation (2.2) is slightly different from p when p is not $\frac{1}{2}$:

$$C = p + \frac{1-2p}{n^*}\,\frac{z_{1-\alpha/2}}{2}.$$

Its length L appears longer than the Wald's:

$$L = z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*} + \frac{z^2_{1-\alpha/2}}{4(n^*)^2}\right)^{1/2} > z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*}\right)^{1/2}.$$

When $P(1-P) \geq B > 0$, however,

$$\begin{aligned}\left(\frac{p(1-p)}{n^*} + \frac{z^2_{1-\alpha/2}}{4(n^*)^2}\right)^{1/2} &= \left(\frac{p(1-p)}{n^*}\right)^{1/2}\left(1 + \frac{\frac{1}{4}z^2_{1-\alpha/2}}{n^* p(1-p)}\right)^{1/2} \\ &= \left(\frac{p(1-p)}{n^*}\right)^{1/2} + o_p\left(\frac{1}{n^*}\right). \qquad (2.3)\end{aligned}$$

3 The logistic transformation

The complex-sampling Wilson coverage interval turns out to be very similar to this two-sided coverage interval derived using a logistic transformation (see Brown et al., 2001):

$$f^{-1}\left\{f(p) - z_{1-\alpha/2}\sqrt{\text{var}[f(p)]}\right\} \leq P \leq f^{-1}\left\{f(p) + z_{1-\alpha/2}\sqrt{\text{var}[f(p)]}\right\}, \qquad (3.1)$$

where $f(p) = \log(p) - \log(1-p)$, and $\text{var}[f(p)] = \left[f'(p)\right]^2 \text{var}(p) = [1/p + 1/(1-p)]^2 p(1-p)/n^* = 1/[n^* p(1-p)]$. The original rationale for this interval appears to be that it has this desirable property: it cannot contain values less than 0 or greater than 1, which would be nonsensical for a proportion.

The left-hand side of equation (3.1) can be rewritten as $g(x-h)$, where

$$g(y) = f^{-1}(y) = [1+\exp(-y)]^{-1},\ x = f(p) = \log\left(\frac{p}{1-p}\right),$$

and

$$h = \frac{z_{1-\alpha/2}}{\sqrt{n^* p(1-p)}}.$$

The first and second derivatives of $g(y)$ are $g'(y) = g(y)[1-g(y)]$, and $g''(y) = g(y)[1-g(y)][1-2g(y)]$. Invoking the mean value theorem, there is an h^* between 0 and h such that

$$\begin{aligned} g(x-h) &= g(x) - g'(x)h + \frac{1}{2}g''(x-h^*)h^2 \\ &= p - p(1-p)\frac{z_{1-\alpha/2}}{\sqrt{n^* p(1-p)}} \\ &+ \frac{1}{2}\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\left\{1-\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\right\}\left\{1-2\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\right\}\frac{z^2_{1-\alpha/2}}{n^* p(1-p)} \\ &= p - p(1-p)\frac{z_{1-\alpha/2}}{\sqrt{n^* p(1-p)}} \\ &+ \frac{1}{2}\frac{p}{1+(1-p)(e^{-h^*}-1)}\frac{(1-p)-(1-p)(e^{h^*}-1)}{1+(1-p)(e^{h^*}-1)}\frac{(1-2p)-(1-p)(e^{h^*}-1)}{1+(1-p)(e^{h^*}-1)}\frac{z^2_{1-\alpha/2}}{n^* p(1-p)}, \end{aligned}$$

using

$$\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1} = \frac{p}{1+(1-p)(e^{h^*}-1)}.$$

An analogous derivation can be made for the right-hand side of equation (3.1).

Consequently,

$$\begin{aligned} p + \frac{1-2p}{n^*}\frac{z^2_{1-\alpha/2}}{2} - z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*}\right)^{1/2} + o_P\left(\frac{1}{n^*}\right) &\le P \\ &\le p + \frac{1-2p}{n^*}\frac{z^2_{1-\alpha/2}}{2} + z_{1-\alpha/2}\left(\frac{p(1-p)}{n^*}\right)^{1/2} + o_P\left(\frac{1}{n^*}\right). \end{aligned}$$

After invoking the asymptotic equality in equation (2.3) and dropping $o_P(1/n^*)$ terms, the last set of inequalities is equivalent to Wilson interval in equation (2.2) so long as n^* is sufficiently large and $P(1-P) > 0$, the latter meaning that the true proportion is neither 0 or 1.

4 Some concluding remarks

The asymptotic equivalence of a coverage interval based on a logistic transformation to the theoretically grounded Wilson interval is the main contribution of this paper. Although in the asymptotic framework, $P(1-P)$ is fixed and positive as n^* grows large, in practice it is the size of $p(1-p)n^*$ that matters when comparing the Wilson-type and logistic-transformation intervals. This requires that $P(1-P)$ not be too small.

Brown et al. (2001) show empirically that under simple random sampling (with $n = 50$), coverage intervals derived from the logistic transformation tend to be larger than corresponding Wilson intervals for small values of $P(1-P)$. Kott and Liu (2009) make the same observation for one-sided intervals based on complex samples, supporting the notion that it is a better choice.

The asymptotic equivalence of the logistic-transformation interval with the Wilson interval explains the former's empirical superiority in the literature (e.g., in Brown et al., 2001) to an analogous interval constructed using an arcsine transformation. Because $\arcsin(p)$ has a constant large-sample variance under simple random sampling no matter the true value of P (so long as $P(1-P) > 0$), it has been hoped that the arcsine transformation would be ideal for interval construction.

Better than a Wilson interval, but not yet incorporated into any software package I know of, is the one-sided coverage intervals for P derived using an Edgeworth expansion on $p - P$ in Kott and Liu (2009). That method produces this two-sided interval:

$$p + \frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right) - z_{1-\alpha/2}\left(\text{var}(p) + \left[\frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right)\right]^2\right)^{1/2} \leq P$$

$$\leq p + \frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right) + z_{1-\alpha/2}\left(\text{var}(p) + \left[\frac{1-2p}{\tilde{n}}\left(\frac{1}{6} + \frac{z_{1-\alpha/2}^2}{3}\right)\right]^2\right)^{1/2},$$

where $\tilde{n} = [(1-2p)\,\text{var}(p)] \,/\, \text{cov}[\text{var}(p),\ p]$, and $\text{cov}[\text{var}(p),\ p]$, a consistent estimator for $\text{Cov}[\text{var}(p),\ p]$, exists and equals a consistent estimator for the third moment of p. Note that $\text{cov}[\text{var}(p),\ p]$ doesn't exist for designs with only two primary sampling units per stratum. Moreover, it is not a consistent estimator for the third moment of p when finite population correction matters.

Observe that $\tilde{n}$ again replaces n^*. In addition, $1/6 + z_{1-\alpha/2}/3$ replaces $z_{1-\alpha/2}/2$, which means that the center will often be closer to the p using this interval rather than the Wilson. The good coverage properties of this interval, like the Wilson, breaks down when the skewness coefficient of $p\left(\text{E}\left[(p-P)^3\right] \,/\, [\text{Var}(p)]^{3/2}\right)$ gets too large in absolute value, how large has yet to be determined.

Finally, SAS/STAT (SAS Institute Inc., 2010) offers a Wilson coverage interval for estimated proportions in its SURVEYFREQ procedure. The procedure's method of adjusting the effective sample size, which can – and should – be turned off, is not related to the $\tilde{n}$ discussed here. Instead, it is based on an ad-hoc t – adjustment that sadly is not related to the variance of the denominator variance of the Wilson pivotal.

Acknowledgements

The author thanks Per Gösta Andersson for introducing me to this area of research and an anonymous referee for correcting errors in a previous version of the manuscript. Remaining errors are my own.

References

Brown, L.D., Cai, T. and Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101-133.

Kott, P.S., and Carr, D.A. (1997). Developing an estimation strategy for a pesticide data program. *Journal of Official Statistics*, 13, 367-383.

Kott, P.S., and Liu, Y.K. (2009). One-sided coverage intervals for a proportion estimated from a stratified simple random sample. *International Statistical Review/Revue Internationale de Statistique*, 77, 251-265.

Kott, P.S., Andersson, P.G. and Nerman, O. (2001). Two-sided coverage intervals for small proportion based on survey data. Presented at Federal Committee on Statistical Methodology Research Conference, Washington, DC. http://fcsm.sites.usa.gov/files/2014/05/2001FCSM_Kott.pdf.

SAS Institute Inc. (2010). *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_surveyfreq_a0000000252.htm.

WesVar (2007). *WesVar® 4.3 Users' Guide*, B28-B29.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2017.

P. Ardilly, *INSEE*
J.-F. Beaumont, *Statistics Canada*
A. Bianchi, *University of Bergamo*
P. Biemer, *RTI International*
H.J. Boonstra, *Statistics Netherlands*
J.M. Brick, *Westat Inc.*
P. Brodie, *Office for National Statistics*
P.J. Cantwell, *U.S. Bureau of the Census*
H. Cardot, *Université de Bourgogne*
K.-C. Chang, *Fu Jen Catholic University*
G. Chauvet, *ENSAI/IRMAR*
B. Chen, *Bureau of Economic Analysis*
J. Chipperfield, *Australian Bureau of Statistics*
T. Deroyon, *INSEE*
J. Dever, *RTI International*
L. Dudoignon, *Médiamétrie*
D. Elliot, *Office of National Statistics*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
S. Er, *University of Cape Town*
S. Falorsi, *ISTAT*
S. Fortier, *Statistics Canada*
C. Franco, *U.S. Census Bureau*
W.A. Fuller, *Iowa State University*
M. Furno, *Università degli Studi di Napoli 'Federico II'*
J. Gambino, *Statistics Canada*
C. Girard, *Statistics Canada*
C. Goga, *Université de Bourgogne*
A. Goia, *Università del Piemonte Orientale*
A. Grafström, *Sweedish University of Agricultural Studies*
E. Gros, *INSEE*
R. Gutman, *Brown University*
D. Haziza, *Université de Montréal*
Y. He, *National Center for Health Statistics*
S.G. Heeringa, *University of Michigan*
M. Hidiroglou, *Statistics Canada*
J. Hua, *Dartmouth College*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
J. Im, *Iowa State University*
D. Judkins, *Abt Associates*
E. Kabzinska, *University of Southampton*
J.K. Kim, *Iowa State University*
N. Kim, *Carnegie Mellon University*
P. Kott, *RTI International*
M. Kozak, *University of Inf. Techmology and Management in Rzeszow*
P. Lahiri, *University of Maryland*
A. Lauger, *U.S. Census Bureau*
P. Lavallée, *Statistics Canada*
C. Léger, *Université de Montréal*
J. Legg, *Amagen Inc.*
E. Lesage, *INSEE*
R. Lethonen, *Université d'Helsinki*
D. Liao, *RTI International*
S. Lohr, *Westat Inc.*
A.G. Matei, *Université de Neuchâtel*
K. McConville, *Swarthmore College*
I. Molina, *Universidad Carlos III de Madrid*
P.K. Mukhopadhyay, *SAS Institute Inc.*
C.O. Nambeu, *Statistics Canada*
E. Neusy, *Statistics Canada*
J. Opsomer, *Colorado State University*
P. Ouwehand, *Statistics Netherland*
D. Pavlopoulos, *Free university of Amsterdam*
D. Pfeffermann, *Hebrew University*
R. Powers, *U.S. Bureau of Labor Statistics*
J.N.K. Rao, *Carleton University*
A. Rebecq, *INSEE*
L.-P. Rivest, *Université Laval*
A. Ruiz-Gazen, *Université Toulouse*
F. Scheuren, *National Opinion Research Center*
T. Schmid, *Freie Universität Berlin*
P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *University of Southampton*
D. Steel, *University of Wollongong*
M. Sverchkov, *U.S. Bureau of Labor Statistics*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R.B. Tiller, *U.S. Bureau of Labor Statistics*
D. Toth, *U.S. Bureau of Labor Statistics*
J. van den Brakel, *Statistics Netherland and Maastricht University*
P. van Kerm, *Luxembourg Institute of Socio-Economic Research*
V. Vehovar, *University of Ljubljana*
B.T. West, *University of Michigan-Ann Arbor*
C. Wu, *University of Waterloo*
S. Yang, *North Carolina State University*
Y. You, *Statistics Canada*
A. Zaslavsky, *Harvard University*
G. Zhang, *University of New Mexico*
L.-C. Zhang, *University of Southampton*
T. Zimmermann, *German Federal Statistical Office*

Acknowledgements are also due to those who assisted during the production of the 2017 issues: Céline Ethier of International Cooperation and Corporate Statistical Methods Division; Joana Bérubé of Business Survey Methods Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Christina Jaworski, Kathy Charbonneau, Jacqueline Luffman, Jenna Waite, Darquise Pellerin and Joseph Prince as well as our partners in the Communications Division.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 33, No. 2, June 2017

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 33, No. 3, September 2017

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 45, No. 3, September/septembre 2017

Issue Information

Review Article

Original Articles

CONTENTS TABLE DES MATIÈRES

Volume 45, No. 4, December/décembre 2017

Issue Information

Original Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (statcan.smj-rte.statcan@canada.ca). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The documents should be divided into numbered sections with suitable verbal titles.
1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.
4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.